Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών

Τομέας Τεχνολογίας Υπολογιστών και Πληροφορικής

Σημασιολογική Αναζήτηση Οπτικοακουστικού Περιεχομένου με βάση τη Γνώση

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

του

ΕΥΑΓΓΕΛΟΥ Χ. ΣΠΥΡΟΥ

Διπλωματούχου Ηλεκτρολόγου Μηχανικού &
Μηχανικού Υπολογιστών Ε.Μ.Π. (2003)

Αθήνα, Νοέμβριος 2009

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Υπολογιστών και Πληροφορικής

Σημασιολογική Αναζήτηση Οπτικοακουστικού Περιεχομένου με βάση τη Γνώση

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

του

ΕΥΑΓΓΕΛΟΥ Χ. ΣΠΥΡΟΥ

Διπλωματούχου Ηλεκτρολόγου Μηχανικού &
Μηχανικού Υπολογιστών Ε.Μ.Π. (2003)

Συμβουλευτική Επιτροπή: Στέφανος Κόλλιας
Ανδρέας Γεώργιος Σταφυλοπάτης
Παναγιώτης Τσανάκας

Εγκρίθηκε από την επταμελή εξεταστική επιτροπή την 20η Νοεμβρίου 2009.

...	...	...
Σ. Κόλλιας	Α.Γ. Σταφυλοπάτης	Π. Τσανάκας
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.
...	...	...
Π. Μαραγκός	Γ. Στάμου	Κ. Καρπούζης
Καθηγητής Ε.Μ.Π.	Λέκτορας Ε.Μ.Π.	Ερευνητής Β' ΕΠΙΣΕΥ-Ε.Μ.Π.

...
Γ. Τζιρίτας
Καθηγητής Παν. Κρήτης

Αθήνα, Νοέμβριος 2009

Η παρούσα διδακτορική διατριβή πραγματοποιήθηκε στα πλαίσια του προγράμματος ΠΕΝΕΔ-2003, της Γενικής Γραμματείας Έρευνας και Τεχνολογίας. Το πρόγραμμα συγχρηματοδοτήθηκε κατά 80% από την Ευρωπαϊκή Ένωση και κατά 20% από το Ελληνικό Δημόσιο.

This Ph.D. thesis was supported by grant PENED-2003 of the Greek Ministry of Development-GSRT and was co-financed by E.U.-European Social Fund (80%) and National Resources (20%).

...

ΕΥΑΓΓΕΛΟΣ Χ. ΣΠΥΡΟΥ
Διδάκτωρ Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

© 2009 - All rights reserved

Περιεχόμενα

Κατάλογος σχημάτων

Κατάλογος πινάκων

ΕΥΧΑΡΙΣΤΙΕΣ

θα ήθελα να εκφράσω τις ευχαριστίες μου καταρχήν προς τον επιβλέποντα της διδακτορικής μου διατριβής Καθηγητή Ε.Μ.Π. κ. Στέφανο Κόλλια για την εμπιστοσύνη που μου έδειξε και την άψογη συνεργασία που είχαμε όλα αυτά τα χρόνια. Επίσης προς τον Δρ Γιάννη Αβρίθη που με καθοδήγησε με τον καλύτερο δυνατό τρόπο και που πάντα ήταν κοντά μου με τις πολύτιμες συμβουλές και τις χρήσιμες παρατηρήσεις του. Επίσης, οφείλω πολλά και στους συναδέλφους και πάνω από όλα φίλους με τους οποίους συνεργάστηκα όλα αυτά τα χρόνια και συγκεκριμένα το Δρ Φοίβο Μυλωνά και τους υποψήφιους διδάκτορες Θάνο Αθανασιάδη, Γιάννη Καλαντίδη, Θεόφιλο Μαΐλη, Νίκο Σίμου και Γιώργο Τόλια. Τέλος, ευχαριστώ την οικογένειά μου και τους φίλους μου που ήταν κοντά μου όλα αυτά τα χρόνια και στάθηκαν πάντα στο πλάι μου όταν τους είχα ανάγκη.

Ευάγγελος Σπύρου
Αθήνα, Νοέμβριος 2009

ΠΕΡΙΛΗΨΗ

Τις τελευταίες δεκαετίες έχει παρατηρηθεί μία τεράστια αύξηση τόσο στην παραγωγή, όσο και στη ζήτηση ψηφιακού οπτικοακουστικού υλικού. Για να ικανοποιηθούν οι ανάγκες των χρηστών του, πρέπει το υλικό αυτό να τεκμηριωθεί, να σχολιαστεί και να ταξινομηθεί σε κατάλληλες θεματικές κατηγορίες, διευκολύνοντας την αναζήτηση και την πρόσβαση σε αυτό. Η παρούσα διατριβή κινείται στο χώρο της ανάλυσης πολυμεσικού υλικού και αντιμετωπίζει ορισμένα από τα βασικότερα ερευνητικά προβλήματα στο χώρο αυτό. Ιδιαίτερο βάρος δίνεται στην ταξινόμηση εικόνων, την ταξινόμηση περιοχών εικόνων καθώς και την ανίχνευση εννοιών σε εικόνες. Για το σκοπό αυτό προτείνονται και αξιολογούνται τεχνικές που αξιοποιούν άμεσα ή έμμεσα υπάρχουσα γνώση για ένα θεματικό πεδίο. Η γνώση αυτή κωδικοποιείται είτε μέσω κατάλληλων οντολογιών, είτε μέσω του εννοιολογικού πλαισίου εικόνων και περιοχών εικόνων, είτε μέσω κατάλληλων τεχνικών μηχανικής μάθησης. Έμφαση δίνεται στη χρήση του μοντέλου bag-of-words για την αναπαράσταση των οπτικών χαρακτηριστικών των εικόνων. Τέλος, οι ιδέες που εφαρμόστηκαν στη ανίχνευση εννοιών σε εικόνες, με τη χρήση του μοντέλου αυτού εφαρμόζονται σε προβλήματα ανάκτησης εικόνων και δημιουργίας περιλήψεων από βίντεο.

ABSTRACT

The growth of production and demand for digital audiovisual content during the last few decades has been overwhelming. To fulfil the needs of it's users, this multimedia content should be annotated, commented and classified into appropriate semantic classes, in order to facilitate search and access to it. This thesis deals with the analysis of multimedia content and faces a few of the most important research problems in the field of multimedia analysis. More specifically, it faces problems such as image classification, image region classification and detection of concepts in images. To achieve this, certain techniques that exploit directly and indirectly the knoweledge of a domain are proposed and evaluated. This knowledge is encoded either in the form of appropriate ontologies, or by modelling the context of the images and their regions, or by applying machine learning techniques. It emphasizes on the use of the bag-of-words model in order to describe the visual features of images. Finally, the techniques applied in high-level concept detection in images are extended in order to be applied to the problems of image retrieval and video summarization.

Κατάλογος Συντμήσεων

AC	:	Alternating Current	Εναλλασσόμενο Ρεύμα
ART	:	Angular Radial Transform	Γωνιακή Ακτινική Μετατροπή
CBSD	:	Contour-Based Shape Descriptor	Περιγραφέας Σχήματος με βάση το Περίγραμμα
CLD	:	Color Layout Descriptor	Περιγραφέας Διάταξης Χρώματος
CSD	:	Color Structure Descriptor	Περιγραφέας Δομής Χρώματος
CSS	:	Curvature Scale Space	Χώρος Κλίμακας Καμπυλότητας
DCD	:	Dominant Color Descriptor	Περιγραφέας Κυρίαρχων Χρωμάτων
DCT	:	Discrete Cosine Transform	Διακριτός Μετασχηματισμός Συνημιτόνου
DoG	:	Difference of Gaussians	Διαφορά Γκαουσιανών
DS	:	Description Scheme	Σχήμα Περιγραφής
EHD	:	Edge Histogram Descriptor	Περιγραφέας Ιστογράμματος Ακμών
EM	:	Expectation Maximization	Μεγιστοποίηση Προσδοκίας
EXIF	:	EXchangeable Image file Format	Ανταλλάξιμο Σχήμα Αρχείου Εικόνων
GMM	:	Gaussian Mixture Model	Γκαουσιανό Μοντέλο Μείξης
HMM	:	Hidden Markov Model	Κρυφό Μαρκοβιανό Μοντέλο
HMMD	:	Hue Max Min Difference	Χροιά Μέγιστη Ελάχιστη Διαφορά
HSV	:	Hue Saturation Value	Χροιά Κορεσμός Τιμή
ISO/IEC	:	International Organization for Standardization/International Electrotechnical Commission	Διεθνής Οργανισμός Πιστοποίησης/Διεθνής Ηλεκτροτεχνική Επιτροπή
LSA	:	Latent Semantic Analysis	Λανθάνουσα Σημασιολογική Ανάλυση
mAP	:	mean Average Precision	-
MCDI	:	Multimedia Content Description Interface	Διεπαφή Περιγραφής Πολυμεσικού Περιεχομένου
MDL	:	Minimum Description Length	Περιγραφή Ελάχιστου Μήκους
MDS	:	Multimedia Description Scheme	Σχήμα Περιγραφής Πολυμέσων
MLP	:	Multi Layer Perceptron	Πολυεπίπεδο Perceptron
MPEG	:	Motion Pictures Experts Group	Ομάδα Ειδικών Κινούμενης Εικόνας
MRF	:	Markov Random Field	Μαρκοβιανό Τυχαίο Πεδίο
MSER	:	Maximally Stable Extremal Regions	Ακραίες Περιοχές Μέγιστης Σταθερότητας
MSO	:	Multimedia Structure Ontology	Οντολογία Δομής Πολυμέσων
NIST	:	National Institute of Standards and Technology	Εθνικό Ινστιτούτο Προτύπων και Τεχνολογίας

OWL	:	Web Ontology Language	Γλώσσα Οντολογιών Ιστού
PCA	:	Principal Component Analysis	Ανάλυση Πρωτογενών Συνιστωσών
pLSA	:	probabilistic Latent Semantic Analysis	πιθανοτική Λανθάνουσα Σημασιολογική Ανάλυση
RBF	:	Radial Basis Function	Ακτινική Συνάρτηση Βάσης
RDF	:	Resource Description Framework	Πλαίσιο Περιγραφής Πόρων
RDFS	:	Resource Description Framework Schema	Σχήμα Πλαισίου Περιγραφής Πόρων
RSD	:	Region Shape Descriptor	Περιγραφέας Σχήματος με βάση το Σχήμα
RSST	:	Recursive Shortest Spanning Tree	Αναδρομικά Ελάχιστο Συνδετικό Δέντρο
SCD	:	Scalable Color Descriptor	Περιγραφέας Κλιμακωτού Χρώματος
SIFT	:	Scale Invariant Feature Transform	Μετασχηματισμός Χαρακτηριστικών Αμετάβλητος σε Κλίμακα
SOM	:	Self-Organized Map	Αυτο-Οργανωνούμενος Χάρτης
SVD	:	Singular Value Decomposition	Αποσύνθεση Ιδιαζουσών Τιμών
SVM	:	Support Vector Machines	Μηχανές Διανυσμάτων Υποστήριξης
TF/IDF	:	Term Frequency/Inverted Document Frequency	Συχνότητα Όρων/Συχνότητα Ανάστροφου Εγγράφου
TREC	:	Text REtrieval Conference	Συνέδριο Ανάκτησης Κειμένου
TRECVID	:	TREC Video retrieval	Ανάκτηση Βίντεο του TREC
VDO	:	Visual Descriptor Ontology	Οντολογία Οπτικών Περιγραφέων
XML	:	eXtensible Markup Language	Εκτατή Γλώσσα Διατύπωσης

Κεφάλαιο 1

Εισαγωγή

1.1 Πρόλογος

Τις τελευταίες δεκαετίες έχει παρατηρηθεί μία τεράστια αύξηση τόσο στην παραγωγή, όσο και στη ζήτηση ψηφιακού οπτικοακουστικού υλικού. Οι πρόσφατες εξελίξεις όσον αφορά τη διάδοση της χρήσης ηλεκτρονικών υπολογιστών, τη διαρκώς αυξανόμενη πρόσβαση στο διαδίκτυο και τη γιγάντωση του Παγκόσμιου Ιστου, έχουν συμβάλλει σε μεγάλο βαθμό στη δημιουργία και τη διάδοση ψηφιακού πολυμεσικού υλικού σε πολύ μεγάλη κλίμακα και έχουν δημιουργήσει νέες ανάγκες τόσο στους χρήστες του, όσο και στο κομμάτι της ερευνητικής κοινότητας που ασχολείται με την παραγωγή, την κωδικοποίηση, τη μετάδοση, την κατανόηση και την ανάκτησή του.

Έτσι, ο μέσος χρήστης του πολυμεσικού υλικού έχει πλέον στην κατοχή του χιλιάδες ψηφιακές φωτογραφίες και δεκάδες ώρες βίντεο. Επιθυμεί την οργάνωση του υλικού αυτού με αποδοτικό τρόπο, ώστε να μπορεί εύκολα να ανακτά φωτογραφίες και βίντεο, σύμφωνα με τις επιθυμίες και τις ανάγκες του. Επίσης, συνηθίζει να μοιράζεται το υλικό αυτό με άλλους χρήστες του διαδικτύου και να αναζητά παρόμοιες φωτογραφίες με τις δικές του, φωτογραφίες από μέρη που επιθυμεί να ταξιδεψει, από προϊόντα που επιθυμεί να αγοράσει κ.ο.κ. Πέρα από αυτό, πολλοί οργανισμοί που διαθέτουν πλούσιο οπτικοακουστικό αρχείο έχουν προβεί στην ψηφιοποίηση και τη διάθεσή του στους χρήστες. Επιπλέον, καθημερινά αναρτώνται στο διαδίκτυο χιλιάδες φωτογραφίες και βίντεο για ενημερωτικούς και ψυχαγωγικούς σκοπούς.

Για να ικανοποιηθούν οι ολοένα αυξανόμενες ανάγκες των χρηστών, αλλά και να διευκολυνθεί η διάθεσή του στους χρήστες σύμφωνα με τις επιθυμίες τους, πρέπει το υλικό αυτό να τεκμηριωθεί, να σχολιαστεί και να ταξινομηθεί σε κατάλληλες κατηγορίες, διευκολύνοντας την αναζήτηση και την πρόσβαση σε αυτό. Ειδάλλως, οι χρήστες κινδυνεύουν να χαθούν στις τεράστιες ποσότητες οπτικοακουστικού υλικού και να σπαταλούν ώρες, δίχως να κατορθώνουν πάντα να ανακτήσουν τις πληροφορίες που επιθυμούν. Εξαιτίας όμως αφενός του ερασιτεχνικού χαρακτήρα που έχει το διαθέσιμο υλικό και αφετέρου εξαιτίας της τεράστιας ποσότητάς του, ο σχολιασμός του είναι πολύ δύσκολο να γίνει σε βάθος από ανθρώπους. Απαιτούνται πολλές ώρες και ιδιαίτερη εμπειρία προκειμένου να γίνει σωστά και με πληρότητα, καλύπτοντας όλες τις παρούσες, αλλά και πιθανές μελλοντικές ανάγκες.

Για το λόγο αυτό, ιδιαίτερο ερευνητικό βάρος έχει δοθεί στην ανάλυση και στην κατανόηση του περιεχομένου του πολυμεσικού υλικού. Το ερευνητικό αυτό πεδίο είναι πολύ ευρύ και ασχολείται με προβλήματα όπως η αναγνώριση αντικειμένων, εν-

νοιών και γενικά της σημασιολογίας του, η ανάκτηση εικόνων από μεγάλες συλλογές, η ανάλυση του ακουστικού περιεχομένου και άλλα συναφή προβλήματα. Δυστυχώς όμως, η ανάπτυξη αυτών των πεδίων δεν έχει ακόμη καταφέρει να επιλύσει τα σημαντικά προβλήματα που απαιτούνται, προκειμένου να καλύψει γρήγορα και αποδοτικά τις ανάγκες των χρηστών. Σαν αποτέλεσμα αυτού, η συνεχώς αυξανόμενη ροή πολυμεσικού υλικού καταλήγει να δυσχεράνει τελικά την πρόσβαση σε αυτό.

Όλα τα προβλήματα ανάλυσης σε σημασιολογικό επίπεδο ξεκινούν με την εξαγωγή περιγραφών των χαρακτηριστικών γνωρισμάτων του υλικού. Η εξαγωγή αυτή γίνεται με τη χρήση υπολογιστών και προσπαθεί γενικά να εξάγει τα χαρακτηριστικά με τρόπο που να μιμείται αυτόν της ανθρώπινη αντίληψης για τον κόσμο. Τα χαρακτηριστικά αυτά αποκαλούνται συνήθως *χαρακτηριστικά χαμηλού επιπέδου*, καθώς αποτελούν μαθηματικές εκφράσεις και είναι δύσκολο για τον μέσο άνθρωπο να αντιληφθεί τυχόν σημασιολογική πληροφορία που περιέχεται σε αυτά. Ωστόσο, είναι ακόμη πιο δύσκολο να αντιστοιχήσει τα χαρακτηριστικά αυτά σε έννοιες τις οποίες μπορεί να αντιληφθεί. Οι έννοιες αυτές αποκαλούνται συνήθως έννοιες *υψηλού επιπέδου* εκφράζοντας έτσι το γεγονός ότι ένας άνθρωπος μπορεί να τις αντιληφθεί. Το πρόβλημα που περιγράφηκε, ονομάστηκε πολύ εύστοχα ως "σημασιολογικό κενό" (semantic gap) από τους Smeulders et al. [193], και χαρακτηρίζει τη δυσκολία αντιστοιχίας των περιγραφών χαμηλού επιπέδου που μπορούν αυτόματα να εξαχθούν από το οπτικοακουστικό περιεχόμενο του πολυμεσικού υλικού προς τις υψηλού επιπέδου έννοιες που το χαρακτηρίζουν.

1.2 Σκοπός της Εργασίας

Η Εργασία αυτή ασχολείται με συγκεκριμένες περιοχές του πεδίου της ανάλυσης και της κατανόησης του πολυμεσικού υλικού. Πιο συγκεκριμένα, τα προβλήματα που αντιμετωπίζονται ανήκουν στο ερευνητικό πεδίο της ταξινόμησης και της ανίχνευσης εννοιών υψηλού επιπέδου σε εικόνες, της δημιουργίας μικρών περιλήψεων από βίντεο και την ανάκτηση εικόνων από μεγάλες συλλογές, με βάση το οπτικό και σημασιολογικό τους περιεχόμενο. Παρότι δεν είναι άμεσα προφανής η σύνδεση των τριών προβλημάτων, πέρα από το ότι εντάσσονται στο γενικότερο πλάισιο της κατανόησης πολυμεσικού υλικού, αυτά μοιράζονται την ίδια βασική ιδέα ως προς την αναπαράσταση των οπτικών χαρακτηριστικών.

Ξεκινώντας από το πρόβλημα της ταξινόμησης και της ανίχνευσης εννοιών που περιέχονται σε αυτές και προκειμένου να γίνουν κατανοητοί οι στόχοι της παρούσης Εργασίας, πρέπει πρώτα να αποσαφηνιστούν οι κατηγορίες στις οποίες μπορούν να χωριστούν οι διάφορες έννοιες που μπορεί να περιέχει μια εικόνα:

- *Σκηνή*. Έτσι μπορεί να χαρακτηριστεί μια έννοια που γενικά βρίσκεται σε σχετικά μεγάλη απόσταση από τον παρατηρητή και συνήθως χαρακτηρίζει συνολικά μια εικόνα. Παραδείγματα σκηνών είναι ο *εξωτερικός* και ο *εσωτερικός χώρος*, η *παραλία*, η *πόλη* κ.ο.κ.

- *Υλικό*. Έτσι χαρακτηρίζονται γενικά οι έννοιες που δεν έχουν καθορισμένο σχήμα, έχουν όμως καθορισμένα οπτικά χαρακτηριστικά. Παραδείγματα υλικών είναι η *θάλασσα*, η *άμμος*, ο *ουρανός*, η *βλάστηση* κ.ο.κ.

- *Αντικείμενο*. Έτσι χαρακτηρίζονται οι έννοιες που έχουν καθορισμένο σχήμα και αυτό είναι συνήθως αρκετό για να τις περιγράψει. Παραδείγματα αντικειμένων

είναι το *δέντρο*, το *αυτοκίνητο*, ο *άνθρωπος* κ.ο.κ

Η παρούσα Εργασία επικεντρώνεται στις δύο πρώτες κατηγορίες και προτείνει διάφορες μεθοδολογίες για την ταξινόμηση εικόνας με βάση τη σκηνή που απεικονίζει, τον χαρακτηρισμό περιοχών της ως προς τις έννοιες που απεικονίζουν, αλλά και την ανίχνευση εννοιών συνολικά σε εικόνες. Μελετώνται και προτείνονται τεχνικές που χρησιμοποιούν τα συνολικά χαρακτηριστικά των εικόνων, τεχνικές που αποθηκεύουν γνώση όσον αφορά τις έννοιες με μορφή πρωτοτύπων, τεχνικές που χρησιμοποιούν περιγραφές εικόνων με βάση τα τοπικά τους χαρακτηριστικά και τέλος, τεχνικές που αξιοποιούν το εννοιολογικό πλαίσιο των εικόνων, αλλά και των περιοχών των εικόνων, σε μια προσπάθεια να βελτιωθούν τα αποτελέσματα των προηγούμενων τεχνικών.

Το πρόβλημα της εξαγωγής περιλήψεων από βίντεο στοχεύει στη δημιουργία με αυτόματο τρόπο μιας σημαντικά ελαττωμένης αναπαράστασης του βίντεο, ικανή να παρέχει όσο το δυνατόν ακριβέστερη πληροφορία σχετικά με το περιεχόμενό του και με τρόπο που να μπορεί εύκολα να την αντιληφθεί ο μέσος χρήστης, αφιερώνοντας ελάχιστο χρόνο σε σχέση με τη συνολική διάρκεια του βίντεο. Η παρούσα εργασία επικεντρώνεται στην κατασκευή περιλήψεων με έναν επιπλέον σκοπό, αυτόν της αποτελεσματικότερης ανίχνευσης εννοιών υψηλού επιπέδου σε ακολουθίες βίντεο.

Τέλος, όσον αφορά το πρόβλημα της ανάκτησης εικόνων, ο στόχος είναι να επιστρέφονται στο χρήστη οι εικόνες που αυτός επιθυμεί με βάση τα ερωτήματά του. Σκοπός της Εργασίας αυτής είναι να προτείνει έναν τρόπο περιγραφής του οπτικού περιεχομένου των εικόνων, με βάση τον οποίο θα διευκολύνεται η ανάκτησή τους σε σημασιολογικό επίπεδο. Δηλαδή, ο στόχος δεν είναι να επιστρέφονται εικόνες που απλά να μοιάζουν οπτικά με την εικόνα που επέλεξε ο χρήστης, αλλά να έχουν και παρόμοιο σημασιολογικό περιεχόμενο.

1.3 Σύνοψη της εργασίας

Το πρώτο πρόβλημα με το οποίο ασχολείται η παρούσα Εργασία και παρουσιάζεται στο Κεφάλαιο 2 είναι η *ταξινόμηση σκηνής*. Στα πλαίσια του προβλήματος αυτού ανιχνεύονται έννοιες που μπορούν να χαρακτηρίσουν συνολικά το οπτικό περιεχόμενο μιας εικόνας, χωρίς να περιγράφουν τις επιμέρους έννοιες από τις οποίες αποτελείται. Για το σκοπό αυτό προτείνονται και δοκιμάζονται πειραματικά 4 μέθοδοι. Ο στόχος είναι να συνδυαστούν κατάλληλα συγκεκριμένοι περιγραφείς του προτύπου MPEG-7 προκειμένου να επιτύχουν τη βέλτιστη δυνατή ακρίβεια. Διερευνάται ο συνδυασμός των περιγραφέων στο στάδιο *πριν* τον υπολογισμό των αποστάσεων ανάμεσα στις εικόνες, αλλά και *μετά* το ταίριασμα, σε κάθε περιγραφέα μεμονωμένα και πιο συγκεκριμένα στις αποστάσεις που υπολογίστηκαν, δύο κατηγορίες τεχνικών γνωστές ως "πρώιμη" και "όψιμη" συγχώνευση. Επίσης διερευνάται κατά πόσο η χρήση νευροασαφών δικτύων μπορεί να οδηγήσει στην εξαγωγή ασαφών κανόνων που να περιγράφουν το μηχανισμό με τον οποίο γίνεται η ταξινόμηση των εικόνων σε κατηγορίες. Οι προτεινόμενες τεχνικές δοκιμάζονται σε πρόβλημα ταξινόμησης σκηνής για τις κατηγορίες *παραλία* και *πόλη*.

Πέρα όμως από την εφαρμογή των τεχνικών αυτών για την αναγνώριση της σκηνής που απεικονίζει μια εικόνα, στο Κεφάλαιο 3 χρησιμοποιούνται οι ιδέες της συγχώνευσης των περιγραφέων και εφαρμόζονται σε πρόβλημα ταξινόμησης περιοχής με βάση τα οπτικά της χαρακτηριστικά. Η γνώση που χρησιμοποιείται για το σκοπό αυτό αποθηκεύεται με τη μορφή πρωτοτύπων σε μια κατάλληλη δομή *οντολογιών*. Ένα

πρωτότυπο για μια έννοια θεωρείται μια χαμηλού επιπέδου περιγραφή που εξήχθη από μια περιοχή που απεικόνιζε αποκλειστικά αυτή την έννοια. Επίσης, διερευνάται η χρησιμοποίηση απλών χωρικών σχέσεων, με σκοπό να διαχωριστούν έννοιες με παρόμοια οπτικά χαρακτηριστικά, τα οποία όμως βρίσκονται συνήθως σε διαφορετικές χωρικές τοποθεσίες στις εικόνες. Έτσι, τελικά κατασκευάζεται ένα ολοκληρωμένο πλαίσιο που αποσκοπεί στη σημασιολογική ανάλυση πολυμέσων, χρησιμοποιώντας μια πολύπλοκη υποδομή οντολογιών. Καθίσταται σαφές και επιβεβαιώνεται πειραματικά ότι η τεχνική συγχώνευσης που παρουσιάστηκε στο Κεφάλαιο 2 μπορεί να εφαρμοστεί και στην περίπτωση των εννοιών που χαρακτηρίζουν τοπικά μια εικόνα. Η τεχνική που παρουσιάζεται, εφαρμόζεται στο θεματικό πεδίο *Παραλία*.

Ωστόσο, όπως γίνεται σαφές στο Κεφάλαιο 3, ένα σημαντικό πρόβλημα για την εφαρμογή της σημασιολογικής ανάλυσης σε επίπεδο περιοχής με τη χρήση πρωτοτύπων συναντά ιδιαίτερη δυσκολία στο στάδιο κατασκευής της γνώσης, καθώς απαιτεί μια ιδιαίτερα χρονοβόρα διαδικασία. Έχει διαπιστωθεί ότι παρότι το διαθέσιμο οπτικοακουστικό υλικό έχει αυξηθεί ραγδαία τα τελευταία χρόνια, η αύξηση αυτή δεν έχει συνοδευτεί από αντίστοιχη αύξηση του αριθμού των διαθέσιμων σχολιασμένων συλλογών εικόνων. Πολύ λίγες είναι οι συλλογές εικόνων που περιέχουν σχολιασμό ανά περιοχή της εικόνας. Αντιθετα, ο καθολικός σχολιασμός μιας εικόνας είναι μια διαδικασία αισθητά πιο εύκολη και πολύ πιο γρήγορη. Έτσι, το πρόβλημα που προσπαθεί να επιλύσει η τεχνική που παρουσιάζεται στο Κεφάλαιο 4 έχει να κάνει με την ανίχνευση εννοιών όταν είναι διαθέσιμες μόνο καθολικά σχολιασμένες εικόνες. Φυσικά, προσδιορίζεται μόνο η ύπαρξη ή όχι μιας έννοιας σε μια εικόνα. Για το σκοπό αυτό, λοιπόν, προτείνεται μια τεχνική που μπορεί να ενταχθεί στην γενικότερη κατηγορία των μεθόδων που χαρακτηρίζονται ως μοντέλα "bag-of-words". Συνοπτικά, οι εικόνες χωρίζονται σε περιοχές και από κάθε περιοχή εξάγονται περιγραφείς χρώματος και υφής. Στη συνέχεια και με τη βοήθεια ενός οπτικού λεξικού που έχει κατασκευαστεί, οι περιοχές αντιστοιχίζονται σε λέξεις του λεξικού, μέσα από μια διαδικασία κβαντοποίησης. Οι λέξεις αυτές ονομάζονται "τύποι περιοχής". Έτσι εκπαιδεύονται κατάλληλοι ανιχνευτές για κάθε μία από τις έννοιες. Διερευνάται επίσης κατά πόσον η εφαρμογή μιας μεθόδου επεξεργασίας φυσικής γλώσσας που ονομάζεται "Λανθάνουσα Σημασιολογική Ανάλυση" και αποσκοπεί στην εκμετάλλευση των λανθάνουσων σχέσεων που υπάρχουν ανάμεσα στις έννοιες. Οι προτεινόμενες τεχνικές εφαρμόζονται στη σχετική διαδικασία αξιολόγησης του TRECVID καθώς και σε σύνολο εικόνων της συλλογής του Corel.

Στο Κεφάλαιο 5 γίνεται προσπάθεια να αντιμετωπιστεί ένα από τα προβλήματα που διαφάνηκαν στο Κεφάλαιο 4. Στην περίπτωση που η ανίχνευση εννοιών γινεται σε μεμονωμένα χαρακτηριστικά καρέ από βίντεο, ενώ οι παραγόμενοι σχολιασμοί αφορούν το συνολικό πλάνο από το οποίο αυτά εξήχθησαν, ενδέχεται να αποδεικνύονται τελικά ελλιπείς. Αυτό, γιατί είναι προφανές ότι και τα υπόλοιπα καρέ περιέχουν διάφορες έννοιες τις οποίες έχει αγνοήσει η ανάλυση. Προσπαθώντας να αντιμετωπίσει το πρόβλημα αυτό, εφαρμόζεται η τεχνική που παρουσιάζεται στο Κεφάλαιο 5. Ο σκοπός είναι η εξαγωγή ενός μικρού αριθμού καρέ από τις ακολουθίες βίντεο, τα οποία και θα αποκαλούνται "χαρακτηριστικά" και το επιθυμητό είναι να έχουν την ιδιότητα να εκφράζουν όσο το δυνατόν πληρέστερα το οπτικό και το σημασιολογικό περιεχόμενο του βίντεο. Έτσι, οι τεχνικές ανάλυσης θα εφαρμόζονται σε αυτά, παράγοντας πληρέστερα αποτελέσματα και οι χρήστες μεγάλων συλλογών βίντεο θα μπορούν εύκολα και γρήγορα να περιηγηθούν σε αυτές, βλέποντας τις περιλήψεις των βίντεο και επιλέγοντας με βάση αυτές τα βίντεο που πραγματικά τους ενδιαφέρουν.

Στο Κεφάλαιο 6 έγινε μια προσπάθεια να αντιμετωπιστεί το πρόβλημα της ανάκτησης εικόνων, το οποίο σχετίζεται άμεσα με την κατανόηση του πολυμεσικού υλικού και την πρόσβαση σε αυτό. Έτσι, παρουσιάζεται μια τεχνική, η οποία επωφελούμενη από την ιδέα περιγραφής μιας εικόνας με βάση τις περιοχές στις οποίες μπορεί αυτή να χωριστεί και τη βοήθεια ενός οπτικού θησαυρού, προσπαθεί να συνεισφέρει στο πρόβλημα της ανάκτησης εικόνων. Στοχεύει σε μια σημασιολογική ανάλυση, όπου σημασία έχουν πρωτίστως οι έννοιες που περιέχουν οι εικόνες που ανακτήθηκαν και έπειτα η εμφάνισή τους σε σχέση με τις εικόνες των ερωτημάτων.

Οι τεχνικές που παρουσιάζονται στο Κεφάλαιο 4 δε λαμβάνουν καθόλου υπόψη τις σημασιολογικές σχέσεις που υπάρχουν μεταξύ των εννοιών. Έτσι, στο Κεφάλαιο 7 επεκτείνονται, προκειμένου να εκμεταλλευθούν το εννοιολογικό πλαίσιο που χαρακτηρίζει ένα θεματικό πεδίο. Αρχικά, εισάγεται μια προσέγγιση για την αναπαράσταση των σχέσεων που διέπουν τις έννοιες ενός θεματικού πεδίου, που έχει τη μορφή μιας οντολογίας. Έπειτα, επιλέγονται κατάλληλες σχέσεις που μπορούν να εφαρμοστούν σε προβλήματα ανάλυσης εικόνας και επαναορίζονται προκειμένου να συμπεριλάβουν *ασάφεια* στον ορισμό τους και προτείνεται ένας αλγόριθμος που επιδρώντας στις αρχικές εκτιμήσεις για την ύπαρξη μιας έννοιας και λαμβάνοντας υπόψη τις σχέσεις αυτών των εννοιών στο συγκεκριμένο θεματικό πεδίο, επαναϋπολογίζει και βελτιστοποιεί τις αρχικές εκτιμήσεις. Στη συνέχεια, προτείνεται μια μεθοδολογία που αποσκοπεί στη βελτίωση των αποτελέσματων της ανίχνευσης των εννοιών υψηλού επιπέδου, αξιοποιώντας το εννοιολογικό πλαίσιο των τύπων περιοχής. Και στην περίπτωση αυτή οι σχέσεις μεταξύ των τύπων περιοχής κωδικοποιούνται σε μια *οντολογία εννοιολογικού πλαισίου*. Η βασική διαφορά στην προσέγγιση αυτή έγκειται στο γεγονός ότι η προτεινόμενη μεθοδολογία αποσκοπεί στο να βελτιώσει την περιγραφή μιας εικόνας με βάση τους τύπους περιοχής που περιέχονται σε αυτή, αντί να επιδρά στους αρχικούς βαθμούς βεβαιότητας που εξήχθησαν για τις έννοιες υψηλού επιπέδου. Οι δύο αυτές τεχνικές αξιολογούνται στο θεματικό πεδίο *Παραλία*.

Τέλος, στο Κεφάλαιο 8 παρουσιάζονται τα συμπεράσματα που προέκυψαν από την τεχνικές που προτάθηκαν, η συνεισφορά της διατριβής αυτής και προτείνονται πιθανές μελλοντικές επεκτάσεις.

Κεφάλαιο 2

Ταξινόμηση Εικόνων με συνδυασμούς Περιγραφέων MPEG-7

2.1 Εισαγωγή

Έχοντας κάποιος επιλέξει τους οπτικούς περιγραφείς που θα χρησιμοποιήσει σε ένα πρόβλημα ανάλυσης εικόνας, το δεύτερο πρόβλημα που πρέπει να αντιμετωπίσει, όσον αφορά τη χρήση τους σε προβλήματα αναγνώρισης ή ταξινόμησης, έχει να κάνει με τον τρόπο με τον οποίο αυτοί θα συνδυαστούν αποτελεσματικά, έτσι ώστε να βελτιώσουν την ακρίβεια που επιτυγχάνεται σε σχέση με τη μεμονωμένη χρήση τους. Σε αυτό το Κεφάλαιο αρχικά μελετάται το πρόβλημα αυτό στο στάδιο *πριν* τον υπολογισμό των αποστάσεων ανάμεσα στις εικόνες. Μια εναλλακτική προσέγγιση, που επίσης μελετάται, είναι ο συνδυασμός να γίνει *μετά* το ταίριασμα σε κάθε περιγραφέα μεμονωμένα και πιο συγκεκριμένα στις αποστάσεις που υπολογίστηκαν. Οι δύο αυτές περιπτώσεις περιγράφονται και συγκρίνονται από τους Snoek et al. [197] και συναντώνται στη βιβλιογραφία ως "πρώιμη συγχώνευση" (early fusion) και "όψιμη συγχώνευση" (late fusion) αντίστοιχα. Γενικά, οι μέθοδοι πρώιμης συγχώνευσης έχουν μικρότερο υπολογιστικό κόστος ενώ οι μέθοδοι όψιμης συγχώνευσης επιτυγχάνουν καλύτερη ακρίβεια. Ωστόσο κάτι τέτοιο δεν μπορεί να θεωρηθεί ως κανόνας, καθώς υπάρχει άμεση εξάρτηση από την τεχνική που χρησιμοποιείται, καθώς και από την εκάστοτε έννοια που αντιμετωπίζεται.

Καθώς ο ρόλος των χαρακτηριστικών χαμηλού επιπέδου είναι πολύ βασικός σε όλα τα προβλήματα που έχουν να κάνουν με την ανάλυση πολυμεσικού υλικού, κρινεται σκόπιμο να παρουσιαστούν διεξοδικά οι οπτικοί περιγραφείς που χρησιμοποιούνται στην παρούσα Εργασία. Έτσι, στο Κεφάλαιο αυτό και συγκεκριμένα στην Ενότητα 2.4 παρουσιάζονται οι οπτικοί περιγραφείς του προτύπου MPEG-7, οι οποίοι και χρησιμοποιήθηκαν για να εξάγουν χαρακτηριστικά χρώματος, υφής και σχήματος, είτε από ολόκληρες εικόνες, είτε από περιοχές εικόνων.

Στη συνέχεια, προτείνονται τέσσερις διαφορετικές προσεγγίσεις μηχανικής μάθησης και αξιολογούνται στο πρόβλημα του συνδυασμού περιγραφέων MPEG-7 για εφαρμογή τους σε πρόβλημα *ταξινόμησης σκηνής*. Πιο συγκεκριμένα, η πρώτη μέθοδος χρησιμοποιεί την τεχνική της πρώιμης συγχώνευσης περιγραφών και μια Μηχανή Διανυσμάτων Υποστήριξης αναλαμβάνει το κομμάτι της ταξινόμησης. Η δεύτερη μέθοδος

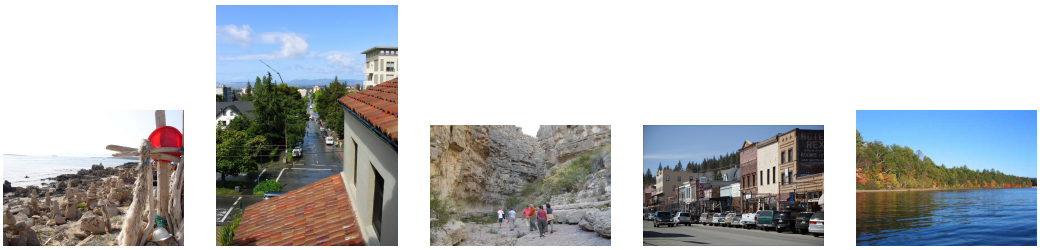
χρησιμοποιεί ένα νευρωνικό δίκτυο για να κάνει την εκτίμηση της απόστασης μεταξύ δύο εικόνων, βασιζόμενο στις συγχωνευμένες περιγραφές τους τόσο αφού εφαρμοστεί η τεχνική της πρώιμης, όσο και της όψιμης συγχώνευσης. Στη συνέχεια, ένα νευροασαφές Falcon-ART δίκτυο και μια Ασαφής Μηχανή Διανυσμάτων Υποστήριξης χρησιμοποιούνται τόσο για την πρώιμη συγχώνευση των περιγραφέων, όσο και για την εξαγωγή γνώσης με τη μορφή ασαφών βασικών κανόνων. Όλα τα παραπάνω δίκτυα εκπαιδεύονται και αξιολογούνται στα ίδια σύνολα εκπαίδευσης και ελέγχου και στο πρόβλημα της ταξινόμησης εικόνων, για τις σκηνές *πόλης*/*παραλίας*.

2.2 Περιγραφή του Προβλήματος

Στα πρώτα χρόνια που η ανίχνευση εννοιών υψηλού επιπέδου άρχισε να μαγνητίζει το ερευνητικό ενδιαφέρον, οι περισσότερες προσπάθειες έδιναν ιδιαίτερη έμφαση σε έννοιες που χαρακτηρίζουν συνολικά τη σημασιολογία της εικόνας, περιγράφουν δηλαδή τη *σκηνή* που περιέχεται σε αυτή. Το πρόβλημα αυτό αφενός ήταν ευκολότερο να αντιμετωπιστεί και αφετέρου τα συμπεράσματα που θα προέκυπταν θα μπορούσαν να οδηγήσουν τη μετέπειτα ανίχνευση τοπικών εννοιών προς σωστότερη κατεύθυνση[1]. Αυτό σημαίνει ότι κάποιες έννοιες που περιέχονται σε μια εικόνα είναι ευκολότερο να ανιχνευθούν, αν είναι εκ των προτέρων γνωστό ποια σκηνή απεικονίζεται συνολικά σε αυτή. Στο πλαίσιο αυτο, λοιπόν, ο σκοπός είναι η ανίχνευση εννοιών που μπορούν να χαρακτηρίσουν συνολικά το οπτικό περιεχόμενο μιας εικόνας, χωρίς να περιγράφουν τις επιμέρους έννοιες από τις οποίες αποτελείται. Για παράδειγμα, μια εικόνα μπορεί να χαρακτηριστεί ως *εσωτερικού χώρου* ή ως *εξωτερικού χώρου*, ανάλογα με την τοποθεσία στην οποία έχει ληφθεί, χωρίς αυτός ο χαρακτηρισμός να περιλαμβάνει ή να υπονοεί τυχόν "τοπικές" έννοιες που περιέχονται, όπως π.χ. *ουρανός*, *θάλασσα* κλπ. Το ίδιο συμβαίνει και για μία ακολουθία βίντεο. Το πρόβλημα αυτό στη σχετική βιβλιογραφία αποκαλείται *ταξινόμηση*, *ανίχνευση* ή *αναγνώριση σκηνής*. Στην παρούσα Εργασία, θα χρησιμοποιείται ο όρος *ταξινόμηση σκηνής*.

Η ταξινόμηση σκηνής ξεκίνησε ως εφαρμογή που αποσκοπούσε στο να διευκολύνει την οργάνωση αρχείων φωτογραφιών, την αποτελεσματικότερη δεικτοδότηση, την επεξεργασία εικόνας και άλλα παρόμοια προβλήματα. Στη συνέχεια όμως, όταν ξεκίνησε η ανάπτυξη τεχνικών ανίχνευσης εννοιών που δε χαρακτηρίζουν συνολικά μια εικόνα αλλά τοπικά, η γνώση για τη σκηνή που απεικονίζεται στην εικόνα φάνηκε σε πολλές μεθόδους ιδιαίτερα χρήσιμη. Αν και αρχικά φαίνεται αρκετά εύκολο πρόβλημα, τελικά ακόμη και πολύ προηγμένες τεχνικές συχνά αποτυγχάνουν. Αυτό οφείλεται στο γεγονός ότι μια σκηνή είναι πολύ δύσκολο να περιγραφεί είτε ποιοτικά, είτε με τη χρήση κάποιων τυποποιημένων περιγραφέων. Η ανομοιογένεια που παρατηρείται ακόμη και σε έναν μικρό αριθμό από εικόνες που ανήκουν στην ίδια σκηνή καθιστά ιδιαίτερα δύσκολη τη μοντελοποίησή τους. Πέρα από αυτό, πολύ συχνά υπεισέρχονται πολλά φαινόμενα, όπως διαφορές στη φωτεινότητα, την κλίμακα και το ζουμ. Έτσι, τα χαρακτηριστικά μιας νυχτερινής σκηνής *εξωτερικού χώρου* που απεικονίζει μια *πόλη* είναι πολύ διαφορετικά από αυτά μιας ηλιόλουστης σκηνής *παραλίας*. Η περιγραφή που εξάγεται από μια σκηνή σε μεγάλη κλίμακα είναι πιο ακριβής και σίγουρα αρκετά διαφορετική από την ίδια σκηνή σε μικρή κλίμακα. Μία εικόνα που απεικονίζει ένα *δάσος*, και μία που απεικονίζει ένα *δέντρο* από την ίδια

[1]Τέτοιες τεχνικές αξιοποιούν το εννοιολογικό πλαίσιο ενός θεματικού πεδίου και αποτελούν αντικείμενο έρευνας του Κεφαλαίου 7



Σχήμα 2.1: *Μερικές εικόνες που αντιστοιχούν στην έννοια εξωτερικός χώρος. Η διαφορά στα οπτικά τους χαρακτηριστικά είναι έντονη.*

εικόνα, έχουν διαφορετικά καθολικά χαρακτηριστικά, ωστόσο και οι δύο μπορούν να χαρακτηριστούν ως *εξωτερικού χώρου* ή *δάσος* ή *βλάστηση* και ούτω καθεξής.

Επιπρόσθετα, σημαντικό ρόλο παίζουν και διάφοροι ανθρώπινοι παράγοντες, όπως για παράδειγμα λανθασμένοι χειρισμοί της κάμερας που ενδέχεται να οδηγήσουν σε υποφωτισμένη εικόνα, η οποία θα δείχνει σκοτεινότερη απ'ό,τι πραγματικά είναι η σκηνή που απεικονίζει, σε θολή εικόνα λόγω αστάθειας ή λόγω κακής εστίασης της κάμερας ώστε να αλλοιώνονται τα χαρακτηριστικά υφής κλπ. Τέλος, πολύ σημαντικό ρόλο παίζουν και διάφορα "εμπόδια" που βρίσκονται στο προσκήνιο (foreground) και κρύβουν το παρασκήνιο (background) της εικόνας που όμως είναι αυτό που συνήθως απεικονίζει ή περιέχει τα χαρακτηριστικά της σκηνής, όπως για παράδειγμα ένας *άνθρωπος* μπροστά από μια εικόνα *παραλίας*.

Στο Σχήμα 2.1 παρουσιάζεται ένα μικρό σύνολο από εικόνες *εξωτερικού χώρου*. Είναι φανερό ότι τα οπτικά χαρακτηριστικά της έννοιας αυτής παρουσιάζουν μεγάλη ανομοιογένεια. Η ποικιλία στα χρώματα και τις υφές είναι πολύ μεγάλη, καθώς συνήθως οι έννοιες που εμφανίζονται σε μια σκηνή είναι ποικίλες και πολύ διαφορετικές μεταξύ τους. Ακόμη και για έννοιες που θα περίμενε κάποιος να είναι απλές, τα φαινόμενα που αναφέρθηκαν προηγουμένως δημιουργούν μεγάλα προβλήματα στην ταξινόμηση. Για παράδειγμα, στο Σχήμα 2.2 παρουσιάζονται σκηνές που αντιστοιχούν στην έννοια *ηλιοβασίλεμα*. Η μεταβολή των χρωμάτων της εικόνας είναι σημαντική, καθώς αυτά γίνονται εντονότερα όσο ο ήλιος πλησιάζει στον ορίζοντα και στη συνέχεια ξεθωριάζουν, καθώς η φωτεινότητα της εικόνας χαμηλώνει, όπως παρατήρησαν οι Boutell et al. [29]. Η ίδια εικόνα στο Σχήμα 2.3 φαίνεται υπερφωτισμένη και υποφωτισμένη, κάτι που οδηγεί σε μεγάλη διαφοροποίηση των οπτικών της χαρακτηριστικών. Έτσι, ένας ταξινομητής σκηνής που θα εκπαιδευτεί με βάση τα χρωματικά χαρακτηριστικά κανονικά φωτισμένων σκηνών, όπως αυτή της μεσαίας εικόνας του Σχήματος 2.3, πιθανώς να αποτύχει να ταξινομήσει σωστά αυτές που έχουν διαφορετικό φωτισμό. Υπάρχουν, ωστόσο κι άλλοι παράγοντες που δυσχεραίνουν το πρόβλημα της σωστής ταξινόμησης, όπως η μερική κάλυψη των εννοιών/περιοχών μιας εικόνας που περιέχουν τα χαρακτηριστικά εκείνα που διαχωρίζουν μια σκηνή από τις υπόλοιπες. Έτσι, στο Σχήμα 2.4 οι *άνθρωποι* που ποζάρουν στο προσκήνιο αλλοιώνουν τα οπτικά χαρακτηριστικά της σκηνής, καθώς με την παρουσία τους κρύβουν το μεγαλύτερο μέρος του παρασκηνίου με βάση το οποίο χαρακτηρίζεται η σκηνή, αλλά ταυτόχρονα προσθέτουν χαρακτηριστικά χρώματος και υφής που ενδέχεται να οδηγήσουν καλά εκπαιδευμένους ταξινομητές σε αποτυχία.

Ψυχοφυσικές και ψυχολογικές μελέτες που έγιναν από τον Biederman [13], [14], αλλά και τους Oliva και Schyns [158], καταλήγουν ότι η ανθρώπινη αντίληψη μπορεί να αναγνωρίσει τη σκηνή που απεικονίζεται σε μια εικόνα χωρίς να είναι αναγκαίο να αναγνωρίσει τις επιμέρους έννοιες οι οποίες περιέχονται μέσα σε αυτή. Ο άνθρωπος,

Σχήμα 2.2: *Μερικές εικόνες που αντιστοιχούν στην έννοια ηλιοβασίλεμα. Η διαφορά στα χρωματικά τους χαρακτηριστικά είναι σημαντική.*

Σχήμα 2.3: *Η ίδια εικόνα εξωτερικού και εσωτερικού χώρου κανονικά φωτισμένη, υπερφωτισμένη και υποφωτισμένη. Τα οπτικά χαρακτηριστικά χρώματος διαφέρουν αισθητά.*

δηλαδή μπορεί για παράδειγμα να αναγνωρίσει τη σκηνή που απεικονίζει μια χαμηλής ευκρίνειας εκδοχή μιας φωτογραφίας, στην οποία τα επιμέρους αντικείμενα δεν είναι αναγνωρίσιμα. Στο Σχήμα 2.5 απεικονίζεται μια εικόνα *παραλίας* και μια εκδοχή της με μειωμένη ευκρίνεια. Είναι φανερό πως ο μέσος άνθρωπος μπορεί εύκολα να αναγνωρίσει τη σκηνή αυτή, παρότι οι έννοιες που τη συνθέτουν (*θάλασσα*, *άμμος*, *ομπρέλα*, *ουρανός*) είναι δυσδιάκριτες και αν του παρουσιάζονταν μεμονωμένα τα μέρη της εικόνας που τις περικλείουν, δε θα ήταν δυνατόν να τις αναγνωρίσει. Κάτι αντίστοιχο φαίνεται και στο Σχήμα 2.6, όπου έχουν δημιουργηθεί τεχνητές σκηνές που αντιστοιχούν σε *πόλη* και *γραφείο*. Τα αντικείμενα που απαρτίζουν τις σκηνές αυτές έχουν αντικατασταθεί απο βασικά γεωμετρικά στερεά με τα οποία εμφανίζουν παρόμοιες οπτικές ιδιότητες. Ωστόσο, η σκηνή παραμένει εύκολα αναγνωρίσιμη από έναν άνθρωπο. Επίσης, έχει διαπιστωθεί ότι όταν ένας άνθρωπος βλέπει μια σκηνή για σύντομο χρονικό διάστημα, μπορεί να εξάγει αρκετή πληροφορία για την ορθή αναγνώρισή της, χωρίς κατ'ανάγκη να προλάβει να εντοπίσει όλα τα αντικείμενα που περιέχονται, αντιλαμβανόμενος μόνο τις λειτουργικές της ιδιότητες και τις ιδιότητες εκείνες που συμβάλλουν στην κατηγοριοποίησή της.

Οι προσεγγίσεις που συναντώνται στον ερευνητικό χώρο της ταξινόμησης σκηνής, συνήθως ξεκινούν με την εξαγωγή οπτικής πληροφορίας με τη μορφή κάποιου οπτικού περιγραφέα σε επίπεδο εικονοστοιχείου, από περιοχές της εικόνας, συνολικά από ολόκληρη την εικόνα, ή ακόμη και από έναν αριθμό χαρακτηριστικών καρέ από ακολουθίες βίντεο. Για την επιλογή του κατάλληλου τρόπου περιγραφής, συνήθως υιοθετείται μια προσέγγιση που χρησιμοποιεί κάποιον αλγόριθμο μηχανικής μάθησης, όπως για παράδειγμα τα νευρωνικά δίκτυα, με σκοπό μέσα από την εκπαίδευση να προκύψει μια αντιστοιχία μεταξύ των χαμηλού επιπέδου περιγραφών που εξάγονται από τις εικόνες και τις έννοιες υψηλού επιπέδου που θα ανιχνευθούν, γεφυρώνοντας έτσι το "Σημασιολογικό Κενό" που αναφέρθηκε στο Κεφάλαιο 1. Οι τεχνικές αυτές παρουσιάζονται αναλυτικά στην Ενότητα 2.3.

Σχήμα 2.4: *Η παρουσία ανθρώπων στο προσκήνιο της εικόνας αλλοιώνει τα συνολικά οπτικά χαρακτηριστικά της σκηνής, η οποία εμφανίζεται στο παρασκήνιο.*

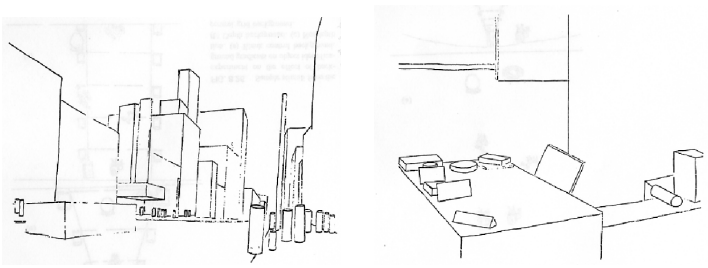
Σχήμα 2.5: *Μια εικόνα που απεικονίζει μια σκηνή παραλίας και μια χαμηλής ευκρίνειας (θολωμένη) εκδοχή της, όπου δε διακρίνονται οι επιμέρους έννοιες, ωστόσο η σκηνή παραμένει αναγνωρίσιμη.*

2.3 Σχετικές Εργασίες

Ξεκινώντας με το πιο συνηθισμένο πρόβλημα, αυτό της ταξινόμησης *εσωτερικού/εξωτερικού χώρου*, στην εργασία των Szummer και Picard [206] καταδεικνύεται ότι για το φαινομενικά απλό αυτό πρόβλημα μια χρωματική περιγραφή με τη μορφή ενός απλού ιστογράμματος δεν είναι επαρκής. Σε συνδυασμό, όμως, με την αξιοποίηση της πληροφορίας που παρέχει η υφή, καθώς και τα χαρακτηριστικά συχνότητας σε έναν DCT χώρο, οδηγεί σε μια περιγραφή που μπορεί να διαχωρίσει τις δύο κατηγορίες. Οι περιγραφές χαμηλού επιπέδου δεν εξάγονται μόνο συνολικά, αλλά και από δομικά στοιχεία (blocks) της εικόνας. Έτσι, γίνεται προσπάθεια για την εκμετάλλευση των χωρικών σχέσεων μεταξύ των δομικών στοιχείων. Για παράδειγμα, συνήθως οι εικόνες *εξωτερικού χώρου* έχουν στο επάνω μέρος τους *ουρανό*. Αν και δεν γίνεται αναζήτηση της έννοιας αυτής, η πληροφορία ότι τα δομικά στοιχεία της εικόνας που βρισκόνται στο επάνω μέρος της είναι *μπλε* βοηθάει για την ορθότερη ταξινόμηση. Για την ταξινόμηση όταν χρησιμοποιούνται συνολικά χαρακτηριστικά, χρησιμοποιείται ένας ταξινομητής K-πλησιέστερων γειτόνων, ενώ στην περίπτωση των τοπικών χαρακτηριστικών χρησιμοποείται ένα νευρωνικό δίκτυο.

Στην εργασία των Luo και Savakis [127] γίνεται ένα επιπλέον βήμα. Με τη χρήση ενός μπεϋσιανού δικτύου γίνεται ενσωμάτωση των χαρακτηριστικών χαμηλού επιπέδου χρώματος και υφής και αυτών του "μεσαίου"[2] επιπέδου τα οποία σύμφωνα με τους συγγραφείς είναι έννοιες όπως *θάλασσα*, *ουρανός* κλπ, που δε χαρακτηρίζουν συνολικά την εικόνα. Οι έννοιες αυτές χρησιμοποιούνται είτε κατευθείαν από το σύνολο δεδομένης αλήθειας (ground truth), ακόμη και για τις εικόνες ελέγχου, είτε από εκπαίδευση διαφορετικών ταξινομητών. Οι Serrano et al. [182], αντιμετωπίζουν

[2]Στην παρούσα εργασία, ως χαρακτηριστικά *μεσαίου* επιπέδου θα αναφέρονται αυτά που βρίσκονται ξεκάθαρα ανάμεσα στα επίπεδα των περιγραφών και των εννοιών. Η χρήση τέτοιου τύπου χαρακτηριστικών αποτελεί αντικείμενο ερευνας του Κεφαλάιου 4



Σχήμα 2.6: *Τεχνητές σκηνές που αντιστοιχούν σε σκηνές πόλης και γραφείου. Τα αντικείμενα έχουν αντικατασταθεί απο βασικά γεωμετρικά στερεά με τα οποία εμφανίζουν παρόμοιες ιδιότητες. Το σχήμα κατασκευάστηκε από τον Biederman [13] για να γίνει κατανοητό ότι ο άνθρωπος μπορεί εύκολα να αντιληφθεί τη σκηνή που περιγράφεται σε μια εικόνα, χωρίς να τη δει ποτέ, αλλά αντί αυτής να δει μια τεχνητή αναπάραστασή της.*

με παρόμοιο τρόπο το ίδιο πρόβλημα, ωστόσο εστιάζουν στη μειώση του υπολογιστικου κόστους, κάτι που επιτυγχάνουν υιοθετώντας ταξινομητές βασισμένους σε SVM.

Ένα αρκετά δημοφιλές πρόβλημα ταξινόμησης σκηνής αποτελεί αυτό της ταξινόμησης μεταξύ *πόλης* και *τοπίου*. Στο πρόβλημα αυτό φαίνεται πολύ καθαρά ότι προσεκτικές παρατηρήσεις των εικόνων προς ταξινόμηση και κατάλληλη επιλογή περιγραφέων μπορεί να οδηγήσει στη λύση του. Οι εικόνες που ανήκουν στην κατηγορία *πόλη* εμφανίζουν συνήθως έντονες οριζόντιες και κάθετες ακμές και γενικά, κατευθυντικότητα στις ακμές τους, οι οποίες οφείλονται κυρίως στα *κτίρια* και άλλες ανθρώπινες κατασκευές. Το αντίθετο συμβαίνει με τις εικόνες *τοπίου*, όπου γενικά οι ακμές είναι λιγότερο κατευθυντικές και απουσιάζουν οι έντονες οριζόντιες και κάθετες ακμές. Η ιδέα αυτή εφαρμόστηκε από τους Vailaya et al. [219], όπου επιβεβαιώθηκε πειραματικά ότι οι χρωματικές περιγραφές έχουν μικρή διαχωριστική ικανότητα στο συγκεκριμένο πρόβλημα. Αυτό επιβεβαιώνεται και διαισθητικά, γιατί για παράδειγμα, μια εικόνα *πόλης* και μια εικόνα *τοπίου* μπορεί να περιέχουν και οι δύο έννοιες όπως *ουρανός*, *βλάστηση*, *δρόμος*, που οδηγούν σε παρόμοια οπτικά χαρακτηριστικά χρώματος και συνεπώς σε παρόμοιες περιγραφές. Αντίθετα, η ύπαρξη ενός κτιρίου στην εικόνα *πόλης* προσθέτει σχεδόν σε κάθε περίπτωση έντονες οριζόντιες και κάθετες ακμές και άρα μια περιγραφή υφής προσφέρει την απαραίτητη διαχωρισιμότητα ανάμεσα στις δύο κατηγορίες. Η ίδια ιδέα ακολουθείται και από τους Le Borgne και O'Connor [110], όπου για την αναπαράσταση των ακμών μιας εικόνας υιοθετείται ο μετασχηματισμός Ridgelet [35]. Με βάση τις περιγραφές που προκύπτουν, εκπαιδεύονται κατάλληλοι ταξινομητές και εφαρμόζονται σε πρόβλημα ταξινόμησης *εσωτερικού χώρου*, *πόλης*, *ανοικτού* και *κλειστού εξωτερικού χώρου* (οι οποίοι ορίζονται ως "με" και "χωρίς ορίζοντα", αντίστοιχα). Οι Gorkani και Pickard [77] αντιμετωπίζουν το δυαδικό πρόβλημα ταξινόμησης *πόλη*/*προάστια*. Παρότι οι δύο κατηγορίες φαίνονται αρκετά όμοιες, το πρόβλημα αντιμετωπίζεται ικανοποιητικά με τη χρήση του προσανατολισμού των ακμών και καταλήγουν στο συμπέρασμα ότι το πλήθος των ακμών, μπορεί από μόνο του να οδηγήσει στη σωστή ταξινόμηση. Αυτό συμβαίνει, καθώς οι εικόνες *πόλης* εμφανίζουν μεγάλο αριθμό οριζόντιων και κάθετων ακμών, σε σχέση με τις εικόνες που χαρακτηρίζονται ως *προάστια*.

Οι Oliva και Torralba [159], [160] εξάγουν μια χαμηλής διάστασης περιγραφή της σκηνής, αποκαλώντας την "χωρικό φάκελο" (spatial envelope) ή "σχήμα της σκηνής", ή "ουσία της σκηνής" (gist of a scene). Ξεκινούν με την παρατήρηση ότι εικόνες που απεικονίζουν την ίδια σκηνή μοιράζονται μια παρόμοια και επαναλήψι-

μη χωρική δομή, η οποία μπορεί να εξαχθεί χωρίς κατάτμηση. Ο χωρικός φάκελος ενός αντικειμένου αποτελείται από ένα σύνολο από όρια. Για παράδειγμα, μια σκηνή *αυτοκινητόδρομου* θα είναι ένα μεγάλο επίπεδο, το οποίο θα συρρικνώνεται καθώς πλησιάζει τη γραμμή του ορίζοντα. Το σχήμα της σκηνής εξαρτάται από τη δομή της και άρα μπορεί να μοντελοποιηθεί και να χρησιμοποιηθεί για την ταξινόμησή της, η οποία και γίνεται σε 8 κατηγορίες.

Οι Vogel και Schiele [225], [224] εισάγουν μια τεχνική που σκοπό έχει την κατασκευή σημασιολογικών μοντέλων από σκηνές *εξωτερικού χώρου*. Στη μέθοδό τους εισάγουν το βήμα της σημασιολογικής μοντελοποίησης (semantic modelling), δηλαδή πρώτα εντοπίζουν "τοπικές" έννοιες και ακολούθως με βάση αυτές καταλήγουν στο συμπέρασμα για τη σκηνή που απεικονίζεται. Για να το πετύχουν αυτό, η εικόνα χωρίζεται σε ορθογώνιες περιοχές με χρήση ενός ορθογωνίου πλέγματος (grid). Κάθε μία από αυτές αντιστοιχείται σε μια έννοια και τελικά η εικόνα περιγράφεται από ένα ιστόγραμμα που δείχνει σε τι (χωρικό) ποσοστό περιέχεται κάθε έννοια στην εικόνα και τελικά, με βάση αυτό γίνεται η ταξινόμηση.

Προκειμένου να ξεπεραστεί το πρόβλημα της ποικιλίας των οπτικών χαρακτηριστικών για μία έννοια, οι Boutell et al. [29] προτείνουν μια τεχνική η οποία αφενός επανασυνθέτει μια εικόνα με σκοπό τα οπτικά της χαρακτηριστικά να πλησιάσουν αυτά των παραδειγμάτων που είναι διαθέσιμα και περιγράφουν τις υπό ανίχνευση έννοιες και αφετέρου αποσκοπεί στο να μεγαλώσει τον αριθμό των παραδειγμάτων, ώστε τελικά να δημιουργηθεί ένα σύνολο δεδομένων με μεγαλύτερη ποικιλία και περισσότερα παραδείγματα. Για να γίνει αυτόματα αυτή η διαδικασία, δηλαδή χωρίς επίβλεψη, οι μέθοδοι που επιλέγονται είναι η αποκοπή του κέντρου της εικόνας, μιας και συνήθως στο κέντρο περιέχεται το θέμα και άρα η απαραίτητη οπτική πληροφορία για τη σκηνή που απεικονίζεται, η συμμετρική εικόνα ως προς τον κάθετο άξονα και η μετατόπιση των χρωμάτων της εικόνας προκειμένου να δημιουργηθούν εικόνες με παραπλήσια οπτικά χαρακτηριστικά και να καλύψουν περισσότερες περιπτώσεις σαν κι αυτές που περιέχονται στη φύση.

Πολυάριθμες τεχνικές έχουν παρουσιαστεί τα τελευταία χρόνια και αποτελούν γενικά παραλλαγές των μεθόδων που ήδη αναφέρθηκαν. Οι Dorado et al. [59] προτείνουν ένα σύστημα που ταξινομεί σκηνές προσπαθώντας να "οδηγήσει" τους περιγραφείς προς τα παραδείγματα που αποτελούν τη γνώση και χρησιμοποιεί την τεχνική συσχετιστικής ανάδρασης για να συνδυάσει εκπαίδευση με και χωρίς την αλληλεπίδραση με το χρήστη. Οι Bosch et al. [19] προτείνουν τη χρήση ενός ταξινομητή που μαθαίνει τις σκηνές και την κατανομή των χαρακτηριστικών τους σε μη σχολιασμένο σύνολο εκπαίδευσης χρησιμοποιώντας την τεχνική της πιθανοτικής Λανθάνουσας Σημασιολογικής Ανάλυσης (probabilistic Latent Semantic Analysis - pLSA) και στη συνέχεια χρησιμοποιεί την κατανομή αυτή στο σύνολο ελέγχου.

Το χρώμα χρησιμοποιείται κατ'εξοχήν στη βιβλιογραφία σαν το απλούστερο (μιας και η εξαγωγή του είναι η απλούστερη) ανάμεσα σε όλα τα οπτικά χαρακτηριστικά. Οι Szummer και Pickard [206] καθώς και οι Serrano et al. [183], [182] χρησιμοποιούν ιστογράμματα χρώματος στους χρωματικούς χώρους Ohta και LST αντίστοιχα. Οι Meine et al. [136] χρησιμοποιούν στατιστικά χαρακτηριστικά πρώτης τάξης που βασίζονται σε ιστογράμματα χρώματος και κλίμακας του γκρι. Μια αρκετά εκτεταμένη μελέτη όσον αφορά τη χρήση των χρωματικών περιγραφέων όπως χρωματικά ιστογράμματα, ιστογράμματα συσχέτισης χρώματος, περιγραφείς χρώματος του MPEG-7 κ.α. έγινε από τους Qiu et al. [174]. Το συμπέρασμα που προέκυψε είναι ότι κανένας από τους περιγραφείς δεν δουλεύει ικανοποιητικά σε όλα τα υπό εξέταση σύνολα

δεδομένων, καθώς και ότι σε όλα τα ιστογράμματα υπάρχει πολλή πλεονάζουσα πληροφορία που μπορεί να αφαιρεθεί και να οδηγήσει σε απλούστερη περιγραφή.

Όσον αφορά τα χαρακτηριστικά υφής, οι Serrano et al. [183], [182] χρησιμοποιούν wavelets, και καταλήγουν στο συμπέρασμα ότι αυτά αποδίδουν καλύτερα από άλλους περιγραφείς υφής. Το πρώτο από τα χαρακτηριστικά καθορίζεται από φιλτράρισμα των συντελεστών χαμηλής συχνότητας με ένα λαπλασιανό φίλτρο, ενώ τα υπόλοιπα χαρακτηριστικά καθορίζονται από την ενέργεια κάτω ζώνης για όλες τις συνιστώσες του wavelet. Οι Guerin-Dugue και Oliva [78] χρησιμοποιούν τοπικές δεσπόζουσες κατανομές προσανατολισμου για όλη την εικόνα. Μέσα από τα πειράματά τους διαπιστώνουν ότι οι εικόνες *εσωτερικού χώρου* έχουν πιο "ομαλούς" προσανατολισμούς στις ακμές σε σχέσεις με τις εικόνες *εξωτερικού χώρου*. Ο Fitzpatrick [69] παρατήρησε ότι στις εικόνες *εσωτερικού χώρου* ο βαθμός που αλλάζει η φωτεινότητα ως προς τον κάθετο άξονα είναι χαμηλός και το ποσοστό των κάθετων ακμών είναι υψηλό, αντίθετα από ότι συμβαίνει στις εικόνες *εξωτερικού χώρου*. Οι Payne και Singh [167] αλλά και οι Traherne και Singh [214] διαπίστωσαν ότι οι εικόνες *εσωτερικού χώρου* (και γενικά οι "συνθετικές" εικόνες) έχουν περισσότερες ευθείες ακμές από τις εικόνες *εξωτερικού χώρου* (και γενικά οι "οργανικές" εικόνες).

Η χρήση της σημασιολογικής πληροφορίας έχει επίσης καθιερωθεί στο πρόβλημα ταξινόμησης σκηνής. Τα χαρακτηριστικά χρώματος και υφής συνδυάζονται με σημασιολογική γνώση (όπως για παράδειγμα ότι ο *ουρανός* βρίσκεται στην κορυφή της εικόνας) και προσπαθούν με αυτόν τον τρόπο να βελτιώσουν τα αποτελέσματα της ταξινόμησης. Για παράδειγμα, οι Serrano et al. [182] αλλά και οι Luo και Savakis [127], εκμεταλλεύονται την ύπαρξη της πληροφορίας αυτής με τη χρήση ενός μπεϋσιανού δικτύου. Επίσης, η χρήση σχολιασμών της εικόνας με τη μορφή κειμένου σε συνδυασμό με τα χαμηλού επιπέδου χαρακτηριστικά έχει αξιοποιηθεί για την ταξινόμηση της εικόνας, όπως για παράδειγμα από τους Paek et al. [162], οι οποίοι ενσωμάτωσαν τεχνικές ανάλυσης κειμένου (TF/IDF) στη διαδικασία της ταξινόμησης.

2.4 Εξαγωγή Χαρακτηριστικών Χαμηλού Επιπέδου

Για την εξαγωγή των χαρακτηριστικών χαμηλού επιπέδου επιλέχθηκαν οι οπτικοί περιγραφείς του προτύπου MPEG-7 [43]. Το πρότυπο αυτό, το οποίο επισήμως ονομάζεται "Διεπαφή Περιγραφής Πολυμεσικού Περιεχομένου" (Multimedia Content Description Interface) αποτελεί ένα από τα νεώτερα ISO/IEC πρότυπα που έχουν αναπτυχθεί από την ομάδα του MPEG και αποσκοπεί στην *περιγραφή* του πολυμεσικού υλικού. Δημιουργεί δομημένες περιγραφές του πολυμεσικού υλικού με τέτοιο τρόπο, ώστε οι χρήστες να μπορούν να πραγμοτοποιήσουν διαδικασίες όπως αναζήτηση, ανάκτηση και φυλλομέτρησή του, αποτελεσματικά και αποδοτικά. Ενώ τα προηγούμενα πρότυπα του MPEG (MPEG-1,2,4) [30],[21],[185] εστίαζαν στην κωδικοποίηση και αναπαράσταση του οπτικοακουστικού υλικού, το MPEG-7 εστιάζει αποκλειστικά στην περιγραφή του. Μπορεί να εφαρμοστεί σε πολυτροπικό υλικό, το οποίο μπορεί να αποτελείται από εικόνα, βίντεο, ήχο, ομιλία, γραφικά καθώς και όλους τους συνδυασμούς τους. Σε ένα τέτοιο περιβάλλον, πλούσιο σε πληροφορία, υπάρχει η ανάγκη για την ανάπτυξη εργαλείων και συστημάτων για δεικτοδότηση, αναζήτηση, φιλτράρισμα και διαχείριση του αποθηκευμένου πολυμεσικού υλικού. Το MPEG-7 αποσκοπεί να επιτύχει τα μέγιστα, όσον αφορά τη διαδραστικότητα μεταξύ

των προαναφερθείσων εφαρμογών, ορίζοντας τη σύνταξη και τη σημασιολογία διαφόρων εργαλείων περιγραφής. Για την παρούσα Εργασία, ιδιαίτερο ενδιαφέρον έχουν οι χαμηλού επιπέδου οπτικοί περιγραφείς που ορίζει το MPEG-7, οι οποίοι χρησιμοποιούνται για την εξαγωγή περιγραφών χρώματος, υφής και σχήματος από εικόνες, ή από περιοχές εικόνων.

2.4.1 Το Οπτικό Μέρος του προτύπου MPEG-7

Στην Ενότητα αυτή περιγράφονται αναλυτικά οι οπτικοί περιγραφείς του προτύπου MPEG-7. Οι περιγραφείς αυτοί χρησιμοποιούνται σε όλα τα ακόλουθα Κεφάλαια για την εξαγωγή των χαρακτηριστικών χαμηλού επιπέδου από εικόνες και χαρακτηριστικά καρέ ακολουθιών βίντεο. Πιο συγκεκριμένα, η Ενότητα αυτή παρουσιάζει αναλυτικά τους περιγραφείς του οπτικού μέρους του προτύπου που εξάγουν χαμηλού επιπέδου περιγραφές του *Χρώματος*, της *Υφής* και του *Σχήματος*. Το οπτικό μέρος του προτύπου περιλαμβάνει και περιγραφείς *Κίνησης*, οι οποίοι δεν περιγράφονται, καθώς δε χρησιμοποιούνται στην παρούσα Εργασία. Παρουσιάζονται συνοπτικά οι αλγόριθμοι εξαγωγής του κάθε περιγραφέα, τα μέτρα που χρησιμοποιούνται για τον υπολογισμό της απόστασης μεταξύ δύο περιγραφών και και τα ποιοτικά τους χαρακτηριστικά τα οποία βοηθούν στην κατανοήση του ποιος περιγραφέας είναι καταλληλότερος για κάθε εφαρμογή.

2.4.2 Περιγραφείς Χρώματος

Το πρότυπο MPEG-7 περιλαμβάνει 4 περιγραφείς χρώματος [131] και πιο συγκεκριμένα, τον Περιγραφέα Κύριων Χρωμάτων, τον Κλιμακωτό Περιγραφέα Χρώματος, τον Περιγραφέα Δομής Χρώματος και τον Περιγραφέα Διάταξης Χρώματος. Κάθε ένας από τους περιγραφείς αυτούς έχει ιδιαίτερα χαρακτηριστικά και έτσι οι πιθανοί χρήστες τους μπορούν με βάση αυτά να επιλέξουν τον κατάλληλο περιγραφέα για την εφαρμογή που επιθυμούν να τον χρησιμοποιήσουν.

2.4.2.1 Περιγραφέας Κύριων Χρωμάτων

Μια αποτελεσματική και ταυτόχρονα συμπαγής περιγραφή μιας εικόνας ή μιας περιοχής της μπορεί να δοθεί από ένα σύνολο από τα αντιπροσωπευτικά της χρώματα. Τέτοιες περιγραφές χρησιμοποιούνται ευρύτατα σε εφαρμογές όπως είναι η ανάκτηση με βάση το χρώμα, όταν πρόκειται για μεγάλες βάσεις δεδομένων, όπου η συμπαγής περιγραφή είναι επιθυμητή για λόγους αποθήκευσης των περιγραφών, αλλά και ταχύτητας. Όπως συμβαίνει στην πλειοψηφία των περιπτώσεων, τα χρώματα σε μια περιοχή μιας εικόνας είναι συγκεντρωμένα γύρω από έναν μικρό αριθμό αντιπροσωπευτικών χρωμάτων. Για να εκμεταλλευθεί αυτή την παρατήρηση, ο Περιγραφέας Κύριων Χρωμάτων (Dominant Color Descriptor) αποτελείται από τα κύρια (αντιπροσωπευτικά) χρώματα, τα ποσοστά τους στην περιοχή, τη χωρική τους συνοχή (spatial coherency), και τη διακύμανση τους (variance). Τα κύρια χρώματα αναπαρίστανται στον τρισδιάστατο χώρο χρώματος και αποφεύγονται έτσι τα υψηλής διάστασης προβλήματα δεικτοδότησης (indexing) που συνδέονται με τα παραδοσιακά ιστογράμματα χρώματος.

Για τον υπολογισμό του περιγραφέα, τα χρώματα που είναι παρόντα σε μια εικόνα ή μια περιοχή ενδιαφέροντος αρχικά συσταδοποιούνται, κάτι που οδηγεί σε

έναν μικρό αριθμό χρωμάτων, τα οποία αντιστοιχούν στα κέντρα των συστάδων που σχηματίστηκαν. Στη συνέχεια υπολογίζονται τα ποσοστά αυτών των χρωμάτων και (προαιρετικά) οι διακυμάνσεις τους. Επίσης υπολογίζεται μια τιμή χωρικής συνοχής που βοηθά στην διαφοροποίηση μεταξύ των μεγάλων ενιαίων περιοχών χρώματος εναντίον των χρωμάτων που είναι εξαπλωμένα σε όλη την εικόνα ή την περιοχή ενδιαφέροντος. Ο αριθμός των κυρίων χρωμάτων μπορεί να ποικίλει από εικόνα/περιοχή σε εικόνα/περιοχή. Ένας μέγιστος αριθμός οκτώ κύριων χρωμάτων μπορεί να εξαχθεί για το σχηματισμό του περιγραφέα. Ο Περιγραφέας Κύριων Χρωμάτων ορίζεται ως

$$DCD = \Big\{(c_i, p_i, v_i), s\Big\}, i = 1, 2, \ldots, N \ , \tag{2.1}$$

όπου c_i είναι το i-οστό κύριο χρώμα, p_i το ποσοστό του, v_i η διακύμανσή του και s η χωρική συνοχή. Για τη σύγκριση μεταξύ δύο περιγραφέων, έστω DCD_1 και DCD_2, το πρότυπο MPEG-7 ορίζει την απόσταση

$$D^2(DCD_1, DCD_2) = \sum_{i=1}^{N_1} p_{1i}^2 + \sum_{j=1}^{N_2} p_{2j}^2 - \sum_{i=1}^{N_1}\sum_{j=1}^{N_2} 2a_{1i,2j}p_{1i}p_{2j} \ , \tag{2.2}$$

όπου οι δείκτες 1 και 2 αντιστοιχούν στους περιγραφείς DCD_1 και DCD_2, $a_{1i,2j}$ είναι οι συντελεστές ομοιότητας ανάμεσα σε δύο χρώματα c_1 και c_2 που ορίζονται ως

$$a_{k,l} = \begin{cases} 1 - \frac{d_{k,l}}{d_{max}}, d_{k,l} \leq T_d \\ 0, d_{k,l} > T_d \end{cases} \ , \tag{2.3}$$

όπου $d_{k,l} = \parallel c_k - c_l \parallel$ είναι η Ευκλείδεια απόσταση μεταξύ δύο χρωμάτων, T_d είναι το κατώφλι κάτω από το οποίο δύο χρώματα θεωρούνται παρόμοια και $d_{max} = a \cdot T_d$. Αυτό σημαίνει ότι δύο χρώματα θεωρούνται παρόμοια αν απέχουν τουλάχιστον T_d. Στην πράξη, μια καλή επιλογή για την απόσταση T_d είναι μεταξύ 10 και 20 στον CIE-LUV χρωματικό χώρο [76]. Αντίστοιχα, μια καλή επιλογή για το a είναι μεταξύ 1-1.5. Στην (2.2) οι διακυμάνσεις και η χωρική συνοχή αγνοούνται.

Μία παραλλαγή της απόστασης που ορίζεται στην (2.2) είναι με χρήση του πεδίου χωρικής συνοχής, όπως έγινε στα πειράματα του MPEG-7, οπότε και χρησιμοποιήθηκε η απόσταση

$$D_S = w_1 \mid s_1 - s2 \mid \cdot D + w_2 D \ , \tag{2.4}$$

όπου τα w_1 και w_2 είναι βάρη με προτεινόμενες τιμές 0.3 και 0.7, αντίστοιχα. Ωστόσο, ο τύπος της απόστασης, προκειμένου να συμπεριλάβει και την προαιρετική διακύμανση, μπορεί να επεκταθεί ως

$$D_V = \sum_{i=1}^{N_1}\sum_{j=1}^{N_2} p_{1i}p_{1j}f_{1i1j} + \sum_{i=1}^{N_1}\sum_{j=1}^{N_2} p_{2i}p_{2j}f_{2i2j} - \sum_{i=1}^{N_1}\sum_{j=1}^{N_2} 2p_{1i}p_{2j}f_{1i2j} \ , \tag{2.5}$$

όπου

$$f_{x_iy_j} = \frac{1}{2\pi\sqrt{{v_{x_iy_j}}^{(l)}{v_{x_iy_j}}^{(u)}{v_{x_iy_j}}^{(v)}}} exp\left[-\frac{1}{2}\left(\frac{{c_{x_iy_j}}^{(l)}}{{v_{x_iy_j}}^{(l)}} + \frac{{c_{x_iy_j}}^{(u)}}{{v_{x_iy_j}}^{(u)}} + \frac{{c_{x_iy_j}}^{(v)}}{{v_{x_iy_j}}^{(v)}}\right)\right] \tag{2.6}$$

και

$${c_{x_iy_j}}^{(l)} = ({c_{x_i}}^{(l)} - {c_{y_j}}^{(l)})^2, {v_{x_iy_j}}^{(l)} = ({v_{x_i}}^{(l)} + {v_{y_j}}^{(l)}) \ . \tag{2.7}$$

Στις (2.6) και (2.7) τα ${c_{x_i}}^{(l)}$ και ${v_{x_i}}^{(l)}$ είναι τιμές των κύριων χρωμάτων και των χρωματικών διακυμάνσεων, τα x και y δεικτοδοτούν τους περιγραφείς του ερωτήματος και τους περιγραφείς του στόχου, τα i, j δεικτοδοτούν τις συνιστώσες του περιγραφέα και l, u, v είναι οι συνιστώσες του χρωματικού χώρου CIE-LUV.

2.4.2.2 Κλιμακωτός Περιγραφέας Χρώματος

Ο Κλιμακωτός Περιγραφέας Χρώματος (Scalable Color Descriptor) εξάγει μια κλιμακωτή αναπαράσταση χρώματος από εικόνες. Η κλιμάκωση αυτή πραγματοποιείται από ένα σχήμα κωδικοποίησης βασισμένο στην μετατροπή Haar [129], που εφαρμόζεται στις τιμές ενός ιστογράμματος χρώματος στον χρωματικό χώρο HSV. Οι τιμές του ιστογράμματος εξάγονται, κανονικοποιούνται, και μετατρέπονται με μη γραμμικό τρόπο σε μία ακέραια αναπαράσταση 4 bits, η οποία δίνει μεγαλύτερο βάρος στις μικρές τιμές.

Ο Κλιμακωτός Περιγραφέας Χρώματος είναι ένα ιστόγραμμα που ορίζεται ως

$$SCD = [c_1, c_1, \ldots, c_N] \ , \tag{2.8}$$

όπου N είναι το μέγεθός του. Το πρότυπο MPEG-7 δεν ορίζει με αυστηρότητα κάποια μετρική υπολογισμού απόστασης για το συγκεκριμένο περιγραφέα. Ωστόσο, στην πράξη εφαρμόζεται με επιτυχία η πολύ γνωστή L1 απόσταση

$$D(SCD_1, SCD_2) = \sum_{i=1}^{N} \mid c_{1i} - c_{2i} \mid \ . \tag{2.9}$$

Η εξαγωγή αποτελείται από τον υπολογισμό ενός ιστογράμματος στον HSV χρωματικό χώρο, ομοιόμορφα κβαντισμένο σε 256 κορυφές. Έπειτα οι τιμές του ιστογράμματος κβαντίζονται μη γραμμικά και εφαρμόζεται σε αυτές ο μετασχηματισμός Haar. Τα πειραματικά αποτελέσματα έχουν δείξει ότι ικανοποιητικά αποτελέσματα μπορούν να επιτευχθούν χρησιμοποιώντας τη μέση ή την πλήρη ανάλυση του περιγραφέα.

2.4.2.3 Περιγραφέας Δομής Χρώματος

Ο Περιγραφέας Δομής Χρώματος (Color Structure Descriptor) περιγράφει μία εικόνα ή μία περιοχή εικόνας με βάση την κατανομή του χρώματος σε αυτήν με παρόμοιο τρόπο όπως ένα ιστόγραμμα χρώματος, αλλά λαμβάνοντας επιπρόσθετα υπόψη και την τοπική χωρική δομή του χρώματος. Έτσι, μπορεί να διαχωρίσει εικόνες που περιλαμβάνουν μεν τα ίδια χρώματα, αλλά αυτά είναι διαφορετικά κατανεμημένα στην εικόνα, κάτι στο οποίο ένα τυπικό ιστόγραμμα χρώματος θα αποτύγχανε. Η δομή ενός επιπέδου είναι ο βαθμός στον οποίο τα εικονοστοιχεία συγκεντρώνονται από κοινού και σχηματίζουν συμπαγή αντικείμενα. Ένα δομικό στοιχείο (block) σαρώνει την εικόνα και παράλληλα καταμετρείται ο αριθμός που κάθε συγκεκριμένο χρώμα περιλαμβάνεται μέσα στο δομικό στοιχείο. Ο περιγραφέας αυτός υπολογίζεται στον χρωματικό χώρο HMMD.

Ο Περιγραφέας Δομής Χρώματος είναι ένα ιστόγραμμα που ορίζεται ως

$$CSD = [h_1, h_1, \ldots, h_N] \ , \tag{2.10}$$

όπου N είναι το μέγεθος του ιστογράμματος, δηλαδή ο αριθμός των χρωμάτων. Το πρότυπο MPEG-7 δεν ορίζει ούτε στην περίπτωση αυτή με αυστηρότητα κάποια

συνάρτηση υπολογισμού απόστασης για τον συγεκριμένο περιγραφέα. Ωστόσο, κι εδώ εφαρμόζεται με επιτυχία η πολύ γνωστή L1 απόσταση

$$D(CSD_1, CSD_2) = \sum_{i=1}^{N} \mid h_{1i} - h_{2i} \mid \ . \tag{2.11}$$

Ο περιγραφέας υπολογίζεται με τη σάρωση όλων (ή ενός υποσυνόλου) των θέσεων στην εικόνα, ανακτώντας τα χρώματα c_m όλων των εικονοστοιχείων που περιλαμβάνονται στο δομικό στοιχείο, το οποίο τοποθετείται σε κάθε θέση και αυξάνοντας τις κορυφές $h(m)$ που αντιστοιχούν στο χρώμα c_m. Στην περίπτωση που λιγότερες από 256 κορυφές είναι επιθυμητές, τότε γειτονικές κορυφές ενοποιούνται για να σχηματίσουν τον επιθυμητό αριθμό. Τέλος οι τιμές των κορυφών κβαντίζονται μη γραμμικά.

2.4.2.4 Περιγραφέας Διάταξης Χρώματος

Ο Περιγραφέας Διάταξης Χρώματος (Color Layout Descriptor) αποτελεί μία πολύ συμπαγή και σταθερή ως προς την ανάλυση αναπαράσταση χρώματος, σχεδιασμένη να αναπαριστά αποτελεσματικά τη χωρική κατανομή του χρώματος. Είναι ιδιαίτερα χρήσιμος σε εφαρμογές ανάκτησης βασισμένες στη χωρική δομή, για παράδειγμα ανάκτηση βασισμένη σε σκίτσο (sketch-based retrieval) και αναγνώριση τμήματος βίντεο. Ο περιγραφέας ορίζεται ως

$$CLD = \left[\left\{ DY_{DC}, DY_{AC_i} \right\}, \left\{ DCr_{DC}, DCr_{AC_j} \right\}, \left\{ DCb_{DC}, DCb_{AC_k} \right\} \right] \ , \tag{2.12}$$

όπου τα i, j, k δηλώνουν τον αριθμό των AC συντελεστών και μπορούν να πάρουν τις τιμές 3, 6, 10, 15, 21, 28, 64. Για τη σύγκριση μεταξύ δύο Περιγραφέων Διάταξης Χρώματος, το MPEG-7 προτείνει τη χρήση της απόστασης

$$D(CLD_1, CLD_2) = \sqrt{\sum w_{yi}(DY_i^1 - DY_i^2)^2 + w_{rj}(DCr_j^1 - DCr_j^2)^2 + w_{bk}(DCb_k^1 - DCb_k^2)^2} \ , \tag{2.13}$$

όπου τα βάρη w_{yi} ,w_{rj} , w_{bk} επιλέγονται από το χρήστη.

Στο πρώτο στάδιο της εξαγωγής, η εικόνα εισόδου διαιρείται σε 64 δομικά στοιχεία και ένα αντιπροσωπευτικό χρώμα επιλέγεται για κάθε ένα. Αυτό μπορεί να γίνει με οποιαδήποτε μέθοδο αλλά συνίσταται να χρησιμοποιείται η μέση τιμή των τιμών των εικονοστοιχείων σαν η αντιπροσωπευτική τιμή για κάθε δομικό στοιχείο. Το αποτέλεσμα είναι μία μικροσκοπική εικόνα μεγέθους 8×8. Στο τρίτο στάδιο η κάθε μία από τις τρεις χρωματικές συνιστώσες μετατρέπεται με έναν 8×8 διακριτό μετασχηματισμό συνημιτόνου (DCT) και έτσι προκύπτουν 3 σύνολα από 64 συντελεστές το κάθε ένα. Σαρώνονται με οδοντωτή τροχιά (zig-zag scanning) και οι πρώτοι συντελεστές κβαντοποιούνται μη ομοιόμορφα. Συνίσταται η χρήση 12 συντελεστών συνολικά, 6 για φωτεινότητα και 3 για κάθε χρωματικό κανάλι.

2.4.3 Περιγραφείς Υφής

Το πρότυπο MPEG-7 ορίζει 3 περιγραφείς υφής [131], τον Περιγραφέα Ομοιογενούς Υφής, τον Περιγραφέα Ιστογράμματος Ακμών και τον Περιγραφέα Αναζήτησης

Υφής. Από αυτούς, πρακτικό ενδιαφέρον έχουν οι 2 πρώτοι, οι οποίοι και παρουσιάζονται. Η εξαγωγή του Περιγραφέα Αναζήτησης Υφής είναι πολύ αργή και για το λόγο αυτό δε χρησιμοποιείται στην παρούσα Εργασία.

2.4.3.1 Περιγραφέας Ομοιογενούς Υφής

Ο Περιγραφέας Ομοιογενούς Υφής (Homogeneous Texture Descriptor) παρέχει έναν ποσοτικό χαρακτηρισμό των περιοχών ομοιογενούς υφής και είναι χρήσιμος σε εφαρμογές ανάκτησης ομοιότητας με βάση την υφή. Βασίζεται στον υπολογισμό των τοπικών χωρικών και συχνοτικών στατιστικών της υφής. Η εικόνα αρχικά φιλτράρεται με μια σειρά ευαίσθητων φίλτρων προσανατολισμού και κλίμακας. Υπολογίζεται η μέση και σταθερή απόκλιση των φιλτραρισμένων αποτελεσμάτων στον χώρο της συχνότητας. Ο περιγραφέας ορίζεται ως

$$HTD = [f_{DC}, f_{SD}, e_1, e_2, \ldots, e_N, d_1, d_2, \ldots, d_N] \ , \tag{2.14}$$

όπου f_{DC} είναι η μέση και f_{SD} η τυπική απόκλιση των τιμών της εικόνας, N ο αριθμός των καναλιών στα οποία χωρίζει ο περιγραφέας το χώρο συχνοτήτων και e_i και d_i η μέση ενέργεια και η απόκλιση ενέργειας του καθενός από αυτά. Η απόσταση μεταξύ δυο περιγραφέων ομοιογενούς υφής ορίζεται ως

$$D(HTD_1, HTD_2) = \sum_{k=1}^{N} \left| \frac{HTD_2(k) - HTD_1(k)}{d_{norm}} \right| . \tag{2.15}$$

Στην (2.15) ο παράγοντας κανονικοποίησης d_{norm} μπορεί να επιλεχθεί ελεύθερα από το χρήστη. Το πρότυπο MPEG-7 προτείνει τη χρήση της τυπικής απόκλισης της βάσης δεδομένων που περιλαμβάνει τους Περιγραφείς Ομοιογενούς Υφής.

Ο Περιγραφέας Ομοιογενούς Υφής χαρακτηρίζει την υφή της εικόνας/περιοχής χρησιμοποιώντας τη μέση ενέργεια και την απόκλιση ενέργειας από ένα σύνολο καναλιών συχνότητας. Το 2-Δ επίπεδο συχνότητας διαμερίζεται σε 30 κανάλια και από κάθε ένα υπολογίζεται η μέση ενέργεια και η απόκλισή της. Η διαμέριση του επιπέδου συχνότητας είναι ομοιόμορφη κατά τη γωνιακή κατεύθυνση (με βήμα μεγέθους 30°) αλλά μη ομοιόμορφη κατά την ακτινική κατεύθυνση, όπου γίνεται με κλίμακα οκτάβας.

2.4.3.2 Περιγραφέας Ιστογράμματος Ακμών

Ο Περιγραφέας Ιστογράμματος Ακμών (Edge Histogram Descriptor) καταγράφει τη χωρική κατανομή των ακμών, η οποία αποτελεί μια καλή αναπαράσταση υφής, χρήσιμη για το ταίριασμα από εικόνα σε εικόνα όταν η υποκείμενη περιοχή δεν είναι ομοιογενής ως προς τις ιδιότητες της υφής. Ο περιγραφέας ορίζεται ως

$$EHD = [e_1, e_2, \ldots, e_{80}] \ , \tag{2.16}$$

όπου φαίνεται ότι και αυτός έχει τη μορφή ιστογράμματος. Το πρότυπο MPEG-7 δεν προτείνει κάποια συγκεκριμένη συνάρτηση για τον υπολογισμό της απόστασης. Ωστόσο, υπονοείται η χρήση της L1 συνάρτησης που χρησιμοποιείται ευρέως για ομοιότητα ιστογραμμάτων, καθώς δίνει ικανοποιητικά αποτελέσματα. Πιο συγκεκριμένα, η απόσταση μεταξύ δύο περιγραφέων EHD_1 και EHD_2 ορίζεται ως

$$D(EHD_1, EHD_2) = \sum_{i=1}^{80} \mid e_{1i} - e_{2i} \mid \ . \tag{2.17}$$

Για την εξαγωγή του απαιτείται ο εντοπισμός μη κατευθυντικών ακμών όπως επίσης και τεσσάρων ειδών κατευθυντικών ακμών (οριζόντιες, κάθετες, διαγώνιες με προσανατολισμό 45°, διαγώνιες με προσανατολισμό 135°). Η εικόνα διαιρείται σε 4×4 υποεικόνες κι έπειτα η κάθε μία από αυτές διαιρείται επιπλέον σε μη επικαλυπτόμενα τετράγωνα δομικά στοιχεία. Το μέγεθος του δομικού στοιχείου εξαρτάται από την ανάλυση της εικόνας, ενώ ο αριθμός τους είναι ανεξάρτητος του μεγέθους της εικόνας. Πειράματα έδειξαν πως ένας αριθμός δομικών στοιχείων περίπου ίσος με 1100 φαίνεται να συλλαμβάνει καλά τα χαρακτηριστικά της κατεύθυνσης των ακμών. Στη συνέχεια κάθε ένα από αυτά ταξινομείται σε μία από τις πέντε κατηγορίες ακμών ή σαν δομικό στοιχείο χωρίς ακμές. Αυτό γίνεται θεωρώντας το κάθε ένα σαν εικόνα 2×2 και εφαρμόζοντας ανιχνευτές ακμών. Αφού γίνει η ταξινόμηση τότε υπολογίζονται τα ιστογράμματα ακμών με 5 κορυφές, μία για κάθε είδος ακμής, για τις 16 υποεικόνες (συνολικά $16 \times 5 = 80$ κορυφές). Έπειτα κάθε τιμή κανονικοποιείται ως προς το συνολικό αριθμό των δομικών στοιχείων στην υποεικόνα κι εφόσον υπάρχουν και δομικά στοιχεία χωρίς ακμές, το άθροισμα των πέντε τιμών για την κάθε υποεικόνα είναι μικρότερο ή ίσο του 1.

2.4.4 Περιγραφείς Σχήματος

Το πρότυπο MPEG-7 ορίζει 4 περιγραφείς σχήματος [17], από τους οποίους ενδιαφέρον για την παρούσα εργασία έχουν ο Περιγραφέας Σχήματος με βάση την Περιοχή και ο Περιγραφέας Σχήματος με βάση το Περίγραμμα.

2.4.4.1 Περιγραφέας Σχήματος με βάση την Περιοχή

Ο Περιγραφέας Σχήματος με βάση την Περιοχή (Region-Based Shape Descriptor) εκφράζει τη κατανομή των εικονοστοιχείων μέσα σε μια 2-Δ περιοχή ενός αντικειμένου. Μπορεί να περιγράψει σύνθετα αντικείμενα που αποτελούνται από περισσότερες από μία ασύνδετες περιοχές καθώς επίσης και απλά αντικείμενα με ή χωρίς τρύπες. Δεν επηρεάζεται από περιστροφή και κλιμάκωση. Δίνει έναν συμπαγή και αποδοτικό τρόπο περιγραφής των ιδιοτήτων πολλαπλών αποσυνδεδεμένων περιοχών ταυτόχρονα. Μερικές φορές, κατά τη διάρκεια της διαδικασίας της κατάτμησης, ένα αντικείμενο μπορεί να χωριστεί σε υποπεριοχές που δε συνδέονται. Ένα τέτοιο αντικείμενο μπορεί ακόμα να ανακτηθεί, υπό τον όρο ότι οι πληροφορίες στις οποίες οι περιοχές αυτές χωρίστηκαν διατηρούνται και χρησιμοποιούνται κατά τη διάρκεια της εξαγωγής του περιγραφέα. Ο περιγραφέας είναι εύρωστος στο θόρυβο κατάτμησης, π.χ. στον κρουστικό θόρυβο (salt and pepper noise). Ο περιγραφέας ορίζεται ως:

$$RSD = [M_i], i = 1, 2, \ldots, 35 \ , \tag{2.18}$$

όπου M_i είναι οι κανονικοποιημένοι και κβαντοποιημένοι ART συντελεστές του περιγραφέα. Για τον υπολογισμό της απόστασης μεταξύ δύο περιγραφέων, χρησιμοποιείται η απόσταση

$$D(RSD_1, RSD_2) = \sum_{i=1}^{35} \parallel M_{1i} - M_{2i} \parallel \ . \tag{2.19}$$

Ο περιγραφέας δουλεύει αποσυνθέτοντας το σχήμα σε έναν αριθμό ορθογώνιων 2-Δ συναρτήσεων βάσης (μιγαδικές) ορισμένες από την Γωνιακή Ακτινική Μετατροπή (ART). Η τεχνική αυτή βασίζεται σε μια ορθογώνια μοναδιαία μετατροπή σε έναν

μοναδιαίο δίσκο αποτελούμενο από τις πλήρεις ορθοκανονικές συναρτήσεις βάσης σε πολικές συντεταγμένες. Ο περιγραφέας αποτελείται από ένα σύνολο κανονικοποιημένων μετρών των μιγαδικών συντελεστών ART. Η σταθερότητα ως προς την περιστροφή αποκτάται με την χρήση του μέτρου των συντελεστών.

2.4.4.2 Περιγραφέας Σχήματος με βάση το Περίγραμμα

Ο Περιγραφέας Σχήματος με βάση το Περίγραμμα (Contour-Based Shape Descriptor) εκφράζει τις ιδιότητες που έχει το σχήμα του περιγράμματος του αντικειμένου. Τα αντικείμενα για τα οποία τα χαρακτηριστικά γνωρίσματα σχήματός περιλαμβάνονται στο περίγραμμά τους, περιγράφονται αποτελεσματικά από αυτόν τον περιγραφέα. Έχει διάφορες ενδιαφέρουσες ιδιότητες όπως ευθυγράμμιση με την ανθρώπινη αντίληψη για την ομοιότητα μορφής, ευρωστία στις σημαντικές εύκαμπτες παραμορφώσεις και υποστήριξη ταιριάσματος κάτω από τους μετασχηματισμούς προοπτικής.

Ο περιγραφέας σχήματος με βάση το περίγραμμα ορίζεται ως

$$CBSD = \left[NoP, C, PrC, HPY, pX[\,], pY[\,]\right] . \tag{2.20}$$

Η εξαγωγή του περιγραφέα βασίζεται στη CSS αναπαράσταση του περιγράμματος ενός σχήματος [141]. Η CSS αναπαράσταση "αποσυνθέτει" το περίγραμμα σε κοίλα και κυρτά μέρη, ορίζοντας τα σημεία καμπής (μηδενικής καμπυλότητας). Γίνεται επαναληπτικά μία εξομάλυνση του περιγράμματος μέχρι να προκύψει σαν αποτέλεσμα ένα κυρτό περίγραμμα. Η CSS εικόνα δείχνει πώς τα σημεία καμπής αλλάζουν όσο το περίγραμμα εξομαλύνεται και τείνει να γίνει κυρτό. Ο περιγραφέας αποτελείται από την κυκλικότητα και την εκκεντρότητα του αρχικού αλλά και του εξομαλυμένου περιγράμματος, έναν δείκτη που δηλώνει τον αριθμό των κορυφών στην CSS εικόνα, το ύψος της υψηλότερης κορυφής και τις συνιστώσες x και y των κορυφών που μένουν.

Στην (2.20), το NoP δηλώνει τον αριθμό των κορυφών της CSS αναπαράστασης, το C περιλαμβάνει την κυκλικότητα (circularity) και την εκκεντρότητα (eccentricity) του περιγράμματος, το PrC περιλαμβάνει την κυκλικότητα και την εκκεντρότητα του εξομαλυμένου περιγράμματος, HPY είναι η απόλυτη τιμή του ύψους της υψηλότερης κορυφής, $pX[\,]$ είναι ένας πίνακας που περιλαμβάνει τις X-τιμές των θέσεων στο περίγραμμα ενός σχήματος και $pY[\,]$ τα ύψη των αντίστοιχων κορυφών. Για την απόσταση μεταξύ δύο περιγραφέων, το MPEG-7 προτείνει τη χρήση της απόστασης

$$D(CBSD_1, DCBSD_2) = 0.4\frac{\mid C_1[0] - C_2[0] \mid}{max\{C_1[0], C_2[0]\}} + 0.3\frac{\mid C_1[1] - C_2[1] \mid}{max\{C_1[1], C_2[1]\}} + M_{CSS} , \tag{2.21}$$

όπου $C_1[0]$ και $C_2[0]$ είναι οι εκκεντρότητες των δύο περιγραφέων και $C_1[1]$ και $C_2[1]$ είναι οι κυκλικότητες. Η ποσότητα M_{CSS} είναι μια L2 απόσταση ανάμεσα στις κορυφές, η οποία "τιμωρεί" τις κορυφές που δεν ταιριάζουν και δίνεται από τη (2.22). Στη Σχέση αυτή, η άθροιση $\sum_1$ πραγματοποιείται για όλες τις κορυφές που ταιριάζουν, ενώ η άθροιση $\sum_2$ για όλες τις κορυφές που δεν ταιριάζουν. Κριτήριο για το αν δύο κορυφές ταιριάζουν ή όχι είναι η L2 απόσταση μεταξύ τους, η οποία δεν πρέπει να υπερβαίνει την τιμή 0.1 και υπολογίζεται ως

$$M_{CSS} = \sum_1((pX[i] - pX[j])^2 + (pY[i] - pY[j])^2) + \sum_2(pY[i])^2 . \tag{2.22}$$

2.5 Τεχνικές Μηχανικής Μάθησης

Για την αντιστοίχηση των χαρακτηριστικών χαμηλού επιπέδου με τις έννοιες υψηλού επιπέδου, πολύ συνήθης είναι η υιοθέτηση τεχνικών μηχανικής μάθησης, οι οποίες εκπαιδεύονται κατάλληλα και "μαθαίνουν" να αναγνωρίζουν έννοιες με βάση τα οπτικά τους χαρακτηριστικά. Οι πιο συνήθεις τεχνικές που εφαρμόζονται, συνοψίζονται στην παρούσα Ενότητα.

2.5.1 Νευρωνικά Δίκτυα

Τα Τεχνητά Νευρωνικά Δίκτυα, ή απλά Νευρωνικά Δίκτυα [83], αποτελούν υπολογιστικά μοντέλα, τα οποία κατασκευάστηκαν με βάση την παρατήρηση ότι ο ανθρώπινος εγκέφαλος λειτουργεί τελείως διαφορετικά από τον Ηλεκτρονικό Υπολογιστή. Τα βασικά δομικά συστατικά τους μιμούνται αυτά του εγκεφάλου και γι'αυτό το λόγο ονομάζονται νευρώνες. Όπως και στον ανθρώπινο εγκέφαλο, οι νευρώνες είναι οργανωμένοι με τέτοιο τρόπο ώστε να εκτελούν διάφορους υπολογισμούς. Όπως είναι γνωστό, ο τρόπος με τον οποίο λειτουργεί ο ανθρώπινος εγκέφαλος βασίζεται στην ικανότητα που έχει να οργανώνεται με βάση την εμπειρία. Αυτή ακριβώς την ικανότητα του εγκεφάλου μιμούνται τα νευρωνικά δίκτυα, στους νευρώνες των οποίων αποθηκεύεται γνώση μέσα από μια διαδικασία μάθησης. Κατά την διαδικασία αυτή, τα βάρη ανάμεσα στις συνδέσεις των νευρώνων, οι οποίες και αποκαλούνται "συνάψεις" προσαρμόζονται προκειμένου να αποθηκεύσουν τη γνώση. Υπάρχουν, ωστόσο και περιπτώσεις δικτύων που αλλάζουν την τοπολογία τους, κάτι που απορρέει από την ιδιότητα του εγκεφάλου να φτιάχνει νέους νευρώνες και συνάψεις. Έτσι, ένας νευρώνας αποτελεί μια μονάδα επεξεργασίας πληροφορίας. Η βασική διαφορά με τους Η/Υ είναι ότι ενώ αυτοί χρειάζονται μια περιγραφή της λύσης του προβλήματος, τα νευρωνικά δίκτυα εκπαιδεύονται συνήθως με την χρήση ενός συνόλου δεδομένων (σύνολο εκμάθησης), δηλαδή αποκτούν "εμπειρία" μέσω παραδειγμάτων, ακριβώς όπως και ο ανθρώπινος εγκέφαλος. Συνδυάζουν τα εξής πλεονεκτήματα:

- *Μη-Γραμμικότητα*: Οι τεχνητοί νευρώνες μπορούν να είναι και μη γραμμικοί. Η μη-γραμμικότητα είναι κατανεμημένη σε όλο το δίκτυο και είναι σημαντική μιας και συνήθως ο φυσικός μηχανισμός που είναι υπεύθυνος για την δημιουργία των εισόδων του δικτύου είναι μη γραμμικός.

- *Προσαρμογή*: Έχουν την ικανότητα να προσαρμόζουν τα βάρη τους σε αλλαγές στο περιβάλλον.

- *Αυτο-Οργάνωση*: Ένα νευρωνικό δίκτυο μπορεί να δημιουργήσει τη δική του οργάνωση κατά την διαδικασία της εκπαίδευσης.

- *Ανοχή Σφαλμάτων*: Ένα νευρωνικό δίκτυο μπορεί να συνεχίσει να λειτουργεί ακόμη και αν "καταστραφούν" κάποιοι από τους νευρώνες του λόγω της κατανεμημένης γνώσης που υπάρχει στο δίκτυο.

Η δομή των νευρωνικών δικτύων είναι παρόμοια με αυτήν του ανθρώπινου εγκεφάλου, έτσι αυτά είναι οργανωμένα σε επίπεδα. Κάθε επίπεδο αποτελείται από ένα σύνολο νευρώνων, οι οποίοι συνδέονται μεταξύ τους με συνάψεις. Κάθε σύναψη έχει ένα βάρος. Υπάρχει πάντα ένα επίπεδο εισόδου που επικοινωνεί με το περιβάλλον και

ένα επίπεδο εξόδου που παρέχει την απόκριση του δικτύου στη διέγερση που οδηγείται στην είσοδο. Τα υπόλοιπα επίπεδα συνήθως είναι κρυφά, δηλαδή δεν επικοινωνούν με το περιβάλλον. Όσον αφορά τη μάθηση, τα είδη της μπορούν να χωριστούν σε δύο μεγάλες κατηγορίες:

- *Επιβλεπόμενη* (Supervised): Στην περίπτωση αυτή, η γνώση αναπαρίσταται με ένα σύνολο από παραδείγματα εισόδων-εξόδων. Τα βάρη του δικτύου προσαρμόζονται με τρόπο ώστε να ελαχιστοποιείται το σφάλμα μεταξύ της επιθυμητής απόκρισης και της απόκρισης του δικτύου. Όταν "μεταφερθεί" η γνώση αυτή στο δίκτυο, κάτι που θα επιτευχθεί με την ελαχιστοποίηση του σφάλματος, τότε η εκπαίδευση του δικτύου σταματά.

- *Μη Επιβλεπόμενη* (Unsupervised/Self Organized): Στην περίπτωση αυτή παρέχονται μόνο κάποιες είσοδοι στο σύστημα. Δεν δίνεται εκ των προτέρων ο τρόπος κατηγοριοποίησης των εισόδων αλλά αυτό το αναλαμβάνει μόνο του το σύστημα, το οποίο "μαθαίνει" από τις διάφορες στατιστικές κανονικότητες του συνόλου των δεδομένων.

- *Ενισχυτική* (Reinforcement): Στη μάθηση αυτή, υπάρχει συνεχής αλληλεπίδραση του δικτύου με το περιβάλλον προκειμένου το δίκτυο να "μάθει" μια απεικόνιση εισόδου/εξόδου. Δεν παρέχεται στο δίκτυο η επιθυμητή απόκριση, αλλά μόνο η δυαδική πληροφορία αν η απόκρισή του είναι σωστή ή λάθος.

Τα νευρωνικά δίκτυα που χρησιμοποιούνται στην παρούσα Εργασία χρησιμοποιούν επιβλεπόμενη μάθηση και εκπαιδεύονται από σύνολα παραδειγμάτων. Για την προσαρμογή των βαρών χρησιμοποιούν τον αλγόριθμο *οπίσθιας ανατροφοδότησης* [84].

2.5.2 Ασαφή Συστήματα

Η θεωρία των ασαφών συνόλων παρουσιάστηκε από τον Zadeh [239] το 1965 και αποτελεί το υπόβαθρο της ασαφούς λογικής και των ασαφών συστημάτων [103]. Στηρίζεται στην αμφισβήτηση της κλασσικής συνολοθεωρίας, όπου ένα στοιχείο μπορεί να ανήκει **ή** να μην ανήκει σε ένα σύνολο. Έτσι, σύμφωνα με την ασαφή θεωρία, ένα στοιχείο μπορεί να ανήκει **και** να μην ανήκει σε ένα σύνολο. Η ασαφής λογική, που στηρίζεται στην θεωρία των ασαφών συνόλων, αποτελεί ένα μέσο μεταφοράς της εμπειρικής γνώσης στα συστήματα που βασίζονται στη γνώση. Αυτό, γιατί οι προσεγγίσεις που βασίζονται στην κανονική λογική δεν μπορούν να αναπαραστήσουν την γνώση με επιτυχία, ιδιαίτερα σε πολύπλοκα συστήματα, αφού οι περισσότερες έννοιες είναι ασαφείς, καθώς η γνώση από την φύση της παρουσιάζει λεκτικές ανακρίβειες. Τις ανακρίβειες αυτές έρχεται να αντιμετωπίσει η ασαφής λογική, η οποία και παρέχει το μαθηματικό πλαίσιο για τον χειρισμό των ανακριβειών αυτών και έτσι επιτυγχάνεται η μεταφορά εμπειρίας στο σύστημα. Έτσι, ένα σύστημα που χρησιμοποιεί ασαφή λογική μπορεί να λάβει τη σωστή απόφαση βασιζόμενο σε ελλιπή και αβέβαια δεδομένα. Ένας τρόπος για την οργάνωση ενός συστήματος που χρησιμοποιεί ασαφή λογική είναι με την χρήση ασαφών κανόνων. Οι κανόνες αυτοί είναι είναι ένα σύνολο δηλώσεων της μορφής AN-TOTE (if-then). Με αυτόν τον τρόπο η γνώση που υπάρχει κωδικοποιείται και έτσι δημιουργείται ένα "έμπειρο σύστημα". Ένα τέτοιο σύστημα όταν εκτεθεί στα ίδια δεδομένα με κάποιον που κατέχει την γνώση αυτή και καλείται "εμπειρογνώμονας", λειτουργεί με παρόμοιο τρόπο με αυτόν.

Γενικά, ως Ασαφές Σύστημα ορίζεται κάθε σύστημα του οποίου οι μεταβλητές (όχι απαραίτητα όλες) έχουν ως πεδίο ορισμού καταστάσεις οι οποιες είναι ασαφή σύνολα. Για κάθε μεταβλητή, τα ασαφή σύνολα είναι ορισμένα σε ένα σύνολο το οποίο είναι συνήθως ένα υποσύνολο των πραγματικών αριθμών. Στην περίπτωση αυτή, τα ασαφή σύνολα είναι ασαφείς αριθμοί και οι σχετιζόμενες μεταβλητές είναι γλωσσικές μεταβλητές (linguistic terms).

2.5.3 Μηχανές Διανυσμάτων Υποστήριξης

Οι Μηχανές Διανυσμάτων Υποστήριξης (SVM) [221] είναι δίκτυα εμπρόσθιας ανατροφοδότησης που μπορούν να χρησιμοποιηθούν για προβλήματα ταξινόμησης και εκτίμησης συναρτήσεων. Η βασική τους ιδέα είναι η κατασκευή ενός υπερεπιπέδου, το οποίο χρησιμοποιείται για την απόφαση της κατηγορίας στην οποία ανήκουν τα παραδείγματα που οδηγούνται στην είσοδο. Το υπερεπίπεδο αυτό κατασκευάζεται έτσι ώστε να μεγιστοποιεί το περιθώριο διαχωρισμού ανάμεσα σε θετικά και αρνητικά παραδείγματα. Πρέπει να σημειωθεί ότι το υπερεπίπεδο αυτό δεν κατασκευάζεται στο χώρο εισόδου, όπου το πρόβλημα πιθανόν να μην επιλύεται γραμμικά, αλλά στο *χώρο χαρακτηριστικών*, όπου έχει οδηγηθεί.

Ένα υπερεπίπεδο που έχει κατασκευαστεί με αυτόν τον τρόπο γενικά αναφέρεται σαν "Βέλτιστο Υπερεπίπεδο", μια ιδιότητα που την αποκτά καθώς οι μηχανές διανυσμάτων υποστήριξης είναι μια ακριβής υλοποίηση της μεθόδου της ελαχιστοποίησης του δομικού ρίσκου. Παρά το γεγονός ότι δεν ενσωματώνουν γνώση προσαρμοσμένη στο θεματικό πεδίο στο οποίο εφαρμόζονται, παρέχουν αξιοσημείωτη ικανότητα γενίκευσης, μια ιδιότητα που διαθέτουν σε μεγαλύτερο βαθμό από τους συνήθεις τύπους νευρωνικών δικτύων.

Το βασικό τους χαρακτηριστικό αποτελεί ένας πυρήνας εσωτερικού γινομένου μεταξύ ενός διανύσματος εισόδου και ενός από τα *διανύσματα υποστήριξης*. Το σύνολο των τελευταίων προσδιορίζεται ως ένα υποσύνολο των διανυσμάτων που απαρτίζουν το σύνολο εκπαίδευσης και εξάγονται με βάση έναν αλγόριθμο βελτιστοποίησης. Με βάση τα διανύσματα αυτά κατασκευάζεται το βέλτιστο υπερεπίπεδο. Ο πυρήνας των μηχανών διανυσμάτων υποστήριξης κατασκευάζεται με διάφορους τρόπους, οδηγώντας σε διάφορους τύπους μη γραμμικών μηχανών. Οι πιο σημαντικές από αυτές είναι οι πολυωνυμικές, τα δίκτυα πυρήνων ακτινικών συναρτήσεων και τα perceptrons ενός κρυφού επιπέδου, όπου αντίστοιχα ο πυρήνας ορίζεται σαν μια πολυωνυμική, μια ακτινική και μια συνάρτηση υπερβολικής εφαπτομένης, αντίστοιχα.

2.5.4 Νευροασαφή Δίκτυα

Στο πεδίο της τεχνητής νοημοσύνης, τα νευροασαφή δίκτυα αναφέρονται σε συνδυασμούς τεχνητών νευρωνικών δικτύων και ασαφούς λογικής. Το αποτέλεσμα είναι ένα υβριδικό σύστημα που συνδυάζει τα πλεονεκτήματα και των δύο τεχνικών. Συνδυάζει τη συλλογιστική με ασαφείς κανόνες που ευθυγραμμίζονται με την ανθρώπινη αντίληψη μαζί με τη μάθηση και τη δομή του νευρωνικού δικτύου. Έτσι, ένα νευροασαφές σύστημα ενσωματώνει τη συλλογιστική των ασαφών δικτύων που εκφράζεται με ασαφείς κανόνες AN-TOTE. Το βασικό πλεονέκτημα που εμφανίζουν είναι ότι διαθέτουν την παγκόσμια προσεγγιστική ιδιότητα μαζί με τη δυνατότητα να δημιουργούν κανόνες AN-TOTE, οι οποίοι γίνονται εύκολα αντιληπτοί από τον άνθρωπο.

Όσον αφορά την ασαφή μοντελοποίηση, αυτή έχει δύο όψεις. Από τη μία είναι η

ερμηνευσιμότητα που πρέπει να διαθέτουν και από την άλλη η ακρίβεια που επιτυγχάνουν στη λύση των προβλημάτων. Έτσι, η έρευνα στο πεδίο των νευροασαφών δικτύων στρέφεται στις δύο αυτές βασικές κατευθύνσεις.

Παρότι γενικά τα νευροασαφή δίκτυα θεωρούνται ως υλοποιήσεις ασαφών συστημάτων μέσω δικτύων με πολλές συνδέσεις, ο όρος χρησιμοποιείται πολύ συχνά και για άλλες περιπτώσεις, όπως:

- Την εξαγωγή ασαφών κανόνων από εκπαιδευμένα RBF δίκτυα,
- Τη ρύθμιση παραμέτρων εκπαίδευσης των νευρωνικών δικτύων με βάση την ασαφή λογική,
- Κριτήρια με βάση την ασαφή λογική για τον καθορισμό του μεγέθους ενός δικτύου
- Υλοποίηση ασαφών συναρτήσεων συμμετοχής μέσω αλγορίθμων συσταδοποίησης, σε περιπτώσεις μάθησης χωρίς επιτήρηση σε SOM και νευρωνικά δίκτυα,
- Αναπαράσταση ασαφοποίησης, ασαφούς συλλογιστικής και αποασαφοποίησης μέσω δικτύων εμπρόσθιας ανατροφοδότησης πολλών επιπέδων.

2.6 Συνδυασμός Περιγραφέων MPEG-7 με χρήση Τεχνικών Μηχανικής Μάθησης

Στην Ενότητα αυτή παρουσιάζονται αναλυτικά οι τεχνικές που υλοποιήθηκαν για τη συγχώνευση περιγραφέων και εφαρμόστηκαν στο πρόβλημα ταξινόμησης σκηνής. Χρησιμοποιήθηκαν ταξινομητές SVM, νευρωνικά και νευροασαφή δίκτυα και μελετήθηκαν τόσο η πρώιμη, όσο και η όψιμη συγχώνευση.

2.6.1 Ταξινόμηση με χρήση Συγχωνευμένης Περιγραφής

Η πρώτη μέθοδος που υλοποιήθηκε και παρουσιάζεται προκύπτει από την ένωση όλων των περιγραφέων σε ένα και μοναδικό διάνυσμα. Η μέθοδος αυτή θα αποκαλείται στο εξής "*συγχώνευση περιγραφέων*". Έστω $\{f_i^k\},\ k = 1, 2, \ldots, N$ οι N διαθέσιμοι περιγραφείς για μια εικόνα I_i. Τότε, ο συγχωνευμένος περιγραφέας $\mathbf{f}_i$ ορίζεται ως:

$$\mathbf{f}_i = \left[f_i^1 \ f_i^2 \ldots \ f_i^N \right] \ . \tag{2.23}$$

Η τεχνική αυτή απαιτεί όλες οι συνιστώσες των διανυσμάτων να έχουν περίπου τις ίδιες αριθμητικές τιμές για να μην προκαλούνται φαινόμενα κλιμάκωσης. Για να επιτευχθεί αυτό, είναι δυνατή η χρήση τεχνικών όπως η Ανάλυση Πρωτογενών Συνιστωσών (PCA) [16]. Στην παρούσα περίπτωση, ωστόσο, κάτι τέτοιο δεν είναι αναγκαίο, μιας και οι περιγραφείς MPEG-7 που θα χρησιμοποιηθούν είναι ήδη κανονικοποιημένοι, κβαντισμένοι και τα μέτρα τους έχουν παρόμοιες τιμές. Για την αξιολόγηση της μεθόδου αυτής, χρησιμοποιήθηκαν ταξινομητές βασισμένοι σε SVM. Καθώς τα SVM έχουν ιδιαίτερα μεγάλη διαχωριστική ικανότητα, επιλέχθηκαν για να διαφανεί πόσο μεγάλη ακρίβεια μπορεί να επιτευχθεί με μια τέτοια απλή τεχνική, θεωρώντας ότι αυτή αποτελεί το καλύτερο δυνατό που μπορεί να επιτευχθεί χωρίς σύνθετες μεθόδους όπως οι υπόλοιπες που θα παρουσιαστούν στην παρούσα Ενότητα.

2.6.2 Ταξινόμηση με χρήση Νευρωνικών Δικτύων και πρώιμη συγχώνευση

Η δεύτερη μέθοδος βασίζεται στη χρήση ενός νευρωνικού δικτύου οπίσθιας ανατροφοδότησης[3]. Στην είσοδο του δικτύου δίνεται ένα διάνυσμα $\mathbf{f}_{ij}$ που σχηματίζεται από τις συγχωνευμένες χαμηλού επιπέδου περιγραφές $\mathbf{f}_i$ και $\mathbf{f}_j$ των δύο εικόνων I_i και I_j, των οποίων και ζητείται ο υπολογισμός της απόστασης. Η απόκριση του δικτύου t_{ij} βασίζεται έτσι σε όλους τους περιγραφείς και αντιστοιχεί στην κανονικοποιημένη απόσταση των δύο εικόνων εισόδου.

Αν ψ_i και ψ_j οι κατηγορίες στις οποίες ανήκουν οι εικόνες I_i και I_j αντίστοιχα, για την εκπαίδευση του δικτύου, γίνεται η παραδοχή ότι

$$d(P_i, P_j) = \begin{cases} 1 & , \text{αν } \psi_i = \psi_j \\ 0 & , \text{αν } \psi_i \neq \psi_j \end{cases} . \tag{2.24}$$

Έτσι προκύπτεί ότι δύο εικόνες που ανήκουν στην ίδια κατηγορία έχουν την ελάχιστη δυνατή απόσταση, δηλαδή 0, ενώ μεταξύ δύο εικόνων που ανήκουν σε διαφορετική κατηγορία η απόσταση μεγιστοποιείται, δηλαδή γίνεται ίση με 1. Όταν ζητείται η απόσταση μεταξύ δύο εικόνων, παρουσιάζονται οι συγχωνευμένοι περιγραφείς τους $\mathbf{f}_1$ και $\mathbf{f}_2$ στο δίκτυο και αυτό αποκρίνεται με την εκτίμηση της απόστασής τους $\hat{d}_{ij}$. Με τον τρόπο αυτό κατασκευάζεται ένας "πίνακας αταξίας" (confusion matrix) που περιέχει τις αποστάσεις ανάμεσα σε όλες τις εικόνες του συνόλου ελέγχου. Ο πίνακας αταξίας $\mathbf{D}$ ορίζεται ως

$$\mathbf{D} = \{\hat{d}_{ij}\} = \{\hat{d}(I_i, I_j)\} \in [0, 1] . \tag{2.25}$$

Η έξοδος του νευρωνικού δικτύου μπορεί στην πράξη να πάρει τιμές λίγο μεγαλύτερες από 1 ή λίγο μικρότερες από 0. Αυτό είναι απολύτως αναμενόμενο αν αναλογιστεί κανείς τον τρόπο εκπαίδευσής τους με τον αλγόριθμο οπίσθιας διάδοσης (back propagation) [83]. Πρακτικά αυτό αντιμετωπίζεται με κατωφλίωση και δεν επηρεάζει τα αποτελέσματα της ταξινόμησης. Με τη χρήση του πίνακα $\mathbf{D}$ και ενός ταξινομητή K-πλησιέστερων γειτόνων [70], ο οποίος κατατάσσει την εικόνα στην κατηγορία που ανήκουν οι K πιο κοντινοί "γείτονές" της. Πιο συγκεκριμένα, αν $\mathcal{I}_i^j$ το σύνολο των j κοντινότερων εικόνων της I_i και ψ_i η κατηγορία στην οποία αυτή ανήκει, τότε και στην περίπτωση που έχουμε 2 κατηγορίες ψ_1 και ψ_2, η απόφαση για την κατηγορία γίνεται με τον κανόνα

$$P_i \in \begin{cases} \psi_1 & , \text{αν } |I_k \in \mathcal{I}_i^1 : \psi_k = \psi_1| > |I_k \in \mathcal{I}_i^1 : \psi_k = \psi_2| \\ \psi_2 & , \text{αλλιώς} \end{cases} . \tag{2.26}$$

Το βασικό μειονέκτημα που παρουσιάζει αυτή η τεχνική είναι ότι η απόσταση ανάμεσα σε περιγραφές που αντιστοιχουν στην ίδια εικόνα εκτιμάται από το δίκτυο ως μια πολύ μικρή αριθμητική τιμή, αντί της μηδενικής που θα ήταν το επιθυμητό, κάτι

[3]Πρέπει να σημειωθεί ότι στην περίπτωση αυτή θα μπορούσε να χρησιμοποιηθεί ένα SVM, αλλά εδώ το ζητούμενο δεν είναι η κατηγορία στην οποία ανήκει μια εικόνα με βάση την περιγραφή της, αλλά ο υπολογισμός αποστάσεων μεταξύ περιγραφών από δύο εικόνες. Βέβαια, μπορεί να υπολογιστεί η απόσταση των δειγμάτων από το υπερεπίπεδο διαχωρισμού, σε ένα SVM, κάτι που έχει μεν χρησιμοποιηθεί ως απόσταση σε κάποιες εφαρμογές, αλλά δεν εξασφαλίζει ότι μικρές αποστάσεις στο χώρο χαρακτηριστικών θα αντιστοιχούν σε μικρές αποστάσεις στο χώρο εισόδου.

για το οποίο ευθύνεται και πάλι ο αλγόριθμος ανανέωσης των βαρών. Παρολαυτά, οι μικρές αυτές τιμές μπορούν να αντικατασταθούν από μηδενικές, καθώς υπάρχει η a priori γνώση ότι πρόκειται για την ίδια εικόνα. Πέρα όμως από αυτό, η τιμή της απόστασης ανάμεσα σε δύο εικόνες εξαρτάται και από τη σειρά με την οποία παρουσιάζονται στο δίκτυο. Αυτό δε φαντάζει παράξενο καθώς ακόμη και για ένα πολύ καλά εκπαιδευμένο δίκτυο, θα υπάρχει πάντα μια μικρή διαφορά στις δύο αυτές τιμές. Έτσι ο πίνακας αταξίας $\mathbf{D}$ που προκύπτει δε μπορεί να χρησιμοποιηθεί με έναν ταξινομητή K-πλησιέστερων γειτόνων, γιατί παραβιάζει την ιδιότητα της τριγωνικής συμμετρίας. Για να ξεπεραστεί το πρόβλημα αυτό, μια λύση είναι να χρησιμοποιηθεί είτε το άνω, είτε το κάτω τριγωνικό κομμάτι του πίνακα, συμπληρώνοντας συμμετρικά, ως προς την κύρια διαγώνιο τις υπόλοιπες θέσεις του πίνακα. Έτσι, στην πρώτη περίπτωση ο πίνακας αταξίας $\mathbf{D}^u$ ορίζεται ως

$$\mathbf{D}^u = \{\hat{d}^u_{ij}\} = \begin{cases} \hat{d}_{ij} & \text{, αν } i > j \\ \hat{d}_{ji} & \text{, αλλιώς} \end{cases} , \tag{2.27}$$

ενώ στη δεύτερη περίπτωση ο πίνακας αταξίας $\mathbf{D}^l$ ορίζεται ως

$$\mathbf{D}^l = \{\hat{d}^l_{ij}\} = \begin{cases} \hat{d}_{ji} & \text{, αν } i > j \\ \hat{d}_{ij} & \text{, αλλιώς} \end{cases} . \tag{2.28}$$

Ένας άλλος τρόπος για να αποκτήσει ο πίνακας αταξίας την ιδιότητα της τριγωνικής συμμετρίας είναι η αντικατάσταση κάθε τιμής του πίνακα από την μέση απόσταση που αντιστοιχεί στις δύο εικόνες. Μετά την τροποποίηση αυτή, ο πίνακας αταξίας αποκτά την ιδιότητα της συμμετρίας. Έτσι, στην περίπτωση αυτή ορίζεται ο πίνακας αταξίας $\mathbf{D}^{av}$ ως

$$\mathbf{D}^{av} = \{\hat{d}^{av}_{ij}\} = \frac{1}{2} \cdot [(\hat{d}_{ij} + \hat{d}_{ji}] \ . \tag{2.29}$$

2.6.3 Ταξινόμηση με χρήση Νευρωνικών Δικτύων και όψιμη συγχώνευση

Προκειμένου να αξιοποιηθεί η πληροφορία της απόστασης μεταξύ δύο εικόνων με βάση έναν από τους διαθέσιμους περιγραφείς, υλοποιήθηκε μια επιπλέον προσέγγιση για τον συνδυασμό τους, η οποία χρησιμοποιεί όλους τους πίνακες αταξίας, έναν για κάθε περιγραφέα. Για να καταστεί εφικτό να χρησιμοποιηθούν όπως και στην προηγούμενη τεχνική, θα πρέπει αυτοί να συνδυαστούν κατάλληλα και να προκύψει μοναδικός πίνακας αταξίας.

Πιο συγκεκριμένα, έστω $\mathbf{D}_i, i = 1, 2, \ldots, N$ οι πίνακες αταξίες που αντιστοιχούν στους διαθέσιμους περιγραφείς $f^i, i = 1, 2, \ldots, N$. Ο συνδυασμός των πινάκων σε μοναδικό πίνακα $\mathbf{D}_m$ μπορεί να προκύψει ως ένα άθροισμα με βάρη

$$\mathbf{D}_m = \frac{1}{N} \cdot \sum_{i=1}^{N} w_i \mathbf{D}_i \ . \tag{2.30}$$

Ο προσδιορισμός των βαρών w_i γίνεται με τη χρήση ενός νευρωνικού δικτύου. Σε αυτή την περίπτωση, το νευρωνικό δίκτυο που χρησιμοποιείται για τον τελικό υπολογισμό των αποστάσεων έχει τόσες εισόδους όσος είναι ο αριθμός των περιγραφέων που χρησιμοποιούνται στο συγκεκριμένο πρόβλημα. Η έξοδος του δικτύου

αντιπροσωπεύει μια κανονικοποιημένη και με βάρη απόσταση μεταξύ των δύο εικόνων εισόδου. Και πάλι, γίνεται η παραδοχή ότι δύο εικόνες που ανήκουν στην ίδια κατηγορία έχουν απόσταση ίση με 0 και δύο εικόνες που ανήκουν σε διαφορετική κατηγορία έχουν απόσταση ίση με 1. Είναι προφανές ότι τα βάρη w_i είναι τα βάρη του μεταξύ του επίπεδου εισόδου και του 1ου κρυφού επιπέδου του νευρωνικού δικτύου. Μεγαλύτερες τιμές του w_i οδηγούν στο συμπέρασμα ότι ο περιγραφέας f^i που του αντιστοιχεί παίζει σημαντικότερο ρόλο στον υπολογισμό της απόστασης.

Ο πίνακας αταξίας $\mathbf{D}_m$ που προκύπτει χρησιμοποιείται και στην περίπτωση αυτή για να τροφοδοτήσει έναν ταξινομητή K-πλησιέστερων γειτόνων. Και στην περίπτωση αυτή, χρησιμοποιώντας την a priori γνώση ότι η απόσταση μιας εικόνας και του εαυτού της είναι ελάχιστη, η διαγώνιός του τίθεται ίση με 0. Πρέπει να σημειωθεί ότι η μέθοδος αυτή οδηγεί σε συμμετρικό πίνακα αταξίας και εξασφαλίζεται έτσι η απαιτούμενη αυτή ιδιότητα.

2.6.4 Ταξινόμηση και Εξαγωγή Κανόνων με χρήση Falcon-ART Νευροασαφών Δικτύων

Οι μέθοδοι που παρουσιάστηκαν στις Ενότητες 2.6.1, 2.6.2 και 2.6.3 μπορούν να επιλύσουν το πρόβλημα της ταξινόμησης σκηνής σε ικανοποιητικό βαθμό, ωστόσο καμία τους δεν μπορεί να ερμηνευθεί σημασιολογικά. Δηλαδή από καμία δεν μπορεί να εξαχθεί ο "μηχανισμός" που τελικά πραγματοποιεί την ταξινόμηση των εικόνων με τρόπο που να μπορεί να γίνει κατανοητός από τον άνθρωπο. Αυτό είναι ένα από τα προβλήματα που συνήθως αντιμετωπίζονται με τη χρήση νευροασαφών δικτύων. Στην παρούσα Ενότητα, για να εξαχθεί η γνώση με τη μορφή ασαφών κανόνων, προτείνεται η χρήση ενός Falcon-ART νευροασαφούς δίκτυου, το οποίο προτάθηκε από τους Makris και Mailis [130]. Το δίκτυο αυτό βασίζεται στον γνωστό Fuzzy-ART αλγόριθμο ομαδοποίησης των Carpenter et al. [37].

Η εκπαίδευση του δικτύου γίνεται σε δύο φάσεις, τη "φάση μάθησης της δομής" (structure learning phase), όπου ο Fuzzy-ART αλγόριθμος χρησιμοποιείται για να δημιουργήσει τη δομή του δικτύου και τη "φάση εκμάθησης των παραμέτρων" (parameter learning phase), όπου οι παράμετροι του δικτύου βελτιώνονται με τη χρήση του αλγόριθμου οπίσθιας ανατροφοδότησης. Συνοπτικά, ο αλγόριθμος Fuzzy-ART δημιουργεί υπερκύβους στο χώρο εισόδου. Ένας υπερκύβος είναι η επέκταση ενός 3-Δ κύβου σε έναν N-διάστατο χώρο. Κάθε υπερκύβος αντιστοιχεί σε μια ομάδα και ο όγκος του αυξάνεται προς τη διεύθυνση των δειγμάτων εκπαίδευσης που παρουσιάζονται στ δίκτο και ανήκουν στην κατηγορία στην οποία αντιστοιχεί. Είναι πιθανό να δημιουργηθούν και νέοι υπερκύβοι, σύμφωνα με το "κριτήριο επιτήρησης" (vigilance criterion). Μετά την εφαρμογή του αλγορίθμου Fuzzy-ART, ο χώρος εισόδου συσταδοποιείται και ένα δίκτυο 5 επιπέδων δημιουργείται από τους υπερκύβους εισόδου και τις μεταξύ τους σχέσεις. Οι παράμετροι και οι σχέσεις μεταξύ των κόμβων κάθε επιπέδου προκύπτουν από τον αλγόριθμο Fuzzy-ART και το δίκτυο των 5 επιπέδων που προκύπτει εκπαιδεύεται με τον αλγόριθμο οπίσθιας ανατροφοδότησης.

Η είσοδος του δικτύου είναι και πάλι μια συγχωνευμένη περιγραφή $\mathbf{f}_i$ μιας εικόνας I_i, όπως περιγράφεται στην (2.23). Μετά τη φάση εκπαίδευσης, η απάντηση του δικτύου θα πρέπει να είναι η κατηγορία ψ_i στην οποία ανήκει η εικόνα στην οποία ανήκει η συγχωνευμένη περιγραφή $\mathbf{f}_i$ που οδηγήθηκε στην είσοδο. Αν και μέχρι τώρα η διαδικασία δεν φαίνεται να παρουσιάζει κάποιο όφελος σε σχέση με τη χρήση παραδοσιακών νευρωνικών δικτύων, η χρήση του νευροασαφούς δικτύου θα επιτρέψει

την εξαγωγή σημασιολογικού περιεχομένου με τη μορφή ασαφών κανόνων. Έτσι, ο τρόπος με τον οποίο τα χαμηλού επιπέδου χαρακτηριστικά της εικόνας καθορίζουν την κατηγορία στην οποία ανήκει θα γινει πλέον πιο προφανής και σε κάποιες περιπτώσεις θα μπορέσει να περιγραφεί με φυσική γλώσσα.

Έστω $\mathbf{f}_i = [f_i^1\ f_i^2\ \ldots, f_i^N]$ η συγχωνευμένη περιγραφή μιας εικόνας I_i. Ο σκοπός είναι να εξαχθούν ασαφείς κανόνες της μορφής:

> AN f_i^1 είναι K_1 ΚΑΙ f_i^2 *είναι* K_1 ΚΑΙ ... ΚΑΙ D_i^N είναι K_N, ΤΟΤΕ ηεικόναI_i *ανήκει στην κατηγορία* ψ_i ,

όπου ως K_i ορίστηκε μια από τις καταστάσεις που μπορεί να λάβει μια συνιστώσα ενός περιγραφέα, η οποία μπορεί να περιγραφεί από ένα ασαφές σύνολο A_i. Αν ο χώρος των περιγραφέων χωριστεί σε τρία ασαφή σύνολα $A_i, i = 1, 2, 3$ και τα σύνολα αυτά χαρακτηριστούν ως:

$$A_1\text{: } \textit{χαμηλός},\ \ A_2\text{: } \textit{μέτριος}\ ,\ A_3\text{: } \textit{υψηλός}\ \ ,$$

τότε, για μια κατάσταση K_i, θα ισχύει ότι

$$K_i \in A_j,\ i = 1, 2,\ \ \ldots, N \text{ και } j = 1, 2, 3\ , \tag{2.31}$$

οπότε ένα παράδειγμα ασαφούς κανόνα για το πρόβλημα της ταξινόμησης της εικόνας I_i θα μπορούσε να διατυπωθεί ως:

> AN f_i^1 είναι *χαμηλός* ΚΑΙ f_i^2 είναι *μέτριος* ΚΑΙ … ΚΑΙ D_i^N είναι *υψηλός*, ΤΟΤΕ η εικόνα I_i ανήκει στην κατηγορία ψ_i .

Στη συνέχεια και για να καταστεί σαφέστερος ο τρόπος με τον οποίο μπορεί να προκύψει μια περιγραφή του τρόπου με τον οποίο γίνεται η απόφαση για την ταξινόμηση κοντά στην ανθρώπινη αντίληψη και με χρήση των κανόνων που εξήχθησαν από τον αλγόριθμο Falcon-Art, παρατίθεται ένα παράδειγμα. Κάθε διάσταση της εικόνας χωρίζεται σε τρία ίσα μέρη, με κάθε ένα από αυτά να αντιστοιχεί σε *χαμηλές*, *μέτριες*, *υψηλές* τιμές. Έτσι, μια εικόνα χωρίζεται σε 9 υποεικόνες. Κάθε υπερκύβος που κατασκευάστηκε από το δίκτυο οδηγεί σε έναν κανόνα που χρησιμοποιεί *χαμηλές*, *μέτριες* και *υψηλές* τιμές. Η εξαγωγή των ασαφών κανόνων γίνεται με τέτοιο τρόπο ώστε αυτοί να είναι απλοί και κατανοητοί. Στη συνέχεια παρουσιάζεται ένα παράδειγμα ταξινόμησης με τη χρήση μόνο του Περιγραφέα Ιστογράμματος Ακμών. Οι υποεικόνες χωρίζονται σε αυτές που περιγράφουν τα άνω, τα μεσαία και τα κάτω μέρη της εικόνας και μια ποιοτική περιγραφή (*χαμηλή*, *μέτρια* ή *υψηλή*) εξάγεται από τον περιγραφέα για κάθε τύπο ακμών. Έτσι, ένα παράδειγμα ασαφούς κανόνα είναι το ακόλουθο:

> AN ο αριθμός των 0° ακμών στο *άνω* μέρος της εικόνας είναι *χαμηλός* ΚΑΙ ο αριθμός των 45° ακμών στο *άνω* μέρος της εικόνας είναι *μέτριος* ΚΑΙ …ΚΑΙ ο αριθμός των μη κατευθυντικών ακμών στο *κάτω* μέρος της εικόνας είναι *υψηλός*, ΤΟΤΕ η εικόνα ανήκει στην κατηγορία *Παραλία* .

Πρέπει να τονιστεί ότι το Falcon-ART δίκτυο είναι αυτό που καθορίζει τον αριθμό των κανόνων. Στις περιπτώσεις που χρησιμοποιούνται άλλοι περιγραφείς του MPEG-7, δεν είναι πάντα εύκολο να παρουσιαστούν οι κανόνες που προκύπτουν, καθώς δεν είναι πάντα τόσο κοντά στην ανθρώπινη αντίληψη, όντας μάλλον ποσοτικοί, παρά ποιοτικοί, όπως είναι ο Περιγραφέας Ιστογράμματος Ακμών.

2.6.5 Ταξινόμηση και Εξαγωγή Κανόνων με χρήση Ασαφών Μηχανών Διανυσμάτων Υποστήριξης

Όπως κατεστή σαφές στην Ενότητα 2.6.4, η χρήση ταξινομητών που βασίζονται σε νευροασαφή δίκτυα εμφανίζει το βασικό πλεονέκτημα της εξαγωγής ασαφών κανόνων, μέσω των οποίων γίνεται κατανοητή η διαδικασία της ταξινόμησης. Στην Ενότητα αυτή χρησιμοποιούνται οι Ασαφείς Μηχανές Διανυσμάτων Υποστήριξης (Fuzzy SVM) που προτάθηκαν από τους Spyrou et al. [200], με σκοπό την εφαρμογή τους στο ίδιο πρόβλημα ταξινόμησης εικόνας και στην εξαγωγή ασαφών κανόνων, όσον αφορά το μηχανισμό της ταξινόμησης.

Τα Fuzzy SVM διαμερίζουν το χώρο εισόδου με τον ίδιο τρόπο που θα έκανε ένα ασαφές σύστημα. Η διαφορά τους έγκειται στη χρήση του αλγορίθμου βελτιστοποίησης των κλασικών SVM. Η είσοδος του δικτύου είναι και πάλι μια συγχωνευμένη περιγραφή μιας εικόνας. Μετά τη φάση εκπαίδευσης, η απάντηση του δικτύου θα πρέπει να είναι η κατηγορία στην οποία ανήκει η εικόνα στην οποία ανήκει η περιγραφή που οδηγήθηκε στην είσοδο. Με βάση τη διαμέριση του χώρου που προκύπτει και έστω $\mathbf{f}^i = [f_i^1 \ f_i^2 \ \ldots f_i^N]$ η συγχωνευμένη περιγραφή μιας εικόνας I_i, τελικά δημιουργούνται ασαφείς κανόνες που έχουν τη μορφή:

> ΑΝ το f_i^1 έχει τιμή ΠΕΡΙΠΟΥ ΙΣΗ με D_{i1} ΚΑΙ το f_i^2 έχει τιμή ΠΕΡΙΠΟΥ ΙΣΗ με D_{i2} ΚΑΙ ... ΚΑΙ το f_i^N έχει τιμή ΠΕΡΙΠΟΥ ΙΣΗ με D_{iN}, ΤΟΤΕ η εικόνα I_i ανήκει στην κατηγορία ψ_i

2.7 Πειραματικά Αποτελέσματα

Η βάση δεδομένων που χρησιμοποιήθηκε για τα πειράματα του Κεφαλαιού αυτού αποτελείται από εικόνες που αποτελούν μέρος του συνόλου δεδομένων του ερευνητικού έργου aceMedia[4]. Πιο συγκεκριμένα, πρόκειται για 767 υψηλής ποιότητας εικόνες, χωρισμένες σε δύο κατηγορίες, *παραλία* και *πόλη*. Χαρακτηριστικά παραδείγματα από το σύνολο των εικόνων που χρησιμοποιήθηκε απεικονίζονται στο Σχήμα 2.8. Πρέπει να επισημανθεί, εδώ, ότι το πρόβλημα της αδυναμίας αυστηρού ορισμού για τις σκηνές οδηγεί στην παρουσία "παραπλανητικών" παραδειγμάτων στο σύνολο εικόνων, όπως είναι οι δύο πρώτες εικόνες της δεύτερης σειράς. Η πρώτη απεικονίζει δύο *ανθρώπους* σε *εσωτερικό χώρο*, ενώ η δεύτερη δύο *ανθρώπους* μπροστά από μια *λίμνη*. Τα δύο προβλήματα που μπορεί να δημιουργηθούν από την ύπαρξη τέτοιων παραδειγμάτων είναι η δυσκολία εκπαίδευσης των ταξινομητών και η αδυναμία επίτευξης μεγάλης ακρίβειας. Αν παρουσιαστούν στους ταξινομητές εικόνες τελείως διαφορετικές από αυτές που χρησιμοποιήθηκαν για την εκπαίδευση, τότε το αποτέλεσμα της ταξινόμησης θα είναι πιθανότατα τυχαίο. Επίσης, η ύπαρξη εικόνων που ανήκουν στη μία κατηγορία, αλλά είναι παρόμοιες οπτικά με εικόνες της άλλης κατηγορίας επίσης θα ταξινομηθούν με τυχαίο τρόπο, προκαλώντας πτώση της ακρίβειας.

Οι περιγραφείς του προτύπου MPEG-7 που επιλέχθηκαν είναι ο Περιγραφέας Δομής Χρώματος, ο Περιγραφέας Κλιμακωτού Χρώματος και ο Περιγραφέας Ιστογράμματος Ακμών. Χρησιμοποιήθηκαν όλοι οι περιγραφείς μεμονωμένα, αλλά και όλοι οι πιθανοί συνδυασμοί τους και αξιολογήθηκαν σε όλες τις παραπάνω τεχνικές. Τα αποτελέσματα παρουσιάζονται αναλυτικά στον Πίνακα 2.1. Επιλέχθηκαν 40 εικόνες

[4] http://www.acemedia.org

από την κατηγορία *παραλία* και 20 από την κατηγορία *πόλη* ως αντιπροσωπευτικά παραδείγματα από τις κατηγορίες τους και χρησιμοποιήθηκαν σαν κοινό σύνολο εκμάθησης για όλες τις τεχνικές που περιγράφηκαν. Οι υπόλοιπες 707 (406 από την κατηγορία *παραλία* και 301 από την κατηγορία *πόλη*) εικόνες χρησιμοποιήθηκαν για έλεγχο. Για κάθε κατηγορία υπολογίστηκε το *μέτρο ακρίβειας* (precision) το οποίο θα αποκαλείται εφεξής "ακρίβεια"[5]. Ορίζοντας σαν $|\cdot|$ το πλήθος των στοιχείων ενός συνόλου, η ακρίβεια υπολογίζεται ως

$$P_i = \frac{|D_i \bigcap G_i|}{|D_i|}, \quad i = 1, 2, \ldots, N_C \ , \tag{2.32}$$

όπου D_i είναι το σύνολο των εικόνων για τις οποίες ο ανιχνευτής της έννοιας C_i αποφάσισε ότι την απεικονίζουν, G_i είναι το σύνολο των εικόνων που πραγματικά απεικονίζουν την έννοια, σύμφωνα με το σύνολο δεδομένης αλήθειας και N_C το πλήθος των εννοιών προς ταξινόμηση. Η ακρίβεια αντιστοιχεί στο ποσοστό των εικόνων που πραγματικά απεικονίζουν την έννοια, σύμφωνα πάντα με το σύνολο δεδομένης αλήθειας, ως προς τις εικόνες που το σύστημα αποφάνθηκε ότι την απεικονίζουν. Στη συνέχεια ακολουθούν αναλυτικά τα αποτελέσματα για κάθε μία από τις τεχνικές που προτείνονται σε αυτό το Κεφάλαιο.

2.7.1 Ταξινόμηση με χρήση Συγχωνευμένης Περιγραφής

Τα συγχωνευμένα διανύσματα χρησιμοποιήθηκαν απευθείας στην είσοδο ενός SVM ταξινομητή με πολυωνυμικό πυρήνα 1ου βαθμού (γραμμικό πυρήνα). Πειράματα με πυρήνες μεγαλύτερου βαθμού (εώς 5) έδωσαν παρόμοια αποτελέσματα, χωρίς ουσιαστικές βελτιώσεις και για το λόγο αυτό δεν αναφέρονται. Τα αποτελέσματα που παρουσιάζονται στον Πίνακα 2.1 δείχνουν ότι η ακρίβεια που επιτυγχάνεται αυξάνεται με τη χρήση περισσότερων περιγραφέων. Ενώ, λοιπόν, μεμονωμένοι περιγραφείς οδηγούν σε ακρίβεια από 0.80 εώς 0.84, η συγχώνευση των δύο οδηγεί σε βελτίωση της ακρίβειας από 0.87 εώς 0.88, ανάλογα με το συνδυασμό που επιλέχθηκε και φτάνουν τελικά στη μέγιστη τιμή ακρίβειας που είναι ίση με 0.89 και προέκυψε με τη συγχώνευση και των τριών. Αυτό που πρέπει να επισημανθεί είναι ότι φαίνεται αναγκαίος ο συνδυασμός περιγραφέα χρώματος και περιγραφέα υφής, ο οποίος και επιτυγχάνει τα καλύτερα αποτελέσματα στην περίπτωση δύο περιγραφέων σε σχέση με το συνδυασμό περιγραφέων χρώματος. Η τεχνική αυτή δεν δίνει πληροφορία για την απόσταση μεταξύ δύο εικόνων, αλλά και δεν παρέχει γνώση για το μηχανισμό της ταξινόμησης.

2.7.2 Ταξινόμηση με χρήση Νευρωνικών Δικτύων

Για τις περιπτώσεις της όψιμης και της πρώιμης συγχώνευσης οπτικών περιγραφέων με χρήση νευρωνικών δικτύων, οι οποίες μελετήθηκαν από κοινού, η απόσταση μεταξύ δύο εικόνων καθορίστηκε να είναι ίση με 0 για εικόνες που ανήκουν στην ίδια

[5]Συνήθως, μαζί με την ακρίβεια υπολογίζεται και το "μέτρο ανάκτησης" η απλά, "ανάκτηση", το οποίο ορίζεται στην Ενότητα 3.5. Στο παρόν πρόβλημα δυαδικής ταξινόμησης, τα δύο αυτά μέτρα ταυτίζονται και άρα ο υπολογισμός του μέτρου ανάκτησης δεν είναι απαραίτητος.

Μέθοδος	**Ακρίβεια**						
	EHD	**CLD**	**SCD**	**EHD,CLD**	**EHD,SCD**	**CLD,SCD**	**όλοι**
Γραμμικό SVM	0.80	0.82	0.83	0.87	0.89	0.87	0.89
ΝΔ-πρώιμη	-	-	-	**0.89**	0.89	0.89	**0.93**
ΝΔ-όψιμη	0.82	**0.87**	**0.86**	0.67	**0.90**	**0.91**	0.86
Ασαφές SVM	**0.83**	0.84	0.84	0.83	0.87	0.88	0.91
Falcon-ART	0.81	0.85	0.84	0.82	0.84	0.86	0.88

Πίνακας 2.1: *Αποτελέσματα χρησιμοποιώντας όλες τις μεθόδους για διαφορετικούς MPEG-7 περιγραφείς: Περιγραφέας Ιστογράμματος Ακμών (EHD), Περιγραφέας Δομής Χρώματος (CLD) και Περιγραφέας Κλιμακωτού Χρώματος (SCD)*

κατηγορία και ίση με 1 για εικόνες που ανήκουν σε διαφορετική κατηγορία. Προκειμένου να βελτιωθεί η εκπαίδευση, για κάθε ζευγάρι εικόνων, δύο δείγματα εκπαίδευσης δημιουργούνται με παρουσίαση των εικόνων με διαφορετική σειρά στο δίκτυο. Αυτό γίνεται προκειμένου να βοηθηθεί το δίκτυο να "μάθει" να απαντάει με όσο το δυνατόν παρόμοιο τρόπο για το ίδιο ζεύγος εικόνων, ανεξάρτητα από τη σειρά που του παρουσιάζονται. Έτσι, δημιουργούνται 2800 δείγματα εκπαίδευση και με βάση αυτά εκπαιδεύεται το νευρωνικό δίκτυο. Αφού παρουσιαστούν όλα τα ζευγάρια εικόνων στο δίκτυο, κατασκευάζεται ο αντίστοιχος πίνακας αταξίας. Για την ταξινόμηση με ταξινομητές K-πλησιέστερων γειτόνων, χρησιμοποιούνται συμμετρικοί πίνακες, όπως ορίζονται από την (2.29). Χρησιμοποιήθηκε η L2 (Ευκλίδεια) απόσταση και η τιμή του K τέθηκε ίση με 5.

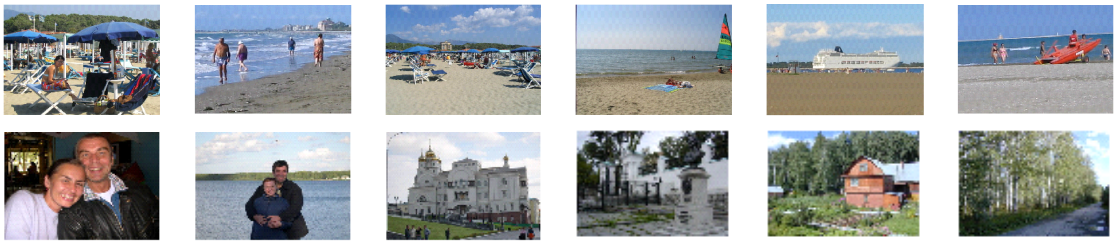
Τα αποτελέσματα της ταξινόμησης δείχνουν ότι η καλύτερη ακρίβεια επιτυγχάνεται γενικά όταν χρησιμοποιούνται και οι τρεις διαθέσιμοι περιγραφείς. Όλα τα αποτελέσματα παρουσιάζονται στον Πίνακα 2.1. Η ακρίβεια που επιτεύχθηκε με όψιμη συγχώνευση ήταν σε γενικές γραμμές η καλύτερη από όλες τις τεχνικές που παρουσιάστηκαν στο Κεφάλαιο αυτό. Ωστόσο, για την περίπτωση της χρήσης όλων των περιγραφέων, με πρώιμη συγχώνευση, επιτεύχθηκε η μέγιστη ακρίβεια.

Πρέπει να σημειωθεί ότι σε αντίθεση με τις περισσότερες περιπτώσεις που συναντώνται στη βιβλιογραφία, η ακρίβεια που επιτεύχθηκε ήταν παρόμοια και στην περίπτωση της όψιμης και σε αυτήν της πρώιμης συγχώνευσης. Ωστόσο η ταχύτητα εφαρμογής της πρώιμης συγχώνευσης ήταν ελαφρώς μεγαλύτερη, όπως ήταν και το αναμενόμενο.

2.7.3 Ταξινόμηση με χρήση Falcon-ART Νευροασαφών Δικτύων

Και στην περίπτωση αυτή χρησιμοποιήθηκαν οι ίδιες 60 εικόνες για την εκπαίδευση του Falcon-ART νευροασαφούς δικτύου. Το δίκτυο εκπαιδεύτηκε με τους συγχωνευμένους περιγραφείς των εικόνων, που παρουσιάστηκαν στο δίκτυο με τυχαία σειρά. Στην περίπτωση του Περιγραφέα Ιστογράμματος Ακμών, το δίκτυο δημιούργησε 5 ασαφείς κανόνες, οι οποίοι παρουσιάζονται αναλυτικά στον Πίνακα 2.2 και μπορούν να ερμηνευθούν σημασιολογικά. Επίσης, στο Σχήμα 2.7 απεικονίζεται μια εικόνα χωρισμένοι σε υποεικόνες. Με βάση τα ποσοτικά χαρακτηριστικά των ακμών σε αυτές, δημιουργούνται οι ασαφείς κανόνες από το δίκτυο Falcon-ART. Η ακρίβεια που επιτεύχθηκε έλαβε τη μέγιστη τιμή της, ίση με 0.88 στην περίπτωση που συγχωνεύθηκαν όλοι οι διαθέσιμοι περιγραφείς, με τον Fuzzy-ART αλγόριθμο να δημιουργεί στην περίπτωση αυτή 8 ασαφείς κανόνες.

Σχήμα 2.7: *Μια εικόνα από την κατηγορία παραλία που έχει χωριστεί με ένα ορθογωνικό πλέγμα σε υποεικόνες. Από κάθε μία από τις υποεικόνες εξάγεται ο Περιγραφέας Ιστογράμματος Ακμών και έπειτα, με βάση τα συνολικά χαρακτηριστικά των ακμών δημιουργούνται ασαφείς κανόνες, όπως περιγράφηκε στην Ενότητα 2.6.4.*



Σχήμα 2.8: *Αντιπροσωπευτικά παραδείγματα του συνόλου εικόνων που χρησιμοποιήθηκε. Πρώτη γραμμή: εικόνες από την κατηγορία παραλία, δεύτερη γραμμή: εικόνες από την κατηγορία πόλη.*

2.7.4 Ταξινόμηση με χρήση Ασαφών Μηχανών Διανυσμάτων Υποστήριξης

Το Ασαφές SVM δίκτυο εκπαιδεύτηκε ακριβώς με τον ίδιο τρόπο όπως και το Falcon-ART δίκτυο. Οι ίδιες 60 εικόνες χρησιμοποιήθηκαν για την εκπαίδευσή του. Οι εξαχθείσες περιγραφές συγχωνεύτηκαν και παρουσιάστηκαν στο δίκτυο. Η ακρίβεια που επιτεύχθηκε ήταν καλύτερη από αυτή του Falcon-ART και έφτασε εώς και τιμή ίση με 0.91, με τη χρήση όλων των διαθέσιμων περιγραφέων. Τα αποτελέσματα της εφαρμογής της μεθόδου αυτής περιέχονται στον Πίνακα 2.1. Το δίκτυο δημιούργησε 32 ασαφείς κανόνες, οι οποίοι είχαν τη μορφή που περιγράφηκε στην Ενότητα 2.6.5.

2.8 Συμπεράσματα

Στο Κεφάλαιο αυτό αντιμετωπίστηκε το πρόβλημα της ταξινόμησης σκηνής. Για το σκοπό αυτό, διερευνήθηκαν τεχνικές που βασίζονται σε αλγορίθμους μηχανικής μάθησης. Με βάση τα πειραματικά αποτελέσματα που παρουσιάστηκαν, κατέστη σαφές ότι και οι τέσσερις τεχνικές αντιμετώπισαν με παρόμοια επιτυχία το πρόβλημα.

Μέρος εικόνας	**Τύπος Ακμής**	K_1	K_2	K_3	K_4	K_5
επάνω	0°	M	X	M-X	M-X	X
	45°	M	X	M-X	M	M
	90°	M	X	M	M	M
	135°	ϒ	M	M	M	M
	nondir.	M	M	M	M	M
μεσαίο	0°	M	X	M	M	M
	45°	M	M-X	ϒ	M	ϒ
	90°	M	M	M	M	ϒ
	135°	ϒ	M	M-X	ϒ	ϒ
	nondir.	ϒ	M	M	ϒ	M
κάτω	0°	M	X	ϒ	M	ϒ
	45°	M	M	ϒ	ϒ	ϒ
	90°	ϒ	M	M	ϒ	ϒ
	135°	M	M-X	M	ϒ	M
	nondir.	M	M	M-X	ϒ	M
Κατηγορία		*πόλη*	*παραλία*	*πόλη*	*παραλία*	*παραλία*

Πίνακας 2.2: *Οι 5 ασαφείς κανόνες* K_i *που εξήχθησαν από το Falcon-ART δίκτυο, το οποίο εκπαιδεύτηκε με τον Περιγραφέα Ιστογράμματος Ακμών. Ως "nondir" αναφέρονται οι μη κατευθυντικές ακμές.*

Η χρήση οπτικών περιγραφέων του προτύπου MPEG-7 και των συνδυασμών τους μπόρεσε με επιτυχία να εφαρμοστεί στην ταξινόμηση *πόλης/παραλίας*. Ωστόσο, κάθε τεχνική έχει τα δικά της πλεονεκτήματα και μειονεκτήματα. Έτσι, κάθε πιθανός χρήστης τους πρέπει να υπολογίσει τι του προσφέρει κάθε τεχνική και με ποιο κόστος και έπειτα να επιλέξει την κατάλληλη, σύμφωνα με το πρόβλημα που θέλει να επιλύσει.

Η τεχνική ταξινόμησης με απλή χρήση συγχωνευμένης περιγραφής και ταξινομητές βασισμένους σε SVM αποτελεί την απλούστερη από όλες όσες διερευνήθηκαν. Η ακρίβεια που επιτυγχάνει είναι η μικρότερη από όλες, στην περίπτωση που χρησιμοποιούνται μεμονωμένοι περιγραφείς, κάτι που βελτιώνεται όσο ανεβαίνει ο αριθμός των περιγραφέων, χωρίς να ξεπεράσει ποτέ όλες τις υπόλοιπες τεχνικές. Παρουσιάζει το πλεονέκτημα ότι είναι πολύ εύκολο να υλοποιηθεί, καθώς οι υλοποιήσεις για δίκτυα SVM είναι πλέον πολλές. Έτσι το μόνο που απαιτείται είναι ένα εργαλείο εξαγωγής περιγραφέων και μια υλοποίηση SVM, δίνοντας τη δυνατότητα για γρήγορη λύση σε πρακτικά προβλήματα ταξινόμησης. Ωστόσο, εξαιτίας των SVM, δεν είναι δυνατός ο υπολογισμός της βεβαιότητας της ταξινόμησης. Αντίθετα, η ταξινόμηση είναι δυαδική.

Οι τεχνικές πρώιμης και όψιμης συγχώνευσης που βασίστηκαν σε νευρωνικά δίκτυα, αφενός βελτιώνουν την ακρίβεια της ταξινόμησης και αφετέρου παρέχουν επιπλέον πληροφορίες για αυτήν. Η μέθοδος με την οποία γίνεται η ταξινόμηση υπολογίζει σε κάποιο στάδιο τις αποστάσεις μεταξύ δύο εικόνων. Αυτό μπορεί να φανεί χρήσιμο σε περιπτώσεις όπου είναι επιθυμητή η ανάκτηση εικόνων, αντί για την ταξινόμηση. Επιπλέον, παρατηρώντας τα βάρη των νευρωνικών δικτύων στην περίπτωση της πρώιμης συγχώνευσης περιγραφέων, είναι δυνατόν να εξαχθούν χρήσιμα συμπεράσματα σχετικά με το ποιοι περιγραφείς είναι πιο σημαντικοί για την ταξινόμηση. Η υλοποίηση των τεχνικών αυτών είναι σχετικά πιο πολύπλοκη και επιπλέον η εφαρμογή τους είναι πιο αργή από τους SVM ταξινομητές, γιατί παρά το γεγονός ότι η εκπαίδευση είναι πιο γρήγορη, ο αλγόριθμος ταξινόμησης πρέπει να υπολογίσει έναν μεγάλο αριθμό από αποστάσεις, κάτι που κάνει αργή την εφαρμογή των τεχνικών σε

μεγάλα σύνολα εικόνων. Πρέπει να σημειωθεί ότι η τεχνική πρώιμης συγχώνευσης με νευρωνικά δίκτυα πέτυχε τη μέγιστη ακρίβεια από ολες τις τεχνικές του Κεφαλαίου, στην περίπτωση κατά την οποία συγχωνεύτηκαν όλοι οι διαθέσιμοι περιγραφείς.

Οι τεχνικές που χρησιμοποίησαν νευροασαφή δίκτυα παρείχαν έναν τρόπο εξαγωγής της γνώσης που απέκτησαν οι ταξινομητές με την εκπαίδευση. Δημιούργησαν ασαφείς κανόνες, μέσω των οποίων μπορεί να εξαχθεί ο μηχανισμός με τον οποίο γίνεται η ταξινόμηση. Το δίκτυο Falcon-ART έδωσε κατανοητούς ασαφείς κανόνες φυσικής γλώσσας. Πέρα από αυτό, η τεχνική που χρησιμοποίησε τα ασαφή SVM πέτυχε αξιόλογα αποτελέσματα, όσον αφορά την ακρίβεια. Βέβαια και τα δύο νευροασαφή δίκτυα αποτελούν ερευνητικές προσπάθειες και όχι καθιερωμένα εργαλεία στο χώρο της μηχανικής μάθησης, κάτι που προφανώς καθιστά την υιοθέτησή τους ιδιαίτερα δύσκολη.

Όσον αφορά, τώρα, το πρόβλημα της ταξινόμησης σκηνής *πόλης/παραλίας*, διαφάνηκε ότι αυτό μπορεί να αντιμετωπισθεί με επιτυχία με τη χρήση καθολικά εξηγμένων περιγραφέων χρώματος και υφής του προτύπου MPEG-7 και ταξινομητές βασισμένους σε τεχνικές μηχανικής μάθησης. Η συγχώνευση των περιγραφέων λειτούργησε αποτελεσματικά και διαφάνηκε ότι όσο μεγαλύτερος είναι ο αριθμός των περιγραφέων που συγχωνεύονται, τόσο βελτιώνεται η ακρίβεια, καθώς διαφορετικοί περιγραφείς συλλαμβάνουν διαφορετικά χαρακτηριστικά των εικόνων.

Κεφάλαιο 3

Ταξινόμηση Περιοχών Εικόνων με χρήση Γνώσης

3.1 Εισαγωγή

Η κατανόηση του οπτικού περιεχομένου των εικόνων παραμένει ένα από τα πιο σημαντικά προβλήματα στο ερευνητικό πεδίο της ανάλυσης πολυμεσικού περιεχομένου. Ο συνεχώς αυξανόμενος όγκος πολυμεσικού υλικού καθιστά αναγκαία την ανάλυσή του, προκειμένου αυτό να μπορέσει να αξιοποιηθεί κατάλληλα, καθώς έτσι θα διευκολυνθεί η πρόσβαση και η ανάκτησή του. Για το σκοπό αυτό, τα τελευταία χρόνια έχουν αυξηθεί σημαντικά οι ερευνητικές προσπάθειες που έχουν ως σκοπό την αυτόματη εξαγωγή περιγραφών του πολυμεσικού υλικού, οι οποίες επιδιώκουν να βρίσκονται όσο το δυνατόν εγγύτερα στην ανθρώπινη αντίληψη. Προσπαθούν, δηλαδή, να περιγράψουν τις εικόνες και τις ακολουθίες βίντεο όπως θα τις περιέγραφε και ένας ανθρώπινος παρατηρητής. Ο απώτερος σκοπός είναι να γεφυρωθεί το σημασιολογικό κενό που χαρακτηρίζει την απόσταση ανάμεσα στις χαμηλού επιπέδου περιγραφές που μπορούν να εξαχθούν αυτόματα, με τις περιγραφές υψηλού επιπέδου που χαρακτηρίζουν τη σημασιολογία του περιεχομένου.

Πολλές είναι οι εφαρμογές που δύνανται να ωφεληθούν από την αυτόματη ανάλυση πολυμεσικού υλικού. Για παράδειγμα, μπορεί να διευκολυνθεί σημαντικά η πρόσβαση σε βάσεις εικόνων, να βελτιωθεί η ανάλυση σε συστήματα παρακολούθησης, το φιλτράρισμα του πολυμεσικού περιεχομένου με βάση τις προτιμήσεις του χρήστη, υπηρεσίες κωδικοποίησης και δημιουργίας περιλήψεων σε βίντεο, βελτίωση στη διαδραστικότητα ανθρώπου-υπολογιστή κ.ο.κ.

Το πρόβλημα που επιδιώκει να αντιμετωπίσει η τεχνική που θα παρουσιαστεί στο Κεφάλαιο αυτό, διαφέρει αρκετά από την τεχνική που παρουσιάστηκε στο Κεφάλαιο 2. Μέχρι τώρα, η συγχώνευση των περιγραφέων έγινε για καθολικά χαρακτηριστικά της εικόνας, με σκοπό την χρήση της περιγραφής αυτής σε πρόβλημα ταξινόμησης σκηνής. Αντίθετα, στο παρόν Κεφάλαιο θα χρησιμοποιηθούν οι ιδέες της συγχώνευσης των περιγραφέων και θα εφαρμοστούν σε πρόβλημα ταξινόμησης περιοχής με βάση τα οπτικά της χαρακτηριστικά. Η γνώση που θα χρησιμοποιηθεί για το σκοπό αυτό θα αποθηκευθεί με τη μορφή *πρωτοτύπων* σε μια κατάλληλη δομή *οντολογιών*. Επίσης, θα χρησιμοποιηθούν απλές χωρικές σχέσεις, με σκοπό να διαχωριστούν έννοιες με παρόμοια οπτικά χαρακτηριστικά, τα οποία όμως βρίσκονται συνήθως σε διαφορετικές χωρικές τοποθεσίες στις εικόνες.

3.2 Περιγραφή του Προβλήματος

Οι προσεγγίσεις που χρησιμοποιούνται για την επίλυση του προβλήματος της ταξινόμησης με χρήση της γνώσης μπορούν γενικά να καταταχθούν σε δύο μεγάλες κατηγορίες. Αυτές είναι από τη μια μεριά οι τεχνικές που βασίζονται στα δεδομένα (data-driven) και από την άλλη αυτές που βασίζονται στη γνώση (knowledge-driven). Η διαφοροποίησή τους αυτή γίνεται με βάση την "κατεύθυνση" που ακολουθεί η κατασκευή των περιγραφών υψηλού επιπέδου. Οι τεχνικές που εντάσσονται στην πρώτη κατηγορία, κατά κύριο λόγο ακολουθούν το κλασικό υπολογιστικό παράδειγμα, δηλαδή κατασκευάζεται μια αντικειμενική συνάρτηση με βάση τα παραδείγματα και για μια νέα είσοδο, η απόκριση θα εξαρτάται από την τιμή της. Δεν κατασκευάζεται καμιά ιεραρχία και καμία ενδιάμεση ερμηνεία της εισόδου. Αντίθετα, οι τεχνικές που εντάσσονται στη δεύτερη κατηγορία δίνουν έμφαση σε ενδιάμεσα επίπεδα περιγραφής. Στηρίζονται στην άποψη ότι το πρόβλημα της όρασης υπολογιστών δεν μπορεί να περιέχει μόνο ένα βήμα για την εύρεση και κατανόηση της αντιστοιχίας ανάμεσα στα χαρακτηριστικά χαμηλού επιπέδου και τις έννοιες υψηλού επιπέδου.

Οι προσεγγίσεις που βασίζονται στα δεδομένα ενεργούν πρώτα εξάγοντας χαμηλού επιπέδου χαρακτηριστικά και έπειτα προσπαθούν να βρουν την αντιστοιχία ανάμεσα σε αυτά και τις έννοιες υψηλού επιπέδου χωρίς να διαθέτουν καμία a priori γνώση, πέρα από αυτή που έχει αυτός που τις αναπτύσσει. Έτσι, οι προσεγγίσεις αυτές επικεντρώνονται στο να συλλέγουν με πλήρως αυτόματο τρόπο περιγραφείς σε μορφή αριθμητικών διανυσμάτων των οπτικών ιδιοτήτων και έπειτα επιδιώκουν να τις αντιστοιχήσουν σε έννοιες, με βάση κριτήρια που κατά κάποιο τρόπο μιμούνται την ανθρώπινη αντίληψη για την οπτική ομοιότητα. Οι προσεγγίσεις αυτές παρουσιάζουν το βασικό μειόνεκτημα ότι αποτυγχανουν να αλληλεπιδράσουν με την αντίληψη των χρηστών, καθώς οι σχέσεις που δημιουργούνται ανάμεσα σε περιγραφείς και έννοιες απέχουν πολύ από την ανθρώπινη λογική και αντίληψη. Συνεπώς, ο μηχανισμός με τον οποίο επιδιώκουν να γεφυρώσουν το σημασιολογικό κενό, τις περισσότερες φορές αντιμετωπίζεται από τους χρήστες σαν ένα "μαύρο κουτί". Στην κατηγορία αυτή ανήκουν συστήματα ανάκτησης που βασίζονται σε ερωτήματα με τη χρήση παραδειγμάτων, αλλά και συστήματα ανάκτησης πολυμεσικού υλικού μέσω λέξεων-κλειδιών. Παρόλο που οι προσεγγίσεις αυτές είναι αποδοτικές σε καλά ορισμένα, εώς και αυστηρά περιορισμένα θεματικά πεδία, παρουσιάζουν ιδιαίτερη δυσκολία να προσαρμοστούν σε νέα θεματικά πεδία.

Από την άλλη μεριά, οι προσεγγίσεις που βασίζονται σε γνώση, χρησιμοποιούν υψηλού επιπέδου γνώση σε επίπεδο θεματικού πεδίου, με σκοπό να εξάγει κατάλληλες περιγραφές του πολυμεσικού υλικού. Οι προσεγγίσεις αυτές σχηματίζουν ένα αυτόνομο ερευνητικό πεδίο, προσπαθώντας να συνδυάσουν και να ωφεληθούν από τις ερευνητικές περιοχές της όρασης υπολογιστών, της επεξεργασίας σήματος και της τεχνητής νοημοσύνης για να επιτύχουν την αυτόματη εξαγωγή της σημασιολογίας του οπτικού περιεχομένου, διαμέσου της εφαρμογής της γνώσης και της ευφυίας. Πιο συγκεκριμένα, ο σκοπός τέτοιων προσεγγίσεων πολυμεσικής ανάλυσης είναι να εξάγουν τις εμπειρίες των χρηστών όσον αφορά το πώς αντιλαμβάνονται την οπτική πληροφορία που λαμβάνουν μέσα από υπολογιστικά μοντέλα. Με άλλα λόγια, επιδιώκουν να μειώσουν τον όγκο των πολυτροπικών δεδομένων σε σύντομες και περιεκτικές περιγραφές που θα μπορούν να συλλαμβάνουν το περιεχόμενο του πολυμεσικού περιεχομένου. Η σχετική βιγλιογραφία μπορεί γενικά να χωριστεί σε δύο κατηγορίες από προσεγγίσεις, με βάση τον τρόπο με τον οποίο αποκτάται και αναπα-

ρίσταται η γνώση. Η πρώτη από αυτές περιλαμβάνει τις τεχνικες στις οποίες η γνώση καθορίζεται από αυστηρά μοντέλα και η δεύτερη τις τεχνικές μηχανικής μάθησης που η γνώση αποκτάται έμμεσα, μέσω παραδειγμάτων.

Το κύριο χαρακτηριστικό των προσεγγίσεων που βασίζονται σε μάθηση είναι η ικανότητά τους να μεταβάλλουν την εσωτερική τους δομή σύμφωνα με τα ζεύγη εισόδου και επιθυμητής εξόδου, με σκοπό να προσεγγίσουν τους κανόνες και τις σχέσεις που υποβόσκουν στο σύνολο εκπαίδευσης, εξομοιώνοντας κατά κάποιο τρόπο μια διαδικασία συλλογιστικής. Συνεπώς, η χρήση τεχνικών μηχανικής μάθησης για την αντιστοίχηση χαμηλού σε υψηλού επιπέδου χαρακτηριστικά αποτελεί μια ισχυρή μεθοδο για την ανακάλυψη πολύπλοκων και κρυφών σχέσεων και έτσι ένας πολύ μεγάλος αριθμός από εφαρμογές έχει προταθεί. Ορισμένες από τις πιο σημαντικές από αυτές θα παρουσιαστούν στην Ενότητα 3.3, όπου και θα καταστεί επιπλέον σαφές ότι οι τεχνικές μηχανικής μάθησης όπως τα νευρωνικά δίκτυα, τα ασαφή συστήματα, οι μηχανές διανυσμάτων υποστήριξης, τα στατιστικά μοντέλα και η συλλογιστική βρίσκονται ανάμεσα στις τεχνικές που έχουν ευρέως χρησιμοποιηθεί στην περιοχή της ταξινόμησης περιοχών, αντικειμένων και σκηνών. Ωστόσο, η λειτουργία των μεθόδων αυτών ως μαύρα κουτιά συχνά καθιστά δύσκολη την αποτελεσματικότητά τους, μιας και αυτή εξαρτάται από ένα πλήθος παραμέτρων που πρέπει να επιλεχθούν από το χρήστη τους, ο οποίος αν δεν είναι εξοικειωμένος τόσο θεωρητικά, όσο και πρακτικά με αυτές, πρέπει να καταφύγει σε μια χρονοβόρα διαδικασία πειραμάτων. Επίσης, προκειμένου οι τεχνικές αυτές να δουλέψουν σωστά σε πρακτικά προβλήματα, απαιτούνται μεγάλα και προσεκτικά κατασκευασμένα σύνολα δεδομένων. Τέλος, όταν κατασκευαστούν οι ταξινομητές, η χρήση τους περιορίζεται συγκεκριμένα στο θεματικό πεδίο για το οποίο σχεδιάστηκαν και η προσαρμογή τους σε νέα πεδία συχνά είναι δύσκολη και χρονοβόρα. Το ίδιο ισχύει και στην περίπτωση που ο χρήστης θέλει να εισάγει νέες λειτουργίες, π.χ. νέες έννοιες προς ταξινόμηση, ή να αυξήσει το σύνολο εκπαίδευσης.

Ακολουθώντας μια εναλλακτική μεθοδολογία, οι μέθοδοι που βασίζονται σε μοντέλα χρησιμοποιούν a priori γνώση με τη μορφή αυστηρά καθορισμένων μοντέλων, κανόνων και περιορισμών. Οι μέθοδοι αυτές προσπαθούν να γεφυρώσουν το σημασιολογικό κενό χρησιμοποιώντας μια συνήθως ιεραρχική αναπαράσταση των αντικειμένων, των γεγονότων, των σχέσεων, των ιδιοτήτων κ.ο.κ. του υπό εξέταση θεματικού πεδίου. Οι όροι που περιέχονται στα μοντέλα που χρησιμοποιούνται για την αναπαράσταση γνώσης (οντολογίες, σημασιολογικά δίκτυα κ.α.), περιέχουν νόημα που σχετίζεται ευθέως με τις "οπτικές" έννοιες που περιέχονται στις εικόνες. Έτσι παρέχουν ένα μοντέλο του θεματικού πεδίου το οποίο μπορεί να υποστηρίξει τη συλλογιστική. Παρολαυτά, η πολυπλοκότητα των συστημάτων αυτών αυξάνει εκθετικά όσο αυξάνονται οι όροι και οι μεταξύ τους σχέσεις, περιορίζοντας την εφαρμογή τους σε περιπτώσεις όπου οι εικόνες περιέχουν μόνο ένα μικρό αριθμό από έννοιες. Σαν αποτέλεσμα αυτού, στις περισσότερες προσεγγίσεις που βασίζονται σε μοντέλα, οι έννοιες πρώτα αναγνωρίζονται χωρίς να χρησιμοποιήσουν τη γνωση του μοντέλου που περιγράφει το θεματικό πεδίο. Έπειτα, με βάση τη γνώση η αρχική υπόθεση ενδεχομένως να αλλάξει, αφού αξιοποιηθούν για παράδειγμα οι διάφορες σχέσεις που το μοντέλο περικλείει.

Συμπερασματικά, η κατανόηση της σημασιολογία του οπτικού περιεχομένου αποτελεί το στόχο για όλες τις τεχνικές ανάλυσης πολυμεσικού υλικού. Οι δυσκολίες που συναντά κανείς σε όλα τα προβλήματα έχουν να κάνουν κατά κύριο λόγο με τον τρόπο με τον οποίο θα γίνει η αντιστοίχηση ανάμεσα στα χαρακτηριστικά χαμη-

λού επιπέδου που εξάγονται αυτόματα από το υλικό και τις έννοιες υψηλού επιπέδου που είναι τελικά αυτές που γίνονται κατανοητές από τους χρήστες. Το Κεφάλαιο αυτό ασχολείται ερευνητικά με τη σημασιολογική ανάλυση εικόνων. Η παρουσίαση των σχετικών εργασιών στην Ενότητα 3.3 φιλοδοξεί να δώσει μια γενική εικόνα των ερευνητικών προσπαθειών που αντιμετωπίζουν τα προβλήματα που αναφέρθηκαν παραπάνω.

3.3 Σχετικές Εργασίες

Στη βιβλιογραφία έχουν προταθεί πολλές μέθοδοι για την εξαγωγή χαμηλού επιπέδου περιγραφών οι οποίες αποσκοπούν στην εξαγωγή χαμηλού επιπέδου χαρακτηριστικών που περιγράφουν το χρώμα, την υφή και το σχήμα. Οι διάφορες προσεγγίσεις που έχουν προταθεί συνδυάζουν αναπαράσταση, διαχείριση γνώσης καθώς και υλοποίηση κάποιας μορφής συλλογιστικής. Για το σκοπό αυτό έχουν χρησιμοποιηθεί τεχνικές όπως νευρωνικά δίκτυα, έμπειρα συστήματα, ασαφή συστήματα, οντολογίες, δέντρα αποφάσεων, στατικά και δυναμικά μπεϋσιανά δίκτυα, μαρκοβιανά τυχαία πεδία και άλλες. Οι τεχνικές αυτές χρησιμοποιούνται για να αποθηκεύσουν γνώση είτε μέσω κάποιας διαδικασίας μάθησης, είτε γνώση που έχει προκύψει από την εμπειρία σχετικά με το θεματικό πεδίο στο οποίο εφαρμόζεται η τεχνική, είτε απλά έναν αριθμό από παραδείγματα κλάσεων ή εννοιών υψηλού επιπέδου.

Στις στοχαστικές προσεγγίσεις ανήκει η μεθοδολογία που προτάθηκε από τους Naphade et al. [150], όπου το πρόβλημα της γεφύρωσης του σημασιολογικού κενού αντιμετωπίστηκε σαν ένα πιθανοτικό πρόβλημα αναγνώρισης προτύπων. Ένας γράφος παραγοντοποίησης (factor graph network) από πιθανοτικά πολυμεσικά αντικείμενα (multijects) ορίζεται με τη χρήση κρυφών μαρκοβιανών μοντέλων (HMM) και μοντέλων μίξης γκαουσιανών (GMM). Τα HMMs συνδυάζονται με κανόνες στο COBRA μοντέλο, που περιγράφεται από τους Petkovic και Jonker [169], όπου αντικείμενα και περιγραφές γεγονότων τυποποιούνται με τη χρήση κατάλληλων γραμματικών. Ταυτόχρονα, η στοχαστική προσέγγιση παρέχει την υποστήριξη για οπτικές δομές που είναι πολύ περίπλοκες για να οριστούν αυστηρά. Ένα ιεραρχικό μοντέλο που βασίζεται σε Τυχαία Μαρκοβιανά Μοντέλα (MRF) χρησιμοποιείται από τους Kato et al. [98] για ταξινόμηση εικόνας με μη επιβλεπόμενη μάθηση.

Οι Chapelle et al. [44] αντιμετωπίζουν το πρόβλημα της ταξινόμησης σκηνής χρησιμοποιώντας SVM. Εξάγουν χρωματικές περιγραφές με τη μορφή ιστογράμματος από τις εικόνες και τα SVM αποδεικνύεται ότι αποδίδουν καλύτερα από άλλες τεχνικές μηχανικής μάθησης. Οι Bose και Grimson [20] ασχολούνται με το πρόβλημα της ταξινόμησης αντικειμένων και χρησιμοποιούν έναν αλγόριθμο "bootstrapping". Οι Wang και Manjunath [229] αρχικά χρησιμοποιούν ένα SVM για την αναπαράσταση των καταστάσεων των κατανομών των διανυσμάτων χαρακτηριστικών που αντιστοιχούν στις έννοιες υψηλού επιπέδου και έπειτα ένα MRF για τη μοντελοποίηση των χωρικών κατανομών των εννοιών. Μέσω της διαδικασίας αυτής, σε κάθε περιοχή αντιστοιχούν τελικά μια έννοια. Ωστόσο, πολλές φορές περισσότερες από μία έννοιες μπορεί να αντιστοιχούν στην ίδια περιοχή ή στην εικόνα καθολικά. Για να αντιμετωπίσουν τέτοιου είδους προβλήματα, οι Li et al. [119] προτείνουν τη χρήση ενός SVM πολλών κατηγοριών. Έτσι, κάθε εικόνα μπορεί να χαρακτηριστεί καθολικά με περισσότερες από μία έννοιες.

Οι Marques και Barman [132] προτείνουν τη χρήση μιας αρχιτεκτονικής 3 επι-

πέδων. Το πρώτο επίπεδο πραγματοποιεί την εξαγωγή της οπτικής πληροφορίας από εικόνες. Στο δεύτερο επίπεδο η πληροφορία αυτή αντιστοιχίζεται με τη χρήση τεχνικών μηχανικής μάθησης σε λέξεις-κλειδιά, οι οποίες τελικά συνδέονται με οντολογίες στο τρίτο επίπεδο. Οι Giro και Marques [74] προτείνουν μια μεθοδολογία για την ανίχνευση αντικειμένων που ανήκουν σε προκαθορισμένες κατηγορίες. Οι κατηγορίες ορίζονται με τη μορφή ενός γράφου, ο οποίος περιλαμβάνει οπτική και δομική γνώση σε σχέση με τις έννοιες που περιέχει. Οι έννοιες αυτές στη συνέχεια οργανώνονται περαιτέρω, με τη χρήση ενός δυαδικού δέντρου. Ένα άλλο παράδειγμα ημι-αυτόματου αλγορίθμου, ο οποίος κατασκευάστηκε για να εφαρμοστεί σε συγκεκριμένο θεματικό πεδίο, παρουσιάστηκε από τους Burghardt et al. [33], όπου ένας οπτικός παρακολουθητής (tracker) *προσώπου ζώων* εκπαιδεύεται με παραδείγματα σχολιασμένα από χρήστες. Ένας ταξινομητής Adaboost εκπαιδεύεται και χρησιμοποιείται η τεχνική οπτικής παρακολούθησης των Kanade-Lucas-Tomasi. Στην εργασία των Shrihari et al. [201] παρουσιάζεται ένα ημιαυτόματο σύστημα που χρησιμοποιεί "ενδείξεις" από χρήστες που δίνονται σε φυσική γλώσσα. Σκοπός είναι να περιοριστεί ο χώρος στον οποίο θα πραγματοποιηθεί η αναζήτηση των αλγορίθμων ανίχνευσης. Για παράδειγμα, ο χρήστης μπορεί να δώσει ενδείξεις όπως "στην πάνω δεξιά γωνία της εικόνας υπάρχει ένα κτίριο σε σχήμα L". Έτσι, το σύστημα χρησιμοποιεί χωρικούς περιορισμούς για να περιορίσει την περιοχή που θα ψάξει για ένα αντικείμενο, καθώς και άλλους περιορισμούς όσον αφορά το σχήμα του.

Ο Kouzani [104] παρουσίασε ένα ευφυές σύστεμα για την εύρεση *ανθρωπίνων προσώπων* σε εικόνες. Το σύστημα αυτό αποτελείται από τρία στάδια. Το πρώτο εκτελεί την προεπεξεργασία, το δέυτερο εξάγει τα συστατικά μέρη των προσώπων, με βάση τα χαρακτηριστικά χαμηλού επιπέδου που εξήχθησαν στο πρώτο, ενώ το τρίτο πραγματοποιεί την τελική απόφαση. Τα συστατικά μέρη εξάγονται με τη χρήση ενός νευροασαφούς δικτύου, ενώ για την τελική απόφαση χρησιμοποιείται μια βάση γνώσης, με την οποία αξιολογούνται τα μέρη που εξήχθησαν. Οι Antunes et al. [3] παρουσίασαν μια μεθοδολογία ταξινόμησης που βασίζεται σε ασαφείς κανόνες. Χρήστες που θεωρούνται ειδικοί του συγκεκριμένου θεματικού πεδίου ορίζουν συγκεκριμένους κανόνες μέσω ενός γραφικού περιβάλλοντος. Το σύστημα χρησιμοποιεί τους κανόνες αυτούς και δημιουργεί αυτόματα σχολιασμούς για τις εικόνες. Μια άλλη προσέγγιση που χρησιμοποιεί συλλογιστική με ασαφείς κανόνες ακολουθείται από τους Dorado και Izquierdo [60] για την ταξινόμηση εικόνων που περιέχουν *κτίρια*. Η αναπαράσταση γνώσης βασίζεται σε ένα μοντελο ασαφούς συλλογιστικής που προσπαθεί να γεφυρώσει τα οπτικά χαρακτηριστικά που εξάγονται και τις έννοιες που απεικονίζουν. Οι Sprague και Luo [199] πρότειναν ένα σύστημα που εκπαιδεύεται να αναγνωρίζει *ντυμένους ανθρώπους*. Ένας γράφος κατασκευάζεται με τους κόμβους του να απεικονίζουν πιθανά μέρη του ανθρώπινου σώματος και τις ακμές του να αντιστοιχούν σε κατανομές των διατάξεων αυτών των μερών. Ο ταξινομητής προσαρμόζεται αυτόματα σε μια σκηνή μαθαίνοντας να αξιοποιεί τα χαρακτηριστικά του εννοιολογικού της πλαισίου. Μια παρόμοια προσέγγιση ακολουθείται στο από του Chopra και Shrihari [47], όπου οι διάφοροι περιορισμοί που αντιστοιχούν στις σχέσεις μεταξύ των εννοιών δημιουργούνται από μια διαδικασία επεξεργασίας φυσικής γλώσσας που εφαρμόζεται στο κείμενο που συνοδεύει την εικόνα.

Μια μέθοδος για την ταξινόμηση εικόνων που βασίζεται σε γνώση που έχει αποκτηθεί από σχολιασμένες εικόνες χρησιμοποιώντας το WordNet [68] παρουσιάστηκε από τους Benitez και Chang [11]. Η αυτόματη κατηγοριοποίηση μέσω του συνδυασμού των ταξινομητών πραγματοποιείται με χρήση εξαχθείσας γνώσης, η οποία αποτε-

λείται από το δίκτυο των εννοιών μαζί με την εικόνα που προκύπτει και παραδείγματα κειμένου. Η προσέγγιση της αυτόματης εξαγωγής σχολιασμών εικόνας μέσω της συσχέτισης λέξεων με εικόνες έχει αποτελέσει αντικείμενο και άλλων ερευνητικών προσπαθειών, όπως αυτές που παρουσιάζονται για παράδειγμα από τους Tansley et al. [210], οι οποίοι κατασκεύασαν ένα σύστημα το οποίο αρχικά ταξινομεί τις έννοιες και έπειτα αξιοποιεί ήδη υπάρχουσα γνώση. Σε παρόμοιο πλαίσιο κινήθηκαν και οι Lavrenko et al. [108], οι οποίοι κατασκεύασαν ένα πιθανοτικό μοντέλο.

Ακολουθώντας τις εξελίξεις του Σημασιολογικού Ιστου (Semantic Web), πολλές προσεγγίσεις έχουν προκύψει, οι οποίες χρησιμοποιούν τις οντολογίες σαν το μέσο αναπαράστασης της γνώσης που είναι διαθέσιμη για το υπό εξέταση θεματικό πεδίο και αξιοποιούν την αυστηρή σημασιολογική αναπαράσταση για να πραγματοποιήσουν συλλογιστική υψηλού επιπέδου. Οι Hudelot και Thonnat [86] παρουσίασαν μια πλατφόρμα που βασίζεται σε οντολογίες, για την αυτόματη αναγνώριση φυσικών και ταυτόχρονα πολύπλοκων εννοιών. Τρία κατανεμημένα συστήματα που βασίζονται στη γνώση πραγματοποιούν την επεξεργασία της εικόνας, την αντιστοίχηση των χαμηλού επιπέδου χαρακτηριστικών σε συμβολικά δεδομένα και τη διαδικασία σημασιολογικής ερμηνείας. Μια παρόμοια προσέγγιση ακολουθήθηκε και από τους Hunter et al. [88], όπου σημασιολογικές περιγραφές εικόνων βασισμένες σε οντολογίες, δημιουργούνται από κατάλληλα καθορισμένους κανόνες που σχετίζουν MPEG-7 περιγραφείς με έννοιες που περιγράφονται στην οντολογία FUSION. Οι Little και Hunter [123] ακολούθησαν μιαπροσέγγιση που χρησιμοποιεί αφενός καθορισμένους κανόνες και αφετέρου επιπρόσθετη βοήθεια από χρήστες. Οι Wallace et al. [226] χρησιμοποίησαν ασαφή άλγεβρα και ασαφή οντολογική πληροφορία, για την εξαγωγή σημασιολογικής πληροφορίας, προκειμένου να τις εφαρμόσουν στο πρόβλημα της θεματικής κατηγοριοποίηση.

Αρκετές είναι και οι προσεγγίσεις που χρησιμοποιούν οντολογίες για την αναπαράσταση γνώσης και πραγματοποιούν ταξινόμηση εικόνας. Για παράδειγμα, οι Wang et al. [228] κατασκεύασαν μια δομή για την αποθήκευση της γνώσης η οποία βασίστηκε σε οντολογίες ανεξάρτητες του θεματικου πεδίου στο οποίο εφαρμόζεται η τεχνική. Οι οντολογίες κατασκευάστηκαν έτσι ώστε να περιέχουν τις περιγραφές από αντικείμενα. Με βάση την γνώση που έχει αποθηκευθεί σε αυτές, οι περιοχές της εικόνας ενώνονται προκειμένου να σχηματιστούν αντικείμενα σαν αυτά που περιγράφονται. Επίσης, οι Breen et al. [31] πρότειναν μια τεχνικη για την ταξινόμηση εικόνων που βασίστηκε σε νευρωνικά δίκτυα και χρησιμοποίησε οντολογίες για την περιγραφή των εννοιών και των σχέσεων μεταξύ τους. Τελος, στην εργασία των Meghini et al. [135], το πρόβλημα της ενσωμάτωσης της σημασιολογίας σε οπτικά δεδομένα προσεγγίζεται με την εισαγωγή ενός μοντέλου δεδομένων που βασίζεται σε Περιγραφικές Λογικές, οι οποίες περιγράφουν τόσο τη δομή, όσο και το περιεχόμενο των πολυμεσικών εγγράφων, επιτρέποντας τόσο δομικά, όσο και εννοιολογικά ερωτήματα.

Παρά τις προσπάθεις που γίνονται στο χώρο της κατανόησης του σημασιολογικού περιεχομένου των εικόνων, η ακρίβεια που επιτυγχάνεται με τις παρούσες τεχνικές παραμένει υποδεέστερη των προσδοκιών που έχουν οι χρήστες και ταυτόχρονα διατηρεί υψηλά επίπεδα πολυπλοκότητας. Παρότι ένας ιδιαίτερα σημαντικός αριθμός από εργασίες με ικανοποιητικά αποτελέσματα έχει αναφερθεί, το πρόβλημα της κατανόησης εξακολουθεί να παραμένει επί της ουσίας άλυτο. Αυτό συμβαίνει καθώς οι περισσότερες τεχνικές αποφεύγουν να ερευνήσουν γενικές στρατηγικές για την επίλυση διαφορετικών προβλημάτων και εστιάζουν τελικά σε εφαρμογές που στοχεύουν σε

συγκεκριμένο θεματικό πεδίο και εκμεταλλεύονται τις ιδιαιτερότητές του και τους κανόνες ή περιορισμούς που εξάγονται από αυτό [193].

3.4 Ταίριασμα Περιοχών Εικόνων

Στην Ενότητα αυτή παρουσιάζεται η προτεινόμενη τεχνική που ασχολείται με το πρόβλημα της ταξινόμησης περιοχών εικόνων, με χρήση γνώσης. Αρχικά παρουσιάζεται ο τρόπος με τον οποίο γίνεται η κατασκευή και η αποθήκευση της γνώσης. Ακολουθεί η τεχνική με την οποία γίνεται το ταίριασμα ανάμεσα σε περιοχές των εικόνων και σε παραδείγματα που έχουν αποθηκευθεί στη βάση. Τέλος, δημιουργούνται μεταδεδομένα, που σκοπό έχουν το διαμοιρασμό των αποτελεσμάτων της ανάλυσης.

3.4.1 Αναπαράσταση Γνώσης

Ανάμεσα στους πιθανούς τρόπους αναπαράστασης και αποθήκευσης της γνώσης, στο πλαίσιο που θα αναπτυχθεί στο κεφάλαιο αυτό έχει επιλεγεί η αναπαράσταση με τη μορφή *οντολογίας* [202]. Οι οντολογίες παρουσιάζουν σημαντικά πλεονεκτήματα, καθώς υποστηρίζουν σαφείς σημασιολογικούς ορισμούς και διευκολύνουν τη συλλογιστική. Μέσω αυτών, διευκολύνεται ακόμη και η εξαγωγή νέας γνώσης, η οποία βασίζεται σε ήδη υπάρχουσα καθώς και σε κανόνες. Έτσι, εμφανίζονται ιδανικές για την αναπαράσταση γνώσης που έχει εξαχθεί από πολυμεσικό υλικό, επιτρέποντας την αυτόματη ανάλυση και την επεξεργασία της εξαχθείσας σημασιολογικής πληροφορίας.

Το σχήμα RDFS είναι μια απλή γλώσσα μοντελοποίησης σε πιο υψηλό επίπεδο από την RDF [32][1]. Και οι δύο γλώσσες εξελίσσονται από το W3C[2]. Πρέπει να σημειωθεί ότι θα μπορούσε να χρησιμοποιηθεί και η OWL [134], η οποία είναι μια γλώσσα που βασίζεται στις περιγραφικές λογικές και επίσης εξελίσσεται από το W3C. Η OWL έχει σχεδιαστεί να χρησιμοποιείται από εφαρμογές που χρειάζονται αυξημένη εκφραστικότητα σε σχέση με την RDFS. Αυτό το επιτυγχάνει με το να παρέχει επιπλέον λεξιλόγιο μαζί με αυστηρή σημασιολογία. Ωστόσο, για τις ανάγκες της παρούσας εργασίας, η RDFS επιλέχτηκε για τη μοντελοποίηση, καθώς η προτεινόμενη συλλογιστική δε θα μπορούσε να αξιοποιήσει την αυξημένη εκφραστικότητα που δύναται να παρέχει η OWL.

Η προτεινόμενη αναπαράσταση γνώσης με χρήση μια υποδομής που βασίζεται σε οντολογίες και αποτελείται από:

1. Μια *Οντολογία Πυρήνα* (Core Ontology), ο ρόλος της οποίας είναι να αποτελέσει ένα σημείο αναφοράς για την κατασκευή νέων οντολογιών.

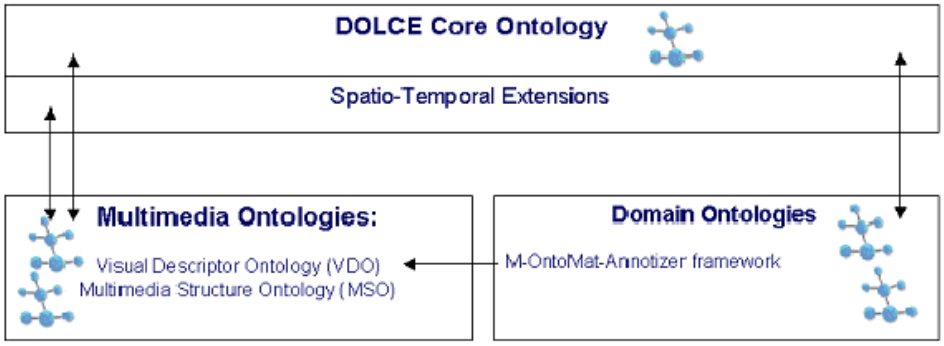
2. Μια *Οντολογία Οπτικών Περιγραφέων* (VDO) που κατασκευάστηκε από τους Simou et al. [186], η οποία περιέχει τις αναπαραστάσεις των MPEG-7 οπτικών περιγραφέων.

3. Μια *Οντολογία Δομής Πολυμέσων* (MSO), η οποία μοντελοποιεί τις βασικές πολυμεσικές οντότητες του MPEG-7 Σχήματος Περιγραφής Πολυμέσων (MDS)[12].

[1] Η RDF δεν αποτελεί σύστημα αναπαράστασης γνώσης αλλά προσπαθεί να βελτιώσει τη διαδραστικότητα μεταξύ των δεδομένων στο WWW. Αυτό επιτυγχάνεται εξειδικεύοντας το μοντέλο της XML μέσω ενός μοντέλου που βασίζεται σε αναπαράσταση ενός γράφου.

[2] WWW Consortium - `http://w3c.org`

4. Τις απαραίτητες *Οντολογίες Θεματικού Πεδίου* (Domain Ontologies), οι οποίες μοντελοποιούν το επίπεδο περιεχομένου του πολυμεσικού υλικού, με βάση τις ιδιότητες του αντίστοιχου θεματικού πεδίου που συναντώνται στον πραγματικό κόσμο.



Σχήμα 3.1: *Η προτεινόμενη υποδομή των οντολογιών που χρησιμοποιείται για την αναπαράσταση της γνώσης. Αποτελείται από την οντολογία πυρήνα, την οντολογία οπτικών περιγραφέων, την οντολογία δομής πολυμέσων και τις οντολογίες θεματικού πεδίου.*

3.4.1.1 Οντολογία Πυρήνα

Γενικά, οι Έννοιες[3] που περιέχονται στις οντολογίες πυρήνα περιέχουν χαρακτηριστικά που δεν έχουν εξάρτηση από κάποιο θεματικό πεδίο και σχέσεις βασιζόμενες σε αυστηρές αρχές που προκύπτουν από τη φιλοσοφία, τα μαθηματικά, τη γλωσσολογία και την ψυχολογία. Ο ρόλος της οντολογίας πυρήνα είναι να εξυπηρετήσει την κατασκευή νέων οντολογιών σαν σημείο αναφοράς, να διευκολύνει συγκρίσεις ανάμεσα σε διαφορετικές προσεγγίσεις και να χρησιμεύσει σαν "γέφυρα" ανάμεσα σε υπάρχουσες οντολογίες. Στο πλαίσιο που παρουσιάζεται, ως οντολογία πυρήνα έχει επιλεχθεί η οντολογία *DOLCE* [71], η οποία έχει σχεδιαστεί αποκλειστικά για χρήση ως οντολογία πυρήνα. Η σχεδίασή της είναι μινιμαλιστική με την έννοια ότι περιέχει αποκλειστικά τις πιο επαναχρησιμοποιήσιμες και εφαρμόσιμες κατηγορίες ανωτάτου επιπέδου, και αξιώματα που έχουν ερευνηθεί και τεμηριωθεί εκτενώς.

Παρότι η οντολογία πυρήνα DOLCE παρέχει τον τρόπο για την αναπαράσταση χωροχρονικών ποσοτήτων, η συλλογιστική με αυτές τις περιγραφές απαιτεί την κωδικοποίηση επιπλέον σχέσεων οι οποίες εκφράζουν τις χωρικές ή/και ποσοτικές σχέσεις μεταξύ περιοχών. Με βάση τις Έννοιες που περιγράφονται από τους Cohn et al. [51] και Allen [1] και τα μοντέλα που προτείνονται από τους Papadias και Theodoridis [165] και Skiadopoulos και Koubarakis [191], η Έννοια Περιοχή που αποτελεί κλαδί της DOLCE επεκτάθηκε για να μπορεί να συμπεριλάβει τοπολογικές και κατευθυντικές σχέσεις μεταξύ περιοχών διαφορετικών τύπων, κυρίως ΧρονικήΠεριοχή και 2ΔΠεριοχή. Οι κατευθυντικές χωρικές σχέσεις περιγράφουν πώς οι περιοχές της εικόνας είναι τοποθετημένες και πώς σχετίζονται μεταξύ τους στον 2-Δ ή στον 3-Δ χώρο (π.χ. *αριστερά και πάνω*). Οι τοπολογικές χωρικές σχέσεις περιγράφουν πώς

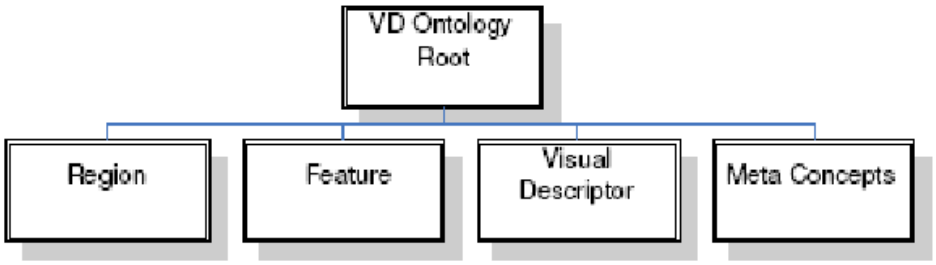
[3]Οι Έννοιες από τις οποίες αποτελείται μια οντολογία θα αναφέρονται με Ε κεφαλαίο, σε αντίθεση με τις έννοιες υψηλού επιπέδου που θα αναφέρονται με ε μικρό.

τα χωρικά όρια των περιοχών σχετίζονται (π.χ. *ακουμπούν*). Με παρόμοιο τρόπο, οι χρονικές σχέσεις μεταξύ περιοχών χρησιμοποιούνται για την αναπαράσταση χρονικών σχέσεων ανάμεσα στις περιοχές, ή γεγονότα.

3.4.1.2 Οντολογία Οπτικών Περιγραφέων

Η Οντολογία Οπτικών Περιγραφέων (VDO) έχει προταθεί από τους Simou et al.[186] και έχει σχεδιαστεί με σκοπό να αναπαραστήσει το οπτικό μέρος του MPEG-7 [43]. Έτσι, περιέχει τις αναπαραστάσεις του συνόλου των οπτικών περιγραφέων που θα χρησιμοποιηθούν στην ανάλυση. Οι μοντελοποιημένες Έννοιες και Ιδιότητες περιγράφουν τα οπτικά χαρακτηριστικά των αντικειμένων. Κατά την κατασκευή της VDO έγινε προσπάθεια να ακολουθηθούν οι προδιαγραφές του οπτικού μέρους του MPEG-7. Κάτι τέτοιο, όμως, κατέστη αδύνατο και για το λόγο αυτό έγιναν διάφορες αναγκαίες τροποποιήσεις με σκοπό να προσαρμοστεί το XML σχήμα του MPEG-7 σε μια οντολογία και στους τύπους δεδομένων που είναι διαθέσιμοι στην RDFS.

Το δέντρο της οντολογίας οπτικών περιγραφέων αποτελείται από τέσσερις κύριες Έννοιες, οι οποίες και είναι `VDO:Region` (περιοχή), `VDO:Feature` (χαρακτηριστικό), `VDO:VisualDescriptor` (οπτικός περιγραφέας) και `VDO:Metaconcepts` (μετα-Έννοιες). Το δέντρο αυτό απεικονίζεται στο Σχήμα 3.2. Οι Έννοιες αυτές δεν περιλαμβάνονται στο XML Σχήμα που ορίζεται από το MPEG-7, αλλά η ύπαρξή τους ήταν απαραίτητη προκειμένου να οριστεί σωστά η οντολογία. Η Έννοια `VDO:VisualDescriptor` περιέχει τους οπτικούς περιγραφείς όπως αυτοί ορίζονται από το MPEG-7. Η Έννοια `VDO:Metaconcepts` περιέχει κάποιες επιπρόσθετες Έννοιες που ήταν απαραίτητες για το σωστό ορισμό της VDO, αλλά δεν έχουν οριστεί ξεκάθαρα στο XML Σχήμα του MPEG-7. Οι άλλες δύο Έννοιες που ορίστηκαν `VDO:Region` και `VDO:Feature`, επίσης δεν περιέχονται στο MPEG-7, αλλά ο ορισμός τους ήταν απαραίτητος για να επιτραπεί η αντιστοίχηση των οπτικών περιγραφέων σε περιοχές εικόνων.



Σχήμα 3.2: *Η Οντολογία Οπτικών Περιγραφέων (VDO). Απεικονίζονται οι Έννοιες που αυτή περιέχει.*

Για παράδειγμα, η Έννοια `VDO:VisualDescriptor` αποτελείται από έξι Υποέννοιες, μία για κάθε κατηγορία των οπτικών περιγραφέων που καθορίζονται από το πρότυπο MPEG-7 καί είναι: *χρώμα*, *υφή*, *σχήμα*, *κίνηση*, *εντοπισμός* και *βασικοί περιγραφείς*. Κάθε μία από αυτές τις Υποέννοιες συμπεριλαμβάνει έναν αριθμό από σχετικούς περιγραφείς, οι οποίοι ορίζονται σαν Έννοιες στην VDO. Σε σχέση με το MPEG-7, μόνο η κατηγορία `VDO:BasicDescriptors` έχει τροποποιηθεί και δεν περιλαμβάνει όλους τους περιγραφείς του προτύπου.

3.4.1.3 Οντολογία Δομής Πολυμέσων

Η Οντολογία Δομής Πολυμέσων (MSO) μοντελοποιεί βασικές πολυμεσικές οντότητες του Σχήματος Περιγραφής Πολυμέσων (MDS) του MPEG-7 [12] και τις αμοιβαίες σχέσεις όπως π.χ. η *αποσύνθεση*. Στο πρότυπο MPEG-7, το πολυμεσικό περιεχόμενο κατηγοριοποιείται σε πέντε κλάσεις: *εικόνα*, *βίντεο*, *ήχος*, *οπτικοακουστικό* και *πολυμεσικό*. Κάθε μία από τις παραπάνω κλάσεις έχει τις δικές της υποκλάσεις. Το MPEG-7 παρέχει έναν αριθμό από εργαλεία για την περιγραφή της δομής πολυμεσικού υλικού σε χρονικό και χωρικό επίπεδο. Το Σχήμα Περιγραφής (Description Scheme - DS) Segment περιγράφει ένα χωρικό ή/και χρονικό τμήμα του πολυμεσικού υλικού. Ένας αριθμός από εξειδικευμένες υποκλάσεις προκύπτουν από το γενικότερο Segment DS. Οι υποκλάσεις αυτές περιγράφουν συγκεκριμένους τύπους από πολυμεσικά τμήματα, όπως τμήματα βίντεο, κινούμενες περιοχές, ακίνητες περιοχές και μωσαϊκά, τα οποία προκύπτουν από χωρική, χρονική και χωροχρονική κατάτμηση των διαφόρων τύπων πολυμεσικού υλικού. Τα πολυμεσικά έγγραφα μπορούν να κατατμηθούν ή να αποσυντεθούν σε υπόκλάσεις με τη χρήση τεσσάρων τύπων αποσύνθεσης: *χωρικής*, *χρονικής*, *χωροχρονικής* και *πηγής πολυμέσων*.

3.4.1.4 Οντολογίες Θεματικού Πεδίου

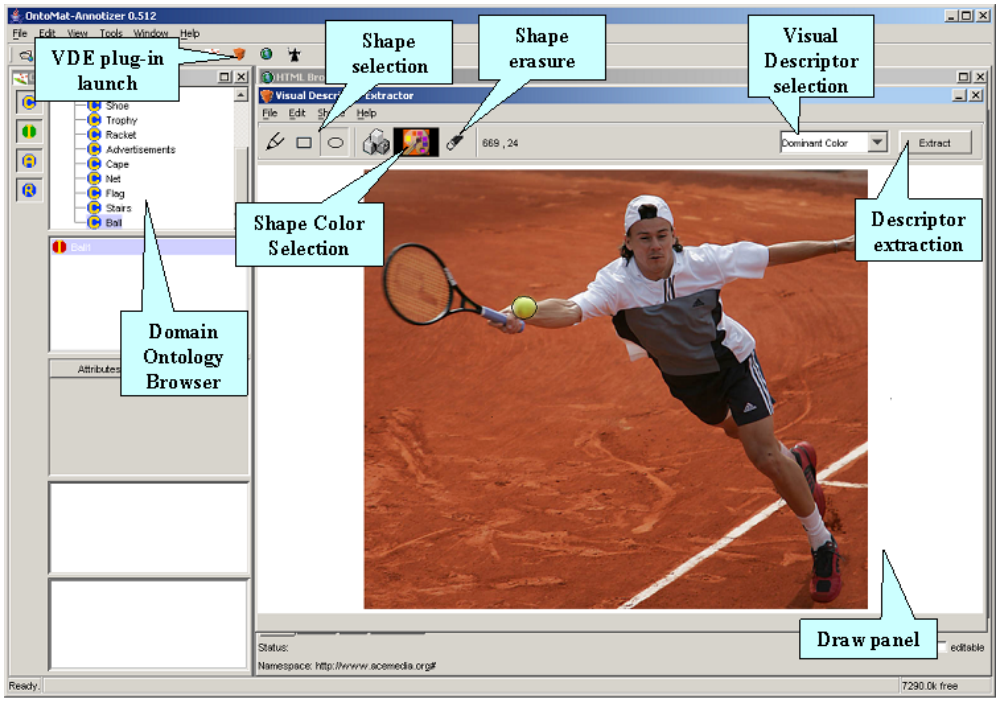
Οι οντολογίες θεματικού πεδίου μοντελοποιούν το επίπεδο περιεχομένου του πολυμεσικού υλικού με βάση συγκεκριμένα θεματικά πεδία του πραγματικού κόσμου, όπως για παράδειγμα "*Τέννις*" ή "*Παραλία*". Επειδή ως οντολογία πυρήνα έχει επιλεχθεί η οντολογία DOLCE, είναι απαραίτητο όλες οι οντολογίες θεματικού πεδίου να βασίζονται αυστηρά σε αυτή, ή έστω να είναι ευθυγραμμισμένες με αυτή και άρα συνδεδεμένες με τις έννοιες υψηλού επιπέδου. Αυτό εξασφαλίζει διαδραστικότητα ανάμεσα σε διαφορετικές οντολογίες θεματικού επιπέδου.

Στην προτεινόμενη μεθοδολογία, οι οντολογίες θεματικού πεδίου ορίζονται με τέτοιο τρόπο ώστε να παρέχουν ένα γενικό μοντέλο του θεματικού πεδίου, με έμφαση στους χρήστες. Πιο συγκεκριμένα, η ανάπτυξη των οντολογιών έγινε με τέτοιο τρόπο ώστε να διευκολύνεται η ανάκτηση από την πλευρά των χρηστών αλλά και από το επίπεδο πολυμέσων. Με άλλα λόγια, οι καθορισμένες έννοιες υψηλού επιπέδου είναι αναγνωρίσιμες από τις μεθόδους αυτόματης ανάλυσης, παραμένοντας παράλληλα κατανοητές στους χρήστες.

3.4.2 Εποικισμός Οντολογιών Θεματικού Πεδίου

Για να αξιοποιηθεί η προτεινόμενη υποδομή οντολογιών, μια οντολογία θεματικού πεδίου θα πρέπει να εποικιστεί από κατάλληλα "πρωτότυπα", δηλαδή οπτικούς περιγραφείς και χωρικές σχέσεις για τις καθορισμένες έννοιες του θεματικού πεδίου, μιας και όπως περιγράφεται στην Ενότητα 3.4.3, οι σχολιασμοί παράγονται μέσω του ταιριάσματος μεταξύ των πρωτοτύπων των εννοιών χαμηλού επιπέδου. Για να επιτευχθεί αυτό, οι χαμηλού επιπέδου περιγραφείς που περιλαμβάνονται στον ορισμό κάθε έννοιας του θεματικού πεδίου θα πρέπει να εξαχθούν για έναν αρκετά μεγάλο αριθμό από δείγματα και να συσχετιστούν με την οντολογία θεματικού πεδίου. Στην πράξη, αυτό γίνεται μέσω ενός γραφικού εργαλείου, με το οποίο ο χρήστης μπορεί να επιλέξει περιοχές που αντιστοιχούν στις έννοιες του θεματικού πεδίου και από αυτές να εξαχθούν οι οπτικοί περιγραφείς και να τοποθετηθούν στην οντολογία.

Το εργαλείο που χρησιμοποιήθηκε ήταν το M-Ontomat[4]. Στο Σχήμα 3.3 φαίνεται ο τρόπος με τον οποίο λειτουργεί το εργαλείο αυτό. Ο χρήστης φορτώνει μια οντολογία θεματικού πεδίου (εδώ, την οντολογία για το θεματικό πεδίο *Τένις*) και επιλέγει μια από τις Έννοιές της (εδώ, την Έννοια *μπάλα*). Στη συνέχεια φορτώνει μια εικόνα που περιέχει την αντίστοιχη έννοια, επιλέγει την περιοχή στην οποία απεικονίζεται, εξάγει τους κατάλληλους περιγραφείς (εδώ, χρώματος, υφής και σχήματος) και τέλος προσθέτει το πρωτότυπο αυτό στην οντολογία θεματικού πεδίου.



Σχήμα 3.3: *Το γραφικό εργαλείο M-Ontomat που χρησιμοποιήθηκε για την κατασκευή των πρωτοτύπων, τα οποία εποικίζουν μια οντολογία θεματικού πεδίου.*

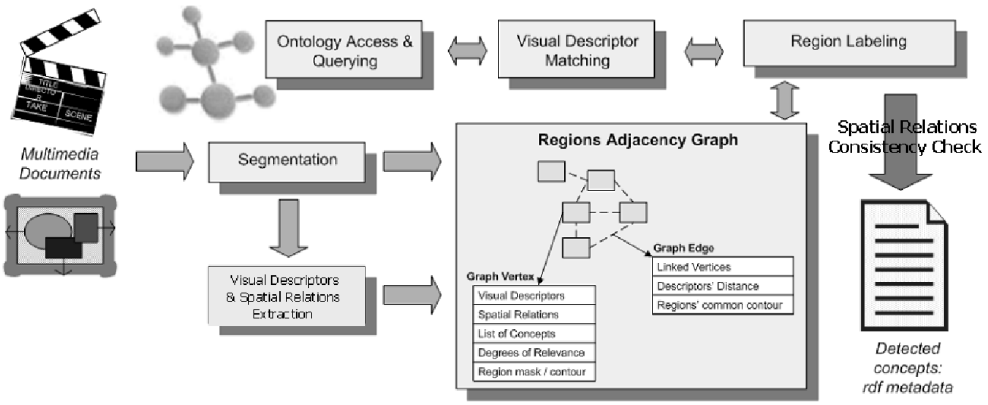
3.4.3 Σημασιολογική Ταξινόμηση Περιοχών Εικόνων

Η αρχιτεκτονική της σημασιολογικής ταξινόμησης παρουσιάζεται στο Σχήμα 3.4. Όπως φαίνεται, η ανάλυση ξεκινά με την κατάτμηση μιας εικόνας και την εξαγωγή των οπτικών περιγραφέων και των χωρικών σχέσεων, σύμφωνα με τους ορισμούς στην οντολογία θεματικού πεδίου. Στη συνέχεια, δημιουργείται ένα πρώτο σύνολο από πιθανές έννοιες που να αντιστοιχούν σε κάθε μία από τις εξαχθείσες περιοχές μετά από ερωτήματα στην οντολογία και ταίριασμα των εξαχθείσων περιγραφών με αυτές των πρωτοτύπων της. Έπειτα για την αξιολόγηση των παραχθείσων υποθέσεων χρησιμοποιείται η χωρική πληροφορία. Με τον τρόπο αυτό, οι έννοιες που περιέχονται στην εικόνα εξάγονται και δημιουργούνται τα μεταδεδομένα που τις περιγράφουν.

3.4.3.1 Αναπαράσταση Εικόνας

Για την αναπαράσταση της εικόνας επιλέχτηκε η δομή του *γράφου γειτονικότητας περιοχών* (region adjacency graph). Κάθε κόμβος του γράφου αντιστοιχεί σε μια

[4] `http://www.acemedia.org/aceMedia/results/software/m-ontomat-annotizer.html`



Σχήμα 3.4: *Η αρχιτεκτονική του συστήματος σημασιολογικής ανάλυσης εικόνας με βάση τη γνώση που παρουσιάζεται σε αυτό το Κεφάλαιο.*

Σχήμα 3.5: *Ένα παράδειγμα εικόνας από το θεματικό πεδίο Παραλία και η αντίστοιχη μάσκα κατάτμησης.*

περιοχή της εικόνας, ενώ κάθε ακμή ενώνει δύο περιοχές που είναι γειτονικές στην εικόνα. Σε κάθε κόμβο αντιστοιχούνται οι MPEG-7 περιγραφείς χρώματος, υφής και σχήματος που εξάγονται, οι χωρικές σχέσεις ανάμεσα στην περιοχή και τις γειτονικές της περιοχές καθώς και οι βαθμοί βεβαιότητας με τους οποίους η συγκεκριμένη περιοχή αντιστοιχεί στις έννοιες του θεματικού πεδίου. Επιπλέον, αποθηκεύεται μια λίστα με όλα τα εικονοστοιχεία που σχηματίζουν την περιοχή, καθώς και όλα τα εικονοστοιχεία που σχηματίζουν το περίγραμμά της. Τέλος, σε κάθε ακμή του γράφου αποθηκεύεται η απόσταση των περιοχών που ενώνει με βάση τους οπτικούς τους περιγραφείς, καθώς και μια λίστα με τα εικονοστοιχεία που σχηματίζουν το κοινό περίγραμμά τους.

3.4.3.2 Κατάτμηση Εικόνας

Για την κατάτμηση μιας εικόνας, χρησιμοποιείται ένας αλγόριθμος κατάτμησης που επεκτείνει τον αλγόριθμο Recursive Shortest Spanning Tree (RSST) και προτάθηκε από τους Avrithis et al. [5]. Ο αλγόριθμος αυτός δημιουργεί έναν μικρό αριθμό περιοχών, με τις οποίες αρχικοποιείται ο γράφος. Ένα παράδειγμα μιας εικόνας εισόδου και μιας αρχικής κατάτμησης φαίνεται στο Σχήμα 3.5.

3.4.3.3 Εξαγωγή MPEG-7 Οπτικών Περιγραφέων

Οι οπτικοί περιγραφείς που εξάγονται στην εφαρμογή που αναπτύχθηκε έχουν επιλεχθεί από το πρότυπο MPEG-7 και πιο συγκεκριμένα είναι ο Περιγραφέας Κύριων

Χρωμάτων, ο Κλιμακωτός Περιγραφέας Χρώματος, ο Περιγραφέας Ομοιογενούς Υφής και ο Περιγραφέας Σχήματος με βάση την Περιοχή του MPEG-7. Οι περιγραφείς αυτοί επιλέχτηκαν καθώς είναι κατάλληλοι για την εξαγωγή των οπτικών χαρακτηριστικών χαμηλού επιπέδου από περιοχές εικόνων.

3.4.3.4 Εξαγωγή Χωρικών Σχέσεων των Περιοχών

Όπως έχει ήδη αναφερθεί, πέρα από τις περιγραφές χαμηλού επιπέδου, είναι απαραίτητο να συμπεριληφθούν στην οντολογία ορισμοί σχετικά με τις χωρικές σχέσεις των εννοιών υψηλού επιπέδου, μιας και αυτός είναι ένας καλός τρόπος για τον διαχωρισμό μεταξύ εννοιών που έχουν παρόμοια οπτικά χαρακτηριστικά. Έννοιες όπως *θάλασσα* και *ουρανός* είναι ένα απλό παράδειγμα που καθιστά φανερή την ανάγκη της επιπλέον χωρικής πληροφορίας. Η πληροφορία της *Γειτονίας* υπάρχει στο γράφο, αφού μία ακμή μεταξύ δύο περιοχών υποδηλώνει ότι οι περιοχές αυτές είναι συνδεδεμένες. άρα γειτονεύουν. Ωστόσο, οι υπόλοιπες τοπολογικές και κατευθυντικές σχέσεις είναι απαραίτητες για την περαιτέρω βελτίωση των αποτελεσμάτων της ανάλυσης και της ακρίβειας του συστήματος. Οι χωρικές σχέσεις που έχουν ενσωματωθεί στην οντολογία είναι οι *Πάνω από*, *Κάτω από*, *Αριστερά από*, *Δεξιά από* και *μέσα σε*. Επιπρόσθετα, εισήχθησαν στην οντολογία οι σχέσεις *Κάτω από όλα* και *Πάνω από όλα*, μιας και φάνηκαν εξαιρετικά χρήσιμες κατά τη διάρκεια των πειραμάτων για έννοιες όπως π.χ. ο *ουρανός*.

3.4.3.5 Ταξινόμηση των Περιοχών των Εικόνων

Μετά την εξαγωγή της πληροφορίας σχετικά με τα οπτικά χαρακτηριστικά των περιοχών της εικόνας και τις χωρικές τους σχέσεις, το επόμενο βήμα είναι ο υπολογισμός του βαθμού ταιριάσματος ανάμεσα στους περιγραφείς που αποτελούν τα πρωτότυπα της οντολογίας καθώς και αυτούς που εξάγονται από τις περιοχές της εικόνας εισόδου με σκοπό να αντιστοιχηθούν πιθανές έννοιες στις περιοχές.

Για να επιτευχθεί αυτό, είναι απαραίτητος ο υπολογισμός της απόστασης μεταξύ δύο περιοχών με βάση όλους τους οπτικούς περιγραφείς που εξήχθησαν. Μια απόσταση για κάθε έναν από τους περιγραφείς μπορεί να υπολογιστεί, με τη χρήση των προτεινόμενων από το MPEG-7 αποστάσεων, αλλά δεν είναι χρήσιμες χωρίς μια μέθοδο που θα τις συνδυάσει και θα παρέχει μια συγχωνευμένη μοναδική απόσταση. Για το σκοπό αυτό ένα *νευρωνικό δίκτυο οπίσθιας ανατροφοδότησης*, όπως περιγράφηκε στην Ενότητα 2.6.2, το οποίο έχει εκπαιδευθεί να κάνει εκτίμηση της ομοιότητας μεταξύ δύο περιοχών. Η είσοδος του δικτύου είναι ένα διάνυσμα που περιέχει τους συγχωνευμένους περιγραφείς από την εικόνα εισόδου και ένα πρωτότυπο της οντολογίας και στην έξοδο παρέχει μια εκτίμηση της κανονικοποιημένης απόστασής τους. Με μια τυπική συνάρτηση κανονικοποίησης, η απόσταση μετατρέπεται σε *βαθμό βεβαιότητας*, ο οποίος και αποτελεί το κριτήριο ομοιότητας για όλες τις διαδικασίες ένωσης.

Από όλη αυτή τη διαδικασία, αποθηκεύονται στο γράφο μια λίστα από πιθανές έννοιες μαζί με ένα βαθμό βεβαιότητας. Στην περίπτωση που για δύο ή περισσότερες γειτονικές περιοχές έχει εξαχθεί η υπόθεση ότι αντιστοιχούν στην ίδια έννοια, ή οι άλλες πιθανές έννοιες έχουν βαθμό βεβαιότητας κάτω από ένα προκαθορισμένο κατώφλι, αυτές οι περιοχές ενώνονται, με το σκεπτικό ότι αποτελούν μέρη μιας μεγαλύτερης περιοχής που δεν έχει κατατμηθεί σωστά, εξαιτίας υπερκατάτμησης. Οι

αντίστοιχοι κόμβοι του γράφου ενώνονται και όλες οι αποστάσεις και οι περιγραφείς επαναϋπολογίζονται.

3.4.3.6 Έλεγχος Χωρικής Ορθότητας της Ταξινόμησης

Το βήμα αντιστοιχίας περιοχών πρωτότυπα εννοιών μόνο με την εξέταση της σχέσης των οπτικών τους περιγραφέων συχνά οδηγεί σε περισσότερες από μία πιθανές έννοιες για κάθε περιοχή της εικόνας. Για την εξαγωγή των τελικών συμπερασμάτων, το τελικό βήμα είναι η αξιοποίηση της γνώσης των χωρικών σχέσεων μεταξύ των εννοιών που αντιστοιχήθηκαν στις περιοχές. Για κάθε μία περιοχή ελέγχεται εάν οι χωρικές της διότητες ταιριάζουν με το χωρικό εννοιολογικό τους πλαίσιο. Για παράδειγμα, αν ABV είναι η σχέση που ορίζει ότι μία έννοια βρίσκεται *πάνω* από μια άλλη έννοια, μια πιθανή επιτρεπτή χωρική σχέση είναι η "*ουρανός* ABV *θάλασσα*". Έτσι, εφόσον βρεθεί μια περιοχή η οποία μπορεί να αντιστοιχεί είτε στην έννοια *ουρανός*, είτε στην έννοια *θάλασσα πάνω* από μια περιοχή που μπορεί να αντιστοιχεί είτε στην έννοια *θάλασσα*, είτε στην έννοια *βράχος*, ο μόνος επιτρεπτός συνδυασμός εννοιών με βάση την υπάρχουσα γνώση είναι η πρώτη περιοχή να χαρακτηριστεί ως *ουρανός* και η δεύτερη ως *θάλασσα*.

3.4.3.7 Ανάκτηση από τη Βάση Γνώσης

Για την ανάκτηση των πρωτοτύπων και των αντίστοιχων περιγραφέων από τη βάση γνώσης, χρησιμοποιείται η μηχανή συλλογιστικής *OntoBroker*[5] προκειμένου να αναλάβει την πραγματοποίηση των σχετικών ερωτημάτων. Το OntoBroker υποστηρίζει το φόρτωμα RDFS οντολογιών και έτσι σε αυτό φορτώνονται οι οντολογίες θεματικού πεδίου στις οποίες και ορίζονται οι έννοιες υψηλού επιπέδου, η VDO που περιέχει τους αντίστοιχους οπτικούς περιγραφείς και τα πρωτότυπα, τα οποία συνιστούν τη βάση γνώσης και παρέχουν την αντιστοιχία μεταξύ των εννοιών του θεματικού πεδίου με τους αντιστοιχους περιγραφείς. Κατάλληλα ερωτήματα ορίζονται, με τα οποία επιτυγχάνεται η ανάκτηση συγκεκριμένων τιμών από διάφορους περιγραφείς και έννοιες. Η γλώσσα με την οποία τίθενται τα ερωτήματα του Ontobroker είναι η *F-Logic*[6]

3.4.3.8 Δημιουργία Μεταδεδομένων

Μετά την εξαγωγή των εννοιών από μια εικόνα, το επόμενο βήμα είναι η δημιουργία μεταδεδομένων σε μια μορφή με την οποία θα είναι εύκολο να διαμοιραστούν σε πολλές διαφορετικές εφαρμογές με σχετική ευκολία. Εξαιτίας της ανάπτυξης του Σημασιολογικού Ιστού (Semantic Web) και των ποικίλων εφαρμογών που χρησιμοποιούν αυτές τις τεχνολογίες, το Σχήμα RDF επιλέχτηκε για τη αναπαράσταση των μεταδεδομένων που περιέχουν τους σχολιασμούς που εξήχθησαν. Έτσι, μια εφαρμογή θα μπορεί να διαβάσει τα RDF αρχεία και να τα χρησιμοποιήσει απευθείας σαν σημασιολογικούς σχολιασμούς, συσχετίζοντας μια εικόνα με έναν αριθμό από έννοιες που ανιχνεύθηκαν.

[5] `http://www.ontoprise.de/products/ontobroker`

[6] `http://www.ontoprise.de/documents/tutorial\_flogic.pdf`. Η F-Logic είναι τόσο μια γλώσσα *αναπαράστασης* που μπορεί να χρησιμοποιηθεί για τη μοντελοποίηση οντολογιών, όσο και μια γλώσσα *ερωτημάτων* με την οποία μπορούν να τεθούν ερωτήματα στην βάση γνώσης στο OntoBroker.

3.5 Πειραματικά Αποτελέσματα

Το πλαίσιο της σημασιολογικής ανάλυσης με βάση τη γνώση που παρουσιάστηκε, δοκιμάστηκε στο θεματικό πεδίο *Παραλία*. Για την ανάλυση κατασκευάστηκε κατάλληλη οντολογία θεματικού πεδίου και εποικίστηκε με κατάλληλα πρωτότυπα από όλες τις έννοιες που επιλέχθησαν. Για το σκοπό αυτό χρησιμοποιήθηκε το εργαλείο M-Ontomat.

Οι έννοιες συνοψίζονται στον Πίνακα 3.1, μαζί με τις επιτρεπτές χωρικές σχέσεις ανάμεσά τους και τον αριθμό των πρωτοτύπων που κατασκευάστηκε από την κάθε μία. Για παράδειγμα, η έννοια *θάλασσα* περιγράφεται με τη χρήση περιγραφέων χρώματος και υφής, ορίζεται ότι βρίσκεται *κάτω* από την έννοια *ουρανός* και *πάνω ή δίπλα* από την έννοια *άμμος*. Με παρόμοιο τρόπο, οι ορισμοί όλων των υπολοίπων εννοιών μπορούν να προκύψουν από τον Πίνακα 3.1. Στο Σχήμα 3.6, έχει δημιουργηθεί μια μάσκα κατάτμησης με βάση την έξοδο του συστήματος, όπου για τις διάφορες περιοχές της κάθε εικόνας έχει σημειωθεί η έννοια που προσδιόρισε το σύστημα. Για τις ανάγκες των πειραμάτων δημιουργήθηκε ένα σύνολο από 191 εικόνες, στις οποίες περιέχονται 4 έννοιες προς ανίχνευση. Το σύνολο των εικόνων αυτών προέρχεται από προσωπικές συλλογές, και από το WWW. Το σύνολο δεδομένης αλήθειας κατασκευάστηκε χειρωνακτικά, στις περιοχές που δημιουργήθηκαν από τον αλγόριθμο κατάτμησης. Το 80% των εικόνων χρησιμοποιήθηκε σαν σύνολο εκπαίδευσης και το 20% σαν σύνολο ελέγχου.

Για την αξιολόγηση της προτεινόμενης μεθοδολογίας υπολογίστηκαν 3 γνωστά μέτρα από το χώρο της ανάκτησης πληροφορίας. Αυτά είναι το *μέτρο ακρίβειας* P (precision), το *μέτρο ανάκτησης* R (recall) και το F-μέτρο (f-measure). Τα δύο πρώτα μέτρα θα αποκαλούνται εφεξής "ακρίβεια" και "ανάκτηση". Ορίζοντας σαν $|\cdot|$ το πλήθος των στοιχείων ενός συνόλου και για κάθε έννοια C_i ως G_i το σύνολο των εικόνων που πραγματικά περιέχουν την έννοια και ως D_i το σύνολο των εικόνων για τις οποίες ο ταξινομητής αποφάσισε ότι περιέχουν την έννοια, η ακρίβεια P_i και η ανάκτηση R_i για την έννοια C_i ορίζονται ως

$$P_i = \frac{|D_i \bigcap G_i|}{|D_i|}, \quad i = 1 \ldots N_C \ , \tag{3.1}$$

$$R_i = \frac{|D_i \bigcap G_i|}{|G_i|}, \quad i = 1 \ldots N_C \ , \tag{3.2}$$

ενώ το F-μετρο υπολογίζεται ως

$$F_i = \frac{2P_iR_i}{P_i + R_i}, \quad i = 1 \ldots N_C \ . \tag{3.3}$$

Η ακρίβεια είναι το ποσοστό των εικόνων που πραγματικά απεικονίζουν την έννοια, σύμφωνα πάντα με το σύνολο δεδομένης αλήθειας, ως προς τις εικόνες που το σύστημα αποφάνθηκε ότι την απεικονίζουν. Η ανάκτηση είναι το ποσοστό των εικόνων που ανιχνεύθηκαν σωστά ότι απεικονίζουν μία έννοια ως προς όλες τις εικόνες που την απεικονίζουν. Το F-μέτρο αποτελεί τον αρμονικό μέσο της ακρίβειας και της ανάκτησης, το οποίο και παίρνει μεγάλες τιμές όταν τα δύο αυτά μέτρα παίρνουν μεγάλες τιμές και είναι χρήσιμο για περιπτώσεις που ένα από τα δύο μέτρα είναι κοντά στη μέγιστη τιμή του και το άλλο κοντά στην ελάχιστη τιμή του.

Έννοια	**Περιγραφείς**	**Χωρικές Σχέσεις**	**Πρωτότυπα**
ουρανός	Χρώμα, Υφή	*ουρανός* ABV *θάλασσα*	30
θάλασσα	Χρώμα, Υφή	*θάλασσα* ABV, ADJ *άμμος*	30
άμμος	Χρώμα, Υφή	*άμμος* BEL, ADJ *θάλασσα*	20
άνθρωπος	Σχήμα	*άνθρωπος* INC *θάλασσα*, *άμμος*	45

Πίνακας 3.1: *Οι έννοιες του θεματικού πεδίου Παραλία, οι περιγραφείς που εξάγονται από την κάθε μία, οι χωρικές τους σχέσεις (ADJ: γειτονία, ABV: πάνω, BEL: κάτω, INC: μέσα) και ο αριθμός των πρωτοτύπων που κατασκευάστηκαν.*

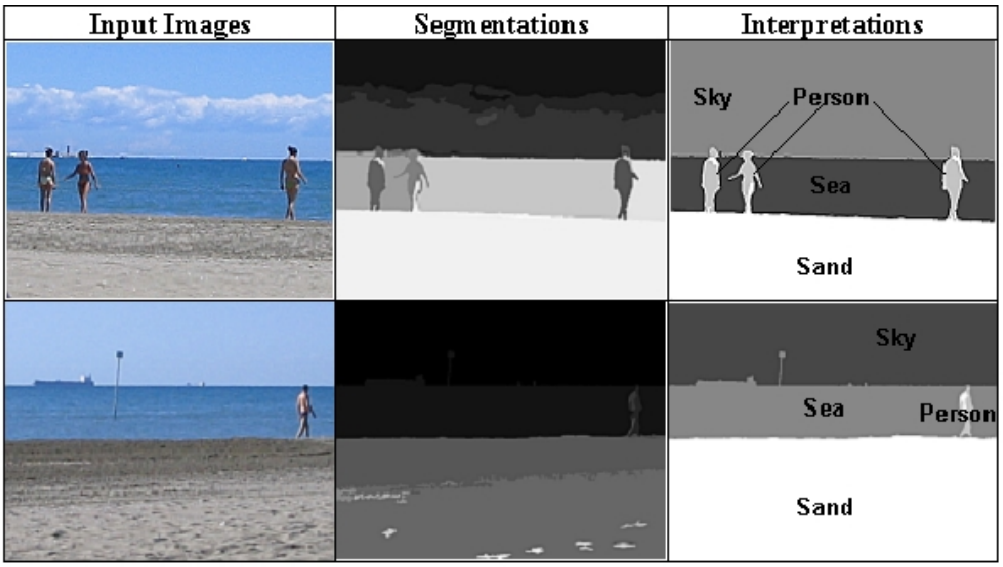
Έννοια	**P**	**R**	**F**
ουρανός	0.77	0.69	0.73
θάλασσα	0.66	0.59	0.62
άμμος	0.75	0.94	0.84
άνθρωπος	0.33	0.65	0.44
Σύνολο	0.69	0.75	0.72

Πίνακας 3.2: *Πειραματικά αποτελέσματα από την εφαρμογή του προτεινόμενου πλαισίου στο θεματικό πεδίο Παραλία και για 4 έννοιες υψηλού επιπέδου. P: ακρίβεια (precision), R: ανάκληση (recall), F: F-μέτρο (F-measure).*

Η χρήση της χωρικής πληροφορίας συλλαμβάνει μέρος του οπτικού εννοιολογικού πλαισίου και τελικά οδηγεί στην εξαγωγή πιο σωστών σημασιολογικών περιγραφών, έχοντας ως προαπαιτούμενο τη σωστή κατάτμηση. Στην περίπτωση που ο αλγόριθμος κατάτμησης ενώσει δύο περιοχές που αντιστοιχούν σε δύο έννοιες σε μία μεγαλύτερη περιοχή, η προτεινόμενη μεθοδολογία θα αδυνατεί να την αναγνωρίσει. Τα πλεονεκτήματα της χρήσης των χωρικών σχέσεων είναι ορατά στα πειραματικά αποτελέσματα του θεματικού πεδίου *Παραλία*, όπου οι έννοιες *θάλασσα* και *ουρανός*, παρότι έχουν παρόμοια οπτικά χαρακτηριστικά, και σε πολλές περιπτώσεις ο ανθρωπος αδυνατεί να τις ξεχωρίσει, αν δει μεμονωμένα τις αντίστοιχες περιοχές, ταξινομούνται σωστά εξαιτίας των διαφορετικών τους οπτικών χαρακτηριστικών. Επίσης, πρέπει να αναφερθεί ότι στην περίπτωση που μια περιοχή δεν έχει ικανοποιητικό βαθμό ομοιότητας με κανένα πρωτότυπο της αντίστοιχης οντολογίας θεματικού πεδίου, σε αυτήν δεν αντιστοιχεί καμία έννοια.

3.6 Συμπεράσματα

Στο Κεφάλαιο αυτό παρουσιάστηκε ένα ολοκληρωμένο πλαίσιο που αποσκοπεί στην ταξινόμηση περιοχών εικόνων με χρήση γνώσης. Αρχικά κατασκευάστηκε μια δομή οντολογιών, η οποία είχε σκοπό να αποτελέσει το μέσο αποθήκευσης της γνώσης. Η γνώση αποθηκεύτηκε με τη μορφή πρωτοτύπων που εξήχθησαν από το διαθέσιμο σύνολο εκπαίδευσης. Ένα πρωτότυπο για μια έννοια θεωρήθηκε μια χαμηλού επιπέδου περιγραφή που εξήχθη από μια περιοχή που απεικόνιζε αποκλειστικά αυτή την έννοια. Κατέστη σαφές και επιβεβαιώθηκε πειραματικά ότι η τεχνική συγχώνευσης που παρουσιάστηκε στο Κεφάλαιο 2 μπορεί να εφαρμοστεί και στην περίπτωση των εννοιών που χαρακτηρίζουν τοπικά μια εικόνα. Επίσης διαφάνηκε ότι οι περιγραφείς του προτύπου MPEG-7 μπορούν να χρησιμοποιηθούν και για την ταξινόμηση περιοχών.



Σχήμα 3.6: *Δύο παραδείγματα ανάλυσης εικόνων από το θεματικό πεδίο Παραλία.*

Το βασικό πλεονέκτημα της μεθοδολογίας που κατασκευάστηκε είναι ότι ο τρόπος με τον οποίο αποθηκεύεται η γνώση καθιστά εύκολο τον εμπλουτισμό της με νέα πρωτότυπα, αλλά και την εισαγωγή νέων περιγραφέων στα ήδη υπάρχοντα πρωτότυπα. Αυτό, καθώς το μόνο που απαιτείται είναι η ενημέρωση των παραδειγμάτων στην υπάρχουσα οντολογία. Δεν απαιτείται κάποιας μορφής εκπαίδευση και έτσι το σύστημα μετά την ενημέρωση μπροεί να χρησιμοποιηθεί άμεσα. Επίσης, η εισαγωγή ενός νέου θεματικού πεδίου στο ήδη υπάρχον σύστημα απαιτεί μόνο την κατασκευή μιας νέας οντολογίας θεματικού πεδίου που να περιέχει τα πρωτότυπα για τις έννοιες που περιλαμβάνει αυτό. Φυσικά, μεγάλο πλεονέκτημα αποτελεί το γεγονός ότι προσδιορίζεται η περιοχή της εικόνας στην οποία απεικονίζεται η έννοια. Τα αποτελέσματα από την εφαρμογή της στο θεματικό πεδίο *Παραλία* έδειξαν ότι η ακρίβεια που επιτεύχθηκε ήταν ικανοποιητική.

Το βασικό μειονέκτημα ωστόσο, είναι ότι η μεθοδολογία αυτή δεν είναι εύκολο να εφαρμοστεί σε μεγάλα σύνολα εικόνων. Αυτό συμβαίνει καθώς όταν ο αριθμός των εικόνων που περιέχουν μια έννοια είναι πολύ μεγάλος, στις περισσότερες περιπτώσεις η έννοια εμφανίζει πολύ διαφορετικές ιδιότητες. Αυτό δημιουργεί την ανάγκη για κατασκευή μεγάλου αριθμού πρωτοτύπων, κάτι που αποτελεί πολύ χρονοβόρα διαδικασία. Δυστυχώς, οι περιπτώσεις των σχολιασμένων συλλογών εικόνας σε επίπεδο περιοχής είναι λίγες, συνήθως αποτελούνται από μερικές εκατοντάδες εικόνες και ο σχολιασμός αφορά έναν μικρό αριθμό εννοιών, κάτι που δυσχεραίνει την αξιοποίησή τους στα περισσότερα θεματικά πεδία. Φυσικά, ένα ακόμη μειονέκτημα της προτεινόμενης μεθοδολογίας έχει να κάνει με το γεγονός ότι εξαρτάται σε μεγάλο βαθμό από τα αποτελέσματα της κατάτμησης της εικόνας. Έτσι, αν ο επιλεγμένος αλγόριθμος κατάτμησης ενώσει δύο περιοχές, η μία από τις οποίες περιέχει μια από τις έννοιες προς ανίχνευση, ο αλγόριθμος θα αποτύχει να την αναγνωρίσει, ακόμη κι αν το τμήμα που της αντιστοιχεί μοιάζει αρκετά με ένα από τα πρωτότυπα. Τέλος στα μειονεκτήματά της μπορεί να θεωρηθεί το γεγονός ότι η κατασκευή της γνώσης γίνεται από τους χρήστες. Στην περίπτωση που η κατασκευή δεν γίνει από κάποιο ειδήμονα του θεματικού πεδίου, αλλά από κάποιο χρήστη είτε με περιορισμένες γνώσεις, είτε που

έχει λάβει ελλιπείς οδηγίες, η γνώση που θα κατασκευαστεί ενδέχεται να δυσκολέψει την ταξινόμηση.

Κεφάλαιο 4

Ανίχνευση Εννοιών σε Εικόνες με χρήση Τεχνικών Οπτικού Θησαυρού

4.1 Εισαγωγή

Καθώς η ανάγκη για δημιουργία αυτόματα σχολιασμένου πολυμεσικού υλικού συνεχώς αυξάνεται, η ανίχνευση εννοιών υψηλού επιπέδου συνεχίζει να αποτελεί ένα από τα πιο ενδιαφέροντα ερευνητικά προβλήματα . Εξ ορισμού, το πρόβλημα αυτό απαιτεί μία "από κάτω προς τα πάνω" (bottom-up) προσέγγιση που εστιάζει σε τοπική ανάλυση για να ανιχνεύσει και να αναγνωρίσει συγκεκριμένες έννοιες σε μια εικόνα. Στην πιο απλοποιημένη του μορφή, το πρόβλημα αντιμετωπίζεται χωρίς αξιοποίηση του εννοιολογικού πλαισίου που τις περιβάλλει. Αναγνωρίζοντας τη σημαντικότητα του προβλήματος, πολλές ερευνητικές προσπάθειες έχουν επικεντρωθεί στην επίλυσή του. Παρολαυτά, το Σημασιολογικό Κενό [193] πολύ συχνά χαρακτηρίζει τις διαφορές μεταξύ των περιγραφών μιας έννοιας με διάφορες αναπαραστάσεις και την απεικόνιση από τα χαρακτηριστικά χαμηλού στις έννοιες υψηλού επιπέδου.

Η ταξινόμηση περιοχών με χρήση της γνώσης που αποτέλεσε το ερευνητικό αντικείμενο του Κεφαλαίου 3 παρουσιάζει το μειονέκτημα της αδυναμίας εφαρμογής της σε μεγάλα σύνολα δεδομένων. Ο βασικός λόγος είναι ότι απαιτεί ιδιαίτερη ανθρώπινη προσπάθεια για την κατασκευή της γνώσης με τη μορφή πρωτοτύπων, καθώς πρέπει να προηγηθεί σχολιασμός ενός μεγάλου αριθμού εικόνων ανά περιοχή. Καθώς το πρόβλημα της κατάτμησης εικόνων δεν είναι πλήρως επιλυμένο, η επιλογή των περιοχών θα πρέπει να γίνει με τη χρήση κατάλληλων εργαλείων, χειρωνακτικά. Ωστόσο, ο καθολικός σχολιασμός εικόνων, δηλαδή ο προσδιορισμός των εννοιών που περιέχονται σε αυτές, χωρίς παράλληλα να προσδιορίζεται η ακριβής περιοχή της εικόνας, αποτελεί μια διαδικασία αισθητά ευκολότερη και γρηγορότερη.

Στο πλαίσιο αυτό κινείται το παρόν Κεφάλαιο. Ο σκοπός είναι η ανίχνευση εννοιών σε εικόνες, χωρίς να προσδιορίζεται η ακριβής περιοχή της εικόνας στην οποία εμφανίζονται. Για το σκοπό αυτό εκπαιδεύονται ανιχνευτές που βασίζονται σε τεχνικές μηχανικής μάθησης. Η εκπαίδευση των ανιχνευτών γίνεται με βάση καθολικά σχολιασμένες εικόνες. Οι τεχνικές που προτείνονται βασίζονται στο ευρέως διαδεδομένο ερευνητικά μοντέλο "bag-of-words. Για την αξιολόγησή τους έχει επιλεχθεί ένα σύνολο αξιολόγησης που προέρχεται από την ευρέως γνωστή δοκιμασία αξιολόγησης

TRECVID.

4.2 Περιγραφή του Προβλήματος

Εστιάζοντας αποκλειστικά στο οπτικό μέρος της φάσης της ανάλυσης, είναι αλήθεια ότι η ανίχνευση εννοιών υψηλού επιπέδου παραμένει ακόμη ένα άλυτο ερευνητικό πρόβλημα. Οι δύο πιο ενδιαφέρουσες όψεις του είναι αφενός η εξαγωγή των χαρακτηριστικών χαμηλού επιπέδου που θα περιγράψουν το οπτικό περιεχόμενο των εικόνων ή των περιοχών εικόνων και αφετέρου ο τρόπος με τον οποίο τα χαρακτηριστικά αυτά θα αντιστοιχηθούν σε έννοιες υψηλού επιπέδου. Η τελευταία αποτελεί και το αντικείμενο έρευνας του παρόντος Κεφαλαίου. Υπενθυμίζεται εδώ ότι ο όρος "έννοια υψηλού επιπέδου" χρησιμοποιείται για να αποδώσει τον ευρέως χρησιμοποιούμενο όρο "high-level concept", εννοώντας αποκλειστικά έννοιες όπως *ουρανός*, *βλάστηση* κ.α. με τις οποίες μπορούν να περιγραφούν περιοχές που περιέχονται σε εικόνες. Στη συνέχεια, όταν αναφέρεται ο όρος "έννοια" θα αναφέρεται σε τέτοιου είδους περιοχές και την απεικόνιση τους ή όχι σε εικόνες[1].

Η πιο συνηθισμένη προσέγγιση που ακολουθείται στα προβλήματα ανίχνευσης και αναγνώρισης εννοιών ξεκινά με την εξαγωγή μιας περιγραφής του οπτικού περιεχομένου της έννοιας. Έπειτα, ένας ανιχνευτής εκπαιδεύεται με βάση τις εξαχθείσες περιγραφές από πολλά παραδείγματα μιας έννοιας έτσι ώστε να την αναγνωρίζει. Η χρήση τεχνικών μηχανικής μάθησης σε προβλήματα ανίχνευσης και αναγνώρισης εννοιών είναι ίσως η πιο συνηθισμένη προσέγγιση που ακολουθείται για την υλοποίηση των ανιχνευτών. Τεχνικές όπως Νευρωνικά Δίκτυα, Ασαφή Συστήματα, Γενετικοί Αλγόριθμοι [140] και Μηχανές Διανυσμάτων Υποστήριξης έχουν με επιτυχία εφαρμοστεί σε τέτοιου είδους προβλήματα, αναλαμβάνοντας την αντιστοίχηση από τα χαρακτηριστικά χαμηλού στα χαρακτηριστικά υψηλού επιπέδου.

Για την εκπαίδευση ανιχνευτών εννοιών χαμηλού επιπέδου, ειναι απαραίτητη η ύπαρξη κάποιου συνόλου δεδομένης αλήθειας, δηλαδή ενός συνόλου από εικόνες ή/και βίντεο που συνοδεύονται από κατάλληλους σχολιασμούς σχετικά με τις έννοιες που περιέχουν. Οι σχολιασμοί αυτοί μπορεί να είναι *καθολικοί ή τοπικοί*. Ως καθολικοί θεωρούνται οι σχολιασμοί είτε σε επίπεδο εννοιών που χαρακτηρίζουν ολόκληρη την εικόνα, όπως για παράδειγμα η σκηνή που απεικονίζεται, είτε σε επίπεδο εννοιών που περιέχονται στην εικόνα, χωρίς όμως να προσδιορίζεται και η αντίστοιχη περιοχή που τις απεικονίζει. Αντίστοιχα, ως τοπικοί χαρακτηρίζονται οι σχολιασμοί στους οποίους προσδιορίζονται και οι περιοχές που απεικονίζουν τις έννοιες, είτε αυτό γίνεται με τη χρήση ορθογωνίων ή πολυγώνων που τις περικλείουν, είτε με σχολιασμό ανά περιοχές που προέκυψαν από κατάτμηση της εικόνας.

Ωστόσο, παρότι το διαθέσιμο οπτικοακουστικό υλικό έχει αυξηθεί ραγδαία τα τελευταία χρόνια, η αύξηση αυτή δεν έχει συνοδευτεί από αντίστοιχη αύξηση του αριθμού των διαθέσιμων σχολιασμένων συλλογών εικόνων. Πολύ λίγες είναι οι συλλογές εικόνων που περιέχουν σχολιασμό ανά περιοχή της εικόνας, όπως είναι για παράδειγμα η συλλογή του LabelMe [178], η οποία έχει προκύψει από συλλογική προσπάθεια. Πρόκειται για μια διαδικτυακή συλλογή από εικόνες και σχολιασμούς που έχει δημιουργηθεί από τους Russell και Torralba και μεγαλώνει μέσω της προσφοράς των

[1]Στα πειραματικά αποτελέσματα και λόγω της φύσης της διαδικασίας αξιολόγησης που χρησιμοποιείται, ο όρος "έννοια" περιλαμβάνει και έννοιες που χαρακτηρίζουν συνολικά τις εικόνες, δηλαδή σκηνές.

επισκεπτών της. Ο χρήστης μπορεί να ανεβάσει μια εικόνα και σχεδιάζοντας πολύγωνα μέσω ενός γραφικού εργαλείου μπορεί να παρέχει τον κατάλληλο σχολιασμό για τις έννοιες που περιέχονται σε επίπεδο περιοχής. Αν και προσφέρει μεγάλο αριθμό από σχολιασμένες περιοχές εικόνων, παρουσιάζει το μειονέκτημα ότι η αξιοπιστία του εξαρτάται από το χρήστη και οι σχολιασμοί που παρέχονται δεν ελέγχονται. Έτσι, υπάρχουν πολλά παραδείγματα που ο χρήστης επέλεξε πολύγωνο μεγαλύτερο ή μικρότερο από την έννοια που ήθελε να προσδιορίσει, με αποτέλεσμα τη δημιουργία παραπλανητικών παραδειγμάτων. Πρέπει να σημειωθεί ότι το LabelMe δεν αποσκοπεί στο να αποτελέσει μια νέα διαδικασία αξιολόγησης στα ερευνητικά πεδία της όρασης υπολογιστών και την αναγνώρισης προτύπων. Αντίθετα, επιδιώκει να αποτελέσει ένα δυναμικά αυξανόμενο σύνολο δεδομένων που θα διευκολύνει την ανάπτυξη νέων τεχνικών. Ιδιαίτερο ενδιαφέρον παρουσιάζει επίσης η συλλογή της διαδικασίας αξιολόγησης PASCAL [63]. Στο πλαίσιο αυτό, διατίθενται τοπικά σχολιασμένες εικόνες με τις οποίες οι συμμετέχοντες μπορούν να εκπαιδεύσουν τα συστήματά τους, τα οποία θα κάνουν ταξινόμηση και ανίχνευση εννοιών, καθορίζοντας την ύπαρξη μιας έννοιας σε μια εικόνα και την ακριβή της τοποθεσία σε αυτή, αντίστοιχα.

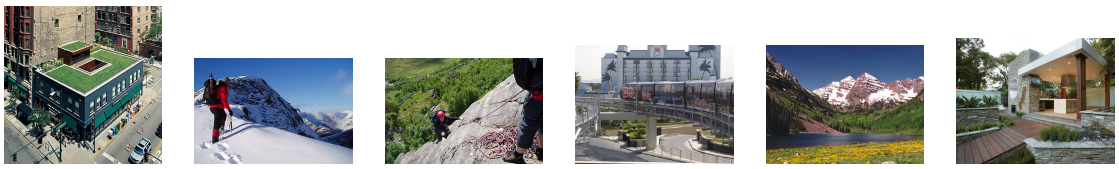
Από την άλλη, ο καθολικός σχολιασμός μιας εικόνας είναι μια διαδικασία αισθητά πιο εύκολη και σίγουρα πολύ πιο γρήγορη. Ένα τέτοιο παράδειγμα αποτελεί ο σχολιασμός που έγινε στα πλαίσια του LSCOM workshop [100], όπου ένας πάρα πολύ μεγάλος αριθμός από πλάνα από βίντεο δελτίων ειδήσεων σχολιάστηκε καθολικά και για έναν πολύ μεγάλο αριθμό από έννοιες υψηλού επιπέδου. Αρχικά κατασκευάστηκε ένα εκτενέστατο λεξικό εννοιών που αποτελείται από περισσότερες από 2000 έννοιες, από τις οποίες για περιπου 400 δημιουργήθηκαν καθολικοί σχολιασμοί σε 80 ώρες βίντεο του συνόλου ανάπτυξης της διαδικασίας αξιολόγησης TRECVID 2005, για την οποία θα γίνει λόγος στη συνέχεια, στην Ενότητα 4.5. Πιο συγκεκριμένα, ελέγχθηκε η ύπαρξη των εννοιών στα χαρακτηριστικά καρέ που εξήχθησαν από τα 61901 πλάνα των βίντεο. Πρέπει να επισημανθεί ότι για πρακτικά προβλήματα ταξινόμησης και ανίχνευσης ιδιαίτερο ενδιαφέρον παρουσιάζει ένα υποσύνολο του παραπάνω λεξικού, το οποίο αποκαλείται "LSCOM-Lite" [148]. Στην περίπτωση αυτή διατίθενται σχολιασμοί για 39 έννοιες υψηλού επιπέδου[2]. Μια ακόμη προσπάθεια σχολιασμού στο πλαίσιο του TRECVID 2007 διοργανώθηκε από τους Ayache και Quenot [6]. Όλες οι ομάδες που συμμετείχαν στην ανίχνευση εννοιών υψηλού επιπέδου σχολίασαν συλλογικά πάνω από 30000 χαρακτηριστικά καρέ από βίντεο και για 36 έννοιες υψηλού επιπέδου.

Θα πρέπει να δοθεί ιδιαίτερη έμφαση στο γεγονός ότι οι δύο πρώτες συλλογές είναι μεν σε επίπεδο περιοχής, αλλά προσφέρουν μόνο από μερικές εκατοντάδες εώς λίγες χιλιάδες από σχολιασμένες εικόνες, ενώ οι τελευταίες προσφέρουν δεκάδες χιλιάδες καθολικά σχολιασμένες εικόνες. Πέρα από αυτό, θα πρέπει να τονιστεί ότι τα τελευταία χρόνια, ιδιαίτερη άνθηση έχουν γνωρίσει οι διαδικτυακές συλλογές φωτογραφιών μέσα από ιστοσελίδες κοινωνικής δικτύωσης (social networks). Παραδείγματα υπάρχουν πολλά, ξεχωρίζει ίσως το Flickr[3], το οποίο περιέχει πολλά εκατομμύρια από φωτογραφίες. Σε τέτοιου τύπου ιστοσελίδες, οι χρήστες ανεβάζουν τις προσωπικές τους φωτογραφίες και βίντεο και τις περισσότερες φορές τις σχολιάζουν σε καθολικό επίπεδο, συνήθως παρέχοντας έναν αριθμό από λέξεις-κλειδιά (keywords). Έτσι δημιουργείται ένα μεγάλο σύνολο από καθολικά σχολιασμένο πολυμεσικό υλι-

[2]Για την ακρίβεια, κάποιες έννοιες υψηλού επιπέδου του LSCOM-Lite δεν υπάρχουν στο πλήρες LSCOM.

[3]`http://www.flickr.com`

Σχήμα 4.1: *Αντιπροσωπευτικές εικόνες που περιέχουν την έννοια θάλασσα.*



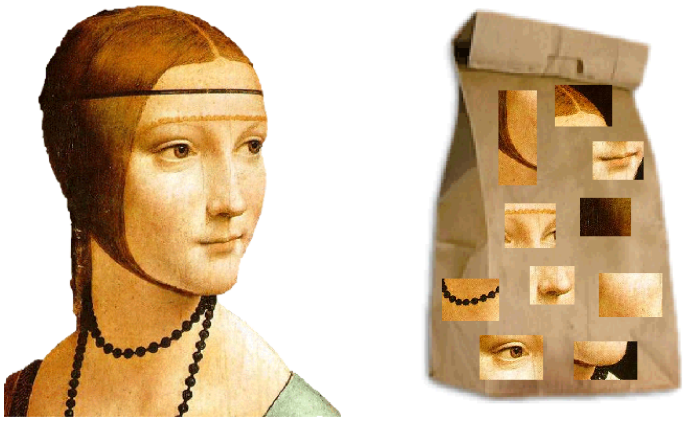
Σχήμα 4.2: *Αντιπροσωπευτικές εικόνες που απεικονίζουν την σκηνή εξωτερικός χώρος.*

κό, το οποίο δύναται να χρησιμοποιηθεί σε προβλήματα αναγνώρισης/ταξινόμησης.

Είναι φανερό από τα παραπάνω ότι πρέπει να δοθεί ιδιαίτερη σημασία στην ανίχνευση εννοιών υψηλού επιπέδου με βάση καθολικά χαρακτηριστικά της εικόνας ή του βίντεο, προκειμένου να γίνει εκμετάλλευση του μεγάλου αριθμού των διαθέσιμων καθολικών σχολιασμών. Ωστόσο, η διαθεσιμότητα του καθολικού σχολιασμού δεν πρέπει να περιορίζει την επιλογή των εννοιών μόνο σε αυτές που χαρακτηρίζουν συνολικά μια εικόνα, όπως για παράδειγμα τα προβλήματα ταξινόμησης σκηνής, αλλά και σε έννοιες που απεικονίζονται σε περιοχές της εικόνας. Σε ένα τέτοιο ερευνητικό πλαίσιο, η ανίχνευση εννοιών θα πραγματοποιείται χωρίς να είναι γνωστά τα χαρακτηριστικά τους, αλλά αντίθετα να είναι γνωστά τα χαρακτηριστικά των εικόνων στις οποίες περιέχονται. Με άλλα λόγια, ένας ανιχνευτής π.χ. για την έννοια *θάλασσα* δε θα εκπαιδευτεί με πρωτότυπα από περιοχές που να περιγράφονται από την έννοια αυτή, όπως για παράδειγμα έγινε στο Κεφάλαιο 2. Δηλαδή δε θα μάθει ποτέ πώς μοιάζει η έννοια την οποία προσπαθεί να ανιχνεύσει! Αν και εκ πρώτης όψεως αυτό φαντάζει περίεργο και ίσως δύσκολο να επιτευχθεί, αν αναλογιστεί κανείς τον τρόπο με τον οποίο εμφανίζονται οι έννοιες στο πολυμεσικό υλικό, δηλαδή το πώς μοιάζουν οι εικόνες που περιέχουν συγκεκριμένες έννοιες, θα γίνει αντιληπτό ότι συγκεκριμένες έννοιες περιέχονται σχεδόν αποκλειστικά σε εικόνες με παραπλήσια οπτικά χαρακτηριστικά.

Στο Σχήμα 4.1 παρουσιάζονται μερικές εικόνες που περιέχουν την έννοια *θάλασσα*. Όπως φαίνεται, στην πλειοψηφία τους οι εικόνες αυτές αποτελούνται από παρόμοιες περιοχές, κάτι λίγο εώς πολύ αναμενόμενο, μιας και συνήθως μια έννοια συνυπάρχει με τις ίδιες έννοιες και γενικά οι έννοιες εμφανίζουν παραπλήσια οπτικά χαρακτηριστικά. Αντίστοιχες παρατηρήσεις γίνονται και για την περίπτωση εννοιών που περιγράφουν καθολικά το περιεχόμενο μιας εικόνας. Για παράδειγμα, στο Σχήμα 4.2 παρουσιάζονται εικόνες που απεικονίζουν τη σκηνή *εξωτερικός χώρος*. Και στην περίπτωση αυτή, οι εικόνες εμφανίζουν ομοιότητες ως προς τις επιμέρους περιοχές από τις οποίες αποτελούνται. Γίνεται προφανές ότι μια προσέγγιση για να γίνει εκμετάλλευση αυτής της παρατήρησης θα ήταν η ακόλουθη: αρχικά, θα πρέπει με κάποιον τρόπο να χωριστεί η εικόνα σε περιοχές και να εξαχθούν από αυτές χαρακτηριστικά χαμηλού επιπέδου. Έπειτα να συνδυαστούν οι περιγραφές αυτές με κάποιο τρόπο που να εξασφαλίζει ότι παρόμοιες σημασιολογικά εικόνες θα έχουν παρόμοιες περιγραφές και τέλος να εκπαιδευτούν κατάλληλοι ανιχνευτές για κάθε έννοια.

Πολλές ερευνητικές προσπάθειες έχουν στραφεί προς παραλλαγές μιας αρκετά γενικής κατηγορίας τεχνικών που συνδυάζει όλα τα παραπάνω βήματα και αποκαλεί-



Σχήμα 4.3: *Η έννοια πρόσωπο που αποτελεί ένα σνθετο αντικείμενο, μπορεί να αποσυντεθεί σε ένα σύνολο από έννοιες, όπως μάτια, στόμα, μύτη κλπ. που αποτελούν πιο απλά αντικείμενα. Το σχήμα είναι από το [66].*

ται "*bag-of-words*". Σύμφωνα με τη βασική αρχή του μοντέλου αυτού, περιγραφείς εξάγονται τοπικά, δηλαδή από υποσύνολα των εικονοστοιχείων της εικόνας. Με τη βοήθεια κατάλληλου λεξικού, οι περιγραφές αυτές κβαντοποιούνται και κωδικοποιούνται. Το σύνολο των περιγραφών αυτών χωρίς να λαμβάνεται υπόψην η χωρική τους κατανομή και οι διάφορες σχέσεις που ενδεχομένως τις συνδέουν, χρησιμοποιείται για την περιγραφή μιας εικόνας. Τελικά, κατάλληλοι ανιχνευτές εκπαιδεύονται με τη χρήση των περιγραφών αυτών. Η λέξη "bag" (σάκος) υποννοεί ότι οι τεχνικές αυτές δεν λαμβάνουν υπόψη τυχόν σχέσεις μεταξύ των λέξεων, οι οποίες αποτελούν ένα σύνολο, τα στοιχεία του οποίου δεν έχουν κατάταξη.

4.3 Το μοντέλο bag-of-words

Σύμφωνα με τη γενικότερη μεθοδολογία των τεχνικών bag-of-words, ένα σύνθετο αντικείμενο μπορεί να αποσυντεθεί σε ένα σύνολο από "κομμάτια", τα οποία μπορεί να αντιπροσωπεύουν έννοιες, μπορεί όμως να μην έχουν καμία σημασιολογία. Ένα παράδειγμα της αποσύνθεσης μιας έννοιας φαίνεται στο Σχήμα 4.3, Ένα σύνθετο αντικείμενο, το *πρόσωπο* αποσυντίθεται σε *μάτια*, *μύτη*, *στόμα* κλπ.

Ένας πιο χαλαρός ορισμός της μεθόδου αρκείται στην αναπαράσταση μιας έννοιας απλά σαν ένα σύνολο από ανεξάρτητα χαρακτηριστικά. Το προφανές πρόβλημα που προκύπτει είναι ότι ανάμεσα σε δύο πρωτότυπα της ίδιας έννοιας, τα χαρακτηριστικά που προκύπτουν από την αποσύνθεσή τους ενδέχεται να είναι μεν σημασιολογικά παρόμοια, αλλά οπτικά να διαφέρουν σε μεγάλο βαθμό. Ένα τέτοιο παράδειγμα παρατίθεται στο Σχήμα 4.4. Για να αντιμετωπιστεί αυτό το πρόβλημα, η λύση που ακολουθείται είναι η εισαγωγή ενός βιβλίου κωδικοποίησης (codebook). Ο κώδικας αυτός κατασκευάζεται με τέτοιο τρόπο ώστε να περιέχει τα πιο συνηθισμένα χαρακτηριστικά που συναντώνται στις έννοιες που πρόκειται να ανιχνευθούν και προκύπτει συνήθως από κάποια διαδικασία ομαδοποίησης ή συσταδοποίησής τους. Στη συνέχεια, κάθε αντικείμενο περιγράφεται από ένα ιστόγραμμα, το οποίο δείχνει τη συχνότητα εμφάνισης των λέξεων του κώδικα. Η διαδικασία αυτή περιγράφεται στο Σχήμα 4.5. Συνεπώς, οι ανιχνευτές εκπαιδεύονται με βάση τις τιμές των ιστογραμμάτων. Η εφαρμογή του μοντέλου bag-of-words στην ανίχνευση εννοιών σε εικόνες απαιτεί την προσαρμογή του στην ιδιαίτερη φύση του προβλήματος.

Σχήμα 4.4: *Δύο παραδείγματα της αποσύνθεσης με τη μέθοδο bag-of-words, για δύο πρωτότυπα που αντιστοιχούν στην έννοια μοτοσυκλέτα. Είναι φανερό ότι παρότι πρόκειται για την ίδια έννοια, τα χαρακτηριστικά διαφέρουν αρκετα. Μια σύγκριση των οπτικών χαρακτηριστικών των λέξεων δε θα κατέληγε σε ασφαλές συμπέρασμα ότι πρόκειται για αποσύνθεση της ίδιας έννοιας.*

Σχήμα 4.5: *Η χρήση ενός βιβλίου κωδικοποίησης για τα παραδείγματα του Σχήματος 4.4 οδηγεί σε παρόμοια περιγραφή. Αριστερά: αντιστοιχία χαρακτηριστικών με λέξεις του βιβλίου κωδικοποίησης. Δεξιά: αναπαράσταση με βάση το βιβλίο κωδικοποίησης. Είναι φανερή η ομοιότητα των δύο ιστογραμμάτων που περιγράφουν τις εικόνες.*

Για την υλοποίηση ενός μοντέλου bag-of-words, το πρώτο που θα πρέπει να καθορισθεί είναι ο τρόπος με τον οποίο μια εικόνα θα αποσυντεθεί σε μέρη. Οι πρώτες προσεγγίσεις χρησιμοποίησαν ένα κανονικό πλέγμα για να χωρίσουν την εικόνα σε περιοχές. Οι Vogel και Schiele [224], αλλά και οι Fei-Fei και Perona [67] επέλεξαν την ιδέα αυτή και χώρισαν την εικόνα σε έναν συνήθως μεγάλο αριθμό από δομικά στοιχεία (blocks). Τα στοιχεία αυτά επιλέχθηκαν να είναι τετράγωνα, ίσου μεγέθους σε όλες τις περιοχές της. Το μοναδικό πλεονέκτημα της μεθόδου αυτής είναι η ταχύτητα με την οποία γίνεται η αποσύνθεση της εικόνας, αφού το μόνο που χρειάζεται είναι ο υπολογισμός των συντεταγμένων του κάθε δομικού στοιχείου. Ωστόσο, η αποσύνθεση αυτή αν και γρήγορη, αδυνατεί να δώσει περιοχές με "σημασιολογικό" περιεχόμενο. Αυτό σημαίνει ότι πολλά από τα δομικά στοιχεία που θα εξαχθούν δε θα αντιστοιχούν σε χαρακτηριστικά όπως αυτά ορίστηκαν παραπάνω. Επίσης, είναι σχεδόν σίγουρο ότι αρκετά από τα δομικά στοιχεία θα αποτελούν "θόρυβο", δηλαδή δε θα ταιριάζουν ικανοποιητικά με καμία από τις λέξεις του κώδικα. Ωστόσο, θα αντιστοιχούνται σε κάποια λέξη και τελικά θα αυξάνεται η συχνότητα εμφάνισης ορισμένων λέξεων του ιστογράμματος, χωρίς αυτό να είναι επιθυμητό. Για να γίνει αυτό κατανοητό, στο Σχήμα 4.6 παρουσιάζονται δύο εικόνες, με την δεύτερη να αποτελεί μέρος της πρώτης (crop). Το πλέγμα που προκαλεί την αποσύνθεση έχει την ίδια διάσταση. Είναι φανερό ότι τα χαρακτηριστικά που εξάγονται είναι πολύ διαφορετικά, ουσιαστικά λόγω της διαφορετικής κλίμακας και θα οδηγήσουν σε διαφορετικά ιστογράμματα περιγραφής, παρόλο που και οι δύο εικόνες απεικονίζουν π.χ. τις έννοιες *εξωτερικός χώρος* και *βλάστηση*. Φαίνεται, όμως, ότι η αριστερή εικόνα περιέχει κυρίως *πράσινες* και *γκρι* περιοχές, ενώ η δεξιά *πράσινες* και *γαλάζιες*. Σε μια προσπάθεια να ξεπεραστεί το πρόβλημα αυτό, οι Ullman et al. [218] εφάρμοσαν τυχαία

Σχήμα 4.6: *Η χρήση ενός τετραγωνικού πλέγματος για την αποσύνθεση μιας εικόνας και ενός μέρους αυτής οδηγεί σε πολύ διαφορετικά χαρακτηριστικά.*

Σχήμα 4.7: *Η εξαγωγή σημείων αμετάβλητων σε γεωμετρικούς μετασχηματισμούς και αλλαγές κλίμακας για την αποσύνθεση των εικόνων οδηγεί σε παρόμοια χαρακτηριστικά. Στο συγκεκριμένο παράδειγμα χρησιμοποιούνται τα χαρακτηριστικά SURF [9].*

δειγματοληψία (random sampling). Επιπλέον, για να ξεπεραστεί το πρόβλημα της κλίμακας, επέλεξαν δομικά στοιχεία της εικόνας με διαφορετικά μεγέθη.

Καθώς το παραπάνω πρόβλημα αποδείχτηκε ιδιαίτερα δύσκολο, οι τεχνικές που ακολούθησαν χρησιμοποίησαν ανιχνευτές σημείων ενδιαφέροντος. Οι Csurka et al. [53] και οι Sivic et al. [189] επέλεξαν την εξαγωγή Harris affine σημείων και εξαγωγή χαρακτηριστικών από μια ελλειψοειδή γειτονιά τους, χρησιμοποιώντας τον ευρέως διαδεδομένο αλγόριθμο των Mikolajczyk και Schmid [137]. Οι Fei-Fei και Perona [67] επέλεξαν την εξαγωγή σημείων με βάση τον ανιχνευτή DoG και την εξαγωγή χαρακτηριστικών σε διάφορες κλίμακες. Στις περιπτώσεις αυτές, η λογική είναι ότι εξάγονται σημεία ενδιαφέροντος αμετάβλητα στην κλίμακα και κάποιους γεωμετρικούς μετασχηματισμούς. Τα χαρακτηριστικά που επιλέγονται είναι τελικά κάποιες κυκλικές ή ελλειπτικές ή ορθογώνιες περιοχές γύρω από τα σημεία αυτά, με κατάλληλη κλίμακα. Έτσι, όπως φαίνεται και στο Σχήμα 4.7, τα χαρακτηριστικά που εξάγονται από τις δύο εικόνες που περιέχουν το ίδιο *κτίριο* είναι πλέον περίπου ίδια και συνεπώς οι δύο εικόνες θα αποκτήσουν παραπλήσιες περιγραφές. Φυσικά η τεχνική αυτή οδηγεί σε αυξημένη πολυπλοκότητα και γενικά σε πολύ μεγάλο αριθμό απο χαρακτηριστικά ανά εικόνα. Στην περίπτωση που οι έννοιες που πρόκειται να ανιχνευτούν μπορούν να χαρακτηριστούν ως "αντικείμενα", σύμφωνα με τον ορισμό που δόθηκε στο Κεφάλαιο 1, οδηγεί σε καλύτερα αποτελέσματα.

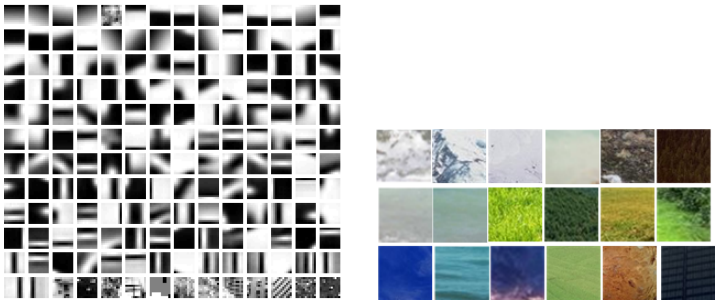
Ο τελευταίος τρόπος που μπορεί να χρησιμοποιηθεί είναι με την εφαρμογή μιας μεθόδου κατάτμησης στα εικονοστοιχεία της εικόνας. Οι Barnard et al. [8] χρησιμοποίησαν έναν αλγόριθμο κατάτμησης για να χωρίσουν μια εικόνα σε περιοχές με κριτήριο το χρώμα ή/και την υφή τους. Αυτές οι περιοχές αποτελούν τελικά τα χαρακτηριστικά που αποσυνθέτουν την εικόνα και από αυτές εξάγονται οι χαμηλού επιπέδου περιγραφές. Ένα τέτοιο παράδειγμα αποσύνθεσης απεικονίζεται στο Σχήμα

Σχήμα 4.8: *Ένα παράδειγμα αποσύνθεσης εικόνας από τις περιοχές που εξήχθησαν με την εφαρμογή ενός αλγορίθμου κατάτμησης.*

4.8. Η μέθοδος αυτή δίνει πλεονέκτημα σε περιπτώσεις που οι έννοιες που ανιχνεύονται ανήκουν στην κατηγορία των "υλικών" ή των "σκηνών". Ωστόσο, λόγω του ότι η κατάτμηση εικόνας δεν είναι ένα πλήρως επιλυμένο πρόβλημα, η εφαρμογή κάποιων αλγορίθμων κατάτμησης σε παρόμοιες εικόνες ενδέχεται πολλές φορές να οδηγήσουν σε αρκετά διαφορετικά αποτελέσματα, όπως για παράδειγμα διαφορετικό αριθμό περιοχών. Κάποιες φορές ενδέχεται να ενωθούν περιοχές και να προκύψουν χαρακτηριστικά που δεν αντιστοιχούν σε λέξεις του κώδικα, εισάγοντας το σχετικό θόρυβο στο ιστόγραμμα που τελικά δημιουργείται.

Στη συνέχεια και από τα χαρακτηριστικά που αποσυνθέτουν την εικόνα εξάγονται οπτικοί περιγραφείς. Μέσω των περιγραφέων αυτών γίνεται η σύγκριση μεταξύ των χαρακτηριστικών της εικόνας και των λέξεων του κώδικα. Στην περίπτωση του πλέγματος εξάγονται συνήθως περιγραφείς χρώματος και υφής, όπως για παράδειγμα είναι οι περιγραφείς του προτύπου MPEG-7 [43]. Για την περίπτωση των χαρακτηριστικών που αντιστοιχούν σε περιοχές γύρω από σημεία ενδιαφέροντος, κατάλληλοι περιγραφείς είναι συνήθως οι SIFT [126], SURF [9] και άλλοι παρόμοιοι. Τέλος, για την περίπτωση της κατάτμησης, επιπλέον των περιγραφέων χρώματος και υφής, σε κάποιες περιπτώσεις εξάγονται και περιγραφείς σχήματος (όταν αυτό έχει νόημα για τις έννοιες υπό ανίχνευση).

Το έτερο αρκετά σημαντικό πρόβλημα που καθορίζει σε μεγάλο βαθμό την επιτυχία ή την αποτυχία της εφαρμογής των διάφορων παραλλαγών του bag-of-words, έχει να κάνει με τη μέθοδο που ακολουθείται για το σχηματισμό του βιβλίου κωδικοποίησης, σύμφωνα με το οποίο μία περιοχή ανατίθεται σε μια λέξη. Δυστυχώς, στον τομέα αυτό δεν έχουν σημειωθεί αξιόλογες ερευνητικές προσπάθειες. Οι περισσότερες τεχνικές υιοθετούν τυπικούς αλγορίθμους συσταδοποίησης, όπως για παράδειγμα ο παραδοσιακός K-means που πρότειναν οι Hartigan και Wong [82], είτε κάποια ιεραρχική παραλλαγή του, όπως αυτή των Nister και Stewenius [154]. Στις τεχνικές αυτές συνήθως η τιμή του K επιλέγεται αυθαίρετα. Στη δευτερη περίπτωση, για λόγους ταχύτητας και ευελιξίας, ο αλγόριθμος K-means εφαρμόζεται αρχικά σε ένα σύνολο από διανύσματα χαρακτηριστικών και στη συνέχεια εφαρμόζεται σε κάθε μία από τις συστάδες που αυτός δημιουργησε. Η διαδικασία αυτή σταματάει μετά από L επίπεδα. Η τιμή του K επιλέγεται εμπειρικά, έτσι ώστε να είναι αρκετά μικρή σε σχέση με το προσδοκώμενο μέγεθος του λεξικού, το οποίο τελικά θα αποτελείται από $K * L$ λέξεις. Η ιεραρχία που κατασκευάζεται έτσι βοηθάει στο να περιοριστούν στο μέγιστο δυνατό βαθμό οι συγκρίσεις των περιοχών των εικόνων με τις λέξεις του λεξικού. Όσο για τον αριθμό των συστάδων που θα δημιουργηθούν, η πιο συνηθισμένη προσέγγιση είναι να γίνονται εκτενείς δοκιμές με διαφορετικά μεγέθη του κώδικα, και να επιλέγεται έτσι αυτό που έχει την μεγαλύτερη ακρίβεια σε κάποιο σύνολο επιβεβαίωσης (validation set). Ωστόσο, έχουν βρει εφαρμογή και κάποιες



Σχήμα 4.9: *Ένα οπτικό λεξικό που κατασκευάστηκε από κβαντοποίηση σημείων ενδιαφέροντος και ένα οπτικό λεξικό που κατασκευάστηκε από κβαντοποίηση περιοχών που προέκυψαν από κατάτμηση.*

μέθοδοι όπως το κριτήριο περιγραφής ελάχιστου μήκους (MDL), όπως είναι για παράδειγμα η παραλλαγή της που βασίζεται στη συνολική εντροπία και προτάθηκε από τους Kim και Kweon [102]. Με χρήση της τεχνικς αυτής μπορεί να γίνει επιλογή των πιο σημαντικών λέξεων, έτσι ώστε τελικά να γίνει ελαχιστοποίηση του μεγέθους της περιγραφής. Τυπικά μεγέθη κώδικα είναι από μερικές δεκάδες εώς λίγες εκατοντάδες για την περίπτωση του πλέγματος ή της κατάτμησης και από μερικές χιλιάδες εώς αρκετές δεκάδες χιλιάδες για την περίπτωση των σημείων ενδιαφέροντος.

Ο ρόλος του βιβλίου κωδικοποίησης, που θα αποκαλείται εφεξής "οπτικό λεξικό" ή "οπτικός θησαυρός" είναι να χρησιμεύσει στη δημιουργία του ιστογράμματος που θα περιγράψει τελικά μια εικόνα, καθώς το σύνολο των χαρακτηριστικών που θα εξαχθεί, θα κβαντοποιηθεί με βάση τις λέξεις του. Ένα παράδειγμα οπτικού λεξικού για την περίπτωση των σημείων ενδιαφέροντος, αλλά και ένα άλλο για την περίπτωση της κατάτμησης παρουσιάζονται στο Σχήμα 4.9. Είναι φανερό οτι τα χαρακτηριστικά που αποτελούν τους δύο αυτούς τύπους οπτικού λεξικού διαφοροποιούνται σε μεγάλο βαθμό. Επίσης, πρέπει να σημειωθεί ότι γενικά, ένα οπτικό λέξικό που βασίζεται σε σημεία ενδιαφέροντος έχει πολύ μεγαλύτερο μέγεθος από κάποιο που βασίζεται σε περιοχές.

Το τελευταίο βήμα πριν την ανίχνευση εννοιών είναι η επιλογή της παραλλαγής του μοντέλου bag-of-words με την οποία θα εκπαιδευθεί ο ανιχνευτής. Πολλές τεχνικές, κάποιες από τις οποίες έχουν γίνει γνωστές και από την ταξινόμηση κειμένου σε θεματικές κατηγορίες έχουν προταθεί και εφαρμοστεί. Οι πιο σημαντικές από τις τεχνικές αυτές παρουσιάζονται στην Ενότητα 4.4

4.4 Σχετικές Εργασίες

Η ιδέα της χρησιμοποίησης ενός λεξικού με βάση το οποίο θα περιγραφούν τα μέρη που αποσυντέθηκαν από μια εικόνα είτε μετά από συσταδοποίηση των εικονοστοιχείων της, είτε μετά από κατάτμηση έχει γίνει αντικείμενο έρευνας πολλών προσπαθειών. Σε μία από τις πιο πρώιμες ερευνητικές προσπάθειες, οι Cula και Dana [55] εργάζονται στο πεδίο της αναγνώρισης και της ταξινόμησης εικόνων με βάση την υφή. Αρχικά, κατασκευάζουν μια "βιβλιοθήκη" από τεξτόνια (textons), χρησιμοποιώντας μια μεγάλη ποικιλία από δείγματα, με τέτοιο τρόπο ώστε να αναπαρίστανται καλά τα βασικότερα χαρακτηριστικά χαμηλού επιπέδου. Στη συνέχεια γίνεται εκπαίδευση ταξινομητών με βάση ιστογράμματα τεξτονίων. Στο ίδιο πλαίσιο κινούνται και οι Leung και Malik [114], οι οποίοι κατασκευάζουν ένα λεξικό από πρωτότυπες μικρές

περιοχές, οι οποίες έχουν γεωμετρικές και φωτομετρικές ιδιότητες. Οι περιοχές αυτές αποκαλούνται "3Δ-τεξτόνια" και κάθε ένα από αυτά περιγράφεται με ένα διάνυσμα που περιγράφει τα χαρακτηριστικά της υφής του. Έτσι, κάθε "υλικό" που συναντάται σε μια εικόνα, θα περιγράφεται με τη βοήθεια του λεξικού. Οι Le Saux και Amato [111] χρησιμοποίησαν τον αλγόριθμο κατάτμησης mean-shift. Με τον τρόπο αυτό χώρισαν την εικόνα σε περιοχές, με βάση το χρώμα. Οι περιοχές αυτές δεν είναι ενωμένες μεταξύ τους, αλλά αντιστοιχούν σε περιοχές με παρόμοιο κύριο χρώμα, το οποίο και αποτελεί το χαρακτηριστικό που θα χρησιμοποιηθεί για την περιγραφή των περιοχών. Για την κατασκευή του οπτικού λεξικού χρησιμοποιούνται τεχνικές ασαφούς συσταδοποίησης. Στη συνέχεια, η εικόνα περιγράφεται με ένα διάνυσμα, το οποίο και δείχνει την ύπαρξη ή όχι των λέξεων του οπτικού λεξικού. Η τεχνική αυτή εφαρμόζεται σε πρόβλημα αναγνώρισης θεματικού πεδίου σε δελτία ειδήσεων. Οι Gokalp και Aksoy [75] έκαναν την παραδοχή ότι για την ταξινόμηση σκηνής δεν είναι απαραίτητη η ακριβής κατάτμηση. Έτσι, για να χωρίσουν την εικόνα σε περιοχές, χρησιμοποίησαν ταξινομητές εννοιών, όπως *ουρανός*, *νερό*, *δέντρο* κλπ και τους εφάρμοσαν σε κάθε εικονοστοιχείο μεμονωμένα. Με τον τρόπο αυτό κάνουν τη συσταδοποίηση της εικόνας σε περιοχές και από κάθε περιοχή εξάγουν χαρακτηριστικά και έπειτα μοντελοποιούν τις χωρικές σχέσεις τους. Για την κατασκευή του οπτικού λεξικού χρησιμοποιούν τον αλγόριθμο K-means. Οι Csurka et al. [53] χρησιμοποιούν την ιδέα του bag-of-words και αντικαθιστούν τις οπτικές λέξεις με χαρακτηριστικά σημεία (bag-of-keypoints). Επιλέγουν περιοχές της εικόνας με τη χρήση των χαρακτηριστικών SIFT. Κατασκευάζουν ένα οπτικό λεξικό επιλέγοντας και αυτοί τον αλγόριθμο K-means. Στη συνέχεια επιλέγουν τον αλγόριθμο Naive Bayes καθώς και SVM προκειμένου να επιτύχουν αναγνώριση εννοιών. Η τεχνική τους αποδείχτηκε πειραματικά ότι δεν παρουσιάζει ευαισθησία σε θόρυβο που βρίσκεται στο παρασκήνιο και παρουσιάζει καλύτερα αποτελέσματα με τη χρήση των SVM. Οι Smith et al. [196] ακολούθησαν μια διαφορετική προσέγγιση, η οποία ωστόσο μπορεί να ενταχθεί στην κατηγορία των bag-of-words τεχνικών. Αντί για οπτικούς περιγραφείς χαμηλού επιπέδου, επέλεξαν να χρησιμοποιήσουν σημασιολογικές περιγραφές. Χρησιμοποιούν εκπαιδευμένους ανιχνευτές για έννοιες όπως *πρόσωπο*, *εσωτερικός χώρος* και άλλους, και οι βαθμοί βεβαιότητας που προκύπτουν σχηματίζουν το διάνυσμα αναπαράστασης. Εφαρμόζουν την τεχνική τους στο σύνολο των βίντεο του TRECVID 2002 και πέρα από την ανίχνευση, τη χρησιμοποιούν και για δεικτοδότηση. Οι Opelt et al. [161] υιοθετούν την ιδέα του bag-of-words και την προσαρμόζουν στο πρόβλημα αναγνώρισης αντικειμένων με βάση το περίγραμμά τους. Κατασκευάζουν ένα οπτικό λεξικό χρησιμοποιώντας τμήματα από καμπύλες και κάνουν την παραδοχή ότι κάθε σχήμα μπορεί να περιγραφεί σαν ένα σύνολο από τέτοιες καμπύλες. Στη συνέχεια και για κάθε μία έννοια εκπαιδεύουν έναν ταξινομητή. Χρησιμοποιούν αδύναμους ταξινομητές και τους συνδυάζουν χρησιμοποιώντας την γνωστή τεχνική Adaboost [83]. Οι Lazebnik et al. [109] αντιμετωπίζουν το πρόβλημα της αναγνώρισης σκηνής χρησιμοποιώντας μια τεχνική που επιδιώκει να προσεγγίσει γεωμετρικές αντιστοιχίες. Η εικόνα αρχικά χωρίζεται σε πολύ λεπτομερείς περιοχές (τις οποίες και χαρακτηρίζουν "υπο-περιοχές", λόγω του μικρού τους μεγέθους) και υπολογίζονται τοπικά ιστογράμματα χαρακτηριστικών εντός των περιοχών αυτών. Ακολουθείται και πάλι μια τεχνική της κατηγορίας bag-of-words και επιτυγχάνεται, έτσι, αυξημένη ακρίβεια στην ταξινόμηση. Οι Zhao et al. [241] προσπαθούν να εμπλουτίσουν τον κλασική τεχνική bag-of-words εισάγοντας κάποιους περιορισμούς όσον αφορά τις σχέσεις μεταξύ σημείων. Η τεχνική που προτείνουν εφαρμόζεται στο πλαίσιο του TRECVID

για τις χρονιές 2003 και 2005, στην ανάκτηση σχεδόν ταυτόσημων χαρακτηριστικών καρέ και στην ανίχνευση εννοιών υψηλού επιπέδου. Πιο συγκεκριμένα, προτείνουν ταίριασμα συμμετρικών σημείων σε μια προσπάθεια να εξαλείψουν τις παρενέργειες του θορύβου κατά τη διαδικασία σύγκρισης των εικόνων. Τα πειραματικά τους αποτελέσματα δείχνουν ότι επιτυγχάνεται αξιοπρόσεκτη βελτίωση στην ακρίβεια της ανίχνευσης εννοιών. Οι Zhu και Zhang [242] υιοθετούν τον όρο "keyblocks" για να αποδώσουν το ανάλογο των λέξεων-κλειδιά στο πεδίο των εικόνων. Κατασκευάζουν ένα λεξικό το οποίο περιέχει τέτοια keyblocks σε διαφορετικές αναλύσεις. Αντιστοιχώντας τις περιοχές της εικόνας, όπως αυτές προκύπτουν έπειτα από την εφαρμογή ενός πλέγματος, σε οπτικές λέξεις, φτιάχνουν μια περιγραφή της εικόνας. Οι Fei-Fei και Perona [67] ερευνούν το πρόβλημα της αναγνώρισης φυσικών σκηνών. Στην τεχνική της, κάθε εικόνα αναπαρίσταται σαν μια συλλογή από περιοχές και κάθε μία αντιστοιχείται σε μια λέξη από ένα οπτικό λεξικό. Για τον προσδιορισμό των περιοχών πειραματίζονται με 4 διαφορετικές τεχνικές και πιο συγκεκριμένα χρησιμοποιούν ένα πλέγμα, ένα σύνολο από τυχαίες περιοχές και 2 τεχνικές που εξάγουν περιοχές γύρω από σημεία ενδιαφέροντος. Για την ταξινόμηση χρησιμοποιούν ένα Μπεϋσιανό ιεραρχικό μοντέλο. Καταλήγουν στο συμπέρασμα ότι παρά το γεγονός ότι το μοντέλο τους αποδίδει ικανοποιητικά, υπάρχει περαιτέρω χώρος για βελτίωση, με τη χρήση πιο "πλούσιων" χαρακτηριστικών και μιας ιεραρχίας από οπτικές λέξεις, αντί για ένα απλό οπτικό λεξικό.

Τα τελευταία χρόνια έχουν εμφανιστεί και κάποιες ερευνητικές προσπάθειες που προσπαθούν να ενσωματώσουν χωρική πληροφορία στο μοντέλο bag-of-words. Οι Savarese et al. [179] εργάζονται στο πιο χαμηλό επίπεδο, αυτό της περιγραφής των οπτικών χαρακτηριστικών. Στην εργασία τους χρησιμοποιούν προσαρμοζόμενα κβαντισμένα ιστογράμματα συσχέτισης τα οποία και αποκαλούν "correlatons" Το μοντέλο που προτείνουν παρουσιάζει ευρωστία σε γεωμετρικούς μετασχηματισμούς καθώς και σε περιπτώσεις που δεν είναι ορατό ολόκληρο το αντικείμενο. Τα πειράματα δείχνουν ότι η βελτίωση της απόδοσης των ταξινομητών είναι αισθητή. Οι Sudderth et al. [203] προτείνουν ένα ιεραρχικό πιθανοτικό μοντέλο για την ανίχνευση και την αναγνώριση αντικειμένων σε φυσικές σκηνές. Το μοντέλο αυτό βασίζεται σε ένα σύνολο από μέρη, τα οποία περιγράφουν την αναμενόμενη εμφάνιση και θέση των αντικειμένων. Η τεχνική αυτή χρησιμοποιεί την παραδοχή ότι τα αντικείμενα βρίσκονται στο κέντρο της εικόνας. Έτσι, κάθε κατηγορία θα έχει τη δική της κατανομή στα μέρη από τα οποία θα αποτελείται. Οι Niebles kai Fei-Fei [153] δουλεύουν σε ακολουθίες βίντεο. Κάθε ακολουθία αναπαρίσταται σαν μια συλλογή από χωρικά και χωρο-χρονικά χαρακτηριστικά, μετά την εξαγωγή στατικών και δυναμικών σημείων ενδιαφέροντος. Προτείνουν τη χρήση ενός ιεραρχικού μοντέλου, το οποίο θα μπορούσε να χαρακτηριστεί σαν μια ομοιογενής ομάδα από χαρακτηριστικά, όπως ακριβώς και στις παραδοσιακές τεχνικές bag-of-words. Το μοντέλο αυτό συνδυάζει χωρικά και χωρο-χρονικά χαρακτηριστικά και χρησιμοποιείται για την κατηγοριοποίηση πράξεων. Καταλήγουν στο συμπέρασμα ότι το μοντέλο τους βελτιώνει την ακρίβεια του απλού bag-of-words. Οι Leibe και Schiele [113] προτείνουν έναν αλγόριθμο που επιτυγχάνει κατάτμηση εικόνας, ως μια επέκταση τεχνικών αναγνώρισης αντικειμένων. Η μέθοδος αυτή μοντελοποιεί τη γνώση σχετικά με τις ιδιότητες των κατηγοριών. Κατασκευάζεται ένα οπτικό λεξικό το οποίο αποτελείται από μέρη των αντικειμένων που θα περιέχονται στις εικόνες. Έπειτα, σε κάθε εικόνα προς κατάτμηση, αναζητώνται αντιστοιχίες των περιοχών της με τις λέξεις του οπτικού λεξικού.

Οι εργασίες που ασχολούνται με το πρόβλημα του μεγέθους και της κατασκευής

του οπτικού λεξικού είναι λίγες σε αριθμό. Χαρακτηριστική είναι η εργασία των Jurie και Triggs [95], οι οποίοι προτείνουν έναν απλό αλγόριθμο για την κατασκευή οπτικών λεξικών. Η μέθοδος που προτείνουν βασίζεται σε δύο παρατηρήσεις. Πρώτον, στην περίπτωση που τα σημεία ενδιαφέροντος από τα οποία εξάγονται οι περιοχές της εικόνας είναι αραιά, χάνεται αρκετή πληροφορία, η οποία θα μπορούσε να βοηθήσει στη διαχωρισιμότητα μεταξύ των κατηγοριών. Άρα κρίνεται αναγκαίο τα σημεία αυτά να είναι όσο το δυνατόν πιο πυκνά. Δεύτερον, στην περίπτωση που τα χαρακτηριστικά είναι πολύ πυκνά, οι περιγραφείς είναι κατανεμημένοι με όχι ομοιόμορφο τρόπο στο χώρο τους, κάτι που οδηγεί μεθόδους συσταδοποίησης όπως ο αλγόριθμος K-means να συγκεντρώνουν την πλειοψηφία των κέντρων τους σε περιοχές υψηλής πυκνότητας, αγνοώντας αυτές με μέτρια πυκνότητα, κάτι που οδηγεί στο σχηματισμό λεξικών που δεν είναι πλήρη. Για να ξεπεράστούν αυτά τα προβλήματα, προτείνουν έναν δικό τους αλγόριθμο συσταδοποίησης, ο οποίος βασίζεται στον mean-shift. Χρησιμοποιώντας το λεξικό που παράγεται κατ'αυτόν τον τρόπο δείχνουν ότι επιτυγχάνεται μεγαλύτερη ακρίβεια σε σχέση με ένα που έχει κατασκευαστεί με τον αλγόριθμο K-means σε πρόβλημα ανίχνευσης εννοιών με την τεχνική bag-of-words. Οι Jiang et al. [90] προσπαθούν να μελετήσουν διεξοδικά το πρόβλημα της ανίχνευσης εννοιών με χρήση των μεθόδων bag-of-words, καθώς μελετούν και αξιολογούν διάφορους παράγοντες οι οποίοι επηρεάζουν σε μικρότερο ή μεγαλύτερο βαθμό την απόδοση των σχετικών τεχνικών. Οι παράγοντες αυτοί περιλαμβάνουν την μέθοδο με την οποία θα εξαχθούν οι περιοχές της εικόνας, το μέγεθος του οπτικού λεξικού, τον τρόπο με τον οποίο ανατίθενται βάρη στις διάφορες οπτικές λέξεις και τις συναρτήσεις πυρήνα που χρησιμοποιούνται στην επιβλεπόμενη μάθηση. Στην εργασία τους αυτή καταλήγουν στα συμπεράσματα ότι το μέγεθος του λεξικού παίζει μικρό και ασαφή ρόλο όταν δεν χρησιμοποιούνται βάρη στις διάφορες οπτικές λέξεις. Επίσης, συμπεραίνουν ότι οι τεχνικές αυτές μπορούν να λειτουργήσουν συμπληρωματικά στα χαρακτηριστικά που εξάγονται καθολικά, από ολόκληρη την εικόνα.

Καθώς το μοντέλο bag-of-words έχει τις ρίζες του στην επεξεργασία φυσικού κειμένου, είναι προφανές ότι μπορεί να εμπλουτιστεί και με άλλες τεχνικές που προέρχονται και αυτές από τον ίδιο ερευνητικό χώρο. Μια από τις πιο γνωστές τεχνικές είναι η Λανθάνουσα Σημασιολογική Ανάλυση (LSA) που προτάθηκε από τους Deerwester et al. [57]. Η τεχνική αυτή πρωτοχρησιμοποιήθηκε στην επεξεργασία φυσικής γλώσσας και εκμεταλλευόμενη τις σχέσεις ανάμεσα σε ένα σύνολο εγγράφων και τους όρους τους οποίους αυτά περιέχουν, δημιουργεί ένα σύνολο εννοιών συσχετισμένες με τα έγγραφα και τους όρους. Τα τελευταία χρόνια εμφανίστηκαν αρκετές ερευνητικές προσπάθειες που την εφαρμόζουν σε προβλήματα ανάλυσης πολυμέσων. Οι Souvannavong et al. [198] ξεκίνησαν με κατάτμηση της εικόνας και αντίστοιχησαν τις περιοχές που προέκυψαν με τις λέξεις ενός οπτικού λεξικού. Στη συνέχεια εφάρμοσαν την τεχνική LSA με σκοπό την αξιοποίηση των σχέσεων μεταξύ των τύπων περιοχής. Οι Bosch et al. [19] προτείνουν τη χρήση ενός ταξινομητή που μαθαίνει τις σκηνές και την κατανομή των χαρακτηριστικών τους σε μη σχολιασμένο σύνολο εκπαίδευσης χρησιμοποιώντας την τεχνική της πιθανοτικής Λανθάνουσας Σημασιολογικής Ανάλυσης (pLSA) που πρότεινε ο Hoffman [85] και στη συνέχεια χρησιμοποιεί την κατανομή αυτή στο σύνολο ελέγχου. Οι Sivic et al. [189] χρησιμοποιούν ένα μοντέλο που βασίζεται επίσης στην pLSA και θεωρούν τις έννοιες ως θέματα και μοντελοποιούν τις εικόνες σαν μια μείξη από θέματα, με το μοντέλο bag-of-words. Το μοντέλο που προτείνουν εφαρμόζεται στις εικόνες αφού έχουν καθοριστεί σε αυτές περιοχές με την εξαγωγή SIFT χαρακτηριστικών. Το μοντέλο τους λειτουργεί χωρίς

επίβλεψη και μπορεί να εντοπίσει αντικείμενα και τη θέση τους σε εικόνες.

4.5 Η Διαδικασία Αξιολόγησης TRECVID

Τα τελευταία χρόνια, το Εθνικό Ινστιτούτο Προτύπων και Τεχνολογίας των Η-ΠΑ (National Institute of Standards and Technology - NIST)[4] έχει ξεκινήσει την ανάπτυξη μιας πρωτοβουλίας που αποσκοπεί στην από κοινού αξιολόγηση τεχνικών ανάκτησης πληροφορίας. Από τα διάφορα έργα αξιολόγησης που διοργανώνει σε ετήσια βάση, ιδιαίτερη προσοχή αξίζει να δοθεί στην προσπάθεια που γίνεται για την αξιολόγηση των τεχνικών ανάκτησης πληροφορίας από ακολουθίες βίντεο. Το εργό αξιολόγησης που ασχολείται με αυτές τις τεχνικές ονομάζεται TRECVID [192]. Οι αλγόριθμοι και οι τεχνικές εξαγωγής εννοιών υψηλού επιπέδου που παρουσιάζονται στο Κεφάλαιο αυτό, έχουν αξιολογηθεί στη δοκιμασία του TRECVID που ασχολείται με την "Εξαγωγή Εννοιών Υψηλού Επιπέδου" (High-level concept extraction). Για το λόγο αυτό γίνεται εκτενής αναφορά στο TRECVID και τη διαδικασία που ακολουθείται στο πλαίσιό του.

Πιο λεπτομερώς, το TREC (Text REtrieval Conference)[5] είναι ένα συνέδριο με θέμα την ανάκτηση κειμένου, χρηματοδοτούμενο από το NIST. Ξεκίνησε το 1992 με σκοπό να υποστηρίξει την κοινότητα ανάκτησης πληροφορίας παρέχοντας την απαραίτητη υποδομή για αξιολόγηση μεθοδολογιών ανάκτησης κειμένου ευρείας κλίμακας. Επιδίωξη του είναι να ενθαρρύνει την έρευνα πάνω στην ανάκτηση πληροφορίας που βασίζεται σε μεγάλες συλλογές ελέγχου και να βελτιώσει την επικοινωνία ανάμεσα στην βιομηχανική και την ακαδημαϊκή κοινότητα, ανοίγοντας ανοιχτή συζήτηση για την ανταλλαγή καινούριων ερευνητικών ιδεών. Επίσης, επιδιώκει να επιταχύνει την μετατροπή της τεχνολογίας από τα εργαστήρια έρευνας σε εμπορικά προϊόντα, προωθώντας ουσιαστικές βελτιώσεις σε μεθοδολογίες ανάκτησης για προβλήματα που συναντώνται στον πραγματικό κόσμο, αλλά και να αυξήσει τη διαθεσιμότητα των κατάλληλων τεχνικών αξιολόγησης για χρήση από βιομηχανίες και ακαδημαϊκούς, συμπεριλαμβανομένης της ανάπτυξης νέων τεχνικών αξιολόγησης, καλύτερα εφαρμόσιμων στα τωρινά συστήματα. Για κάθε συνέδριο, το NIST παρέχει ένα σύνολο ελέγχου από έγγραφα και ερωτήματα. Οι συμμετέχοντες χρησιμοποιούν πάνω σε αυτά τα δεδομένα τα δικά τους συστήματα ανάκτησης και επιστρέφουν τη λίστα των εγγράφων που ανακτήθηκαν. Έπειτα, τα ανεκτημένα έγγραφα κρίνονται όσον αφορά την ορθότητα τους και γίνεται αξιολόγηση των αποτελεσμάτων. Ο κύκλος του συνεδρίου τελειώνει με μία ανοιχτή συζήτηση όπου οι συμμετέχοντες μοιράζονται τις εμπειρίες τους.

Το TREC το 2001 άρχισε να χρηματοδοτεί και μια νέα ενότητα με βίντεο, αφιερωμένη στην έρευνα στην αυτόματη κατάτμηση, δεικτοδότηση και ανάκτηση ψηφιακού βίντεο βασισμένη στο περιεχόμενο. Ξεκινώντας το 2003 αυτή η ενότητα έγινε ένα ανεξάρτητο έργο αξιολόγησης (benchmark) και ονομάστηκε TRECVID. Κύριος στόχος του είναι να προωθήσει την πρόοδο σε ανάκτηση με βάση το περιεχόμενο από ψηφιακό βίντεο μέσω ανοιχτής και βασισμένης σε συγκεκριμένα μέτρα αξιολόγησης. Πρόκειται για μία αξιολόγηση εργαστηριακής τεχνοτροπίας που προσπαθεί να μοντελοποιήσει καταστάσεις πραγματικού κόσμου ή ενδεικτικές εργασίες που εμπλέκονται σε τέτοιου είδους καταστάσεις. Το 2006 συμπληρώθηκε για το TRECVID ο δεύτερος

[4] `http://www.nist.gov`

[5] `http://trec.nist.gov/`

κύκλος αφιερωμένος σε ψηφιακό βίντεο ειδήσεων στα αγγλικά, αραβικά και κινέζικα. Συμπληρώθηκαν επίσης δύο χρόνια πιλοτικών μελετών στην εκμετάλλευση ακατέργαστου (raw) υλικού βίντεο. Το 2007 εισήλθε σε μία παρόμοια εκ πρώτης όψεως, αλλά ταυτόχρονα άγνωστη και δύσκολη περιοχή. Μετά από τα τέσσερα χρόνια σε βίντεο ειδήσεων το TRECVID 2007/2008 ελέγχει τις τρεις θεμελιώδεις εργασίες του (καθορισμός συνόρου στιγμιότυπου, εξαγωγή χαρακτηριστικών υψηλού επιπέδου, αναζήτηση) σε βίντεο που προέρχεται από αρχείο με τηλεφημερίδες, επιστημονικά νέα, ντοκιμαντέρ, εκπαιδευτικά προγράμματα κ.α. με σκοπό την εφαρμογή προϋπαρχουσών τεχνολογιών και συστημάτων σε νέα είδη δεδομένων.

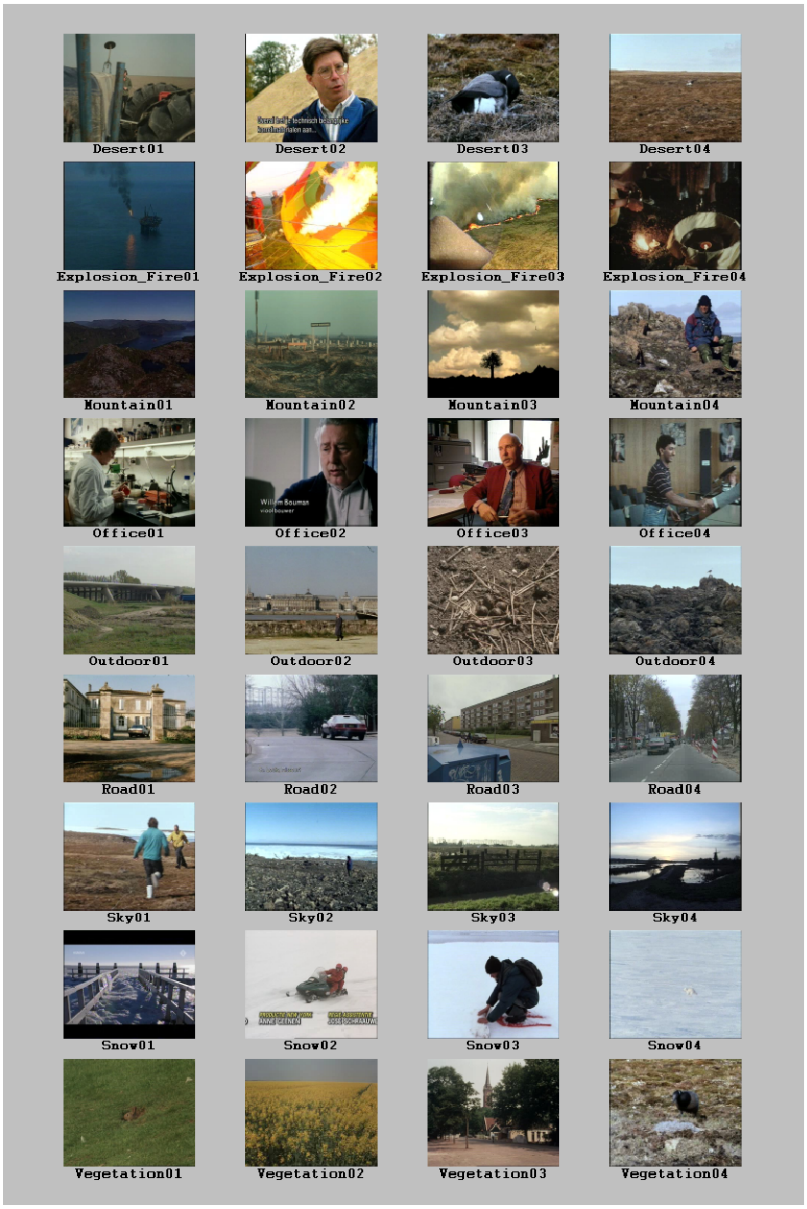
Το TRECVID παρέχει αρχικά στους συμμετέχοντες ένα σύνολο από βίντεο τα οποία είναι χωρισμένα σε πλάνα (shots), το οποίο και θα αποκαλείται εφεξής *σύνολο ανάπτυξης* (development set), όπως αποκαλείται και από το NIST. Ο αλγόριθμος με τον οποίο γίνεται η κατάτμηση του βίντεο σε πλάνα έχει κατασκευαστεί από τον Petersohn [168]. Οι συμμετέχοντες στη συνέχεια πρέπει να δημιουργήσουν συνολικούς σχολιασμούς για τα πλάνα όσον αφορά τις έννοιες με τις οποίες ασχολείται το TRECVID. Αυτό σημαίνει ότι πρέπει να σημειωθεί ο βαθμός βεβαιότητας που εκφράζει κατά πόσο ένα πλάνο απεικονίζει κάποια συγκεκριμένη έννοια ή όχι. Στη συνέχεια, βάσει αυτού του συνόλου και των σχολιασμών τους δημιουργούν συστήματα τα οποία πρέπει να είναι ικανά να αποφανθούν για το επόμενο σύνολο αν οι συγκεκριμένες έννοιες απεικονίζονται ή όχι. Το σύνολο αυτό είναι το σύνολο ελέγχου, αποτελούμενο από βίντεο χωρισμένα και πάλι σε πλάνα. Οι συμμετέχοντες χρησιμοποιώντας τα συστήματα που ανέπτυξαν με τις διάφορες προσεγγίσεις ο καθένας, παραδίδουν αποτελέσματα στο NIST για το αν το κάθε πλάνο απεικονίζει τις επιλεγμένες έννοιες και ποιες από αυτές απεικονίζει. Καθώς η συντριπτική πλειοψηφία των συμμετεχόντων δουλεύει σε μεμονωμένα καρέ των βίντεο, εξάγεται ένα χαρακτηριστικό καρέ από όλα τα πλάνα που έχουν διάρκεια μεγαλύτερη από 2 sec. Το καρέ αυτό είναι το μεσαίο κάθε πλάνου, γεγονός που όπως θα εξηγηθεί στην Ενότητα 4.7.4 δημιουργεί αρκετά προβλήματα. Τέλος το TRECVID αξιολογεί αντικειμενικά τις μεθόδους όλων των συμμετεχόντων υπολογίζοντας μια εκτίμηση της μέσης ακρίβειας σύμφωνα με τον αλγόριθμο των Yilmaz και Aslam [236] πάνω στα μεταδεδομένα που υποβλήθηκαν και αποτελούνται από όλα τα καρέ που εκτιμήθηκαν ότι περιέχουν κάθε έννοια, ταξινομημένα ως προς το βαθμό βεβαιότητας. Περισσότερες πληροφορίες μπορούν να βρεθούν στην ιστοσελίδα του TRECVID[6].

Οι έννοιες υψηλού επιπέδου, οι οποίες και περιλαμβάνονται στο TRECVID και για τις χρονιές 2007 και 2008 είναι οι εξής: *sports*, *government-leader*, *entertainment*, *corporate-leader*, *weather*, *police_security*, *court*, *military*, *office*, *prisoner*, *meeting*, *animal*, *studio*, *computer_tv-screen*, *outdoor*, *flag-US*, *building*, *airplane*, *desert*, *car*, *vegetation*, *bus*, *mountain*, *truck*, *road*, *boat_ship*, *sky*, *walking_running*, *snow*, *people-marching*, *urban*, *explosion_fire*, *waterscape_waterfront*, *natural-disaster*, *crowd*, *maps*, *face*, *charts*, *person*.

Σε αυτήν την εργασία και για λόγους που έχουν γίνει σαφείς στο Κεφάλαιο 1, επιλέγονται 9 από τις έννοιες προς εντοπισμό οι οποίες πλέον θα αναφέρονται ως: *βλάστηση* (*vegetation*), *χιόνι* (*snow*), *βουνό* (*mountain*), *δρόμος* (*road*), *γραφείο* (*office*), *εξωτερικός χώρος* (*outdoor*), *έκρηξη_φωτία* (*explosion_fire*), *έρημος* (*desert*) και *ουρανός* (*sky*). Στο Σχήμα 4.10 φαίνονται 4 παραδείγματα εικόνων από κάθε έννοια. Από τις εικόνες αυτές αναδεικνύεται επίσης και η δυσκολία του συνόλου

[6] `http://www-nlpir.nist.gov/projects/trecvid/`

των βίντεο αλλά και γενικότερα οι έννοιες με τις οποίες ασχολείται το TRECVID, καθώς υπάρχουν εικόνες από ίδιες έννοιες που διαφέρουν σε πολύ μεγάλο βαθμό στα οπτικά τους χαρακτηριστικά.



Σχήμα 4.10: *Παραδείγματα εικόνων (χαρακτηριστικών καρέ) που απεικονίζουν τις 9 έννοιες που επιλέχθηκαν από το σύνολο των εννοιών του TRECVID και επιλέχθηκαν για την αξιολόγηση των προτεινόμενων τεχνικών.*

4.6 Ανίχνευση Εννοιών Υψηλού Επιπέδου σε Εικόνες

Σε αυτό το Κεφάλαιο, ο εντοπισμός εννοιών υψηλού επιπέδου πραγματοποιείται σε ακίνητες εικόνες ή σε ακολουθίες βίντεο. Σην περίπτωση του βίντεο, κάθε μία ακολουθία χωρίζεται σε πλάνα (shots) και στη συνέχεια από κάθε πλάνο εξάγεται ένα χαρακτηριστικό καρέ (keyframe). Στο υπόλοιπο Κεφάλαιο και όπου γίνεται αναφορά

Σύμβολο	**Σημασία**	**Σχέσεις**	**Πλήθος**
V	το σύνολο των διαθέσιμων βίντεο	-	-
v_i	ένα τυχαίο βίντεο	$v_i \in V$	N_V
S	το σύνολο των εξαχθέντων πλάνων	-	-
s_i	ένα τυχαίο πλάνο	$s_i \in S$	N_S
$S(v_i)$	το σύνολο των πλάνων του v_i	$S(v_i) \subset S$	-
K	το σύνολο των εξαχθέντων χαρακτηριστικών καρέ	-	-
k_i	ένα τυχαίο χαρακτηριστικό καρέ	$k_i \in K$	$N_K = N_S$
R	το σύνολο των εξαχθέντων περιοχών από το K	-	-
r_i	μια τυχαία περιοχή	$r_i \in R$	N_R
$R(k_i)$	το σύνολο των περιοχών του k_i	$R(k_i) \subset R$	-
F	το σύνολο των διανυσμάτων χαρακτηριστικών από το R	-	-
f_i	ο συγχωνευμένος περιγραφέας της περιοχής r_i	$f_i \in F$	$N_F = N_R$

Πίνακας 4.1: *Οι συμβολισμοί που θα χρησιμοποιηθούν σε αυτό το Κεφάλαιο.*

για "εικόνα", θεωρείται ότι αυτή ενδέχεται να προέρχεται από ακολουθία βίντεο και να έχει επιλεχθεί σαν χαρακτηριστικό καρέ.

Από κάθε περιοχή εξάγονται οπτικά χαρακτηριστικά χαμηλού επιπέδου. Τα χαρακτηριστικά αυτά είναι οι MPEG-7 περιγραφείς χρώματος και υφής που έχουν περιγραφεί στο Κεφάλαιο 2. Λόγω της φύσης των εννοιών που θα ανιχνευθούν, οι περιγραφείς σχήματος, κίνησης και ήχου αποκλείονται. Με τον κατάλληλο συνδυασμό των περιγραφέων αυτών, σύμφωνα με την τεχνική συγχώνευσης που έχει περιγραφεί στο Κεφάλαιο 2, σχηματίζεται ένα *διάνυσμα χαρακτηριστικών* (feature vector) για κάθε περιοχή. Το διάνυσμα αυτό περιέχει την εξαχθείσα χαμηλού επιπέδου περιγραφή. Στη συνέχεια, και με την εφαρμογή μιας μεθόδου συσταδοποίησης σε ένα μεγάλο υποσύνολο των χαρακτηριστικών καρέ, κατασκευάζεται ένας *οπτικός θησαυρός περιοχών* (visual region thesaurus) ο οποίος στη συνέχεια θα αποκαλείται απλούστερα *οπτικός θησαυρός ή θησαυρός*. Σε αυτόν περιέχονται οι πιο χαρακτηριστικοί τύποι περιοχής (region types) που συναντώνται στο σύνολο ανάπτυξης και αναμένεται ότι με βάση αυτές θα καταστεί εφικτή η αποτελεσματική περιγραφή εικόνων από το σύνολο ελέγχου.

Πιο συγκεκριμένα, ο οπτικός θησαυρός κατασκευάζεται για να αποθηκεύσει τη γνώση που εξάγεται από το σύνολο ανάπτυξης και να διευκολύνει τη συσχέτιση των οπτικών χαρακτηριστικών χαμηλού επιπέδου, με τις έννοιες υψηλού επιπέδου. Οι τύποι περιοχής είναι οι περιοχές αυτές που βρίσκονται πιο κοντά στα κέντρα της κάθε συστάδας. Κάθε τύπος περιοχής περιγράφεται από το διάνυσμα χαρακτηριστικών του. Υπολογίζοντας τις αποστάσεις των περιοχών της εικόνας από τους τύπους περιοχής του οπτικού θησαυρού, σχηματίζεται ένα *διάνυσμα αναπαράστασης* (model vector) το οποίο κωδικοποιεί το οπτικό περιεχόμενο της εικόνας με βάση το θησαυρό.

Το επόμενο (προαιρετικό) στάδιο της μεθόδου αποτελεί η εφαρμογή της τεχνικής LSA, προκειμένου να γίνει εκμετάλλευση των σχέσεων μεταξύ των τύπων περιοχών. Το πρόβλημα έτσι οδηγείται στο "χώρο εννοιών" (concept space), ο οποίος και έχει μικρότερη διάσταση σε σχέση με το χώρο εισόδου.

Τέλος, για κάθε μία έννοια, εκπαιδεύεται ένα νευρωνικό δίκτυο, το οποίο και πραγματοποιεί τελικά την ανίχνευσή της. Η είσοδος ενός τέτοιου δικτύου, είναι ένα διάνυσμα αναπαράστασης είτε στον χώρο των τύπων περιοχών (αν δεν έχει εφαρμοστεί η τεχνική LSA), είτε στον χώρο των εννοιών (αν εφαρμοστεί η τεχνική LSA). Στις παρακάτω Ενότητες περιγράφονται αναλυτικά τα βήματα του αλγορίθμου ανί-

χνευσης των εννοιών υψηλού επιπέδου και στον Πίνακα 4.1 συνοψίζονται οι διάφοροι συμβολισμοί που θα χρησιμοποιηθούν στις επόμενες Ενότητες.

4.6.1 Τα Χαρακτηριστικά Χαμηλού Επιπέδου

Για την αναπαράσταση των οπτικών χαρακτηριστικών χαμηλού επιπέδου ενός χαρακτηριστικού καρέ, χρησιμοποιούνται περιγραφείς χρώματος και υφής. Πιο συγκεκριμένα, για τα χαρακτηριστικά χρώματος χρησιμοποιούνται και οι τέσσερις περιγραφείς χρώματος του MPEG-7 που έχουν περιγραφεί στο Κεφάλαιο 2, δηλαδή ο Κλιμακωτός Περιγραφέας Χρώματος, ο Περιγραφέας Κύριων Χρωμάτων, ο Περιγραφέας Δομής Χρώματος και ο Περιγραφέας Διάταξης Χρώματος. Για τα χαρακτηριστικά της υφής χρησιμοποιείται μόνο ο Περιγραφέας Ιστογράμματος Ακμών. Το κάθε χαρακτηριστικό καρέ χωρίζεται σε χονδροειδείς περιοχές με την εφαρμογή ενός εργαλείου κατάτμησης με βάση το χρώμα και έπειτα για κάθε περιοχή εξάγονται οι 5 περιγραφείς με τη μορφή 5 διανυσμάτων, όπως αυτά έχουν περιγραφεί στο Κεφάλαιο 1. Λόγω του μεταβλητού μήκους του περιγραφέα κύριων χρωμάτων χρησιμοποιείται μόνο το χρώμα με το μεγαλύτερο ποσοστό εμφάνισης. Τα διανύσματα αυτά συγχωνεύονται τελικά σε ένα μοναδικό διάνυσμα, το οποίο θα αποκαλείται εφεξής *διάνυσμα χαρακτηριστικών*.

Για μια τυχαία περιοχή r_i, το διάνυσμα χαρακτηριστικών f_i που της αντιστοιχεί περιγράφεται ως

$$f_i = f(r_i) = \Big[CLD(r_i), DCD(r_i), CSTD(r_i), SCD(r_i), EHD(r_i)\Big], \qquad (4.1)$$

όπου $CLD(r_i)$ ο Περιγραφέας Διάταξης Χρώματος, $DCD(r_i)$ ο Περιγραφέας Κύριων Χρωμάτων, $CSTD(r_i)$ ο Περιγραφέας Δομής Χρώματος, $SCD(r_i)$ ο Περιγραφέας Κλιμακωτού Χρώματος και $EHD(r_i)$ ο Περιγραφέας Ιστογράμματος Ακμών για την περιοχή r_i.

4.6.2 Κατασκευή Οπτικού Θησαυρού

Στην περίπτωση που η ανίχνευση εννοιών γίνεται σε ακολουθίες βίντεο, εαν παρατηρήσει κανείς το σύνολο των χαρακτηριστικών καρέ που εξήχθησαν, μπορεί εύκολα να διαπιστώσει την ύπαρξη πολλών παρόμοιων καρέ, καθώς το σύνολό τους προέρχεται από την ίδια συλλογή ακολουθιών βίντεο, αλλά και παρόμοιων περιοχών. Επίσης μπορεί να παρατηρηθεί ότι εικόνες που περιέχουν την ίδια έννοια υψηλού επιπέδου αποτελούνται συνήθως και από παρόμοιες περιοχές και συνεπώς έχουν και παρόμοιες περιγραφές χαμηλού επιπέδου.

Με βάση τις παραπάνω παρατηρήσεις, διαφαίνεται ότι θα ήταν αποδοτικό να οργανωθούν όλες οι περιοχές όλων των εικόνων που θα χρησιμοποιηθούν σαν σύνολο εκπαίδευσης με κάποιον τρόπο που να διευκολύνει την περιγραφή μιας εικόνας με βάση τις περιοχές από τις οποίες αυτή αποτελείται. Για να επιτευχθεί αυτό, ένας αλγόριθμος συσταδοποίησης εφαρμόζεται στις περιοχές του συνόλου εκπαίδευσης.

Μετά τη συσταδοποίηση των περιοχών, μία πρώτη παρατήρηση είναι ότι κάθε συστάδα μπορεί να περιέχει περιοχές που ανήκουν σε διαφορετικές έννοιες υψηλού επιπέδου. Αντίθετα, περιοχές που ανήκουν στην ίδια έννοια υψηλού επιπέδου μπορεί να έχουν καταταχθεί σε διαφορετικές συστάδες. Για παράδειγμα, περιοχές που ανήκουν στην έννοια *βλάστηση* μπορεί να ανήκουν σε περισσότερες από μία συστάδες

διαφέροντας για παράδειγμα στο στο χρώμα των *φύλλων* των δέντρων. Επίσης, μια συστάδα μπορεί να αποτελείται από *μπλε* περιοχές, οι οποίες να προέρχονται από έννοιες όπως *θάλασσα*, *ουρανός* κλπ, οι οποίες μπορεί να είναι τόσο όμοιες, ώστε και ένας άνθρωπος να αδυνατεί να τις ξεχωρίσει. Από κάθε συστάδα επιλέγεται η περιοχή εκείνη που βρίσκεται πιο κοντά στο κέντρο της και θα αναφέρεται σαν *τύπος περιοχής*. Πρέπει εδώ να σημειωθεί ότι οι περιοχές αναπαρίστανται με το διάνυσμα χαρακτηριστικών τους.

Τελικά, ο οπτικός θησαυρός T που κατασκευάστηκε μπορεί να περιγραφεί ως

$$T = \{w_i\}, \quad i = 1, 2, \ldots, N_T, \quad w_i \subset R \ . \tag{4.2}$$

Για το σύνολο των σχέσεών του ισχύει ότι

$$\bigcup_i w = R, \quad i = 1 \ldots N_{\mathrm{T}} \tag{4.3}$$

και

$$\bigcap_{i,j} w = \emptyset, \quad i \neq j \ , \tag{4.4}$$

όπου $\mathrm{N_T}$ ο αριθμός των τύπων περιοχής του θησαυρού, (και προφανώς και των συστάδων), w_i είναι η i-οστή συστάδα, η οποία μπορεί να θεωρηθεί ως ένα σύνολο από περιοχές εικόνων που ανήκουν στο R, όπως φαίνεται και στην (4.2). Τύπος περιοχής ή αλλιώς *οπτική λέξη* (visual word) του οπτικού θησαυρού ονομάζεται αυτή που βρίσκεται πιο κοντά στο κέντρο της κάθε συστάδας και επιλέγεται για να την αντιπροσωπεύσει. Επίσης, από τις (4.3) και (4.4) φαίνεται ότι η ένωση όλων των συστάδων είναι όλο το σύνολο R, εφόσον όλες οι περιοχές έχουν χρησιμοποιηθεί για τη συσταδοποίηση και διαφορετικές συστάδες δεν έχουν κοινές περιοχές (δηλαδή η συσταδοποίηση είναι σαφής).

Γενικά, ένας "θησαυρός" συνδυάζει μία λίστα με τους όρους ενός συγκεκριμένου πεδίου γνώσης και ένα σύνολο από σχετικούς όρους για κάθε έναν από τους όρους της λίστας, οι οποίοι αποκαλούνται ως τα *συνώνυμά* του. Ο οπτικός θησαυρός που κατασκευάστηκε με τη διαδικασία που παρουσιάστηκε περιέχει όλους τους τύπους περιοχών που συναντώνται στο σύνολο εκπαίδευσης. Το κάθε κέντρο $z(w_i)$ υπολογίστηκε σαν το "κέντρο μάζας" όλων των διανυσμάτων χαρακτηριστικών που ανήκουν στην ίδια συστάδα, δηλαδή ως

$$z(w_i) = \frac{1}{|w_i|} \sum_{r \in w_i} f(r) \ , \tag{4.5}$$

όπου με $|w_i|$ συμβολίζεται το πλήθος των στοιχείων του w_i. Έτσι, το διάνυσμα χαρακτηριστικών που επιλέγεται να αντιπροσωπεύει τον κάθε τύπο περιοχής υπολογίζεται ως

$$f(w_i) = f\Big(\arg\min_{r \in w_i} \Big\{ d\big(f(r), z(w_i)\big) \Big\} \Big) \ . \tag{4.6}$$

Ο κάθε τύπος περιοχής αναπαρίσταται από ένα διάνυσμα χαρακτηριστικών το οποίο περιέχει όλη την εξαχθείσα πληροφορία χαμηλού επιπέδου για την περιοχή. Όπως είναι προφανές, μία περιγραφή χαμηλού επιπέδου δεν περιέχει καμία σημασιολογική πληροφορία και αποτελεί απλώς μία τυποποιημένη αναπαράσταση των οπτικών χαρακτηριστικών. Μία έννοια υψηλού επιπέδου περιέχει μόνο σημασιολογική πληροφορία.

Με το σκεπτικό αυτό, ένας τύπος περιοχής θα μπορούσε να περιγραφεί ως κάτι ενδιάμεσο στα χαρακτηριστικά υψηλού και χαμηλού επιπέδου. Περιέχει τις απαραίτητες πληροφορίες για να περιγράψει τυποποιημένα τα χαρακτηριστικά υφής και χρώματος αλλά μπορεί επίσης να περιγράφει με μία περιγραφή *χαμηλότερου* επιπέδου από τις έννοιες υψηλού επιπέδου, η οποία διαθέτει σημασιολογία. Για παράδειγμα κάποιος μπορεί να αντιληφθεί διαισθητικά έναν τύπο περιοχής σαν *πράσινη περιοχή με τραχιά υφή*.

Για το σχηματισμό του οπτικού θησαυρού είναι απαραίτητος ένας μεγάλος αριθμός από διανύσματα χαρακτηριστικών από περιοχές εικόνων. Σε πρακτικά προβλήματα που το σύνολο όλων των περιοχών δεν είναι πολύ μεγάλο, τότε όλες οι περιοχές θα χρησιμοποιηθούν για τον σχηματισμό του θησαυρού. Συγκεκριμένα ο αλγόριθμος συσταδοποίησης θα εφαρμοστεί στα διανύσματα χαρακτηριστικών όλων των χαρακτηριστικών καρέ. Αντίθετα, σε προβλήματα που το διαθέσιμο σύνολο εκπαίδευσης είναι πολύ μεγάλο (όπως το σύνολο ανάπτυξης του TRECVID), δεν είναι πάντα εφικτό να χρησιμοποιηθεί ολόκληρο, μιας και αυτό συνήθως απαιτεί πολλούς και χρονοβόρους υπολογισμούς. Έτσι επιλέγεται ένα υποσύνολο αυτών. Ένας πρακτικός τρόπος με τον οποίο μπορεί να γίνει αυτό είναι να συλλέγονται περιοχές από εικόνες οι οποίες απεικονίζουν τις επιλεγμένες προς εντοπισμό έννοιες και στην συνέχεια με τυχαίο τρόπο κάποιες ακόμα περιοχές για χάρη μεγαλύτερης γενίκευσης. Οι περιοχές αυτές εισάγουν τον απαραίτητο "θόρυβο" στον οπτικό θησαυρό, κάτι που θα βοηθήσει στη γενίκευση. Αυτό το υποσύνολο θα χρησιμοποιηθεί για να κατασκευαστεί ο θησαυρός στην περίπτωση ενός μεγάλου αριθμού συνολικών περιοχών. Στην περίπτωση αυτή δεν ισχύει η (4.3) αλλά ότι

$$\bigcup_i w \subset R, \quad i = 1 \dots N_{\mathrm{T}} \ . \tag{4.7}$$

4.6.3 Κατασκευή Διανυσμάτων Αναπαράστασης Εικόνων

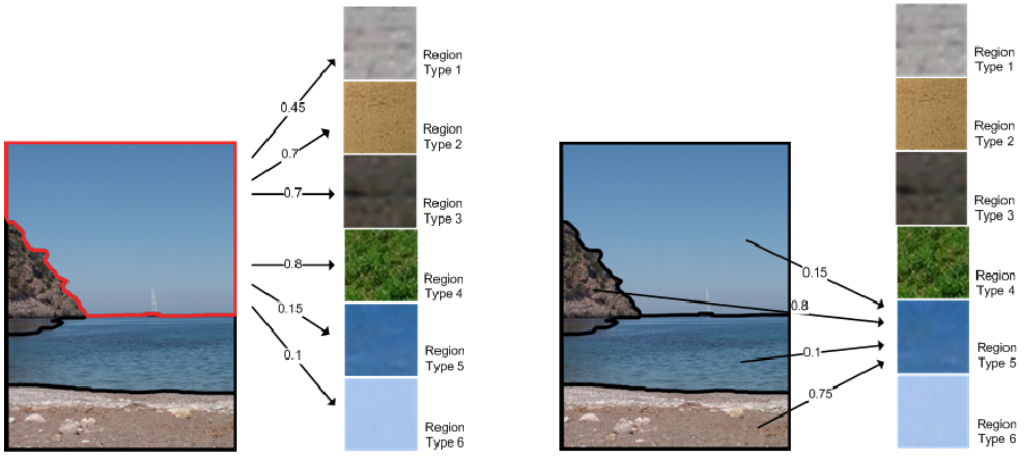
Σε αυτή την Ενότητα παρουσιάζεται ο αλγόριθμος που χρησιμοποιήθηκε για την κατασκευή της περιγραφής μια εικόνας που έχει κατατμηθεί σε περιοχές, με τη βοήθεια ενός οπτικού θησαυρού. Για τον υπολογισμό της απόστασης μεταξύ δύο διανυσμάτων χαρακτηριστικών επιλέγεται η Ευκλίδεια απόσταση που έχει χρησιμοποιηθεί με επιτυχία σε πολλά προβλήματα ταιριάσματος περιγραφεών. Έτσι, για δύο τυχαίες περιοχές r_1, r_2 , με διανύσματα χαρακτηριστικών f_1, f_2, αντίστοιχα, η απόστασή τους $d(f_1, f_2)$ θα υπολογίζεται ως

$$d(f_1, f_2) = \sqrt{\sum_{i=1}^{n} (f_1^i - f_2^i)^2} \ , \tag{4.8}$$

όπου f_i^j είναι η j-οστή συνιστώσα του διανύσματος χαρακτηριστικών f_i.

Πρέπει να σημειωθεί ότι η επιλογή της Ευκλίδειας απόστασης δεν έρχεται σε αντίθεση με όσα ορίζει το MPEG-7, μιας και αυτό δεν ορίζει αυστηρά μέτρα απόστασης για τους περιγραφείς, παρά μόνο προτείνει ενδεικτικά αυτούς τους τύπους που περιγράφηκαν στο Κεφάλαιο 2. Έτσι, επιτρέπεται η χρήση και άλλων αποστάσεων οι οποίες μπορεί να είναι πιο αποτελεσματικές, ανάλογα και με την εφαρμογή στην οποία χρησιμοποιούνται.

Στη συνέχεια και έχοντας υπολογίσει την απόσταση της κάθε περιοχής της εικόνας από όλους τους τύπους περιοχής του οπτικού θησαυρού, το διάνυσμα αναπαράστασης του κάθε χαρακτηριστικού καρέ που σημασιολογικά περιγράφει το οπτικό



Σχήμα 4.11: *Αποστάσεις ανάμεσα σε περιοχές εικόνας και τύπους περιοχής. Αριστερά: αποστάσεις ανάμεσα σε μία περιοχή της εικόνας και σε όλους τους τύπους περιοχής. Δεξιά: αποστάσεις ανάμεσα σε όλες τις περιοχές μίας εικόνας και ενός τύπου περιοχής.*

περιεχόμενο της εικόνας σχηματίζεται κρατώντας τη μικρότερη απόσταση από κάθε τύπο περιοχής του θησαυρού. Πιο συγκεκριμένα, το διάνυσμα αναπαράστασης m_i που περιγράφει το χαρακτηριστικό καρέ k_i ορίζεται ως

$$m_i = \Big[m_i(1)\ m_i(2)\ \ldots\ m_i(j)\ \ldots\ m_i(N_T)\Big], \quad i = 1, 2, \ldots, N_K\ , \tag{4.9}$$

όπου

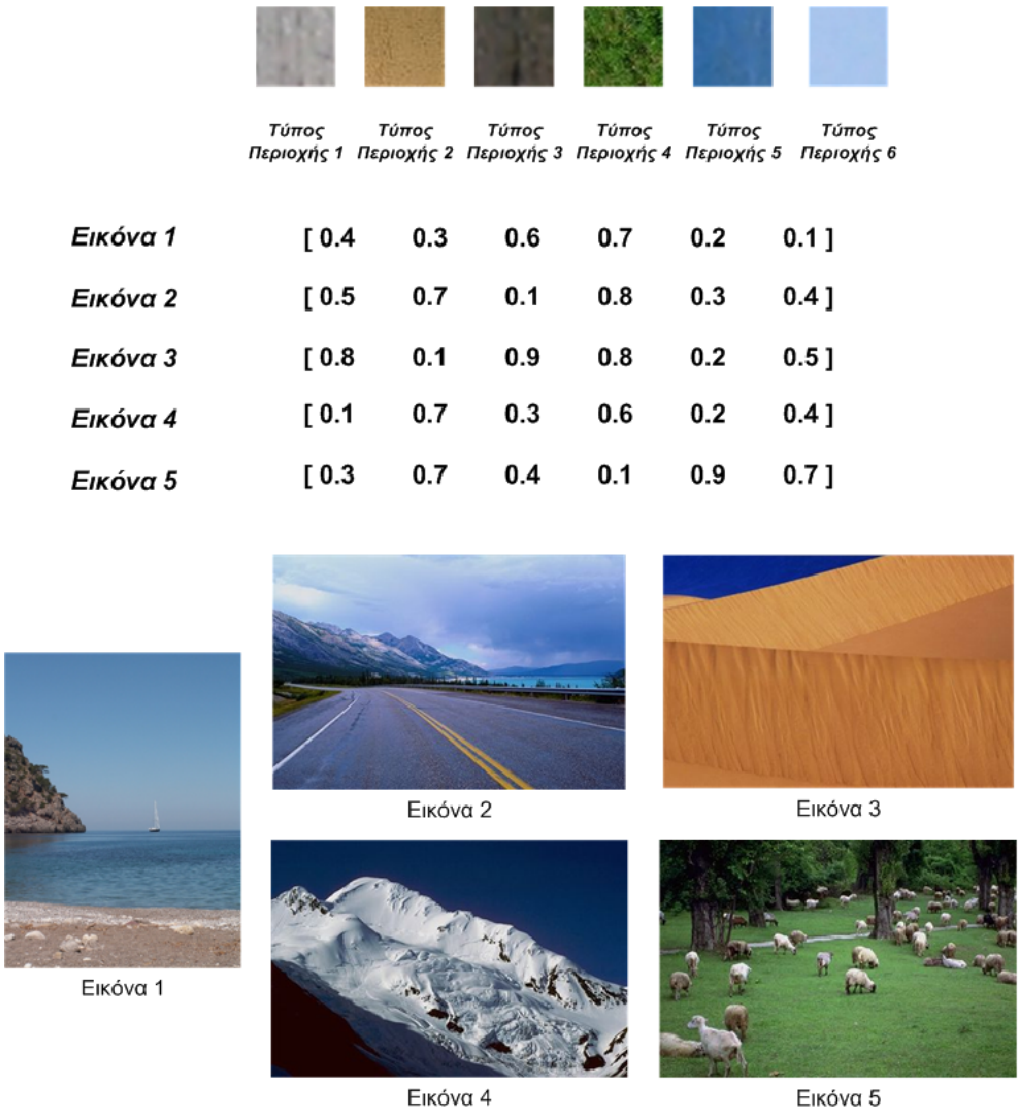
$$m_i(j) = \min_{r \in R(k_i)} \Big\{ d\Big(f(w_j), f(r) \Big) \Big\}, \quad i = 1, 2 \ldots, N_K, \quad j = 1, 2, \ldots, N_T\ . \tag{4.10}$$

Στο Σχήμα 4.11 και για λόγους απλότητας της παρουσίασης, απεικονίζεται ένα τεχνητό παράδειγμα που περιλαμβάνει μία εικόνα χωρισμένη σε χονδροειδείς περιοχές, καθώς και ένας απλός οπτικός θησαυρός που αποτελείται από 6 τύπους περιοχών. Αριστερά παρουσιάζονται οι αποστάσεις για την περιοχή του ουρανού από κάθε τύπο περιοχής και δεξιά οι αποστάσεις κάθε περιοχής της εικόνας από τον *τύπο περιοχής 5*. Το διάνυσμα αναπαράστασης σχηματίζεται από τις μικρότερες αποστάσεις για κάθε τύπο περιοχής. Στην προκειμένη περίπτωση για τον *τύπο περιοχής 5*, η ελάχιστη απόσταση είναι ίση με 0.1. Το διάνυσμα αναπαράστασης για αυτή την εικόνα και για το συγκεκριμένο θησαυρό περιγράφεται ως

$$m = [m(1)\ m(2)\ m(3)\ m(4)\ m(5)\ m(6)]\ , \tag{4.11}$$

άρα το στοιχείο $m(5)$ που αντιστοιχεί στη σχέση της εικόνας με τον 5ο τύπο περιοχής του θησαυρού θα πάρει τιμή ίση με 0.1 κ.ο.κ. Υπολογίζοντας όλες τις αποστάσεις ανάμεσα σε όλες τις περιοχές της εικόνας και σε όλους τους τύπους περιοχής, συνολικά δηλαδή $4 \times 6 = 24$ αποστάσεις, σχηματίζεται το διάνυσμα αναπαράστασης.

Στο Σχήμα 4.12 φαίνονται τα διανύσματα αναπαράστασης για 5 εικόνες και για τον απλό οπτικό θησαυρό 6 περιοχών, ο οποίος επίσης απεικονίζεται στο ίδιο Σχήμα.



Σχήμα 4.12: *Παραδείγματα διανυσμάτων αναπαράστασης με χρήση του οπτικού θησαυρού του Σχήματος 4.11*

4.6.4 Λανθάνουσα Σημασιολογική Ανάλυση (Latent Semantic Analysis - LSA)

Στη συνέχεια η τεχνική της Λανθάνουσας Σημασιολογικής Ανάλυσης υιοθετείται για να διερευνηθεί εφόσον μπορεί να ενισχύσει το μοντέλο που παρουσιάστηκε εώς τώρα. Η τεχνική αυτή εκμεταλλεύεται τις συσχετίσεις ανάμεσα σε έννοιες (εδώ σε τύπους περιοχής) και μεταφέροντας το πρόβλημα σε έναν νέο χώρο, μειώνοντας ταυτόχρονα τη διάστασή του, προσπαθεί να δημιουργήσει μια βελτιωμένη περιγραφή του οπτικού περιεχομένου των εικόνων.

Για να γίνει η χρήση της τεχνικής αυτής η οποία εφαρμόζεται πάνω σε πίνακες που περιέχουν τον αριθμό των εμφανίσεων κάθε λέξης σε κάθε έγγραφο, γίνεται ένας απαραίτητος μετασχηματισμός των διανυσμάτων αναπαράστασης. Υπενθυμίζεται ότι τα διανύσματα αναπαράστασης περιέχουν για κάθε τύπο περιοχής τις ελάχιστες αποστάσεις από όλες τις περιοχές της εικόνας. Έτσι ένας τύπος περιοχής που εμφανίζει οπτική ομοιότητα με έναν τύπο περιοχής θα έχει χαμηλή τιμή απόστασης με αυτόν.

Όλες οι αποστάσεις διαιρούνται με την μέγιστη απόσταση και αφαιρούνται από το 1. Έτσι κανονικοποιούνται στο διάστημα [0,1], με τιμές κοντά στο 1 να έχουν εκείνες οι περιοχές οι οποίες είναι πολύ όμοιες με τους τύπους περιοχής. Πλέον τα διανύσματα αναπαράστασης m_i μετασχηματίζονται στα m'_i, τα οποία ορίζονται ως

$$m'_i = \Big[m'_i(1)\ m'_i(2)\ \ldots\ m'_i(j)\ \ldots m'_i(N_T)\Big], \quad i = 1, 2, \ldots, N_K\ , \tag{4.12}$$

όπου

$$m'_i(j) = 1 - \frac{m_i(j)}{d_{max}} \tag{4.13}$$

και

$$d_{max} = \max_{r \in R} \Big\{ d\Big(f(w_j), f(r) \Big) \Big\}, \quad j = 1, 2, \ldots, N_T\ . \tag{4.14}$$

Πιο συγκεκριμένα, έστω $\mathcal{M}$ ένας πίνακας του οποίου το (i, j) στοιχείο περιγράφει την ύπαρξη του όρου i στο έγγραφο j. Σε αυτή την προσέγγιση, οι όροι αντιστοιχούν στους τύπους περιοχής του θησαυρού και το έγγραφο σε μία δεδομένη εικόνα, καθώς μία εικόνα θεωρείται σαν ένα "πολυμεσικό έγγραφο", του οποίου οι λέξεις είναι οι τύποι περιοχής που αντιστοιχούν στις περιοχές της. Αν ο Πίνακας $\mathcal{M}$ οριστεί ως

$$\mathcal{M} = \begin{pmatrix} m'_1(1) & \ldots & m'_{N_K}(1) \\ . & & . \\ . & & . \\ . & \ddots & . \\ m'_1(N_T) & \ldots & m'_{N_K}(N_T) \end{pmatrix}\ , \tag{4.15}$$

τότε μία γραμμή του πίνακα $q_i^T = [m'_1(i), \ldots, m'_{N_K}(i)]$ είναι ένα διάνυσμα που περιγράφει τη σχέση των όρων (τύποι περιοχής), συγκεκριμένα του τύπου περιοχής w_i, με το κάθε έγγραφο (εικόνα). Τα στοιχεία δηλαδή του διανύσματος γραμμής i είναι τα $m'_j(i)$ στοιχεία των μετασχηματισμένων διανυσμάτων αναπαράστασης $m'_j, \quad j = 1 \ldots N_K$. Κάθε στήλη του $\mathcal{M}$, που περιγράφεται ως

$$m'_j = \Big(\ m'_j(1) \ldots m'_j(N_T)\ \Big)^T \tag{4.16}$$

είναι ένα διάνυσμα το οποίο αντιστοιχεί σε ένα συγκεκριμένο έγγραφο (εικόνα) και περιγράφει τη σχέση του με κάθε όρο (τύπος περιοχής). Αυτή η σχέση αντιστοιχεί στη βεβαιότητα για την ύπαρξη του είδους περιοχής στη δεδομένη εικόνα. Είναι λοιπόν προφανές ότι το m'_j είναι το μετασχηματισμένο διάνυσμα αναπαράστασης της εικόνας αυτής. Επομένως ισχύει ότι

$$\mathcal{M} = [m'^T_1, \ldots, m'^T_{N_K}]\ . \tag{4.17}$$

Το εσωτερικό γινόμενο $q_i^T q_p$ ανάμεσα σε δύο διανύσματα όρων δίνει την συσχέτιση τους και ο πίνακας $\mathcal{M}\mathcal{M}^T$ περιέχει όλα αυτά τα εσωτερικά γινόμενα. Επιπλέον ο πίνακας $\mathcal{M}^T\mathcal{M}$ περιέχει όλα τα εσωτερικά γινόμενα ανάμεσα στα διανύσματα εγγράφων $m'^T_i m'_p$, περιγράφοντας τη συσχέτιση τους πάνω στους όρους. Υποθέτοντας ότι υπάρχει μία αποσύνθεση του $\mathcal{M}$ που δίνεται από τη σχέση $\mathcal{M} = \mathbf{U}\Sigma\mathbf{V}^T$, όπου $\mathbf{U}$ και $\mathbf{V}$ είναι ορθοκανονικοί πίνακες και ο Σ διαγώνιος. Αυτή η αποσύνθεση είναι γνωστή ως *αποσύνθεση ιδιαζουσών τιμών* (SVD) και περιγράφεται ως

$$\mathcal{M} = \left(\left(\mathbf{u_1}\right)\ldots\left(\mathbf{u_{N_T}}\right)\right)\begin{pmatrix} \sigma_1 & \ldots & 0 \\ . & & . \\ . & & . \\ . & \ddots & . \\ 0 & \ldots & \sigma_{N_T} \end{pmatrix}\left(\left(\mathbf{v_1}\right)\ldots\left(\mathbf{v_{N_T}}\right)\right)^{\mathcal{T}}, \tag{4.18}$$

όπου σ_i είναι οι ιδιοτιμές και $\mathbf{u}_i$,$\mathbf{v}_i$ είναι τα αριστερά και δεξιά ιδιοδιανύσματα αντίστοιχα.

Τώρα οι N_L μεγαλύτερες ιδιοτιμές μπορούν να κρατηθούν από τον Σ μαζί με τις αντίστοιχες στήλες του $\mathbf{U}$ και σειρές του $\mathbf{V}$. Μία εκτίμηση του $\mathcal{M}$ είναι η

$$\hat{\mathcal{M}} = \mathcal{M}_{N_L} = \mathbf{U}_{N_L}\Sigma_{N_L}\mathbf{V}_{N_L}^T \ . \tag{4.19}$$

Με αυτόν τον τρόπο επιτυγχάνεται μία απεικόνιση των διανυσμάτων όρων και εγγράφων στον χώρο των "εννοιών". Αν $\hat{q}_i$ και $\hat{m}_i$ είναι τα διανύσματα που δηλώνουν την ύπαρξη των όρων στις έννοιες και την σχέση ανάμεσα σε έγγραφα και έννοιες αντίστοιχα, τότε η μετατροπή ενός διανύσματος εγγράφων στον χώρο των εννοιών πετυχαίνεται με την χρήση του πίνακα Σ και $\mathbf{U}$ με την μετατροπή

$$\hat{m}_i = \Sigma_{N_L}^{-1}\mathbf{U}_{N_L}^T m'_i \ . \tag{4.20}$$

Στην προσέγγιση που παρουσιάζεται στο κεφάλαιο αυτό, τα διανύσματα αναπαράστασης του συνόλου εκπαίδευσης εξάγονται με τη χρήση του οπτικού θησαυρού όπως εξηγήθηκε στην Ενότητα 4.6.2. Ακολούθως, εφαρμόζεται η LSA κι έτσι ορίζονται οι πίνακες Σ και $\mathbf{U}$. Τελικά κάθε διάνυσμα αναπαράστασης μεταφέρεται στο χώρο των εννοιών.

4.6.5 Ανίχνευση Εννοιών Υψηλού Επιπέδου σε Εικόνες

Για κάθε έννοια ένα νευρωνικό δίκτυο τύπου πολυεπίπεδου perceptron (MLP), εκπαιδεύτηκε να λύνει το (δυαδικό) πρόβλημα, της ύπαρξης ή όχι της έννοιας υπό εξέταση. Η είσοδος του δικτύου είναι ένα διανύσμα m_i (ή $\hat{m}_i$, στην περίπτωση που εφαρμόζεται η τεχνική LSA). Η έξοδός του q_i είναι κανονικοποιημένη στο διάστημα $[0, 1]$ και αντιστοιχεί στο βαθμό βεβαιότητας της ύπαρξης της έννοιας στην εικόνα. Τιμές κοντά στο 1 δηλώνουν υψηλή βεβαιότητα απεικόνισης ενώ τιμές κοντά στο 0 δηλώνουν υψηλή βεβαιότητα μη απεικόνισης. Η ύπαρξη του βαθμού βεβαιότητας αντί για μια δυαδική απάντηση βοηθάει την αξιολόγηση του συστήματος με τη χρήση συγκεκριμένων μέτρων, όπως θα φανεί στην Ενότητα 4.7. Προκειμένου να επιλεχθούν μόνο τα καρέ που θεωρείται ότι περιέχεται η έννοια υπό εξέταση, είναι απαραίτητη η χρήση ενός κατωφλίου, μέσω του οποίου να μετασχηματίζεται η έξοδος σε δυαδική απάντηση της ύπαρξης ή μη της κάθε έννοιας. Όπως θα φανεί ακολούθως, είναι απαραίτητο για κάθε έννοια να επιλεχθεί κατάλληλη τιμή κατωφλίου ώστε να επιτυγχάνεται τελικά το καλύτερο δυνατό αποτέλεσμα.

Έστω c_i η κάθε έννοια που έχει επιλεχθεί να εντοπιστεί, C το σύνολο όλων των εννοιών και N_C το πλήθος τους. Η δυαδική απόφαση για ένα χαρακτηριστικό καρέ $k \in K$ για το αν απεικονίζει ή όχι την έννοια c_i συμβολίζεται με $b_i(k)$. Το σύνολο

των χαρακτηριστικών καρέ τα οποία τελικά εντοπίζονται πως απεικονίζουν την έννοια c_i συμβολίζονται με D_i. Δηλαδή

$$b_i(k) \in \{0,1\}, \quad k \in K, \quad i = 1 \dots N_C \tag{4.21}$$

και

$$D_i = \{k \in K : \quad b_i(k) = 1, \quad i = 1 \dots N_C\} \ . \tag{4.22}$$

Με τη χρήση δηλαδή ενός κατωφλίου h_j για την έννοια c_j πάνω στις τιμές q_i θα προκύπτει $b_j(k_i) = 0$ εάν $q_i \leq h_j$, ενώ $b_j(k_i) = 1$ εάν $q_i > h_j$.

4.7 Πειραματικά Αποτελέσματα

Στην Ενότητα αυτή παρουσιάζονται τα πειραματα που πραγματοποιήθηκαν προκειμένου να αξιολογηθούν οι τεχνικές ανίχνευσης εννοιών υψηλού επιπέδου που παρουσιάστηκαν στο Κεφάλαιο αυτό. Η εφαρμογή των τεχνικών έγινε στα πλαίσια της δοκιμασίας ανίχνευσης εννοιών υψηλού επιπέδου του TRECVID. Εφαρμόστηκε ο αλγόριθμος ανίχνευσης που περιγράφηκε στην Ενότητα 4.6 και επιλέχθηκαν κατάλληλα μέτρα για την αξιολόγηση της αποτελεσματικότητας του αλγορίθμου στην ανίχνευση των εννοιών. Πέρα από τη συλλογή βίντεο του TRECVID 2005-2006, ο προτεινόμενος αλγόριθμος δοκιμάστηκε και σε ένα σύνολο εικόνων από τη συλλογή του Corel[7].

4.7.1 Κριτήρια Αξιολόγησης

Το σύνολο δεδομένης αλήθειας περιλαμβάνει την πληροφορία σχετικά με το ποια χαρακτηριστικά καρέ απεικονίζουν τις επιλεγμένες προς εντοπισμό έννοιες. Έτσι για κάθε έννοια c_i το σύνολο των χαρακτηριστικών καρέ μπορεί να χωριστεί σε δύο υποσύνολα, με το πρώτο να αποτελείται από τις εικόνες που απεικονίζουν την έννοια c_i και το δεύτερο από εκείνες που δεν την απεικονίζουν. Τα δύο αυτά υποσύνολα θα συμβολίζονται με G_i και $\bar{G}_i$ αντίστοιχα και ορίζονται ως

$$G_i : \{k \in K : \quad c_i \in C(k)\} \tag{4.23}$$

και

$$\bar{G}_i : \{k \in K : \quad c_i \notin C(k)\} \ , \tag{4.24}$$

όπου $C(k)$ είναι το σύνολο των εννοιών εκείνων οι οποίες απεικονίζονται στο χαρακτηριστικό καρέ k. Από την έξοδο του ανιχνευτή, για κάθε έννοια c_i σχηματίζεται το σύνολο $D_i = \{k \in K : \quad b_i(k) = 1\}$, το οποίο και περιέχει τα χαρακτηριστικά καρέ που δόθηκαν σαν είσοδο, στα οποία οι ανιχνευτές ανίχνευσαν την έννοια c_i.

Για την αξιολόγηση της απόδοσης των προτεινόμενων τεχνικών, είναι αναγκαίο να επιλεχθούν τα κατάλληλα αντικειμενικά μέτρα. Ορισμένα από αυτά εφαρμόζονται στη δυαδική έξοδο D_i του ανιχνευτή, ενώ άλλα σε μία ταξινομημένη λίστα των ποσοστών πεποίθησης q_i. Η ταξινόμηση γίνεται με φθίνουσα σειρά και σε κάθε ένα από αυτά αντιστοιχεί μία δυαδική τιμή x_i, η οποία είναι ίση με 1 εάν για το χαρακτηριστικό καρέ k στην i-οστή θέση της ταξινομημένης λίστας για την έννοια c_j ισχύει $k \in G_j$, αλλιώς είναι ίση με 0. Στην περίπτωση που η τιμή του x_i είναι ίση με 1, τότε το χαρακτηριστικό

[7] `http://www.corel.com`

καρέ στην i-οστή θέση της ταξινομημένης λίστας θα καλείται "σχετικό" ή "θετικό" ως προς την απεικόνιση της έννοιας. Στην αντίθετη περίπτωση, το καρέ θα καλείται "αρνητικό".

Από τα πιο συνήθη μέτρα που εφαρμόζονται σε αντίστοιχα προβλήματα είναι το *μέτρο ακρίβειας* (precision) και το *μέτρο ανάκτησης* (recall), τα οποία θα αποκαλούνται εφεξής "ακρίβεια" και "ανάκτηση". Ορίζοντας σαν $|\cdot|$ το πλήθος των στοιχείων ενός συνόλου, η ακρίβεια και η ανάκτηση περιγράφονται αντίστοιχα ως

$$P_i = \frac{|D_i \bigcap G_i|}{|D_i|}, \quad i = 1 \dots N_C \tag{4.25}$$

και

$$R_i = \frac{|D_i \bigcap G_i|}{|G_i|}, \quad i = 1 \dots N_C \ . \tag{4.26}$$

Η ακρίβεια είναι το ποσοστό των εικόνων που πραγματικά απεικονίζουν την έννοια, σύμφωνα πάντα με το σύνολο δεδομένης αλήθειας, ως προς τις εικόνες που το σύστημα αποφάνθηκε ότι την απεικονίζουν. Η ανάκτηση είναι το ποσοστό των εικόνων που ανιχνεύθηκαν σωστά ότι απεικονίζουν μία έννοια ως προς όλες τις εικόνες που την απεικονίζουν. Επίσης, όπως και στο Κεφάλαιο 3, υπολογίζεται και το F-μέτρο, το οποίο ορίζεται ως

$$F_i = \frac{2P_iR_i}{P_i + R_i}, \quad i = 1 \dots N_C \tag{4.27}$$

και αποτελεί τον αρμονικό μέσο της ακρίβειας και της ανάκτησης.

Στη συνέχεια, στην ταξινομημένη λίστα των χαρακτηριστικών καρέ που αναφέρθηκε προηγουμένως, εφαρμόζεται το μέτρο της *μέσης ακρίβειας* (Average Precision). Το μέτρο αυτό ορίζεται στην (4.28) και δηλώνει την ακρίβεια που υπολογίζεται από τα m χαρακτηριστικά καρέ με τα μεγαλύτερα q, πάνω σε αυτά. p_m είναι η μέση τιμή των $x_1, x_2, \dots, x_m$, και συμβολίζεται με $\bar{x}_m$, ενώ υπολογίζεται ως

$$p_m = \frac{1}{m} \sum_{k=1}^{m} x_k \ . \tag{4.28}$$

Έπειτα, η μέση ακρίβεια ορίζεται σαν η μέση τιμή των ακριβειών μετά από κάθε σχετικό χαρακτηριστικό καρέ που συναντιέται στην λίστα. Η μέση ακρίβεια για την έννοια c_i με παράθυρο n και υπολογίζεται ως

$$AP_i^n = \frac{1}{|G_i|} \sum_{j=1}^{n} I(x_j) p_j = \frac{1}{|G_i|} \sum_{j=1}^{n} I(x_j) \bar{x_j} \ , \tag{4.29}$$

όπου n είναι το πλήθος των χαρακτηριστικών καρέ από τη λίστα με τα μεγαλύτερα q που επιλέγεται για να υπολογιστεί η μέση ακρίβεια, το παράθυρο δηλαδή πάνω στο οποίο υπολογίζεται. $I(x_i)$ είναι μία συνάρτηση για την οποία ισχύει $I(x_i) = x_i$, εφόσον $x_i \in \{0, 1\}$[8]. Άρα τελικά η μέση ακρίβεια ορίζεται ως

$$AP_i^n = \frac{1}{|G_i|} \sum_{j=1}^{n} x_j p_j = \frac{1}{|G_i|} \sum_{j=1}^{n} \frac{x_j}{j} \sum_{k=1}^{j} x_k \ . \tag{4.30}$$

[8]Το $I(x_i)$ ορίζεται με διαφορετικό τρόπο σε περιπτώσεις όπου το x_i είναι συνεχές.

Το μέτρο της μέσης ακρίβειας μοιάζει με το μέτρο που χρησιμοποιείται κατά τη διαδικασία αξιολόγησης των συμμετοχών που έχουν υποβληθεί στο TRECVID για τη δοκιμασία ανίχνευσης εννοιών υψηλού επιπέδου. Το τελευταίο αποκαλείται *συνηγμένη μέση ακρίβεια* (inferred average precision) και προτάθηκε από τους Yilmaz και Aslam [236]. Πρόκειται για μια εκτίμηση της μέσης ακρίβειας, η οποιά δεν απαιτεί την αξιολόγηση του συνόλου των αποτελεσμάτων, αλλά ενός μέρους τους. Χρησιμοποιείται στο TRECVID λόγω του μεγάλου αριθμού των πλάνων που αξιολογούνται, κάτι που σε συνδυασμό με τον μεγάλο αριθμό συμμετεχόντων καθιστά την πλήρη αξιολόγηση ανέφικτη.

4.7.2 Πειράματα στο πλαίσιο του TRECVID

Τα διαθέσιμα βίντεο από τη συλλογή του TRECVID κατά την αρχική φάση της εκπαίδευσης ήταν 110. Το κάθε ένα από αυτά ήταν χωρισμένο σε πλάνα. Από κάθε πλάνο επιλέχθηκε το μεσαίο καρέ ως χαρακτηριστικό και προέκυψε έτσι ένα σύνολο από 18113 χαρακτηριστικά καρέ. Ο σχολιασμός των καρέ που χρησιμοποιήθηκε οργανώθηκε από τους Ayache και Quenot [6] και δημιουργήθηκε συλλογικά από έναν μεγάλο αριθμό συμμετεχόντων στη δοκιμασία ανίχνευσης εννοιών υψηλού επιπέδου του TRECVID και έτσι προέκυψε το σύνολο δεδομένης αλήθειας. Ιδιαίτερη έμφαση πρέπει να δοθεί ότι ο σχολιασμός έγινε συνολικά σε επίπεδο της εικόνας και όχι σε επίπεδο περιοχών. Στον Πίνακα 4.15 φαίνεται ο αριθμός των χαρακτηριστικών καρέ που απεικονίζουν την κάθε έννοια.

Έννοια c_i	**θετικά χαρ. καρέ** $\|G_i\|$
Έρημος (Desert)	52
Δρόμος (Road)	923
Ουρανός (Sky)	2146
Χιόνι (Snow)	112
Βλάστηση (Vegetation)	1939
Γραφείο (Office)	1419
Εξωτερικός χώρος (Outdoor)	5185
Έκρηξη_ Φωτιά (Explosion_ Fire)	29
Βουνό (Mountain)	97

Πίνακας 4.2: *Αριθμός χαρακτηριστικών καρέ που απεικονίζουν την κάθε έννοια*

Η εξαγωγή περιοχών από τα χαρακτηριστικά καρέ, έγινε με ένα εργαλείο κατάτμησης με βάση το χρώμα των Avrithis et al. [5] ρυθμισμένο ώστε να δίνει έναν σχετικά μικρό αριθμό από περιοχές ανά χαρακτηριστικό καρέ. Έτσι προέκυψε ένα σύνολο από 345994 "χονδροειδείς" περιοχές. Ο αριθμός αυτός των περιοχών μπορεί να χαρακτηριστεί ως πολύ μεγάλος όσον αφορά την κατασκευή του θησαυρού. Τελικά, οι περιοχές που επιλέχθηκαν να χρησιμοποιηθούν για την κατασκευή του θησαυρού είναι ένα υποσύνολο των συνόλων εκπαίδευσης που πρόκειται να δημιουργηθούν για την εκπαίδευση των ανιχνευτών και περιγράφονται στη συνέχεια. Αρχικά συμπεριλήφθηκαν περιοχές από εικόνες που ανήκουν στο G_i σύνολο για κάθε έννοια C_i. Επιλέχθηκαν οι περιοχές από όλα τα χαρακτηριστικά καρέ για έννοιες με μικρό $|G_i|$, ενώ από μέρος των χαρακτηριστικών καρέ για μεγάλο $|G_i|$. Στη συνέχεια συμπεριλήφθηκε και ίσος αριθμός περιοχών με τον αριθμό των περιοχών από τα θετικά καρέ,

Συνολικός Αριθμός Βίντεο	N_V	110
Συνολικός Αριθμός Πλάνων	N_S	18113
Συνολικός Αριθμός Χαρ. Καρέ	N_K	18113
Συνολικός Αριθμός Περιοχών	N_R	345994
Συνολικός Αριθμός Διαν. Χαρακτηριστικών	N_F	345994
Μέγεθος Οπτικού Θησαυρού	N_T	100
Μέγεθος Διαν. Μετά Από LSA	N_L	70

Πίνακας 4.3: *Στοιχεία πειράματος TRECVID*

οι οποίες αντιστοιχούν σε εικόνες που επιλέχθηκαν τυχαία για να αναπαραστήσουν την ποικιλομορφία των περιοχών που συναντάται στις υπόλοιπες έννοιες.

Το μέγεθος του θησαυρού N_T επιλέχθηκε εμπειρικά ίσο με 100 τύπους περιοχής, λόγω του μεγάλου συνόλου εικόνων και ο αριθμός των ιδιοτιμών N_L της τεχνικής LSA επιλέχθηκε ίσος με 70, μειώνοντας έτσι αντίστοιχα την διάσταση των διανυσμάτων. Πρέπει να σημειωθεί ότι πειράματα με μεγαλύτερο μέγεθος θησαυρού δεν έδειξαν αξιοσημείωτη μεταβολή στην ακρίβεια που επιτυγχάνεται. Συνεπώς η διάσταση των διανυσμάτων αναπαράστασης είναι ίση με 100, καθώς 100 είναι και οι τύποι περιοχής, ενώ στα πειράματα που χρησιμοποιήθηκε η τεχνική LSA η διάσταση αυτή μειώθηκε σε 70 καθώς τα διανύσματα μεταφέρθηκαν στον χώρο των εννοιών που κατασκευάστηκε από τις 70 μεγαλύτερες ιδιοτιμές. Για να γίνει μία αξιολόγηση της τεχνικής με το δεδομένο σύνολο ανάπτυξης για το οποίο δημιουργήθηκε και σχολιασμός σε επίπεδο ολόκληρης της εικόνας, αυτό χωρίσθηκε σε δύο υποσύνολα για κάθε έννοια τα οποία και αποτελούν το σύνολο εκπαίδευσης και το σύνολο ελέγχου για κάθε ανιχνευτή. Έστω ότι τα δύο αυτά σύνολα συμβολίζονται με $\mathrm{T}R_i$ και TE_i για την έννοια c_i, αντίστοιχα. Ισχύει ότι

$$TR_i \cup TE_i \subseteq K \tag{4.31}$$

και

$$TR_i \cap TE_i = \emptyset \ . \tag{4.32}$$

Συνολικά διεξήχθησαν τρεις σειρές πειραμάτων για την κάθε έννοια χωρίς την χρήση της τεχνικής LSA:

Πειράματα με Σύνολο Εκπαίδευσης: Χρησιμοποιούνται σύνολα εκπαίδευσης με διαφορετικό αριθμό στοιχείων από το σύνολο $\bar{G}_i$ για κάθε έννοια c_i και τελικά επιλέγεται το καταλληλότερο.

Πειράματα με Τιμή Κατωφλίωσης: Για κάθε ανιχνευτή μεταβάλλεται η τιμή του κατωφλίου h_i και επιλέγεται εκείνο το οποίο δίνει βέλτιστες τιμές στα μέτρα ακρίβειας και ανάκτησης.

Πειράματα με Σύνολο Ελέγχου: Ο κάθε ανιχνευτής χρησιμοποιείται σε σύνολα εκπαίδευσης με συνεχώς αυξανόμενο πλήθος στοιχείων, προκειμένου να καταστεί εμφανής η μεταβολή της απόδοσής του.

Για την κατασκευή των συνόλων εκπαίδευσης και ελέγχου το 70% του G_i συμπεριλήφθηκε στο TR_i. Χρησιμοποιήθηκαν 5 διαφορετικά σύνολα εκπαίδευσης. Έστω λ ο λόγος του πλήθους των στοιχείων του TR_i που προέρχονται από το $\bar{G}_i$ προς

το πλήθος των στοιχείων από το από το G_i. Τότε, στα σύνολα εκπαίδευσης αντιστοιχούν οι τιμές $\lambda = 1$, $\lambda = 2$, $\lambda = 3$, $\lambda = 4$ και $\lambda = 5$. Αυτό σημαίνει ότι τα σύνολα αυτά περιέχουν 50%, 66.6%, 75%, 80% και 83.3% αντίστοιχα, των στοιχείων του συνόλου $\bar{G}_i$, δηλαδή αρνητικών χαρακτηριστικών καρέ στην απεικόνιση των εννοιών. Τα σύνολα ελέγχου TE_i που χρησιμοποιήθηκαν αποτελούνται από το 30% του G_i και από ισάριθμα στοιχεία του $\bar{G}_i$, δηλαδή ισχύει $\lambda = 1$. Στην (4.33) φαίνεται ο τρόπος που υπολογίζεται το λ για το σύνολο εκπαίδευσης της έννοιας c_i. Με τον ίδιο τρόπο υπολογίζεται και για το σύνολο ελέγχου ως

$$\lambda_i = \frac{|\{k \in K : \quad k \in TR_i \cap \bar{G}_i\}|}{|\{k \in K : \quad k \in TR_i \cap G_i\}|} \ . \tag{4.33}$$

Το μέτρο που χρησιμοποιήθηκε είναι η μέση ακρίβεια της (4.29), όπου το n ήταν ίσο με ολόκληρο το σύνολο ελέγχου για κάθε έννοια. Έτσι έγινε η επιλογή για κάθε ανιχνευτή του αποδοτικότερου συνόλου εκπαίδευσης. Επιλέχθηκε να κατασκευαστούν αυτά τα 5 σύνολα ελέγχου καθώς για μερικές έννοιες όλο το σύνολο ανάπτυξης του TRECVID δεν θα επέτρεπε περισσότερα χαρακτηριστικά καρέ να χρησιμοποιηθούν στο σύνολο εκπαίδευσης διότι πρέπει να χρησιμοποιηθούν και αρκετά χαρακτηριστικά καρέ στο σύνολο ελέγχου προκειμένου να δημιουργηθεί σύνολο της τάξεως εκείνων που δίνει και το TRECVID για έλεγχο. Τα αποτελέσματα δίνονται στον Πίνακα 4.4, όπου πρέπει να σημειωθεί ότι για την έννοια *εξωτερικός χώρος* χρησιμοποιήθηκαν μόνο 3 σύνολα εκπαίδευσης, με λ=1, λ=2 και λ=3 γιατί τα θετικά χαρακτηριστικά καρέ ήταν πάρα πολλά και δεν κατέστη δυνατόν να δημιουργηθούν τα αντίστοιχα σύνολα εικόνων. Μέσα από αυτή την διαδικασία επιλέγονται τα καταλληλότερα σύνολα

Έννοια c_i	**Μέση Ακρίβεια** AP_i				
	λ=1	**λ=2**	**λ=3**	**λ=4**	**λ=5**
Έρημος	0.66	**0.70**	0.37	0.48	0.66
Δρόμος	0.60	0.61	0.60	0.61	**0.70**
Ουρανός	0.68	0.72	0.69	0.72	**0.74**
Χιόνι	0.91	0.91	0.93	0.92	**0.95**
Βλάστηση	0.72	0.77	0.77	0.75	**0.78**
Γραφείο	0.63	0.71	**0.74**	0.71	0.72
Εξωτερικός χώρος	0.68	0.68	**0.70**	-	-
Έκρηξη_ Φωτιά	0.39	0.37	0.35	**0.65**	0.38
Βουνό	0.69	0.61	0.55	0.63	**0.77**

Πίνακας 4.4: *Μέση ακρίβεια για σύνολα εκπαίδευσης με διαφορετικό λόγο λ, με τονισμένη γραμματοσειρά είναι οι υψηλότερες τιμές*

εκπαίδευσης για κάθε ανιχνευτή. Συνοπτικά τα χαρακτηριστικά καρέ που αποτελούν το κάθε σύνολο εκπαίδευσης και ελέγχου φαίνονται στον Πίνακα 4.5.

Θεωρητικά ένα νευρωνικό δίκτυο που χρησιμοποιείται για δυαδική ταξινόμηση θα πρέπει να εκπαιδευτεί με παρόμοιο αριθμό από πρότυπα από την κάθε κατηγορία, όσο αυτό είναι δυνατό, ώστε να "μαθαίνει" εξίσου και τις δύο. Στην προκειμένη όμως περίπτωση το διαθέσιμο σύνολο δεδομένων εμφανίζει μεγάλη ποικιλομορφία. Έτσι για κάθε έννοια υπάρχει η κατηγορία εκείνη με τα πρότυπα (εικόνες) που απεικονίζουν την έννοια και η κατηγορία με εκείνες που δεν την απεικονίζουν. Η δεύτερη όμως

Έννοια c_i	$\mathbf{TR_i}$		$\mathbf{TE_i}$	
	G_i	$\bar{G}_i$	G_i	$\bar{G}_i$
Έρημος	36	72	16	16
Δρόμος	646	3230	277	277
Ουρανός	1502	7510	644	644
Χιόνι	78	390	34	34
Βλάστηση	1357	6785	582	582
Γραφείο	993	2979	426	426
Εξωτερικός χώρος	3629	10887	1556	1556
Έκρηξη_ Φωτιά	20	80	9	9
Βουνό	68	340	29	29

Πίνακας 4.5: *Αριθμός χαρακτηριστικών καρέ που απεικονίζουν ή όχι την κάθε έννοια για τα σύνολα εκπαίδευσης και ελέγχου.*

κατηγορία περιέχει εικόνες με ποικίλα θέματα άρα και με ποικίλα οπτικά χαρακτηριστικά. Συνεπώς απαιτούνται περισσότερα πρότυπα από την δεύτερη κατηγορία για να "μάθει" και εκείνη ο ανιχνευτής. Αυτό αποδεικνύεται και στην πράξη με τα παραπάνω πειράματα όπου τα σύνολα εκπαίδευσης που αποδίδουν καλύτερα είναι εκείνα με περισσότερα αρνητικά χαρακτηριστικά καρέ από ότι θετικά.

Μία παρατήρηση που αξίζει να γίνει για το μέτρο της μέσης ακρίβειας που χρησιμοποιήθηκε είναι η εξής: Κοιτώντας το διάγραμμα του Σχήματος 4.13, όπου απεικονίζεται η μέση ακρίβεια ως προς το παράθυρο πάνω στο οποίο υπολογίζεται, δηλαδή το n φαίνεται ότι καθώς αυξάνεται το n. Φαίνεται ότι η μέση ακρίβεια επίσης θα αυξάνεται, διότι θα συναντώνται όλο και περισσότερες σχετικές εικόνες. Μεγάλη κλίση καμπύλης υποδηλώνει πως εντοπίζονται πολλές σχετικές εικόνες σε συνεχόμενες θέσεις, ενώ πιο μικρή ότι μεταξύ των σχετικών παρεμβάλλονται και πολλές μη σχετικές.

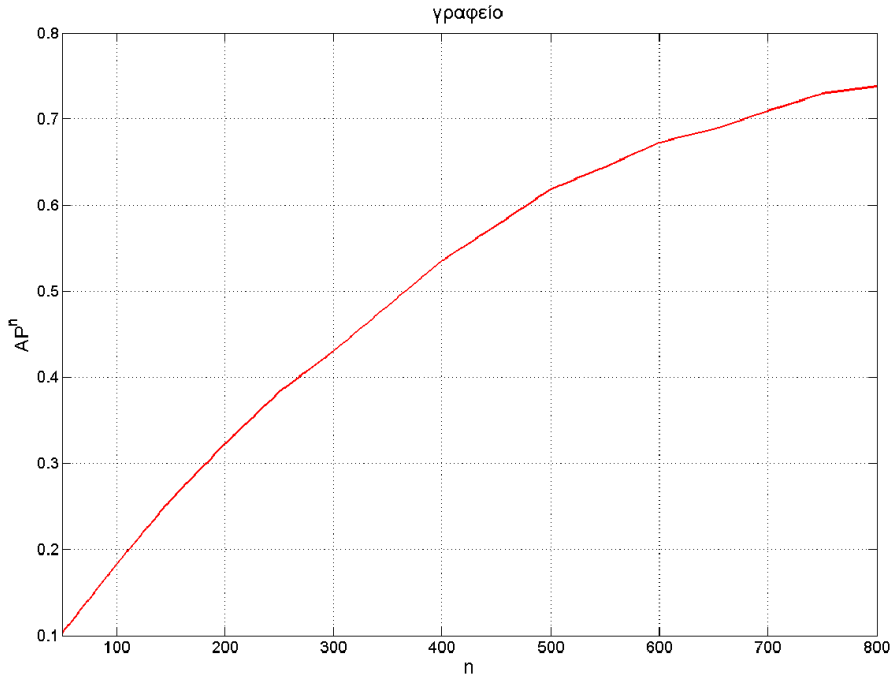
Χρησιμοποιούνται τα σύνολα εκπαίδευσης που επιλέχθηκαν από την πρώτη σειρά πειραμάτων και σαν σύνολα ελέγχου για κάθε έννοια χρησιμοποιούνται όλα τα υπόλοιπα χαρακτηριστικά καρέ του K. Δηλαδή ισχύει ότι

$$TR_i \cup TE_i = K \tag{4.34}$$

και

$$TR_i \cap TE_i = \emptyset \ . \tag{4.35}$$

Οι ανιχνευτές εκπαιδεύονται με τα επιλεγμένα σύνολα εκπαίδευσης και σαν είσοδος μετά δίνεται το σύνολο ελέγχου. Πάνω στα ποσοστά πεποίθησης τα οποία για κάθε χαρακτηριστικό καρέ εξάγονται γίνεται μετατροπή με χρήση κατωφλίου σε δυαδικό {0,1}. Πάνω στην δυαδική έξοδο εφαρμόζονται τα μέτρα ακρίβειας-ανάκτησης. Το πείραμα γίνεται για τιμές κατωφλίου από 0 έως 0.9 με βήμα ίσο με 0.1. Έτσι δημιουργούνται διαγράμματα σαν αυτά του Σχήματος 4.14 και τελικά η τιμή κατωφλίου που επιλέγεται είναι εκείνη η οποία δίνει τιμές πιο κοντά στο σημείο της καμπύλης όπου ακρίβεια=ανάκληση. Είναι επιθυμητό και τα δύο μέτρα αυτά να είναι όσο το δυνατό καλύτερα με την ίδια σημασία στο κάθε ένα και για αυτό το κατώφλι επιλέγεται με τον συγκεκριμένο τρόπο. Επίσης στον Πίνακα 4.6 παρουσιάζονται οι τιμές για τα δύο αυτά ρα όσο το κατώφλι αυξάνεται για ον ανιχνευτή της έννοια *ουρανός*. Τελικά η τιμή κατωφλίου που επιλέγεται για τον ανιχνευτή της έννοιας *βλάστηση* είναι ίση



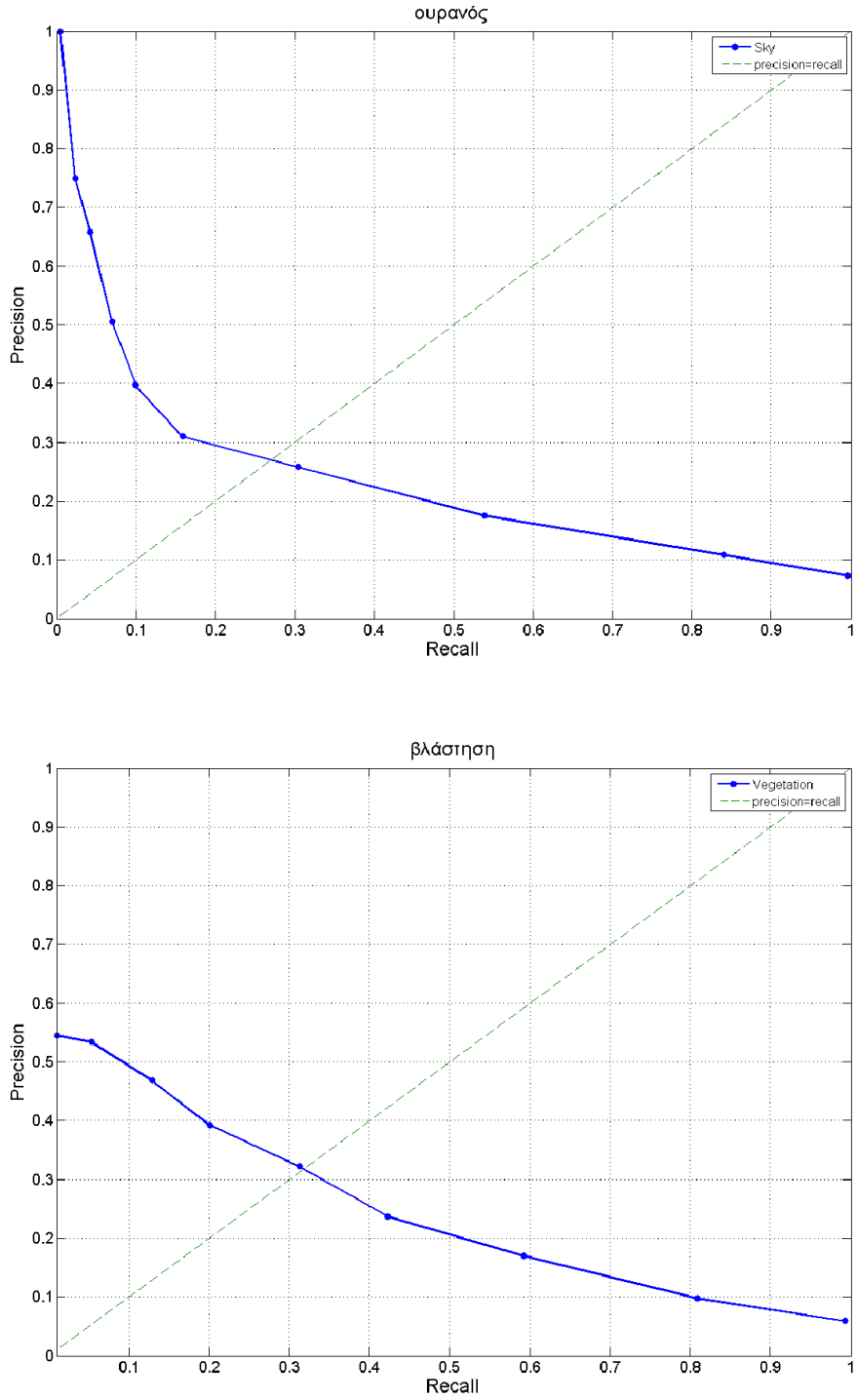
Σχήμα 4.13: *Διάγραμμα μέσης ακρίβειας καθώς μεταβάλλεται το παράθυρο υπολογισμού της για την έννοια γραφείο.*

με 0.4, το οποίο φαίνεται και από το σχήμα αλλά και από τον πίνακα των τιμών. Για την έννοια *ουρανός* επιλέγεται κατώφλι ίσο με 0.3 ενώ τα κατώφλια που τελικά επιλέχθηκαν για όλες τις έννοιες φαίνονται στον Πίνακα 4.7.

Κατώφλι	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
Ακρίβεια	0.05	0.09	0.17	0.23	0.32	0.39	0.46	0.53	0.54	0.33
Ανάκτηση	0.99	0.80	0.59	0.42	0.31	0.20	0.12	0.05	0.01	0.01

Πίνακας 4.6: *Ζεύγη ακρίβειας-ανάκτησης μεταβάλλοντας το κατώφλι, για την έννοια βλάστηση.*

Όσο πιο μεγάλη (δηλαδή όσο πιο κοντά στο 1) τιμή έχει ένα κατώφλι, τόσο πιο μεγάλο είναι το μέτρο ακρίβειας και πιο μικρό το μέτρο ανάκτησης. Αυτό συμβαίνει διότι επιλέγονται σαν θετικά χαρακτηριστικά καρέ ως προς την απεικόνιση της κάθε έννοιας εκείνα τα οποία έχουν υψηλά ποσοστά πεποίθησης κι έτσι υπάρχει μεγαλύτερη σιγουριά ότι εκείνα θα απεικονίζουν όντως την έννοια, για αυτό προκύπτει η υψηλή τιμή του μέτρου ακρίβειας. Με αυτόν όμως τον τρόπο υπάρχει περισσότερη αυστηρότητα ως προς το ποια παραδείγματα θεωρούνται θετικά και συνεπώς ένας μικρός μόνο αριθμός εικόνων επιλέγεται. Αυτό οδηγεί σε χαμηλό μέτρο ανάκτησης εφόσον μέσα σε αυτόν τον μικρό αριθμό ανακτημένων εικόνων θα υπάρχει και μικρός αριθμός σχετικών εικόνων ως προς όλο το σύνολο των σχετικών εικόνων. Για μικρές (κοντά στο 0) τιμές κατωφλίου συμβαίνει το αντίθετο, δηλαδή χαμηλό μέτρο ακρίβειας και υψηλό μέτρο ανάκτησης. Συνεπώς, η αποτελεσματικότητα του κάθε ανιχνευτή εξαρτάται άμεσα και σε μεγάλο βαθμό από την τιμή αυτή του κατωφλίου. Ο τρόπος επιλογής τιμής κατωφλίου που δίνει τιμές κοντά στο σημείο τομής της καμπύλης *ανάκτησης-ακρίβειας* με την καμπύλη *ανάκτηση=ακρίβεια* εξασφαλίζει την



Σχήμα 4.14: *Διαγράμματα ακρίβειας-ανάκτησης καθώς η τιμή κατωφλίου μεταβάλλεται, για τις έννοιες ουρανός και βλάστηση.*

καλύτερη απόδοση "συνυπολογίζοντας" κατά κάποιο τρόπο από κοινού και εξίσου τα δύο μέτρα.

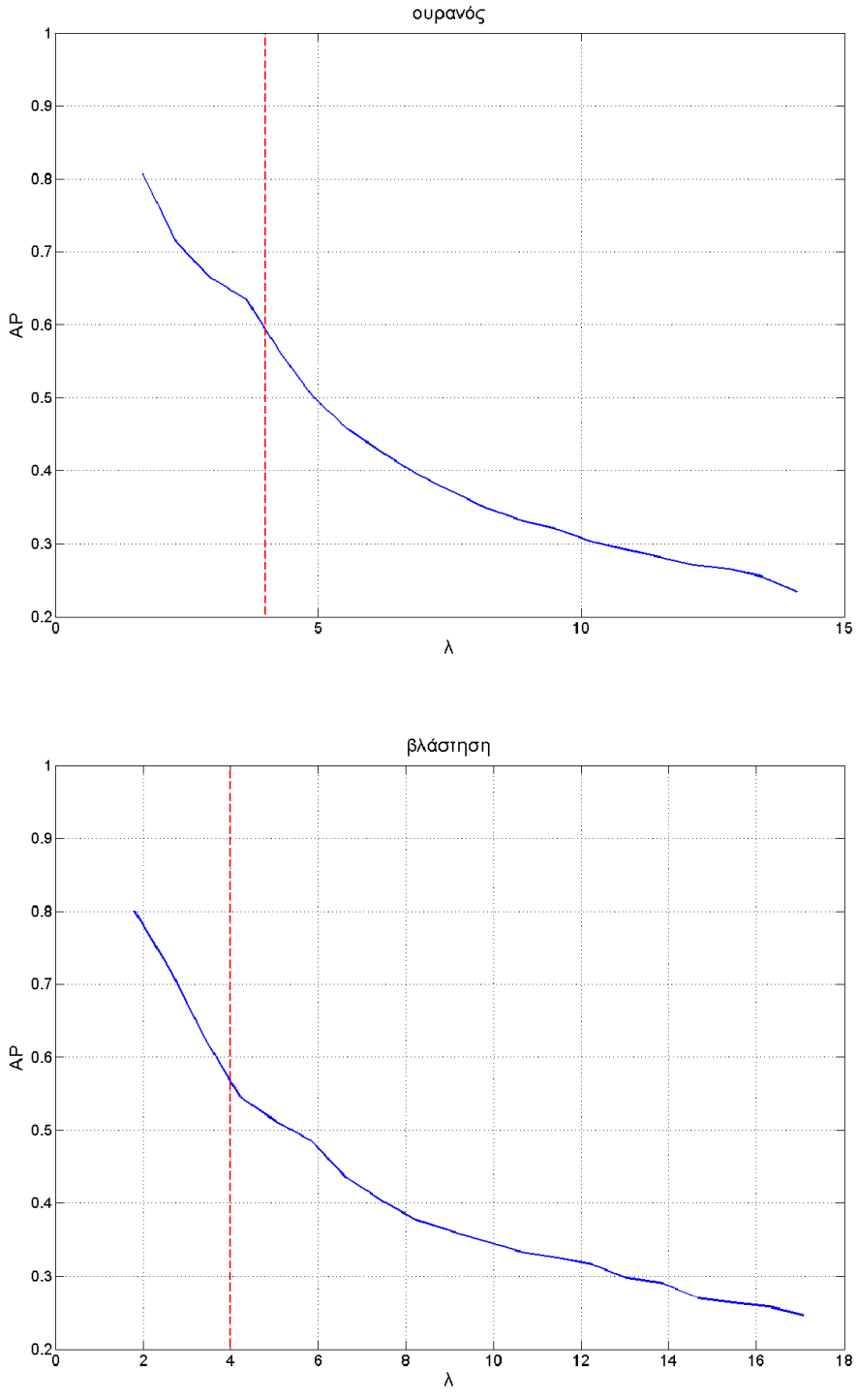
Έρημος	*Δρόμος*	*Ουρανός*	*Χιόνι*	*Βλάστηση*	*Γραφείο*	*Εξ.Χώρος*	*Έκ.-Φωτιά*	*Βουνό*
0.8	0.5	0.3	0.6	0.4	0.5	0.3	0.2	0.8

Πίνακας 4.7: *Κατώφλια για όλους τους ανιχνευτές, χωρίς LSA.*

Τα σύνολα ελέγχου τώρα αποτελούνται όπως και νωρίτερα από το 30% του G_i και συμπληρώνονται κάθε φορά με ένα μέρος των αρνητικών χαρακτηριστικών καρέ ($\bar{G}_i$) που δεν ανήκουν στο TR_i. Αυτό γίνεται με ένα βήμα μέχρι να δημιουργηθεί το σύνολο ελέγχου που θα περιέχει και όλα τα k για τα οποία ισχύει ότι: $\{k \in \bar{G}_i, \quad k \notin TR_i\}$. Κάθε φορά δηλαδή αυξάνεται ο λόγος λ για το σύνολο ελέγχου. Όλα τα ενδιάμεσα σύνολα ελέγχου δίνονται σαν είσοδος στους ανιχνευτές και στην έξοδο τους υπολογίζεται κάθε φορά η μέση ακρίβεια με παράθυρο όλο το σύνολο. Κατασκευάζονται έτσι διαγράμματα μέσης ακρίβειας ως προς το λόγο λ του συνόλου ελέγχου. Στα διαγράμματα αυτά φαίνεται πως η μέση ακρίβεια μειώνεται γενικά καθώς μεγαλώνει το σύνολο ελέγχου, αλλά πιο συγκεκριμένα καθώς αυξάνεται ο λόγος των αρνητικών χαρακτηριστικών καρέ προς τα θετικά.

Με τη χρήση ενός ανιχνευτή εννοιών πάνω σε ένα σύνολο χαρακτηριστικών καρέ αντιστοιχίζεται κάθε καρέ με μία δυαδική τιμή που αντιστοιχεί σε απεικόνιση ή μη της έννοιας. Από εκείνα τα καρέ τα οποία ανιχνεύθηκαν ότι εντοπίζουν την έννοια, δηλαδή αυτά για τα οποία ο ανιχνευτής έδωσε έξοδο με τιμή 1, είναι αναμενόμενο ότι μερικά από αυτά στην πραγματικότητα δεν θα απεικονίζουν τη συγκεκριμένη έννοια. Αυτό τελικά οδηγεί σε χαμηλότερη μέση ακρίβεια, καθώς οι λανθασμένοι εντοπισμοί γίνονται με ποσοστό πεποίθησης μεγαλύτερο από καρέ που πραγματικά απεικονίζουν την έννοια. Έτσι, τα θετικά αυτά καρέ εντοπίζονται σε χαμηλότερες θέσεις στην ταξινομημένη λίστα με βάση το ποσοστό πεποίθησης και συνεισφέρουν λιγότερο στο άθροισμα όπως φαίνεται και στην (4.30). Όταν τα αρνητικά χαρακτηριστικά καρέ είναι περισσότερα αυτό που συμβαίνει είναι να είναι και περισσότεροι αυτοί οι λανθασμένοι εντοπισμοί μεταξύ των πραγματικών εντοπισμών και η μέση ακρίβεια να πέφτει, γεγονός το οποίο φαίνεται στα διαγράμματα του Σχήματος 4.15.

Τελικά ο ανιχνευτής για κάθε έννοια εκπαιδεύτηκε με το σύνολο εκπαίδευσης το οποίο επιλέχθηκε στην πρώτη σειρά πειραμάτων. Χρησιμοποιήθηκε επίσης για τον κάθε ένα το κατώφλι που επιλέχθηκε στην δεύτερη σειρά πειραμάτων προκειμένου να υπολογιστούν τα μέτρα ακρίβειας-ανάκτησης πάνω στη δυαδική έξοδο. Πάνω σε σύνολο ελέγχου το οποίο αποτελείται από όλα τα χαρακτηριστικά καρέ τα οποία δεν ανήκουν στο σύνολο εκπαίδευσης υπολογίστηκε και το μέτρα μέσης ακρίβειας. Το μέτρο αυτό υπολογίστηκε πάνω σε παράθυρο ίσο με το μισό του μεγέθους του συνόλου ελέγχου, διότι σε περίπτωση που υπολογιζόταν σε σταθερό παράθυρο αυτό θα έφερνε σε μειονεκτική θέση ανιχνευτές για έννοιες με λίγα θετικά καρέ. Όλα τα αποτελέσματα συνοψίζονται στον Πίνακα 4.8. Η χρήση ενός τέτοιου συνόλου είναι λογικό να έχει χαμηλά αποτελέσματα καθώς στα περισσότερα σύνολα έχουμε ένα μεγάλο αριθμό αρνητικών χαρακτηριστικών καρέ με λίγα θετικά. Σε μερικές περιπτώσεις (όπως για τις έννοιες *έκρηξη_φωτία*, *έρημος*, *χιόνι* και *βουνό*) αυτή η διαφορά είναι πολύ μεγάλη, εφόσον τα συνολικά θετικά χαρακτηριστικά καρέ ήταν πολύ λίγα. Για αυτό τον λόγο χρησιμοποιήθηκε κι ένα διαφορετικό σύνολο ελέγχου το οποίο αποτελείται από τον ίδιο αριθμό θετικών χαρακτηριστικών καρέ με το προηγούμενο σύνολο ελέγχου (τα οποία είναι και τα θετικά που φαίνονται στον Πίνακα 4.15) και από αρ-



Σχήμα 4.15: *Διαγράμματα μέσης ακρίβειας καθώς μεγαλώνει το σύνολο ελέγχου και μεγαλώνει το λ, μέχρι την τιμή λ=4 που θεωρείται λογική για σύνολο ελέγχου ενός συστήματος και για τις έννοιες ουρανός και βλάστηση.*

νητικά χαρακτηριστικά καρέ τετραπλάσια σε αριθμό από τα θετικά. Χρησιμοποιήθηκε δηλαδή τελικά σύνολο ελέγχου με $\lambda = 4$. Έτσι ο Πίνακας 4.13 για το συγκεκριμένο σύνολο είναι και πιο κατάλληλος για να αξιολογηθεί η απόδοση του κάθε ανιχνευτή. Για να κατασκευαστεί ένα τέτοιο σύνολο ελέγχου για την έννοια *εξωτερικός χώρος* της οποίας τα θετικά καρέ ήταν πολλά και το τετραπλάσιο τους ξεπερνά το σύνολο όλων των καρέ, χρησιμοποιήθηκαν λιγότερα θετικά καρέ από τα συνολικά και τελικά τετραπλάσιος αριθμός αρνητικών κι έτσι δημιουργείται το σύνολο με λ=4. Ενώ για όλες τις έννοιες τα αποτελέσματα του Πίνακα 4.8 είναι με σύνολα ελέγχου όπου το λ είναι πολύ μεγάλο, μόνο στην συγκεκριμένη έννοια *εξωτερικός χώρος* το λ είναι και μικρότερο του 4, λόγω της έλλειψης αρνητικών χαρακτηριστικών καρέ. Έτσι φαίνεται τελικά στον Πίνακα 4.13 τα ποσοστά να πέφτουν αντί να ανεβαίνουν, διότι τα θετικά εδώ είναι λιγότερα, κάτι που συμβαίνει μόνο στην περίπτωση αυτής της έννοιας.

Έννοια c_i	**Μέτρα Αξιολόγησης**			
	P	**R**	**AP**	**F**
Βλάστηση	0.32	0.31	0.23	0.32
Δρόμος	0.05	0.05	0.04	0.05
Έκρηξη_ Φωτιά	0.00	0.00	0.00	0.00
Ουρανός	0.26	0.30	0.21	0.28
Χιόνι	0.01	0.41	0.01	0.02
Γραφείο	0.12	0.16	0.07	0.14
Έρημος	0.00	0.31	0.06	0.00
Εξωτερικός χώρος	0.68	0.51	0.52	0.58
Βουνό	0.00	0.38	0.04	0.00

Πίνακας 4.8: *Τελικά αποτελέσματα πειραμάτων χωρίς την χρήση LSA, με σύνολο ελέγχου όλες τις εικόνες.*

Έννοια c_i	**Μέτρα Αξιολόγησης**			
	P	**R**	**AP**	**F**
Βλάστηση	0.64	0.31	0.46	0.42
Δρόμος	0.30	0.05	0.28	0.09
Έκρηξη_ Φωτιά	0.29	0.78	0.18	0.42
Ουρανός	0.57	0.30	0.44	0.39
Χιόνι	0.78	0.41	0.46	0.54
Γραφείο	0.45	0.16	0.32	0.24
Έρημος	0.33	0.31	0.29	0.32
Εξωτερικός χώρος	0.43	0.51	0.36	0.47
Βουνό	0.44	0.14	0.24	0.21

Πίνακας 4.9: *Τελικά αποτελέσματα πειραμάτων χωρίς την χρήση LSA, με σύνολο ελέγχου αποτελούμενο από 20% θετικά και 80% αρνητικά χαρακτηριστικά καρέ (λ=4).*

4.7.3 Πειράματα στο πλαίσιο του TRECVID με τη χρήση της τεχνικής LSA

Τα παραπάνω πειράματα διεξήχθησαν με σύνολα όπου κάθε χαρακτηριστικό καρέ αναπαρίσταται από το διάνυσμα αναπαράστασης το οποίο και έχει διάσταση όση και το μέγεθος του θησαυρού, δηλαδή 100 στην προκειμένη περίπτωση, όσοι είναι και οι τύποι περιοχής. Στα σύνολα εκπαίδευσης που επιλέχθηκαν εφαρμόζεται η τεχνική LSA και τα διανύσματα αναπαράστασης μεταφέρονται έτσι σε έναν άλλο χώρο, τον χώρο των "εννοιών" και η διάσταση τους μειώνεται από 100 σε 70 στοιχεία. Η ίδια τεχνική εφαρμόστηκε και στο σύνολο ελέγχου το οποίο αποτελείται από όλα τα υπόλοιπα χαρακτηριστικά καρέ. Έτσι στην έξοδο του κάθε ανιχνευτή που αφορά το αντίστοιχο σύνολο εκπαίδευσης εφαρμόστηκαν τα διάφορα μέτρα για να γίνει η σύγκριση με εκείνα στα οποία δεν εφαρμόστηκε η τεχνική LSA.

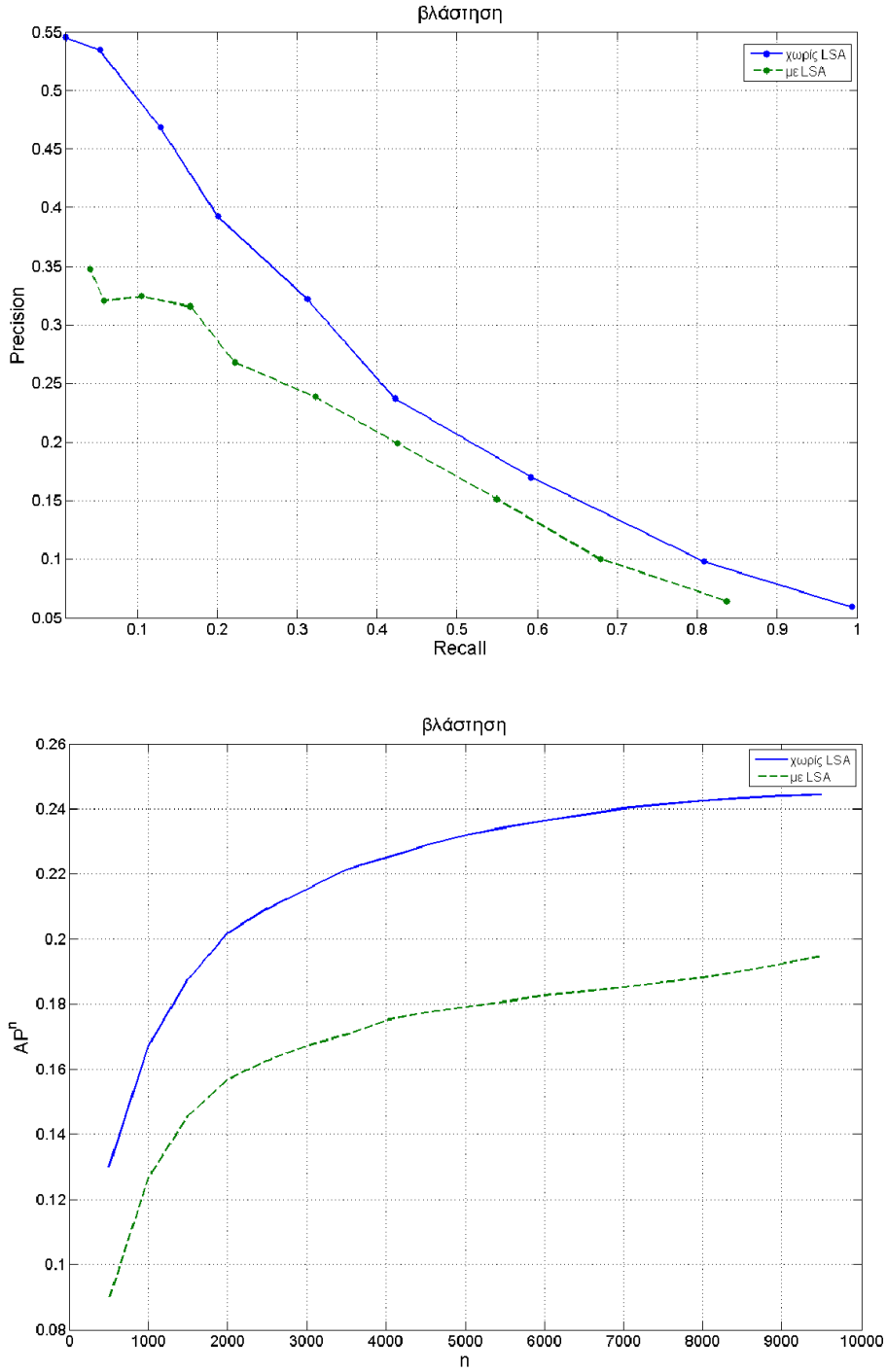
Στο Σχήμα 4.16 (πάνω) φαίνονται οι καμπύλες ακρίβειας-ανάκτησης για την έννοια *βλάστηση* καθώς το κατώφλι μεταβάλλεται. Οι δύο καμπύλες αντιστοιχούν σε αποτελέσματα χωρίς αλλά και με την χρήση LSA. Η καμπύλη που αντιστοιχεί στα αποτελέσματα χωρίς την χρήση της τεχνικής βρίσκεται συνεχώς πάνω από την άλλη γεγονός που αποδεικνύει ότι για όλες τις τιμές του κατωφλίου αποδίδει καλύτερα χωρίς LSA ο ανιχνευτής. Το συμπέρασμα αυτό επιβεβαιώνει και το διάγραμμα του Σχήματος 4.16 (κάτω) για την μέση ακρίβεια καθώς το παράθυρο υπολογισμού μεγαλώνει. Η καμπύλη που αντιστοιχεί στην μη χρήση LSA είναι και πάλι από πάνω. Τα διαγράμματα αυτά συνεπώς δείχνουν ότι η τεχνική LSA δεν δίνει καλύτερα αποτελέσματα για τον ανιχνευτή της συγκεκριμένης όμως έννοιας. Αντίθετα, για τον ανιχνευτή της έννοιας *εξωτερικός χώρος* αυτό που φαίνεται στο διάγραμμα της μέσης ακρίβειας (Σχήμα 4.17) είναι πως οι σωστά εντοπισμένες εικόνες βρίσκονται στις πρώτες θέσεις της διατεταγμένης λίστας και για αυτό η μέση ακρίβεια είναι μεγαλύτερη για τα μικρά παράθυρα με την χρήση LSA. Παρόμοια και στην καμπύλη ακρίβειας-ανάκτησης φαίνεται κάτι παρόμοιο, ότι δηλαδή σωστά εντοπισμένες εικόνες δίνονται με μεγάλα ποσοστά πεποίθησης. Το συμπέρασμα αυτό πως η τεχνική LSA βελτιώνει για μερικούς από τους ανιχνευτές τις θέσεις στις οποίες οι σωστοί εντοπισμοί συμβαίνουν επιβεβαιώνεται από τα διαγράμματα μέσης ακρίβειας του Σχήματος 4.18.

Η τεχνική LSA αποδεικνύεται χρήσιμη για μερικούς από τους ανιχνευτές όσον αφορά το γεγονός ότι ανάμεσα στις πρώτες εικόνες με το μεγαλύτερο βαθμό βεβαιότητας βρίσκονται περισσότεροι σωστοί εντοπισμοί από ότι χωρίς την χρήση LSA. Αυτό συμβαίνει καθώς η τεχνική αυτή λαμβάνει υπόψη τις λανθάνουσες σχέσεις μεταξύ των περιοχών και αυτές οι συσχετίσεις μπορεί να είναι πιο μεγάλες για μερικές έννοιες από κάποιες άλλες. Έτσι για τις έννοιες με υψηλές συσχετίσεις μεταξύ των περιοχών η τεχνική αποδίδει καλύτερα. Αυτό που συμβαίνει με την τεχνική αυτή είναι ενώ πριν υπήρχε η εμφάνιση του κάθε τύπου περιοχής να περιγράφει την εικόνα, τώρα αντί των τύπων περιοχής εντοπίζονται συσχετίσεις μεταξύ των περιοχών και χρησιμοποιούνται για την περιγραφή των εικόνων.

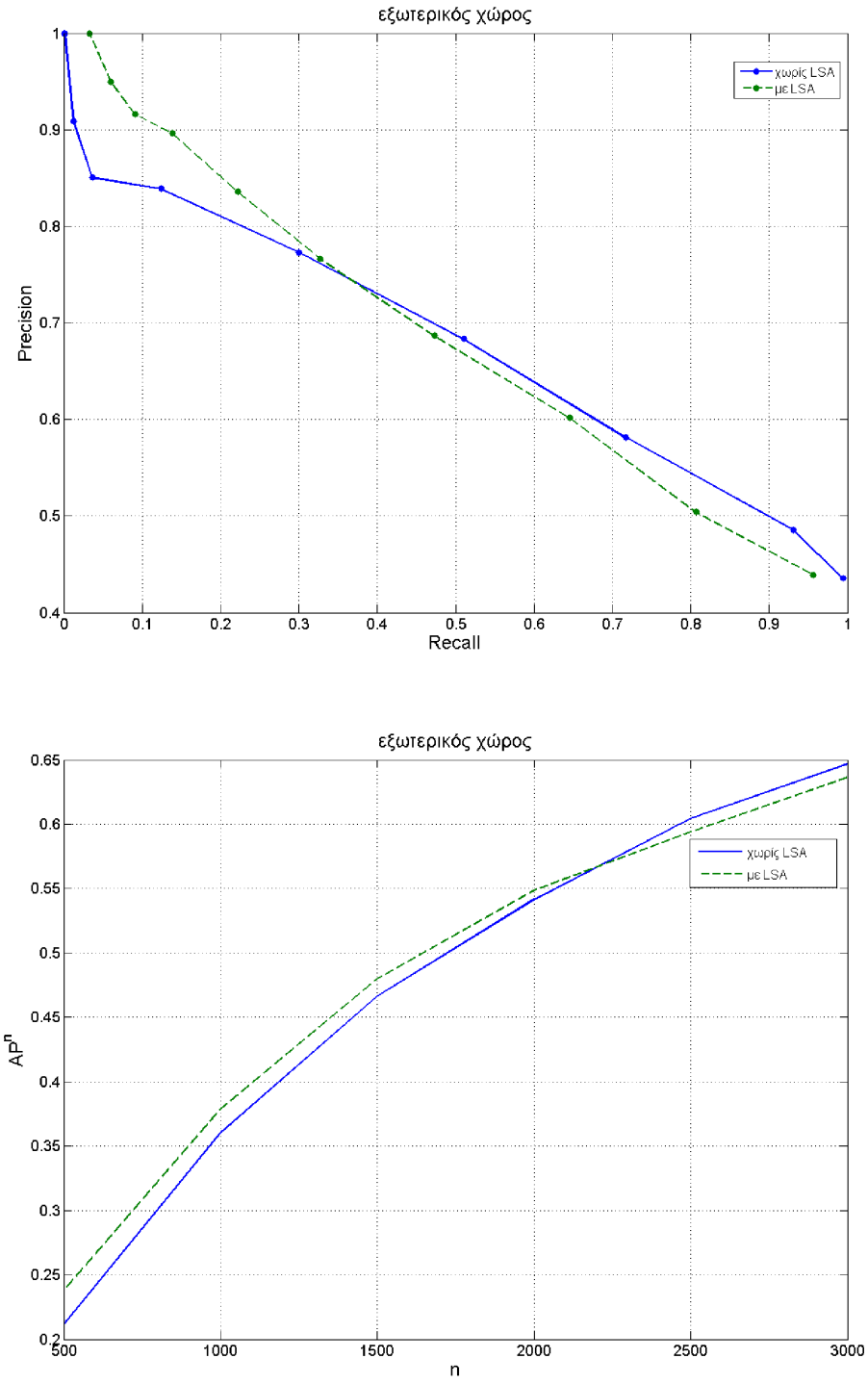
Οι τιμές των κατωφλίων επιλέγονται και πάλι για κάθε ανιχνευτή με τον ίδιο τρόπο και είναι εκείνες που φαίνονται στον Πίνακα 4.10.

Έρημος	*Δρόμος*	*Ουρανός*	*Χιόνι*	*Βλάστηση*	*Γραφείο*	*Εξ.Χώρος*	*Εκ.-Φωτιά*	*Βουνό*
0.8	0.8	0.4	0.8	0.5	0.8	0.2	0.8	0.8

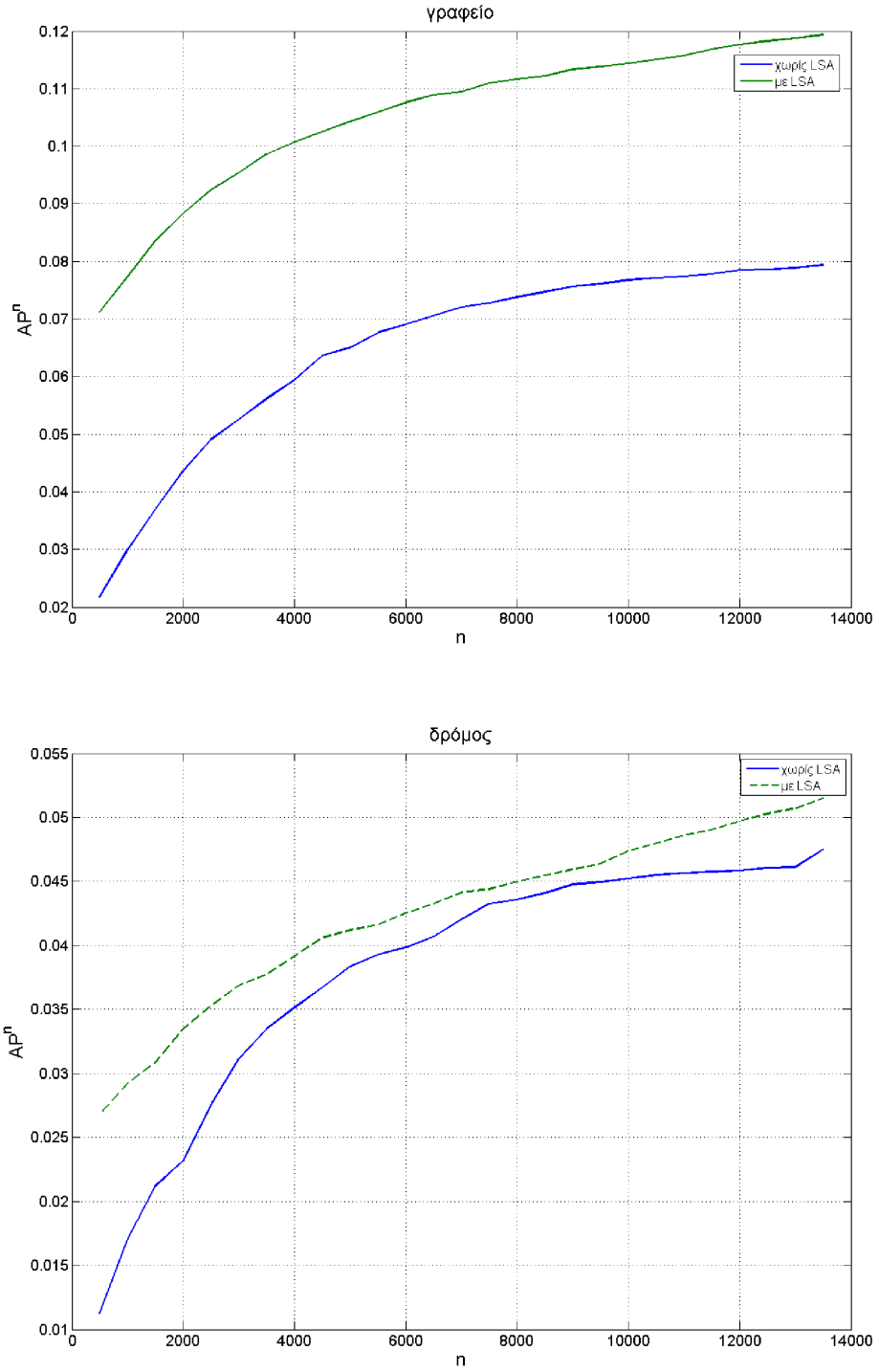
Πίνακας 4.10: *Κατώφλια για όλους τους ανιχνευτές, με LSA*



Σχήμα 4.16: *Διαγράμματα ακρίβειας-ανάκτησης και μέσης ακρίβειας, καθώς αυξάνεται το μέγεθος παραθύρου, χωρίς και με LSA, για την έννοια βλάστηση.*



Σχήμα 4.17: *Διαγράμματα ακρίβειας-ανάκτησης και μέσης ακρίβειας, καθώς αυξάνεται το μέγεθος παραθύρου, χωρίς και με LSA, για την έννοια εξωτερικός χώρος.*



Σχήμα 4.18: *Διαγράμματα μέσης ακρίβειας, καθώς αυξάνεται το μέγεθος παραθύρου, χωρίς και με LSA, για τις έννοιες δρόμος και γραφείο.*

Τέλος, στους Πίνακες 4.11 και 4.12 παρουσιάζονται όλα τα μέτρα τα οποία χρησιμοποιήθηκαν και για τους ανιχνευτές χωρίς LSA, για σύνολο ελέγχου με όλα τα αρνητικά χαρακτηριστικά καρέ και για σύνολο ελέγχου με τετραπλάσια αρνητικά από τα θετικά αντίστοιχα. Μία τελευταία σύγκριση και σε αυτά οδηγεί στο τελικό συμπέρασμα πως η τεχνική LSA που χρησιμοποιήθηκε και αξιολογήθηκε οδηγεί σε καλύτερα αποτελέσματα εντοπισμού από εκείνα των ανιχνευτών χωρίς την χρήση της τεχνικής για μερικές από τις έννοιες. Παρατηρούνται καλύτερα αποτελέσματα σε έννοιες των οποίων οι εικόνες έχουν καλύτερα "συσχετισμένες" περιοχές. Για παράδειγμα η έννοια *χιόνι* αποτελείται από εικόνες, από τις οποίες οι περισσότερες καταλαμβάνονται από κατάλευκες περιοχές, έτσι παρατηρείται αισθητή βελτίωση με την τεχνική LSA. Σε αντίθετη περίπτωση έννοιες όπως *βουνό* όπου στο σύνολο των εικόνων παρατηρείται μεγαλύτερη ανομοιομορφία, καθώς δεν περιλαμβάνονται μόνο και χαρακτηριστικά τοπία βουνών, αλλά αντιθέτως και αρκετές φωτογραφίες οι οποίες απλώς έχουν ληφθεί σε ένα βουνό χωρίς όμως αυτό να γίνεται ιδιαίτερα αντιληπτό από κάποια συγκεκριμένα χαρακτηριστικά. Σαν αποτέλεσμα η τεχνική LSA αποδίδει χειρότερα στην συγκεκριμένη έννοια λόγω έλλειψης υψηλών συσχετίσεων ανάμεσα στις περιοχές του συνόλου των θετικών για την έννοια εικόνων.

Έννοια c_i	**Μέτρα Αξιολόγησης**			
	P	**R**	**AP**	**F**
Βλάστηση	0.27	0.22	0.18	0.24
Δρόμος	0.04	0.05	0.04	0.04
Έκρηξη_ Φωτιά	0.00	0.11	0.00	0.00
Ουρανός	0.29	0.21	0.18	0.24
Χιόνι	0.02	0.27	0.01	0.04
Γραφείο	0.10	0.15	0.11	0.12
Έρημος	0.00	0.44	0.06	0.00
Εξωτερικός χώρος	0.60	0.65	0.52	0.62
Βουνό	0.00	0.17	0.00	0.00

Πίνακας 4.11: *Τελικά αποτελέσματα πειραμάτων με τη χρήση LSA.*

Έννοια c_i	**Μέτρα Αξιολόγησης**			
	P	**R**	**AP**	**F**
Βλάστηση	0.63	0.22	0.40	0.33
Δρόμος	0.40	0.05	0.21	0.09
Έκρηξη_ Φωτιά	0.20	0.11	0.15	0.14
Ουρανός	0.56	0.27	0.37	0.36
Χιόνι	0.82	0.26	0.53	0.40
Γραφείο	0.41	0.15	0.29	0.22
Έρημος	0.22	0.69	0.25	0.33
Εξωτερικός χώρος	0.33	0.63	0.38	0.43
Βουνό	0.11	0.04	0.07	0.06

Πίνακας 4.12: *Τελικά αποτελέσματα πειραμάτων με τη χρήση LSA και σύνολο ελέγχου αποτελούμενο από 20% θετικά και 80% αρνητικά χαρακτηριστικά καρέ (λ=4).*

Έννοια c_i	χωρίς LSA			με LSA		
	P	**R**	**AP**	**P**	**R**	**AP**
Βλάστηση	0.64	0.31	0.46	0.63	0.22	0.40
Δρόμος	0.23	0.05	0.28	0.40	0.05	0.21
Έκρ.Φωτιά	0.29	0.78	0.18	0.20	0.11	0.15
Ουρανός	0.57	0.30	0.44	0.56	0.27	0.37
Χιόνι	0.78	0.41	0.46	0.82	0.26	0.53
Γραφείο	0.45	0.16	0.32	0.41	0.15	0.29
Έρημος	0.33	0.31	0.29	0.22	0.69	0.25
Εξ.χώρος	0.43	0.51	0.36	0.33	0.63	0.38
Βουνό	0.44	0.14	0.24	0.11	0.04	0.07

Πίνακας 4.13: *Σύγκριση αποτελεσμάτων χωρίς και με LSA, με σύνολο ελέγχου αποτελούμενο από 20% θετικά και 80% αρνητικά χαρακτηριστικά καρέ (λ=4).*

4.7.4 Παραδείγματα Σωστών και Λανθασμένων Εντοπισμών

Σε αυτή την Ενότητα σχολιάζονται και δίνονται παραδείγματα εικόνων οι οποίες απεικονίζουν μία έννοια και είτε ανιχνεύθηκαν σωστά είτε όχι, επίσης εικόνων οι οποίες δεν απεικονίζουν μία έννοια αλλά ανιχνεύθηκαν λανθασμένα πως την απεικονίζουν. Οι έννοιες για τις οποίες θα παρουσιαστούν παραδείγματα είναι ο *ουρανός* και η *βλάστηση*.

Στο Σχήμα 4.19 φαίνονται 4 εικόνες που απεικονίζουν την έννοια *ουρανός* (1-4) και 4 εικόνες την έννοια *βλάστηση* (5-8). Αυτές είναι παραδείγματα σωστής ανίχνευσης, δηλαδή το σύστημα για την έννοια *ουρανός* ανίχνευσε πως σις εικόνες (1-4) απεικονίζεται η συγκεκριμένη έννοια. Αντιστοίχως για την έννοια *βλάστηση* και τις εινες (5-8). Στις 4 πρώτες εικόνες παρατηρούνται χαρακτηριστιές περιοχές *ουρανού* από την κοπιά του χρώματος και της υφής, ο οποίες και διευκολύνουν τον σωστό εντοπισμό. Το ίδιο συμβαίνει και στις 4 τελευταίες εικόνες, όπου υπάρχουν χαρακτηριστικές περιοχές πράσινης *βλάστησης*.

Στο Σχήμα 4.20 φαίνονται 4 εικόνες που απεικονίζουν την έννοια *ουρανός* (1-4) και 4 εικόνες την έννοια *βλάστηση* (5-8). Αυτές είναι παραδείγματα λανθασμένης ανίχνευσης, δηλαδή το σύστημα για την έννοια *ουρανός* δεν ανίχνευσε πως στις εικόνες (1-4) απεικονίζεται η συγκεκριμένη έννοια. Αντιστοίχως για την έννοια *βλάστηση* και τις εικόνες (5-8). Για τον *ουρανό* και τις εικόνες 1 και 3 ισχύει πως τα τεχνητά χαρακτηριστικά τους τα οποία προσδίδουν και διαφορετικά χρώματα στον ουρανό συμβάλλει στην λανθασμένη ανίχνευση. Το σύστημα το οποίο έχει εκπαιδευτεί με παραδείγματα *ουρανού* ο οποίος έχει γαλάζιο ή και γαλάζιο με γκρι (σύννεφα) χρώμα, αδυνατεί να ανιχνεύσει έναν *ουρανό* με μωβ (1) ή κίτρινο (3) χρώμα. Για τις εικόνες 1 και 2, ισχύει ότι η κατάτμηση των εικόνων ενώνει την περιοχή του *ουρανού* με άλλες περιοχές, όπως εκείνες των *κλαδιών* οπότε και επηρεάζονται τα οπτικά χαρακτηριστικά χαμηλού επιπέδου. Τέλος στην εικόνα 4 υπάρχει περιοχή *ουρανού* η οποία είναι εντελώς λευκή. Για τις εικόνες της *βλάστησης* και συγκεκριμένα για τις εικόνες 5, 6 και 8 φαίνονται μία πράσινη λωρίδα βλάστησης, ένα δέντρο και μία γλάστρα αντίστοιχα τα οποία εξαιτίας της μικρής και μη συμπαγούς περιοχής που καταλαμβάνουν, στην κατάτμηση ενώνονται με άλλες περιοχές κι έτσι αλλοιώνονται τα οπτικά χαρα-

Σχήμα 4.19: *Εικόνες που απεικονίζουν την έννοια ουρανός(1-4) και την έννοια βλάστηση (5-8) οι οποίες εντοπίστηκαν επιτυχώς.*

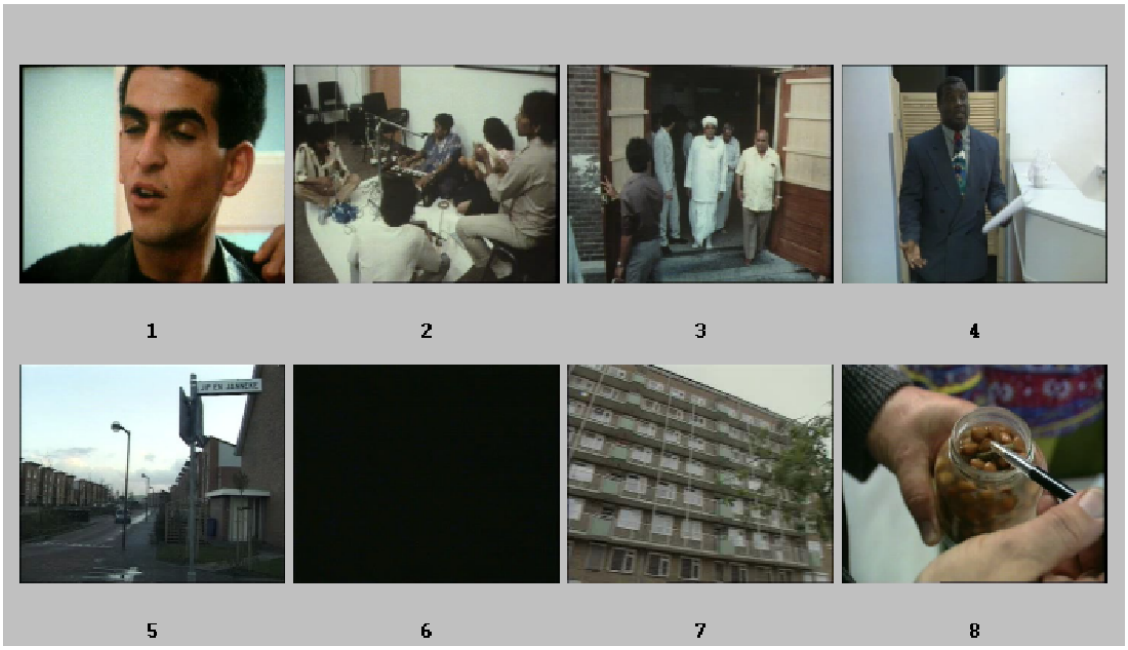
κτηριστικά χαμηλού επιπέδου της περιοχής. Η εικόνα 7 είναι ένα παράδειγμα εικόνας διαφορετικής από εκείνο με το οποίο το σύστημα έχει εκπαιδευτεί, δηλαδή πράσινες περιοχές *βλάστησης* και όχι κίτρινα λουλούδια τα οποία είναι μία εξαίρεση.

Στο Σχήμα 4.21 φαίνονται 4 εικόνες που δεν απεικονίζουν την έννοια *ουρανός* (1-4) και 4 εικόνες που δεν απεικονίζουν την έννοια *βλάστηση* (5-8). Αυτές είναι παραδείγματα λανθασμένης ανίχνευσης, δηλαδή το σύστημα για την έννοια *ουρανός* ανίχνευσε πως στις εικόνες (1-4) απεικονίζεται η συγκεκριμένη έννοια. Αντιστοίχως για την έννοια *βλάστηση* και τις εικόνες (5-8). Οι εικόνες (1-4) φαίνεται πως περιέχουν γαλάζιες περιοχές οι οποίες πλησιάζουν στα χαρακτηριστικά των περιοχών *ουρανού*. Για την *βλάστηση* οι εικόνες 5 και 7 είναι παραδείγματα λάθους στο σύνολο δεδομένης αλήθειας. Αυτό διότι έχουν σημειωθεί ότι δεν απεικονίζουν την έννοια *βλάστηση* ενώ τελικά την απεικονίζουν και το σύστημα ανίχνευσε επίσης πως την απεικονίζουν. Ενώ δηλαδή εδώ υπάρχει ένας σωστός εντοπισμός, στις μετρήσεις υπολογίζεται σαν λανθασμένος. Η εικόνα 8 περιέχει μία πράσινη περιοχή η οποία ενδεχομένως προκαλεί σύγχυση στο σύστημα. Τέλος η εικόνα 6 έχει μία πράσινη απόχρωση καθώς και υφή η οποία θυμίζει αυτή της *βλάστησης*, έτσι κι ανιχνεύεται ως απεικόνιση της (παρότι αυτό ενδέχεται να μην φαίνεται στο χαρτί, παρατίθεται γιατί αποτελεί σύνηθες φαινόμενο σε ακολουθιες βίντεο από τις οποίες έχει εξαχθεί ένα χαρακτηριστικό καρέ).

4.7.5 Πειράματα σε Σύνολα Εικόνων του COREL

Ο ίδιος αλγόριθμος ανίχνευσης εικόνων εφαρμόστηκε και σε ένα σύνολο από εικόνες της βάσης δεδομένων του Corel. Αυτό έγινε για να μπορέσει η μέθοδος να αξιολογηθεί καλύτερα με την χρήση ενός πιο "εύκολου" συνόλου εικόνων, αλλά και με ένα σύνολο από το οποίο θα μπορεί να υπάρχει και σύνολο ελέγχου με γνωστή απόκριση. Επίσης κάτι σημαντικό που προσφέρει αυτό το σύνολο εικόνων είναι ότι λόγω του μικρού του μεγέθους είναι πιο εύκολο και γρήγορο να γίνουν διάφορα πειράματα αλλάζοντας για παράδειγμα το μέγεθος του θησαυρού και να επιλεχθεί

Σχήμα 4.20: *Εικόνες που απεικονίζουν την έννοια ουρανός(1-4) και την έννοια βλάστηση (5-8) οι οποίες δεν εντοπίστηκαν.*



Σχήμα 4.21: *Εικόνες που δεν απεικονίζουν την έννοια ουρανός(1-4) ή την έννοια βλάστηση (5-8) οι οποίες εντοπίστηκαν λανθασμένα πως τις απεικονίζουν.*

έτσι το καταλληλότερο μέγεθος. Στο Σχήμα 4.22 φαίνονται μερικές εικόνες από το σύνολο που χρησιμοποιήθηκε. Στην προκειμένη περίπτωση δεν υπάρχουν βίντεο από τα οποία γίνεται εξαγωγή χαρακτηριστικών καρέ αλλά εικόνες απευθείας. Έτσι κάθε εικόνα εδώ είναι που συμβολίζεται με k_i.

Συνολικός Αριθμός Χαρ. Καρέ	N_K	750
Συνολικός Αριθμός Περιοχών	N_R	18150
Συνολικός Αριθμός Διαν. Χαρακτηριστικών	N_F	18150
Μέγεθος Οπτικού Θησαυρού	N_T	50/40/30/20

Πίνακας 4.14: *Στοιχεία πειράματος COREL*

Οι έννοιες υψηλού επιπέδου που επιλέχθηκαν να εντοπιστούν είναι *ουρανός*, *βλάστηση* και *χιόνι*, έννοιες δηλαδή συμβατές και με εκείνες του TRECVID. Έτσι στο σύνολο των εικόνων υπάρχουν γενικά εικόνες που απεικονίζουν τουλάχιστον μία από αυτές τις έννοιες. Προστέθηκαν επίσης στο σύνολο κάποιες εικόνες οι οποίες δεν απεικόνιζαν καμία από τις παραπάνω έννοιες. Συγκεκριμένα το σύνολο αποτελείται από 600 εικόνες οι οποίες απεικονίζουν τουλάχιστον μία από τις 3 παραπάνω έννοιες (497 *ουρανό*, 126 *χιόνι* και 342 *βλάστηση*) και 150 εικόνες των οποίων η απεικόνιση είναι διαφορετική με τις 3 έννοιες. Οι 750 εικόνες αυτές διαχωρίστηκαν σε σύνολο εκπαίδευσης και ελέγχου, καταχωρώντας 525 εικόνες στο πρώτο και 225 στο δεύτερο. Η σύνθεση των δύο αυτών συνόλων ανάλογα με το πόσες εικόνες περιέχουν που απεικονίζουν την έννοια και πόσες όχι φαίνεται στον Πίνακα 4.16.

Έννοια c_i	**Εικόνες** $\lvert G_i \rvert$
Ουρανός	497
Χιόνι	126
Βλάστηση	342

Πίνακας 4.15: *Αριθμός χαρακτηριστικών καρέ που απεικονίζουν την κάθε έννοια στο σύνολο του Corel.*

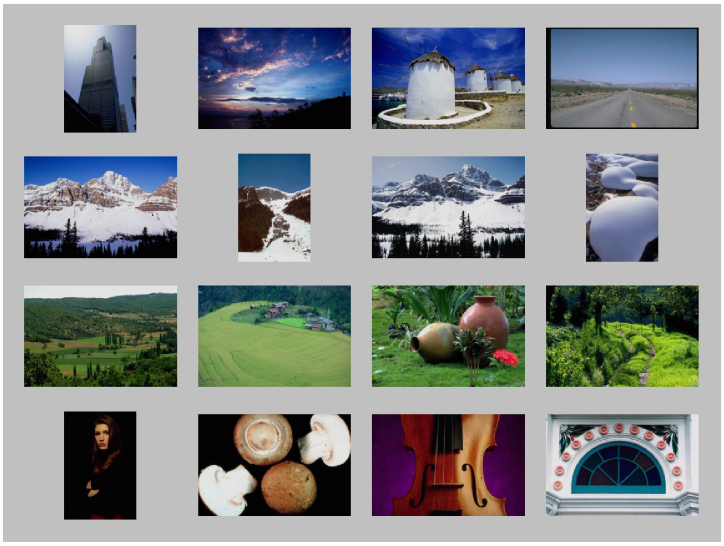
Έννοια c_i	$\mathbf{TR_i}$		$\mathbf{TE_i}$	
	G_i	$\bar{G}_i$	G_i	$\bar{G}_i$
Ουρανός	344	181	153	72
Χιόνι	88	147	38	187
Βλάστηση	236	289	106	119

Πίνακας 4.16: *Σύνολα εκπαίδευσης και ελέγχου για τα πειράματα με το σύνολο του Corel.*

Σε αυτά τα πειράματα το διάνυσμα χαρακτηριστικών αποτελείται από έναν παραπάνω περιγραφέα ο οποίος δεν χρησιμοποιήθηκε στο TRECVID, τον Περιγραφέα Ομοιογενούς Υφής και είναι

$$f_i = f(r_i) = \Big[CLD(r_i)\, DCD(r_i)\, CSTD(r_i)\, SCD(r_i)\, EHD(r_i)\, HTD(r_i)\Big], \quad r_i \in R\,. \tag{4.36}$$

Έγιναν πειράματα με θησαυρό 50, 40, 30 και 20 τύπων περιοχής. Αυτό γίνεται για να βρεθεί ο αριθμός εκείνος των τύπων περιοχών ο οποίος μπορεί καλύτερα να



Σχήμα 4.22: *Μερικές από τις εικόνες που χρησιμοποιήθηκαν από την συλλογή του Corel, εικόνες που απεικονίζουν ουρανό στην πρώτη σειρά, χιόνι στην δεύτερη, βλάστηση στην τρίτη και καμία από τις 3 έννοιες στην τέταρτη.*

περιγράψει τις εικόνες του κάθε συνόλου και για τις κάθε φορά έννοιες. Για την κατασκευή του θησαυρού χρησιμοποιήθηκε όλο το σύνολο των περιοχών καθώς ήταν ικανοποιητικής διάστασης. Έτσι τα νευρωνικά δίκτυα για κάθε μία από τις 3 έννοιες εκπαιδεύτηκαν με 4 διαφορετικά σύνολα εκπαίδευσης. Το διαφορετικό ανάμεσα σε αυτά ήταν το μέγεθος του θησαυρού. Υπολογίστηκε τελικά η μέση ακρίβεια καθώς το n αυξάνεται πάνω στα 4 αντίστοιχα σύνολα ελέγχου. Τα διαγράμματα στο Σχήμα 4.24 δείχνουν της καμπύλες AP^n ως προς το n και από αυτές επιλέγεται και το καταλληλότερο μέγεθος θησαυρού. Τελικά τα μεγέθη θησαυρού τα οποία επιλέγονται είναι 50, 30 και 40 για τις έννοιες *χιόνι*, *ουρανός* και *βλάστηση* αντίστοιχα. Είναι προφανές ότι για κάθε έννοια διαφέρει ο αριθμός των χαρακτηριστικών περιοχών οι οποίες είναι κατάλληλες για να περιγράψουν τις εικόνες της έννοιας κατάλληλα.

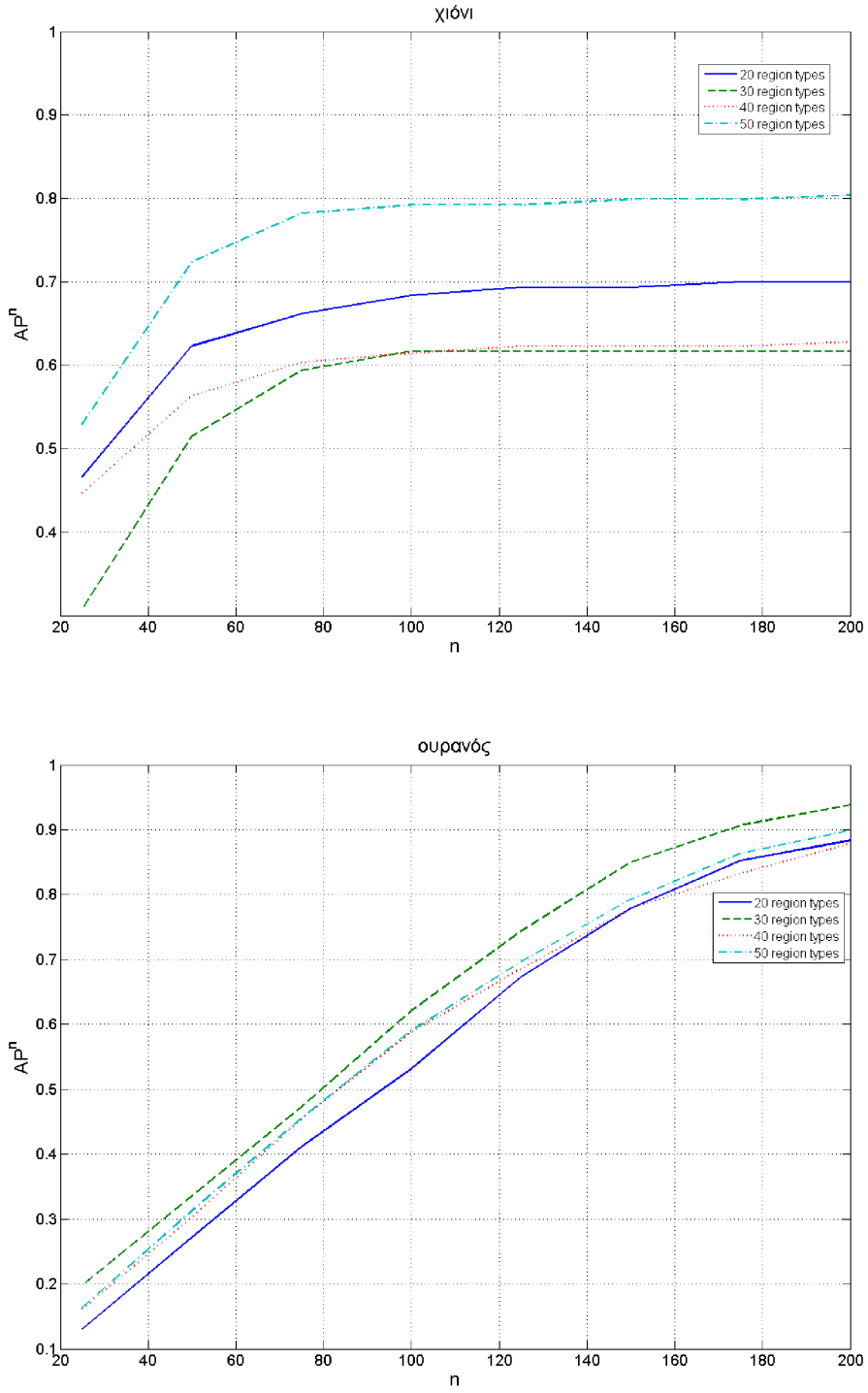
	ουρανός	*χιόνι*	*βλάστηση*
χωρίς LSA	0.5	0.4	0.3
με LSA	0.5	0.4	0.3

Πίνακας 4.17: *Κατώφλια για όλους τους ανιχνευτές που εκπαιδεύτηκαν.*

Τα κατώφλια επιλέχθηκαν για κάθε ανιχνευτή όπως φαίνονται στον Πίνακα 4.17. Τα τελικά αποτελέσματα για τα μέτρα που χρησιμοποιήθηκαν για την αξιολόγηση των ανιχνευτών παρουσιάζονται στους Πίνακες 4.18 και 4.19 χωρίς και με την χρήση LSA αντίστοιχα.

Έννοια c_i	**Μέτρα Αξιολόγησης**			
	P	**R**	**AP**	**F**
Βλάστηση	0.83	0.94	0.64	0.88
Ουρανός	0.68	0.79	0.79	0.73
Χιόνι	0.73	0.74	0.63	0.73

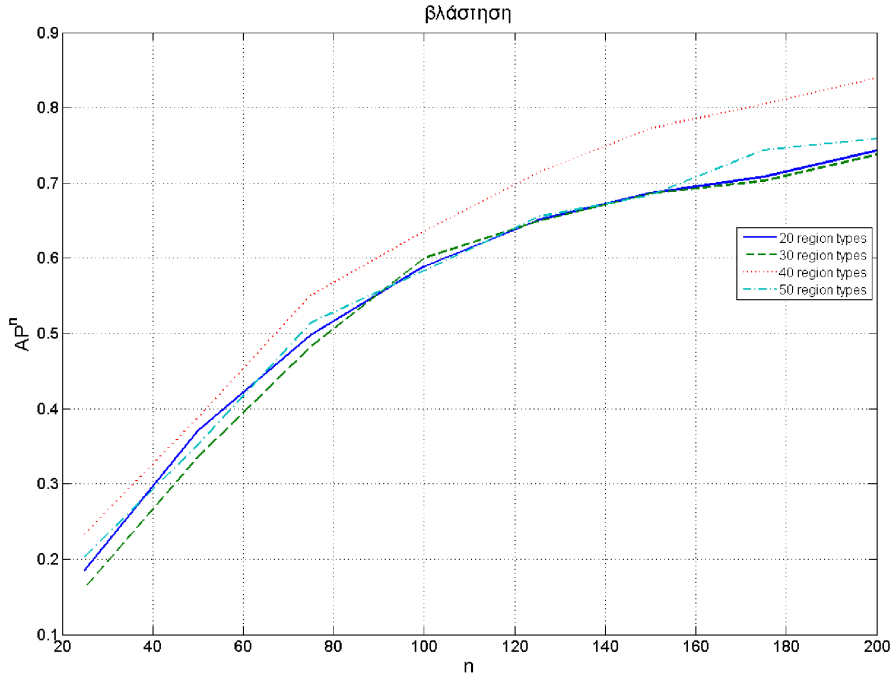
Πίνακας 4.18: *Τελικά αποτελέσματα πειραμάτων στο σύνολο του Corel, χωρίς τη χρήση LSA.*



Σχήμα 4.23: *Διαγράμματα Average Precision καθώς αυξάνεται το μέγεθος παραθύρου για τα τέσσα μγέθη θησαυρού, για τις έννοιες χιόνι και ουρανός.*

4.8 Συμπεράσματα

Στο Κεφάλαιο αυτό αντιμετωπίστηκε το πρόβλημα της ανίχνευσης εννοιών υψηλού επιπέδου σε εικόνες. Για το σκοπό αυτό αναπτύχθηκε μια τεχνική που βασίζεται στο μοντέλο bag-of-words. Η τεχνική αυτή εξάγει περιγραφείς από εικόνες και με



Σχήμα 4.24: *Διάγραμμα Average Precision καθώς αυξάνεται το μέγεθος παραθύρου για τα τέσσερα μεγέθη θησαυρού, για την έννοια βλάστηση.*

Έννοια c_i	**Μέτρα Αξιολόγησης**			
	P	**R**	**AP**	**F**
Βλάστηση	0.81	0.94	0.65	0.87
Ουρανός	0.54	0.74	0.67	0.62
Χιόνι	0.73	0.71	0.63	0.72

Πίνακας 4.19: *Τελικά αποτελέσματα πειραμάτων στο σύνολο του Corel, με τη χρήση LSA*

τη βοήθεια ενός οπτικού θησαυρού, κατασκευάζει περιγραφές των εικόνων με βάση τους πιο συνηθισμένους τύπους περιοχής που αυτές περιέχουν. Επιπρόσθετα, το μοντέλο bag-of-words ενσωματώθηκε στην τεχνική της λανθάνουσας σημασιολογικής ανάλυσης και διερευνήθηκε ποιες έννοιες μπορούν να ωφεληθούν από αυτό. Η τεχνική που παρουσιάστηκε εμφανίζει το πλεονέκτημα ότι παρέχει μια προσέγγιση για την ανίχνευση εννοιών υψηλού επιπέδου σε εικόνες η οποία μπορεί να εφαρμοστεί για την ανίχνευση πολλών εννοιών, αρκεί αυτές να χαρακτηρίζονται ως υλικά ή σκηνές. Δε στηρίζεται σε ευρετικές μεθόδους και δεν εκμεταλλεύεται συγκεκριμένα χαρακτηριστικά κάποιων εννοιών. Επιπρόσθετα, μιας και βασίζεται στην παρουσία των τύπων περιοχής, δεν εξαρτάται από το μέγεθος της περιοχής της εικόνας στην οποία περιέχεται η έννοια υπό ανίχνευση.

Τα μειονεκτήματα της προτεινόμενης τεχνικής είναι κατά πρώτον η εξάρτησή της από τον αλγόριθμο κατάτμησης που χρησιμοποιείται στο αρχικό της στάδιο και κατά δεύτερον η αδυναμία επέκτασής της προκειμένου να ανιχνευθούν έννοιες που μπορούν να χαρακτηριστούν ως αντικείμενα. Έτσι, αν η περιοχή που περιέχει μια από τις έννοιες υπό ανίχνευση είναι αρκετά μικρή και παρόμοια με τις γειτονικές της, μπορεί σε πολλές περιπτώσεις να ενωθεί με αυτές και να κατασκευαστεί ένα διάνυ-

σμα αναπαράστασης που δεν θα οδηγήσει στην ανίχνευσή της. Επίσης, η περιγραφή των ιδιοτήτων της εικόνας με βάση τις περιοχές της δεν μπορεί σε καμία περίπτωση να οδηγήσει στην αναγνώριση αντικειμένων, καθώς κάτι τέτοιο θα απαιτούσε τέλεια κατάτμηση, κατ'αρχήν, κάτι που δεν είναι εφικτό.

Η εφαρμογή της τεχνικής στο σύνολο εικόνων που προέρχεται από το TRECVID οδήγησε σε ορισμένα χρήσιμα συμπεράσματα αφενός για την απόδοσή της και αφετέρου για τον τρόπο με τον οποίο πρέπει να αντιμετωπίζονται μεγάλα σύνολα δεδομένων. Τα αποτελέσματα που παρουσιάστηκαν στην Ενότητα 4.7 καταδεικνύουν ότι η ανίχνευση εννοιών υψηλού επιπέδου σε εικόνες μπορεί να επιτευχθεί σε ικανοποιητικό βαθμό, όταν το περιεχόμενο μιας εικόνας περιγράφεται με ένα διάνυσμα αναπαράστασης το οποίο βασίζεται σε έναν οπτικό θησαυρό. Η ακρίβεια που επιτεύχθηκε είναι άμεσα συγκρίσιμη με αυτές από ανταγωνιστικές τεχνικές, στην ίδια διαδικασία αξιολόγησης, που επιδιώκουν να λύσουν το ίδιο πρόβλημα με τα ίδια μέσα, δηλαδή την οπτική πληροφορία χρώματος και υφής. Η μελέτη του λόγου των αρνητικών ως προς τα θετικά παραδείγματα των εννοιών που ανιχνεύονται δείχνει ότι όσο αυξάνεται ο λόγος αυτός, τόσο δυσκολεύει το έργο των ανιχνευτών, γιατί αφενός εμφανίζονται παραδείγματα με τα οποία δεν έχουν εκπαιδευτεί και αφετέρου γιατί αυξάνεται η πιθανότητα να εμφανιστούν αρνητικά παραδείγματα οπτικά παρόμοια με τα θετικά.

Η τεχνική LSA αποδείχθηκε πειραματικά ότι βελτίωσε την ακρίβεια σε κάποιες έννοιες. Η αξιολόγηση έδειξε ότι ανάμεσα στις πρώτες εικόνες με το μεγαλύτερο ποσοστό πεποίθησης βρίσκονται περισσότεροι σωστοί εντοπισμοί από ότι χωρίς την χρήση LSA. Αυτό συμβαίνει καθώς η τεχνική αυτή λαμβάνει υπόψη τις λανθάνουσες σχέσεις μεταξύ των περιοχών και αυτές οι συσχετίσεις μπορεί να είναι πιο μεγάλες για μερικές έννοιες από κάποιες άλλες. Έτσι για τις έννοιες με υψηλές συσχετίσεις μεταξύ των περιοχών η τεχνική αποδίδει καλύτερα.

Κεφάλαιο 5

Εξαγωγή Χαρακτηριστικών Καρέ από Βίντεο με χρήση Οπτικού Θησαυρού

5.1 Εισαγωγή

Τα τελευταία χρόνια έχει σημειωθεί μια αξιοσημείωτη αύξηση στη δημιουργία αλλά και τη ζήτηση πολυμεσικού υλικού και ιδιαίτερα ψηφιακών βίντεο. Η αύξηση αυτή οφείλεται αφενός στο ότι οι συσκευές καταγραφής και αποθήκευσης εικόνας έχουν πλέον γίνει ιδιαίτερα προσιτές και αφετέρου στη ραγδαία εξάπλωση του διαδικτύου. Έτσι, έχει δοθεί στον καθέναν πρόσβαση σε μεγάλες συλλογές οπτικοακουστικού περιεχομένου. Οι συλλογές αυτές μπορεί να περιέχουν προσωπικά βίντεο από χρήστες του διαδικτύου, ταινίες, οπτικοακουστικά αρχεία, δελτία ειδήσεων και άλλα.

Η άνθηση αυτή του βίντεο έχει προκαλέσει στους χρήστες νέες ανάγκες και συνεπώς έχει οδηγήσει στην ανάπτυξη νέων εφαρμογών, κάτι που συνεπάγεται έρευνα και ανάπτυξη νέων τεχνολογιών. Οι τεχνολογίες αυτές έχουν σκοπό να βοηθήσουν στη δημιουργία αρχείων, καταλόγων και δεικτοδότησης του οπτικοακουστικού υλικού, αλλά και να διευκολύνουν την απόδοση, τη χρηστικότητα και την πρόσβαση στις συλλογές του υλικού. Ανάμεσα στις εφαρμογές αυτές, ιδιαίτερη βαρύτητα έχει η προσπάθεια για εύκολη περιήγηση σε πολυμεσικό υλικό, κάτι που απαιτεί αποδοτική πρόσβαση και αναπαράσταση. Για το σκοπό αυτό έχουν αναπτυχθεί τεχνικές περίληψης βίντεο, οι οποίες έχουν μαγνητίσει ιδιαίτερα το ερευνητικό ενδιαφέρον, τα τελευταία χρόνια.

Στο πλαίσιο αυτό κινείται και η τεχνική που παρουσιάζεται στο παρόν Κεφάλαιο. Ο σκοπός είναι η εξαγωγή ενός μικρού αριθμού καρέ από ακολουθίες βίντεο, τα οποία και θα αποκαλούνται "χαρακτηριστικά" και το επιθυμητό είναι να έχουν την ιδιότητα να εκφράζουν όσο το δυνατόν πληρέστερα το οπτικό και το σημασιολογικό περιεχόμενο του βίντεο. Έτσι, οι χρήστες μεγάλων αρχείων θα μπορούν εύκολα και γρήγορα να περιηγηθούν σε αυτά, βλέποντας τις περιλήψεις των βίντεο και επιλέγοντας με βάση αυτές τα βίντεο που πραγματικά τους ενδιαφέρουν. Ωστόσο, το παρόν Κεφάλαιο επιδιώκει κατά κύριο λόγο να χρησιμοποιήσει τις περιλήψεις για την εξαγωγή εννοιών χαμηλού επιπέδου από βίντεο. Αντί να εφαρμοστούν οι τεχνικές ανίχνευσης εννοιών σε ολόκληρα τα βίντεο, κάτι εξαιρετικά χρονοβόρο, το ιδανικό θα ήταν με την εφαρμογή τους στις περιλήψεις που εξάγονται από αυτά να ανιχνεύονται

οι ίδιες έννοιες.

5.2 Περιγραφή του προβλήματος

Η δημιουργία περιλήψεων σχετίζεται με το πρόβλημα της παραγωγής μιας σημαντικά ελαττωμένης αναπαράστασης ακολουθιών βίντεο. Το πρόβλημα αυτό είναι ιδιαίτερα σημαντικό σε αρκετά πεδία, όπως για παράδειγμα είναι η πλοήγηση σε ψηφιακές βιβλιοθήκες, η δημιουργία περιλήψεων σε εφαρμογές όπως ή διαδραστική τηλεόραση, η δημιουργία προσωποποιημένων εφαρμογών για φορητές συσκευές, αλλά και για ενημερωτικούς τηλεοπτικούς σταθμούς, που λαμβάνουν καθημερινά τεράστιες ποσότητες από ακατέργαστα βίντεο [205].

Ο σκοπός όλων των τεχνικών που εντάσσονται στην ευρύτερη κατηγορία δημιουργίας περιλήψεων είναι να εξάγουν ένα τμήμα του βίντεο που θα αποτελείται είτε από κινούμενες είτε από ακίνητες εικόνες με τέτοιο τρόπο ώστε αυτό να μπορεί να παρέχει όσο το δυνατόν ακριβέστερη πληροφορία σχετικά με το περιεχόμενο του συνολικού βίντεο. Είναι επιθυμητό ο χρήστης που θα δεί την περίληψη που δημιουργήθηκε να αφιερώσει όσο το δυνατόν λιγότερο χρόνο και η πληροφορία που θα λάβει να μην αλλοιώσει το περιεχόμενο της αρχικής, όπως τονίζουν στην εργασία τους οι Pfeiffer et al. [170].

Θεωρητικά, η δημιουργία μιας περίληψης ενός βίντεο μπορεί να γίνει τόσο χειροκίνητα όσο και αυτόματα. Είναι, όμως προφανές ότι ο τεράστιος όγκος οπτικοακουστικού υλικού που έχει συσσωρευτεί τα τελευταία χρόνια καθιστά ανεδαφικές τις όποιες σκέψεις για χειροκίνητη δημιουργία περιλήψεων μεγάλων ποσοτήτων βίντεο, ιδιαίτερα όταν κάτι τέτοιο απαιτείται να γίνει σε σύντομο χρονικό διάστημα. Για το σκοπό αυτό είναι απαραίτητη η ανάπτυξη ευφυών τεχνικών οι οποίες θα πρέπει να ελαχιστοποιούν ή και να μην απαιτούν καθόλου την ανθρώπινη ανάμιξη. Φυσικά, το Κεφάλαιο αυτό εστιάζει σε τεχνικές όπου η δημιουργία περιλήψεων από βίντεο γίνεται τελείως αυτόματα.

Υπάρχουν γενικά δύο κατηγορίες περιλήψεων, όπως έχει ήδη αναφερθεί, αυτές που αποτελούνται από ένα σύνολο από ακίνητες εικόνες και αυτές που αποτελούνται από μικρά κομμάτια βίντεο. Η πρώτη κατηγορία, η οποία είναι γνωστή και ως "μακέτες ιστορίας" (image storyboards). Ένα σύνολο από εικόνες εξάγεται ή δημιουργείται από την ακολουθία βίντεο. Στη συνέχεια, καθώς η τεχνική που θα παρουσιαστεί στην Ενότητα 5.4 ανήκει σε αυτή την κατηγορία, αφενός θα δοθεί ιδιαίτερη έμφαση σε αυτή και αφετέρου όταν αναφέρεται ο όρος "περίληψη", θα υποννοείται μια τεχνική που εντάσσεται σε αυτήν την κατηγορία. Όσον αφορά, τώρα τη δεύτερη κατηγορία, ένα σύνολο από ακολουθίες εικόνων μαζί με το ηχητικό τους κομμάτι εξάγεται από τις ακολουθίες βίντεο. Έτσι και αυτές τελικά είναι ένα μικρότερο βίντεο. Οι τεχνικές που εντάσσονται στην κατηγορία αυτή, συχνά αναφέρονται ως "video skimming".

Υπάρχουν βασικές διαφορές ανάμεσα στις κατηγορίες αυτές. Καταρχάς, μια περίληψη ενός βίντεο μπορεί να κατασκευαστεί αρκετά γρηγορότερα, καθώς συνήθως απαιτείται μόνο η επεξεργασία της οπτικής πληροφορίας. Αντίθετα, στην άλλη περίπτωση απαιτείται επεξεργασία του ήχου που συνοδεύει την εικόνα καθώς και του κειμένου που ενδεχομένως εμφανίζεται στο βίντεο. Επίσης, η παρουσίαση της περίληψης είναι απλούστερη και ευκολότερη και πολύ συχνά επιτυγχάνεται πολύ ικανοποιητικό αποτέλεσμα, χωρίς να φανεί η απουσία ήχου. Επιπρόσθετα, υπάρχουν πολλοί τρόποι παρουσίασης των περιλήψεων και έτσι μπορεί η παρουσίαση να προσαρμοστεί στις

ανάγκες της εκάστοτε εφαρμογής. Τέλος, ένα αρκετά σημαντικό πλεονέκτημα είναι ότι οι περιλήψεις μπορεί να εκτυπωθούν και να διανεμηθούν, άρα και να αξιοποιηθούν χωρίς να χρειάζεται ειδικός εξοπλισμός.

Φυσικά και η έτερη κατηγορία παρουσιάζει κάποια πλεονεκτήματα, τα οποία θα ήταν παράλειψη να μην αναφερθούν. Η αξιοποίηση της πληροφορίας που περιέχει ο ήχος μπορεί να χρησιμοποιηθεί για παράδειγμα για την ανίχνευση σημαντικών στιγμών σε βίντεο που περιέχουν αθλητικά γεγονότα. Επίσης, το περιεχόμενο του ήχου σε κάποιες κατηγορίες περιεχομένου όπως για παράδειγμα σε εκπαιδευτικά βίντεο έχει πληροφορία η οποία δεν υπάρχει στο οπτικό του μέρος. Παρά την αυξημένη υπολογιστική ισχύ που απαιτούν, οι τεχνικές αυτές δίνουν συνήθως αρτιότερα αποτελέσματα, τα οποία μάλιστα είναι πιο ευχάριστα στους τελικούς χρήστες.

5.3 Σχετικές Εργασίες

Όπως έχει γίνει σαφές, οι περιλήψεις βίντεο με τις οποίες ασχολείται το παρόν Κεφάλαιο είναι στην ουσία ένα σύνολο από εικόνες που εξάγονται από ακολουθίες βίντεο, το οποίο προσπαθεί να αναπαραστήσει με τον καλύτερο δυνατό τρόπο το οπτικό περιεχόμενο του βίντεο. Έτσι, οι ερευνητικές προσπάθειες στρέφονται κατά κύριο λόγο στην εξαγωγή των εικόνων αυτών από τις ακολουθίες βίντεο. Επειδή οι εικόνες αυτές στην περίπτωση του βίντεο αποκαλούνται "καρέ", τα καρέ που εξάγονται έχει καθιερωθεί να αποκαλούνται "χαρακτηριστικά καρέ". Στη βιβλιογραφία έχουν προταθεί πολλές τεχνικές που προσπαθούν με ποικίλες μεθοδολογίες να εξάγουν χρήσιμες και αποδοτικές περιλήψεις από ακολουθίες βίντεο. Εκτεταμένες βιβλιογραφικές μελέτες των ερευνητικών προσπαθειών στο χώρο έχουν γίνει από τους Truong και Venkatesh [215], οι οποίοι τις κατέταξαν με βάση τις τεχνικές που χρησιμοποιούν, από τους Li et al. [121], [120], οι οποίοι τις χώρισαν με βάση το μέγεθος του βίντεο στο οποίο εφαρμόζονται, από τους Sundaram και Chang [205], οι οποίοι δίνουν βάρος στην παράλληλη ανάλυση ταυτόχρονα με την εξαγωγή της περίληψης και τέλος από τον Kang [97], ο οποίος αντιμετώπισε σε πιο υψηλό επίπεδο το πρόβλημα, παρουσιάζοντας όλες του τις πλευρές.

Οι τεχνικές εξαγωγής χαρακτηριστικών καρέ, σύμφωνα με τους Li et al. [121], μπορούν να χωριστούν στις παρακάτω κατηγορίες:

- Τεχνικές που βασίζονται σε *δειγματοληψία* (sampling-based)
- Τεχνικές που βασίζονται στα *πλάνα* του βίντεο (shot-based)
- Τεχνικές που βασίζονται σε *τμήματα* του βίντεο (segment-based)
- Λοιπές Τεχνικές

5.3.1 Τεχνικές που βασίζονται σε δειγματοληψία

Στις τεχνικές που εντάσσονται στην κατηγορία αυτή, τα χαρακτηριστικά καρέ εξάγονται με τυχαία ή ομοιόμορφη δειγματοληψία στα καρέ του βίντεο, σε συγκεκριμένα χρονικά διαστήματα. Σε αυτή την κατηγορία εντάσσεται το σύστημα Hierarchical Video Magnifier των Mills et al. [139]. Το σύστημα αυτό επέτρεπε στους χρήστες του να περιηγούνται σε ακολουθίες βίντεο, αλλάζοντας την χρονική τους ανάλυση.

Τα βίντεο χωρίζονταν σε "κεφάλαια" και το πρώτο καρέ κάθε κεφαλαίου χρησιμοποιούταν για τη συνολική περιγραφή του. Έτσι κατασκεύαζαν ιεραρχικές περιγραφές των βίντεο, με βάση τις οποίες οι χρήστες μπορουσαν με γρήγορο τρόπο να αντιληφθούν το περιεχόμενο του βίντεο. Με παρόμοιο τρόπο λειτουργούσε και το σύστημα Mini-Video των Taniguchi et al. [209]. Κάθε βίντεο χωριζόταν σε διαστήματα ίσου μεγέθους και στο τέλος κάθε ενός από αυτά εξαγόταν ένα καρέ. Το χρονικό διάστημα ανάμεσα σε δύο καρέ καθοριζόταν αποκλειστικά από τις ανάγκες της εκάστοτε εφαρμογής, αλλά και το διαθέσιμο χώρο για την αποθήκευσή τους. Οι τεχνικές αυτές, ωστόσο, δεν έχουν κάποιο ερευνητικό ενδιαφέρον. Το μόνο ίσως που προσφέρουν στο αντικείμενο της έρευνας είναι ο τρόπος με τον οποίο παρουσιάζονται οι περιλήψεις στο χρήστη.

Όπως ήδη αναφέρθηκε, ένας από τους σκοπούς της τεχνικής που θα παρουσιαστεί στην Ενότητα 5.4 είναι να εφαρμοστεί στο σύνολο των βίντεο του TRECVID, στο οποίο και αξιολογήθηκαν οι τεχνικές του Κεφαλαίου 4, προκειμένου να διερευνηθεί κατά πόσο η περίληψη ενός πλάνου με περισσότερα από ένα χαρακτηριστικά καρέ βελτιώνει την ακρίβεια της ανάλυσης. Μια απλοϊκή προσέγγιση στο πλαίσιο αυτό έχει πραγματοποιηθεί από την ομάδα του ερευνητικού έργου K-Space στη συμμετοχή της στο TRECVID [230]. Αντί του μοναδικού χαρακτηριστικού καρέ που προσφέρεται από τους διοργανωτές του TRECVID, όπως περιγράφηκε στην Ενότητα 4.5, προτιμήθηκε μια πολύ πυκνή δειγματοληψία των καρέ, με σταθερό χρονικό βήμα.

Συμπερασματικά, οι τεχνικές που εντάσσονται σε αυτή την κατηγορία έχουν το πλεονέκτημα ότι δεν απαιτούν ιδιαίτερη υπολογιστική ισχύ και μπορούν εύκολα και σχετικά γρήγορα να εφαρμοστούν σε μεγάλες συλλογές βίντεο. Βέβαια, η αξιολόγηση του αποτελέσματος είναι καθαρά υποκειμενική και υπόκειται πάντα στην εφαρμογή στην οποία και χρησιμοποιείται, αλλά και στις απαιτήσεις του συγκεκριμένου χρήστη. Το βασικό τους μειονέκτημα είναι ότι αν ο ρυθμός εξαγωγής χαρακτηριστικού καρέ είναι πολύ μικρός, τότε είναι πολύ πιθανό και είναι αυτό που συμβαίνει στην πράξη, να μην εξάγονται καρέ από σημαντικά σημεία του βίντεο, αν αυτά είναι πολύ μικρά σε διάρκεια.

5.3.2 Τεχνικές που βασίζονται στα πλάνα του βίντεο

Προκειμένου να επιτύχουν καλύτερη αναπαράσταση του οπτικού περιεχομένου ενός βίντεο μέσω ενός αριθμού από χαρακτηριστικά καρέ, πολλές τεχνικές προσπαθούν να προσαρμοστούν στο περιεχόμενο της ακολουθίας βίντεο. Ο πιο απλός τρόπος να το επιτύχουν αυτό είναι να εφαρμόζονται αντί σε ολόκληρο το βίντεο, σε πλάνα του. Για να γίνει καλύτερα κατανοητό αυτό, πρέπει να υπενθυμιστεί εδώ ότι ένα πλάνο μπορεί να οριστεί ως ένα κομμάτι ενός βίντεο στο οποίο η λήψη ήταν συνεχόμενη. Έτσι, ένα πλάνο τις περισσότερες φορές έχει ξεκάθαρο περιεχόμενο και μπορεί να περιγραφεί με ανεξάρτητο τρόπο από τα υπόλοιπα.

Όπως και στην περίπτωση των τεχνικών ταξινόμησης σκηνής, οι πρώτες ερευνητικές προσπάθειες ξεκίνησαν βασιζόμενες στο χρώμα. Έτσι, για παράδειγμα, το σύστημα των Zhang et al. [240] συγκρίνει ιστογράμματα χρώματος από διαδοχικά καρέ. Έχοντας θέσει ένα κατάλληλο κατώφλι, όταν η διαφορά ανάμεσα σε δύο καρέ το υπερβεί, το δεύτερο καρέ επιλέγεται ως χαρακτηριστικό. Με παρόμοιο τρόπο δουλεύει και το σύστημα των Yeung et al. [235]. Οι Zhuang et al. [243] υιοθετούν μια μέθοδο συσταδοποίησης. Τα καρέ περιγράφονται με ένα ιστόγραμμα χρώματος.

Στη συνέχεια, αυτά χωρίζονται με έναν αλγόριθμο συσταδοποίησης χωρίς επίβλεψη. Από τις συστάδες που προκύπτουν, επιλέγονται τελικά όσες είναι μεγαλύτερες ως προς τον αριθμό των καρέ που περιέχουν από ένα κατώφλι, το οποίο και καθορίζεται εμπειρικά να είναι ίσο με το μέσο αριθμό των καρέ που περιέχει μια συστάδα. Ως χαρακτηριστικό καρέ, τελικά επιλέγεται αυτό που βρίσκεται πιο κοντά στο κέντρο της κάθε συστάδας. Η τεχνική αυτή είναι σχετικά απλή και γρήγορη, παρέχοντας μια ικανοποιητική περίληψη των βίντεο στα οποία εφαρμόζεται. Οι Hammoud και Mohr [80] δουλεύουν σε επίπεδο πλάνου και χρησιμοποιούν μείξεις Γκαουσιανών για να μοντελοποιήσουν τις χρονικές μεταβολές του οπτικού περιεχομένου τους. Τα οπτικά χαρακτηριστικά περιγράφονται με ιστογράμματα χρώματος. Με τον τρόπο αυτό τα καρέ ομαδοποιούνται και από κάθε ομάδα επιλέγεται το ενδιάμεσο καρέ ως χαρακτηριστικό.

Η χρήση του χρώματος προτιμήθηκε στις πρώτες τεχνικές, καθώς αφενός η εξαγωγή των περιγραφών ήταν απλή και γρήγορη και αφετέρου δεν παρουσιάζει ευαισθησία στον προσανατολισμό της εικόνας και στο θόρυβο. Το βασικό τους, ωστόσο, πρόβλημα έχει να κάνει με το ότι απαιτούν τον προσδιορισμό ενός κατωφλίου από το χρήστη, κάτι που δεν μπορεί να γίνει πάντα με προφανή τρόπο και απαιτεί διαφορετικές τιμές ανάλογα με τις συλλογές βίντεο που θα χρησιμοποιηθεί. Φυσικά αυτό δεν εγγυάται ότι οι αλγόριθμοι θα δουλεύουν πάντα ικανοποιητικά και σε όλα τα βίντεο. Έτσι, το επόμενο και φυσικό βήμα ήταν η εισαγωγή των χαρακτηριστικών κίνησης.

Οι Lagendijk et al. [107] δουλεύουν πάνω σε πλάνα από ακολουθίες βίντεο και χρησιμοποιούν τις διαφορές μεταξύ των εικονοστοιχείων προκειμένου να αποφασίσουν από ποια σημεία της ακολουθίας πρέπει να εξαχθούν τα χαρακτηριστικά καρέ. Δείχνουν πώς μπορεί να επιλεχθούν τα καρέ αυτά, βασιζόμενοι σε μια βελτιστοποίηση μέσω μέτρων τα οποία εκφράζουν την ενέργεια των "πράξεων" στο βίντεο, χωρίς όμως να ορίζουν τα μέτρα αυτά. Ο Wolf [231] υπολόγισε την οπτική ροή ανάμεσα σε διαδοχικά καρέ και πρότεινε την επιλογή ως χαρακτηριστικών των καρέ εκείνων που βρίσκονται στα τοπικά ελάχιστά της. Όπως παρατήρησε πειραματικά, αλλά όπως φαίνεται και διαισθητικά, δεν είναι δυνατόν να περιμένει κανείς από μια τεχνική εξαγωγής χαρακτηριστικών καρέ να αποδίδει παντού. Ωστόσο, ο αλγόριθμός του δεν υπέθεσε ότι κάθε πλάνο θα αναπαρίσταται από ένα καρέ, επιτρέποντας την περιγραφή του από περισσότερα, εφόσον αυτό προέκυπτε από την ανάλυση. Οι Divakaran et al. [58] αξιοποιούν τον Περιγραφέα Δραστηριότητας Κίνησης του MPEG-7. Σε κάθε πλάνο, όταν η συσσωρευμένη δραστηριότητα της κίνησης φτάσει σε κάποιο κατώφλι, επιλέγεται το μεσαίο καρέ από το τρέχον κομμάτι του βίντεο. Επίσης, συνδέουν εμπειρικά την κίνηση με τον αριθμό των χαρακτηριστικών καρέ που απαιτούνται για την περιγραφή του και έτσι ο αριθμός των χαρακτηριστικών καρέ προσαρμόζεται δυναμικά στο περιεχόμενο του βίντεο και δεν απαιτείται παρέμβαση του χρήστη, ούτε προαποφασίζεται κάποιος αριθμός που ενδεχομένως να είναι πλεονάζων σε κάποια βίντεο, ελλιπής σε άλλα.

Οι Fauvet et al. [65] προτείνουν μια μέθοδο που εκμεταλλεύεται τον υπολογισμό της κυρίαρχης κίνησης της εικόνας, η οποία υποθέτουν ότι οφείλεται σε κίνηση της κάμερας και βασίζεται σε γεωμετρικές ιδιότητες που σχετίζονται με τη συνεισφορά του κάθε καρέ. Εφαρμόζουν την τεχνική τους σε αθλητικά βίντεο και σε ντοκυμαντέρ. Οι Hanjalic et al. [81] χρησιμοποιούν τις διαφορές ανάμεσα σε διαδοχικά καρέ των βίντεο. Ορίζουν μια συνάρτηση που υπολογίζει συσσωρευτικά το περιεχόμενο των καρέ εώς ότου αυτό να υπερβεί κάποιο κατώφλι. Η τεχνική τους προσπαθεί να κατανείμει ομοιόμορφα τα καρέ και περιορίζεται μόνο από ένα μέγιστο αριθμό από καρέ

που προκαθορίζεται. Οι Liu et al. [124] κατασκεύασαν ένα μοντέλο με το οποίο περιγράφουν την κίνηση του βίντεο, το οποίο σχετίζεται με την ενέργεια που "λαμβάνει" ο χρήστης από το οπτικό περιεχόμενο. Ο απώτερος σκοπός τους ήταν να μπορούν να εξάγουν χαρακτηριστικά καρέ χωρίς να προκαθορίζουν τον αριθμό τους και χωρίς να χρειάζεται να ορίσουν κατώφλια. Τέλος, οι Chang et al. [41] προτείνουν μια τεχνική που δημιουργεί μια ιεραρχική αναπαράσταση του βίντεο, με βάση τα χαρακτηριστικά χρώματος και από αυτή εξάγουν τα χαρακτηριστικά καρέ.

5.3.3 Τεχνικές που βασίζονται σε τμήματα του βίντεο

Παρότι σε γενικές γραμμές οι τεχνικές που παρουσιάστηκαν στην Ενότητα 5.3.2 για την εξαγωγή χαρακτηριστικών καρέ από πλάνα ακολουθιών βίντεο κρίνεται ότι παρέχουν ικανοποιητικά αποτελέσματα, στην περίπτωση που απαιτείται η δημιουργία περιλήψεων από μεγάλες ακολουθίες βίντεο, όπως π.χ. από μια κινηματογραφική ταινία, ο αριθμός των πλάνων δύναται να είναι πολύ μεγάλος. Κάτι τέτοιο μπορεί να καταστήσει προβληματικές τις περιλήψεις που εξάγονται με τους παραπάνω τρόπους, καθώς αυτές καταλήγουν σε εκατοντάδες ή χιλιάδες χαρακτηριστικά καρέ. Έτσι, εάν κάποιος χρήστης επιθυμεί να αποκτήσει γρήγορα μια ιδέα για το περιεχόμενο του βίντεο, δε θα μπορέσει να το επιτύχει αν πρέπει να παρατηρήσει μεγάλο αριθμό από καρέ. Για το λόγο αυτό έχουν αναπτυχθεί τεχνικές που εφαρμόζονται σε μεγαλύτερες ακολουθίες βίντεο από ότι είναι τα πλάνα.

Οι Uchihachi et al. [217] πρότειναν μια τεχνική η οποία είχε ως στόχο τη δημιουργία περιλήψεων που θα μοιάζουν με εικονογραφημένες ιστορίες. Μέσα από μια διαδικασία ιεραρχικής συσταδοποίησης, χωρίζουν το βίντεο σε τμήματα και για κάθε τμήμα υπολογίζουν ένα μέτρο σημαντικότητας. Τα τμήματα του βίντεο που έχουν μέτρο μικρότερο από ένα προκαθορισμένο κατώφλι αγνοούνται. Από τα υπόλοιπα εξάγονται χαρακτηριστικά καρέ, το μέγεθος των οποίων είναι ανάλογο με το μέτρο σημαντικότητας. Το μέτρο αυτό εξαρτάται από το πόσο σπάνια ή συχνά είναι τα καρέ των τμημάτων συνολικά σε όλο το βίντεο, με βάση τα οπτικά τους χαρακτηριστικά. Οι περιλήψεις που δημιουργούν τονίζουν τα πιο σημαντικά καρέ και δίνουν έτσι στο χρήστη τη δυνατότητα να εστιάσει μόνο σε αυτά σε περίπτωση που επιθυμεί μια γρήγορη ενημέρωση για το βίντεο, ή να μελετήσει και τα λιγότερο σημαντικά, σε περίπτωση που επιθυμεί μια πιο λεπτομερή ενημέρωση. Παρομοίως κινείται και η τεχνική των Girgensohn και Boreczky [73], σύμφωνα με την οποία αρχικά επιλέγονται τα N περισσότερο ανόμοια καρέ από το βίντεο. Αυτό επιτυγχάνεται μέσω ενός μεγάλου αριθμού από συγκρίσεις, ανάμεσα στα καρέ του βίντεο, οπότε και τελικά επιλέγονται τα ζεύγη με τις μεγαλύτερες διαφορές. Στη συνέχεια και με τη χρήση ενός ιεραρχικού αλγορίθμου συσταδοποίησης, δημιουργούνται M συστάδες, αφού εφαρμοστούν μερικοί χρονικοί περιορισμοί, οι οποίοι σκοπό έχουν να κατανείμουν τα καρέ όσο το δυνατόν πιο ομοιόμορφα στο χρόνο. Από κάθε συστάδα, τελικά, επιλέγεται το καρέ που βρίσκεται πιο κοντά στο κέντρο της ως χαρακτηριστικό. Οι Sun και Kankanhalli [204] επιλέγουν να εργαστούν σε ολόκληρο το βίντεο, χωρίς, δηλαδή να απαιτούν το χωρισμό του σε πλάνα. Τα καρέ των βίντεο αντιμετωπίζονται σαν σημεία σε ένα πολυδιάστατο χώρο, ο οποίος κατασκευάζεται με βάση τα χαρακτηριστικά χρώματος, υφής, σχήματος και κίνησης. Ένα βίντεο χωρίζεται πρώτα σε μεγάλες ακολουθίες ομοιόμορφα και στη συνέχεια, για κάθε μία από αυτές υπολογίζεται η διαφορά στα χαρακτηριστικά ανάμεσα στο πρώτο και το τελευταίο καρέ. Στη συνέχεια από τις ακολουθίες με μικρές διαφορές επιλέγονται το πρώτο και το

τελευταίο καρέ, ενώ από τις υπόλοιπες όλα τα καρέ. Η διαδικασία αυτή μπορεί να συνεχιστεί με πολλές επαναλήψεις θεωρώντας κάθε ακολουθία από τη δεύτερη κατηγορία σαν ένα βίντεο και εφαρμόζοντας ξανά τον αλγόριθμο. Περατώνεται όταν τελικά επιλεχθεί ο επιθυμητός αριθμός από καρέ. Φυσικά, το εμφανές μειονέκτημα της τεχνικής αυτής έχει να κάνει με την αδυναμία της τεχνικής να χωρίσει το αρχικό βίντεο σε ακολουθίες με κάποιο ευφυή τρόπο που να εκμεταλλεύεται τα χαρακτηριστικά των καρέ. Ο Ratakonda [176] κατασκεύασε ιεραρχικές περιλήψεις από βίντεο. Έτσι, ένας χρήστης που επιθυμεί να δει την περίληψη ενός βίντεο, θα μπορεί να δει καταρχήν τα πιο σημαντικά χαρακτηριστικά καρέ (σύμφωνα με το σύστημα) και έπειτα, επιλέγοντας αυτά που τον ενδιαφέρουν, να δει πιο ολοκληρωμένη περίληψη του αντίστοιχου μέρους του βίντεο. Για να το επιτύχει αυτό, χρησιμοποίησε τον αλγόριθμο K-means και τον εφάρμοσε διαδοχικά, με μόνο περιορισμό ότι δύο καρέ που απέχουν χρονικά δε θα μπορούν να ενωθούν στην ίδια συστάδα, ακόμη και αν τα οπτικά τους χαρακτηριστικά είναι πολύ κοντά. Τέλος, η τεχνική των Avrithis et al. [5] μετατρέπει την αναπαράσταση των πλάνων από τα καρέ που αυτά περιέχουν, στα χαρακτηριστικά χαμηλού επιπέδου τους. Αρχικά, τα καρέ χωρίζονται σε περιοχές με τη χρήση ενός αλγορίθμου κατάτμησης και έπειτα, από κάθε μία από τις περιοχές αυτές εξάγονται χαρακτηριστικά χαμηλού επιπέδου, τα οποία και συγχωνεύονται σε ένα ασαφές πολυδιάστατο ιστόγραμμα. Η εξαγωγή των χαρακτηριστικών καρέ γίνεται με δύο μεθόδους. Η πρώτη βασίζεται σε παρατηρήσεις τις τροχιάς του διανύσματος χαρακτηριστικών ως προς το χρόνο, ενώ η δεύτερη προσπαθεί να ελαχιστοποιήσει ένα κριτήριο συσχέτισης ανάμεσα σε καρέ, κάτι που γίνεται με λογαριθμική αναζήτηση ή με εφαρμογή ενός γενετικού αλγορίθμου.

5.3.4 Λοιπές Τεχνικές

Οι Shahraray και Gibbon [184] πρότειναν μια τεχνική για τη δημιουργία περιλήψεων από βίντεο τα οποία συνοδεύονται από υπότιτλους. Αρχικά επιλέγεται ένας αριθμός από χαρακτηριστικά καρέ και στη συνέχεια γίνεται επεξεργασία των υποτίτλων, από την οποία προκύπτει επιπρόσθετη πληροφορία σχετικά με το περιεχόμενό του. Τελικά τα καρέ συγχρονίζονται με τους υπότιτλους και κατασκευάζονται αναπαραστάσεις του περιεχομένου που συνδυάζουν κείμενο και εικόνα. Οι Taniguchi et al. [208] προσπαθούν να ξεπεράσουν το πρόβλημα της ελλιπούς αναπαράστασης του οπτικού περιεχομένου από έναν μικρό αριθμό χαρακτηριστικών καρέ, κατασκευάζοντας μικρές πανοραμικές εικόνες. Οι Yu et al. [237] αρχικά εξάγουν ιστογράμματα χρώματος από τα χαρακτηριστικά καρέ. Στη συνέχεια αναπαριστούν το βίντεο σε πολλά επίπεδα, με βάση τις πρωτεύουσες συνιστώσες των χαρακτηριστικών. Για την τελική αναπαράσταση χρησιμοποιούν μια μέδοδο που βασίζεται στην PCA και σε πυρήνες. Τα χαρακτηριστικά καρέ εξάγονται έπειτα από ασαφή συσταδοποίηση στο χώρο που έχουν οδηγηθεί τα χαρακτηριστικά. Η εξαγωγή αντικειμένων από το βίντεο χρησιμοποιήθηκε από τους Kim και Huang [101] για να βοηθήσει στην εξαγωγή των καρέ. Τα αντικείμενα που εξάγονται, περιγράφονται με τη χρήση περιγραφέων σχήματος. Τα καρέ που περιέχουν συγκεκριμένα αντικείμενα επιλέγονται ως χαρακτηριστικά. Οι Panagiotakis et al. [164] πρότειναν 4 τεχνικές για την εξαγωγή χαρακτηριστικών καρέ, οι οποίες εξάγουν τον ίδιο αριθμό από καρέ και βασίζονται σε έναν αλγόριθμο υπολογιστικής γεωμετρίας. Αν και χρησιμοποιούν τον περιγραφέα διάταξης χρώματος του MPEG-7, η τεχνική τους μπορεί να εφαρμοστεί με οποιοδήποτε περιγραφέα. Τέλος, οι Gibson et al. [72] οδήγησαν τα οπτικά χαρακτηριστικά των καρέ σε έναν

Σχήμα 5.1: *Απεικόνιση των χαρακτηριστικών καρέ με διαφορετικά μεγέθη, ανάλογα με τη σημαντικότητά τους. Το σχήμα προέρχεται από τη συμμετοχή της ομάδας του COST292 στο TRECVID 2006 [34].*

χώρο με βάση τα ιδιοδιανύσματά τους και στη συνέχεια χρησιμοποιούν γκαουσιανά μοντέλα για τη συσταδοποίηση τους. Έτσι, τα χαρακτηριστικά καρέ εξάγονται ως αυτά που βρίσκονται κοντύτερα στα κέντρα των συστάδων.

5.3.5 Οπτικοποίηση Αποτελεσμάτων

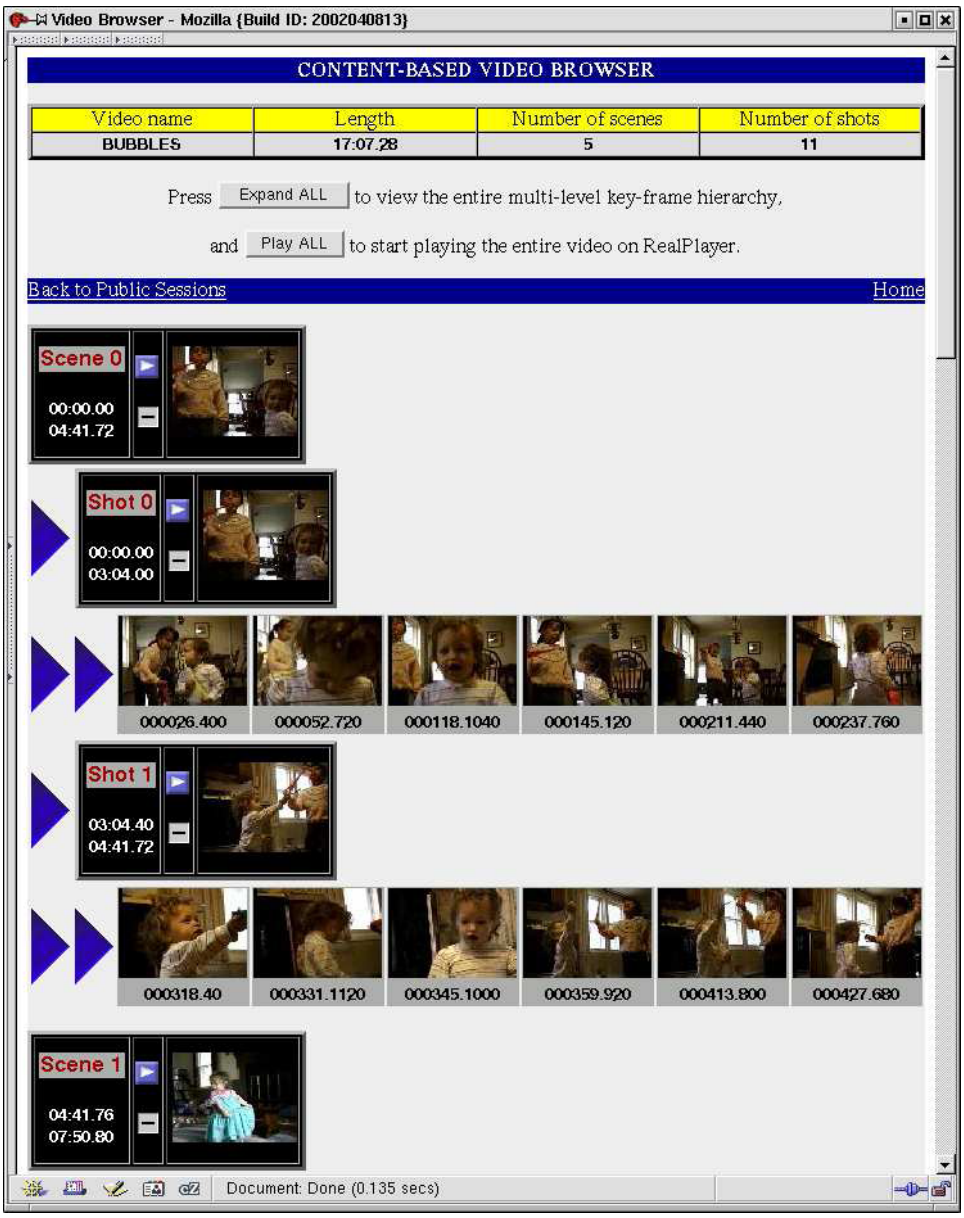
Ο πιο απλός τρόπος παρουσίασης των αποτελεσμάτων της περίληψης είναι με την παράθεση όλων των χαρακτηριστικών καρέ από τα οποία αυτή αποτελείται. Ένα τέτοιο παράδειγμα που δημιουργήθηκε από τον αλγόριθμο που προτείνεται σε αυτό το Κεφάλαιο απεικονίζεται στο Σχήμα 5.7. Μια παραλλαγή της τεχνικής αυτής παρουσιάζει τα καρέ με διαφορετικό μέγεθος, ανάλογα με τη "σημαντικότητά" τους. Ως σημαντικό, μπορεί να θεωρηθεί για παράδειγμα ένα καρέ που το οπτικό του περιεχόμενο συναντάται πιο σπάνια ανάμεσα στα καρέ του βίντεο ή του πλάνου από το οποίο έχει εξαχθεί. Ένα τέτοιο παράδειγμα απεικονίζεται στο Σχήμα 5.1. Μια παραλλαγή αυτού του τρόπου παρουσίασης απεικονίζεται στο Σχήμα 5.2, όπου τα χαρακτηριστικά καρέ σχηματίζουν ένα "μονοπάτι", το μέγεθός τους είναι ανάλογο της σημαντικότητάς τους και ο κενός χώρος ανάμεσά τους προορίζεται για σχολιασμούς. Μια ιεραρχική αναπαράσταση των χαρακτηριστικών καρέ απεικονίζεται στο Σχήμα 5.3. Στο χρήστη παρουσιάζεται το σημαντικότερο χαρακτηριστικό καρέ από ένα πλάνο και αν αυτός το επιθυμεί, μπορεί να περιηγηθεί και σε περισσότερα χαρακτηριστικά καρέ. Στη βιβλιογραφία συναντώνται αρκετοί ακόμη τρόποι παρουσίασης των καρέ, όπως για παράδειγμα τα μωσαϊκά, όπου ενώνονται περιοχές από πολλά διαδοχικά καρέ σε μια εικόνα, με σκοπό να περιγράψουν ένα γεγονός. Ωστόσο, η οπτικοποίηση των αποτελεσμάτων δεν αποτελεί κομμάτι της έρευνας του Κεφαλαίου αυτού και έτσι η παρούσα Ενότητα περιορίζεται στους πιο συνήθεις τρόπους απεικόνισης των περιλήψεων.

Σχήμα 5.2: *Απεικόνιση των χαρακτηριστικών καρέ έτσι ώστε αυτά να σχηματίζουν ένα "μονοπάτι", το μέγεθός τους είναι ανάλογο της σημαντικότητάς τους και ο κενός χώρος ανάμεσά τους προορίζεται για σχολιασμούς. Το Σχήμα προέρχεται από την εργασία των Uchihashi et al. [217].*

5.4 Εξαγωγή Χαρακτηριστικών Καρέ από Ακολουθίες Εικόνων

Στην Ενότητα αυτή, η αναπαράσταση των οπτικών χαρακτηριστικών των εικόνων με τη χρήση του οπτικού θησαυρού χρησιμοποιείται για την εξαγωγή χαρακτηριστικών καρέ από ακολουθίες βίντεο. Στην πλειοψηφία των αλγορίθμων δημιουργίας περιλήψεων, συγκεκριμένα χαρακτηριστικά χαμηλού επιπέδου εξάγονται και με βάση αυτά δημιουργείται μια περιγραφή για κάθε καρέ της ακολουθίας. Ο σκοπός είναι να εξαχθεί ένας σχετικά μικρός αριθμός από αντιπροσωπευτικά καρέ, τα οποία θα περιλαμβάνουν το οπτικό περιεχόμενο ενός πλάνου. Ο αλγόριθμος που θα παρουσιαστεί εντάσσεται στην κατηγορία των *Image Storyboards* προσεγγίσεων και όπως η πλειοψηφία των τεχνικών της κατηγορίας αυτής, μπορεί να χωριστεί στα παρακάτω βήματα, όπως σημειώνουν οι Sundaram και Chang [205]:

- Καθορισμός ενός κατάλληλου χώρου χαρακτηριστικών για την αναπαράσταση των καρέ.
- Συσταδοποίηση της ακολουθίας των εικόνων στο χώρο χαρακτηριστικών.
- Υπολογισμός ενός μέτρου "σημαντικότητας" το οποίο θα αποτελέσει κριτήριο για την επιλογή των χαρακτηριστικών καρέ.



Σχήμα 5.3: *Ιεραρχική απεικόνιση των χαρακτηριστικών καρέ, η οποία επιτρέπει στο χρήστη να περιηγείται στα πλάνα ενός βίντεο είτε παρατηρώντας το σημαντικότερο χαρακτηριστικό καρέ τους, είτε περισσότερα καρέ που έχουν εξαχθεί από το ίδιο πλάνο. Το Σχήμα προέρχεται από την εργασία των Guillemot et al. [79].*

5.4.1 Εξαγωγή χαρακτηριστικών χαμηλού επιπέδου

Για την εξαγωγή χαρακτηριστικών χαμηλού επιπέδου επιλέχθηκαν και στην περίπτωση αυτή οι περιγραφείς χρώματος και υφής του προτύπου MPEG-7. Η εξαγωγή τους γίνεται τοπικά, από περιοχές της εικόνας, οι οποίες έχουν προκύψει έπειτα από κατάτμησή της με τη χρήση του αλγορίθμου των Avrithis et al. [5]. Οι περιγραφείς που εξάγονται στην περίπτωση αυτή είναι οι ίδιοι με αυτούς του Κεφαλαίου 4. Για τα χαρακτηριστικά χρώματος χρησιμοποιούνται ο Κλιμακωτός Περιγραφέας Χρώματος, ο Περιγραφέας Κύριων Χρωμάτων, ο Περιγραφέας Δομής Χρώματος και ο Περιγραφέας Διάταξης Χρώματος. Για τα χαρακτηριστικά της υφής χρησιμοποιείται μόνο ο Περιγραφέας Ιστογράμματος Ακμών.

5.4.2 Κατασκευή Τοπικού Οπτικού Θησαυρού Περιοχών

Μετά την εξαγωγή των περιγραφέων χρώματος και υφής από περιοχές της εικόνας, κατασκευάζεται ένας *τοπικός* οπτικός θησαυρός περιοχών. Ο χαρακτηρισμός αυτός χρησιμοποιείται για να περιγράψει το γεγονός ότι η κατασκευή του θησαυρού γίνεται από τις περιοχές που εξάγονται με κατάτμηση στα καρέ ενός πλάνου, ή από ολόκληρη την ακολουθία του βίντεο. Και στην περίπτωση αυτή χρησιμοποιείται η διαδικασία που περιγράφεται στην Ενότητα 4.6.3. Έτσι, ένα καρέ p_i της ακολουθίας βίντεο θα αναπαρίσταται στο χώρο χαρακτηριστικών με ένα διάνυσμα αναπαράστασης m_i όπως αυτό περιγράφεται από την (4.9) και στην περίπτωση αυτή θα είναι

$$m_i = \Big[m_i(1), m_i(2), \ldots, m_i(j), \ldots, m_i(N_T)\Big] \,, \tag{5.1}$$

όπου N_T είναι το μέγεθος του τοπικού οπτικού θησαυρού. Θα πρέπει να δοθεί έμφαση στο γεγονός ότι η αναπαράσταση που χρησιμοποιείται για το οπτικό περιεχόμενο των καρέ και βασίζεται σε τοπικά χαρακτηριστικά είναι πιο κοντά σε μια σημασιολογική περιγραφή παρά σε μια απλή οπτική. Αυτό γιατί είναι βασισμένη στους τύπους περιοχής που αποτελούν τον οπτικό θησαυρό και όπως έχει αναφερθεί στο Κεφάλαιο 4, αυτοί ενσωματώνουν πληροφορία υψηλότερου επιπέδου από τους απλούς περιγραφείς, χωρίς βέβαια να φτάνουν στο επίπεδο των εννοιών.

5.4.3 Επιλογή Χαρακτηριστικών Καρέ

Όπως έχει διαφανεί εώς τώρα, το σημασιολογικό περιεχόμενο ενός καρέ από βίντεο μοντελοποιείται με το συνδυασμό διαφόρων MPEG-7 περιγραφέων χρώματος και υφής που εξάγονται από περιοχές των καρέ. Στη συνέχεια με χρήση ενός οπτικού θησαυρού που κατασκευάζεται τοπικά, υπολογίζονται οι αποστάσεις τους από τους τύπους περιοχής και σχηματίζεται κατάλληλο διάνυσμα αναπαράστασης. Το πρώτο που θα πρέπει να καθοριστεί έπειτα είναι μια κατάλληλη μετρική, μέ χρήση της οποίας θα συγκρίνονται τα διανύσματα αναπαράστασης από δύο καρέ.

Μια από τις πιο δημοφιλείς μετρικές που χρησιμοποιούνται για σύγκριση τέτοιων περιγραφών που έχουν τη μορφή διανύσματος και περιέχουν σημασιολογική πληροφορία είναι η *συνάρτηση ομοιότητας συνημιτόνου* (cosine similarity function). Συγκεκριμένα, αν m_1 και m_2 είναι δύο διανύσματα αναπαράστασης που αντιστοιχουν σε δύο χαρακτηριστικά καρέ p_1 και p_2, τότε η απόστασή τους $D_{cos}(m_1, m_2)$ υπολογίζεται ως

$$D_{cos}(m_1, m_2) = \arccos \frac{m_1^T m_2}{||m_1|| \cdot ||m_2||} \tag{5.2}$$

όπου $m_1^T m_2$ είναι το εσωτερικό γινόμενο μεταξύ των δύο διανυσμάτων αναπαράστασης και με $||\cdot||$ συμβολίζεται το μέτρο ενός διανύσματος.

Το μέτρο αυτό χρησιμοποιείται στη βιβλιογραφία για τον προσδιορισμό της ομοιότητας μεταξύ δύο n-διάστατων διανυσμάτων, υπολογίζοντας το συνημίτονο της μεταξύ τους γωνίας. Πολύ συχνή είναι η εφαρμογή του σε προβλήματα σύγκρισης εγγράφων για εξόρυξη κειμένου. Σε τέτοια προβλήματα, τα δύο διανύσματα που συγκρίνονται είναι συνήθως τα διανύσματα συχνότητας όρων (term frequency vectors)

των εγγράφων. Η χρήση του μέτρου αυτού μπορεί να θεωρηθεί σαν μια μέθοδος κανονικοποίησης του μεγέθους του εγγράφου κατά τη σύγκριση.

Το αποτέλεσμα της σύγκρισης με το μέτρο αυτό είναι μια τιμή από -1 (που υποννοεί ότι τα δύο έγγραφα είναι τελείως διαφορετικά) μέχρι 1 (που υποννοεί ότι τα δύο έγγραφα είναι ακριβώς ίδια). Τιμή ίση με 0 σημαίνει ότι τα δύο έγγραφα είναι ανεξάρτητα και οι ενδιάμεσες τιμές δείχνουν ομοιότητα ή ανομοιότητα. Στην παρούσα περίπτωση, ωστόσο, όπως συμβαίνει γενικά σε προβλήματα ανάκτησης πληροφορίας, η τιμή του μέτρου είναι από 0 εώς 1, μιας και οι συχνότητες των όρων δεν μπορούν να λάβουν αρνητικές τιμές. Έτσι η γωνία μεταξύ δύο διανυσμάτων δεν μπορεί να είναι μεγαλύτερη από 90°.

Από τα παραπάνω γίνεται εύκολα σαφές ότι η συνάρτηση ομοιότητας συνημιτόνου μπορεί να χρησιμοποιηθεί στο παρόν πρόβλημα, καθώς υπάρχει προφανής αντιστοιχία μεταξύ του διανύσματος αναπαράστασης μιας εικόνας και του διανύσματος συχνότητας όρων ενός εγγράφου. Πιο συγκεκριμένα και όπως εξηγήθηκε στην Ενότητα 4.6.3, το διάνυσμα αναπαράστασης αποτελείται από τους βαθμούς βεβαιότητας των περιοχών της εικόνας, ως προς όλες τις λέξεις ενός οπτικού θησαυρού. Οι λέξεις αυτές αντιστοιχούν σε όρους, και οι βαθμοί βεβαιότητας μπορούν να σχετιστούν με τη συχνότητα εμφάνισης.

Το πρώτο βήμα του αλγορίθμου κατασκευάζει τα διανύσματα αναπαράστασης είτε από όλα τα καρέ που αποτελούν ένα πλάνο είτε από ένα υποσύνολό τους (π.χ. μπορεί να επιλεχθεί ένας χαμηλότερος ρυθμός όπως 10 καρέ/sec). Έστω

$$\mathcal{S} = \{s_i\},\ i = 1, 2, \ldots, N_{\mathcal{S}} \tag{5.3}$$

το σύνολο των καρέ s_i του πλάνου $\mathcal{S}$. Στην περίπτωση που επιλεχθεί ένα υποσύνολο των καρέ $\mathcal{S}'$, με κάποιο ρυθμό δειγματοληψίας r καρέ/sec, τότε η περίληψη θα κατασκευαστεί με βάση το σύνολο

$$\mathcal{S}' = \{s_{ri}\},\ i = 1, 2, \ldots, \lfloor N_{\mathcal{S}}/r \rfloor\ . \tag{5.4}$$

Με βάση τα καρέ αυτά κατασκευάζεται ο οπτικός θησαυρός, με χρήση της τεχνικής της Ενότητας 4.6.3. Στη συνέχεια εφαρμόζεται ο αλγόριθμος *αφαιρετικής συσταδοποίησης* (*subtractive clustering*) του Chiu [46], με σκοπό να ομαδοποιηθούν τα καρέ σε συστάδες από σημασιολογικά παρόμοια καρέ, όπως περιγράφεται στην (5.5). Η μέθοδος αυτή προτιμήθηκε σε σχέση με τον αλγόριθμο K-means που χρησιμοποιήθηκε στο Κεφάλαιο 4 γιατί υπολογίζει ο ίδιος τον αριθμό των συστάδων N_S, καθώς στο συγκεκριμένο πρόβλημα δεν είναι πάντα εύκολο να καθορισθεί αυτός εμπειρικά και γενικά δεν μπορεί να είναι ο ίδιος για κάθε πλάνο. Με αυτόν τον τρόπο, τελικά, επιλέγεται ένα υποσύνολο S των καρέ μέσα στο πλάνο, που ορίζεται ως

$$S = \Big\{w'_i, \quad i = 1, 2, \ldots N_S\Big\}, \quad w'_i \subset K\ , \tag{5.5}$$

όπου ως w_i συμβολίζονται οι συστάδες που δημιουργήθηκαν. Τα χαρακτηριστικά καρέ θα επιλεχθούν τελικά ανάμεσα στο σύνολο S που προέκυψε μετά την εφαρμογή του αλγορίθμου συσταδοποίησης.

Θα πρέπει στο σημείο αυτό να δοθεί έμφαση στο ότι κάποιες εφαρμογές, όπως η ανίχνευση εννοιών υψηλού επιπέδου σε ακολουθίες βίντεο πολλές φορές απαιτούν περισσότερα από ένα καρέ από κάθε πλάνο, προκειμένου να εφαρμοστούν επιτυχώς. Αυτό συμβαίνει γιατί τις περισσότερες φορές η έννοια που αναζητείται δεν είναι παρούσα σε όλα τα καρέ του πλάνου, με συνέπεια πολλές φορές να απουσιάζει από το

καρέ που έχει επιλεχθεί ως χαρακτηριστικό. Όταν η ποσότητα των βίντεο προς ανάλυση είναι πολύ μεγάλη, τότε η εφαρμογή αλγορίθμων ανίχνευσης είναι πολύ αργή στην περίπτωση που εφαρμοστούν σε κάθε ένα από τα καρέ. Το επιθυμητό θα ήταν να εξαχθούν ως χαρακτηριστικά όλα τα καρέ στα οποία περιέχονται οι έννοιες υψηλού επιπέδου, αλλά παράλληλα, ο αριθμός τους να παραμείνει σε χαμηλά επίπεδα, για να μπορεί να γίνει γρήγορη εφαρμογή αλγορίθμων ανίχνευσης.

Επιπλέον, οι εφαρμογές δημιουργίας περιλήψεων και ανάκτησης βίντεο μπορούν να λειτουργήσουν πιο αποδοτικά όταν ένα πλάνο ή μια ακολουθία βίντεο περιγράφεται από έναν ικανό αριθμό από καρέ αντί ενός μοναδικού χαρακτηριστικού καρέ. Έτσι ο πιθανός χρήστης των εφαρμογών αυτών μπορεί να αντιληφθεί καλύτερα το περιεχόμενο του βίντεο, κάτι που γενικά δε συμβαίνει στην περίπτωση του μοναδικού καρέ, το οποίο μπορεί τελικά να μην χαρακτηρίζει καθόλου το περιεχόμενο του βίντεο.

Η μεθοδολογία που θα παρουσιαστεί σε αυτήν την Ενότητα, θα χρησιμοποιηθεί για την επιλογή ενός σχετικά μικρού αριθμού από αντιπροσωπευτικά καρέ μέσα σε ένα πλάνο. Ο ρόλος των καρέ αυτών είναι αφενός να παρέχουν μια περίληψη του οπτικού αλλά και του σημασιολογικού περιεχομένου και παράλληλα να εφαρμοστούν σε αυτά αλγόριθμοι ανίχνευσης εννοιών, αντί να εφαρμοστούν σε ολόκληρο το βίντεο. Για να επιτευχθεί αυτό, γίνεται επιλογή των καρέ εκείνων που τα διανύσματα αναπαράστασής τους βρίσκονται κοντά κοντύτερα στα κέντρα των συστάδων που δημιούργησε ο αλγόριθμος.

Το κέντρο $z(w_i')$ της συστάδας w_i στο χώρο των διανυσμάτων αναπαράστασης προσδιορίζεται ως

$$z(w_i') = \frac{1}{|w_i'|} \sum_{k \in w_i'} m_k \ , \tag{5.6}$$

ενώ τα διανύσματα αναπαράστασης $m(w_i')$ που βρίσκονται κοντύτερα σε αυτά ορίζονται ως

$$m(w_i') = m_k \ , \tag{5.7}$$

όπου

$$k = \arg\min_{k \in w_i'} \Big\{ D_{cos}\big(m_k, z(w_i')\big) \Big\} \ . \tag{5.8}$$

Το σύνολο M_Z όλων των διανυσμάτων αναπαράστασης που αντιστοιχούν στα καρέ που θα επιλεχθούν ορίζεται ως

$$M_z = \{m(w_i')\}, \ i = 1 \ldots N_S \tag{5.9}$$

και έστω K_z το σύνολο των χαρακτηριστικών αυτών καρέ. Το τελευταίο επιτυγχάνεται με την επιλογή του κατάλληλου αριθμού καρέ από το υποσύνολο K_z των προεπιλεγμένων καρέ, τα οποία και περιέχουν όσο το δυνατόν περισσότερη πληροφορία από όλους τους τύπους περιοχής του τοπικά κατασκευασμένου οπτικού θησαυρού.

Η επιλογή των αρχικών (πιο αντιπροσωπευτικών) καρέ ξεκινά με την εύρεση του τύπου περιοχής με το μεγαλύτερο αριθμό από συνώνυμα, δηλαδή αυτόν που η συστάδα που του αντιστοιχεί έχει το μεγαλύτερο πλήθος. Το επόμενο βήμα είναι η επιλογή του διανύσματος αναπαράστασης από το σύνολο M_z, το οποίο έχει το μεγαλύτερο ποσοστό πεποίθησης (δηλαδή τη μικρότερη απόσταση) προς τον αντίστοιχο τύπο περιοχής. Άρα, αν ο τύπος περιοχής που επιλέχτηκε ήταν ο i-οστός του οπτικού θησαυρού, τότε το επιλεχθέν διάνυσμα αναπαράστασης ανάμεσα στο σύνολο M_z είναι εκείνο για το οποίο ελαχιστοποιείται η τιμή του i-οστού στοιχείου. Για

το πιο αντιπροσωπευτικό καρέ και για κάθε επόμενο, ελέγχεται ποιοι τύποι περιοχής περιέχονται σε αυτό. Ένα καρέ θεωρείται ότι περιέχει έναν τύπο περιοχής, αν η απόσταση για αυτόν τον τύπο περιοχής είναι κάτω από ένα προεπιλεγμένο κατώφλι t_s. Έστω $R_{s,i}$ και $M_{s,i}$ τα σύνολα των επιλεγμένων τύπων περιοχής και τα διανύσματα αναπαράστασης, στην i-οστή επανάληψη του αλγορίθμου, αντίστοιχα. Το $R_{s,i}$ περιέχει τους δείκτες των τύπων περιοχής. Ο τύπος περιοχής που επιλέχτηκε και το διάνυσμα αναπαράστασης του επιλεχθέντος καρέ προστίθενται στα αρχικά άδεια σύνολα $R_{s,0}$ και $M_{s,0}$, αντίστοιχα, έτσι ώστε να μην είναι δυνατόν να ξαναεπιλεχθούν. Θα πρέπει να σημειωθεί ότι και όλοι οι υπόλοιποι τύποι περιοχής που περιέχονται στο καρέ που επιλέχτηκε προστίθενται στο $R_{s,i}$ κατά την i-οστή επανάληψη.

Όλα τα υπόλοιπα χαρακτηριστικά καρέ εξάγονται χρησιμοποιώντας τον αλγόριθμο που περιγράφηκε. Το διάνυμα αναπαράστασης με το μέγιστο βαθμό βεβαιότητας για τον τύπο περιοχής με το μεγαλύτερο αριθμό από συνώνυμα, παρλείποντας τους τύπους περιοχής που περιείχαν τα ήδη επιλεγμένα καρέ. Η διαδικασία σταματά όταν όλοι οι τύποι περιοχής του θησαυρού περιέχονται στα επιλεγμένα καρέ, δηλαδή: $|R_{s,i}| = N_T$.

Είναι προφανές ότι ο αριθμός των εξαχθέντων χαρακτηριστικών καρέ για την αναπαράσταση του πλάνου δεν μπορεί να είναι μεγαλύτερος από N_T, το οποίο και αποτελεί τον συνολικό αριθμό των τύπων περιοχής. Το σύνολο $R_{s,k}$ στην k-οστή επανάληψη προσδιορίζεται ως

$$R_{s,k} = \{r_i\} \cup \{\arg_j\{m'_i(j) < t_s\}\},\ i = 0 \dots k-1,\ R_{s,0} = \emptyset\ , \tag{5.10}$$

ενώ ο τύπος περιοχής κάθε φορά προσδιορίζεται ως

$$r_k = \arg\max_i(|w_i|),\ i \notin R_{s,k}\ . \tag{5.11}$$

Επίσης, το σύνολο των επιλεχθέντων διανυσμάτων αναπαράστασης στην k-οστή επανάληψη, $M_{s,k}$ προσδιορίζεται ως

$$M_{s,k} = \{m'_i\},\ i = 0 \dots k-1,\ M_{s,0} = \emptyset\ , \tag{5.12}$$

ενώ ο υπολογισμός του διανύσματος αναπαράστασης που επιλέχθηκε γίνεται ως

$$m'_k = \arg\min_m \Big(m(r_k)\Big),\ m \in M_z,\ m \notin M_{s,k}\ . \tag{5.13}$$

Όταν ο αλγόριθμος που προτάθηκε σε αυτό το Κεφάλαιο χρησιμοποιείται για δημιουργία περίληψης βίντεο, παρέχει έναν μεγάλο αριθμό από χαρακτηριστικά καρέ, τα οποία και αναπαριστούν το περιεχόμενο του βίντεο. Ο αριθμός αυτός είναι ίσος με τις συστάδες που δημιούργησε ο αλγόριθμος αφαιρετικής συσταδοποίησης, εφαρμοζόμενος στο σύνολο των διανυσμάτων αναπαράστασης N_S. Επιπρόσθετα, όταν χρησιμοποιείται για την επιλογή κάποιων χαρακτηριστικών καρέ μέσα σε ένα πλάνο, οδηγεί στην επιλογή ενός αριθμού από καρέ μικρότερο ή ίσο με αυτόν των τύπων περιοχής του θησαυρού. Έτσι, τα χαρακτηριστικά καρέ που εξάγονται περιέχουν συνολικά όσο το δυνατόν περισσότερη πληροφορία του οπτικού θησαυρού. Αυτό μπορεί να φανεί πολύ χρήσιμο στην ανίχνευση εννοιών υψηλού επιπέδου σε ακολουθίες βίντεο, όπου απαιτείται συνήθως μια πλούσια σε σημασιολογικό περιεχόμενο αναπαράσταση του οπτικού περιεχομένου.

5.5 Εφαρμογές και Πειραματικά Αποτελέσματα

Ο αλγόριθμος εξαγωγής χαρακτηριστικών καρέ χρησιμοποιείται σε δύο εφαρμογές. Αρχικά διερευνάται κατά πόσο η εφαρμογή της τεχνικής ανίχνευσης εννοιών υψηλού επιπέδου που παρουσιάστηκε στο Κεφάλαιο 4 σε περισσότερα από ένα χαρακτηριστικά καρέ μπορεί να αυξήσει ουσιαστικά την ακρίβεια που επιτυγχάνεται. Στη συνέχεια, εξετάζεται κατά πόσο οι περιλήψεις που κατασκευάστηκαν μπορούν να συλλάβουν ικανοποιητικά το οπτικό και σημασιολογικό περιεχόμενο των βίντεο, ως προς τις ανάγκες του χρήστη.

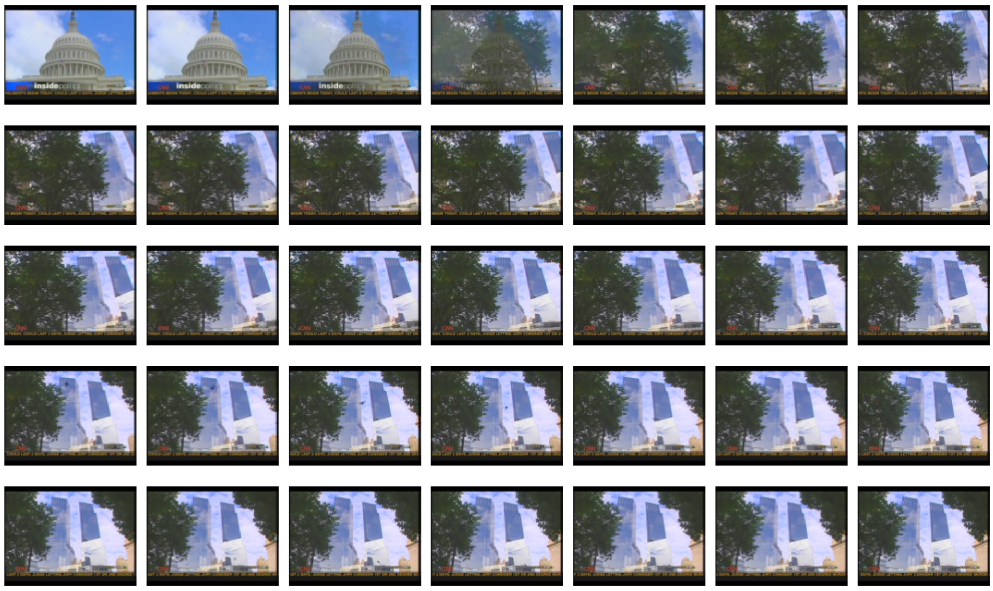
5.5.1 Ανίχνευση Εννοιών Υψηλού Επιπέδου σε πολλά Χαρακτηριστικά Καρέ

Στην Ενότητα αυτή χρησιμοποιείται ο αλγόριθμος που περιγράφηκε στο Κεφάλαιο 4 για την ανίχνευση εννοιών υψηλού επιπέδου και εφαρμόζεται σε έναν αριθμό από εξαχθέντα χαρακτηριστικά καρέ για κάθε πλάνο, μέσω του αλγορίθμου που περιγράφηκε στην Ενότητα 5.4. Η αξιολόγηση γίνεται σε 6 από τα βίντεο της συλλογής του TRECVID και συγκρίνεται με την περίπτωση που η ανίχνευση πραγματοποιείται σε μοναδικό χαρακτηριστικό καρέ. Τα αποτελέσματα φαίνονται στον Πίνακα 5.1. Τα αποτελέσματα καταδεικνύουν ακριβώς το αναμενόμενο, δηλαδή μεγαλύτερες τιμές ανάκτησης και παρόμοιες τιμές ακρίβειας.

Έννοια	**Μοναδικό Καρέ**			**Περίληψη**			
	P	**R**	**AP**	**P**	**R**	**AP**	**Διαφορά AP**
βλάστηση	0.58	0.40	0.48	0.63	0.45	0.52	8.3%
δρόμος	0.28	0.12	0.29	0.30	0.06	0.32	10.3%
έκρηξη-φωτιά	0.20	0.81	0.14	0.22	0.78	0.18	28.5%
ουρανός	0.47	0.45	0.46	0.53	0.52	0.59	28.2%
χιόνι	0.50	0.32	0.42	0.64	0.41	0.45	7.1%
γραφείο	0.35	0.12	0.33	0.43	0.17	0.34	3.0%
έρημος	0.27	0.30	0 .29	0.31	0.38	0.35	20.7%
εξωτερικός χώρος	0.40	0.60	0.41	0.41	0.66	0.45	9.7%
βουνό	0.31	0.09	0.19	0.42	0.14	0.23	21.1%

Πίνακας 5.1: *Αποτελέσματα ανίχνευσης σε ένα χαρακτηριστικό καρέ και σε περίληψη που αποτελείται από περισσότερα χαρακτηριστικά καρέ. P: ακρίβεια, R: ανάκτηση, AP: μέση ακρίβεια. Επσης υπολογίστηκε και η διαφορά στη μέση ακρίβεια. Σε κάθε περίπτωση η μέση ακρίβεια αυξήθηκε.*

Για να γίνει καλύτερα κατανοητό γιατί η αναπαράσταση ενός πλάνου με περισσότερα από ένα χαρακτηριστικά καρέ οδηγεί σε καλύτερα αποτελέσματα στην ανίχνευση εννοιών υψηλού επιπέδου σε πλάνα από βίντεο, παρουσιάζεται ένα παράδειγμα από ένα πλάνο, ένα χαρακτηριστικό καρέ που εξήχθη, καθώς και ένας μεγαλύτερος αριθμός από εξαχθέντα χαρακτηριστικά καρέ. Ένας σχετικά μεγάλος αριθμός από καρέ του πλάνου εξάγεται για να αναπαραστήσει το οπτικό αλλά και σημασιολογικό του περιεχόμενο. Αυτά παρουσιάζονται στο Σχήμα 5.4. Η διάρκεια αυτού του πλάνου είναι



Σχήμα 5.4: *Ένα σχετικά μεγάλο υποσύνολο των καρέ από τα οποία αποτελείται ένα πλάνο. Το πλάνο έχει διάρκεια 5 sec και τα καρέ έχουν εξαχθεί με ρυθμό 9 καρέ/sec*

(α′) Μοναδικό χαρακτηριστικό καρέ

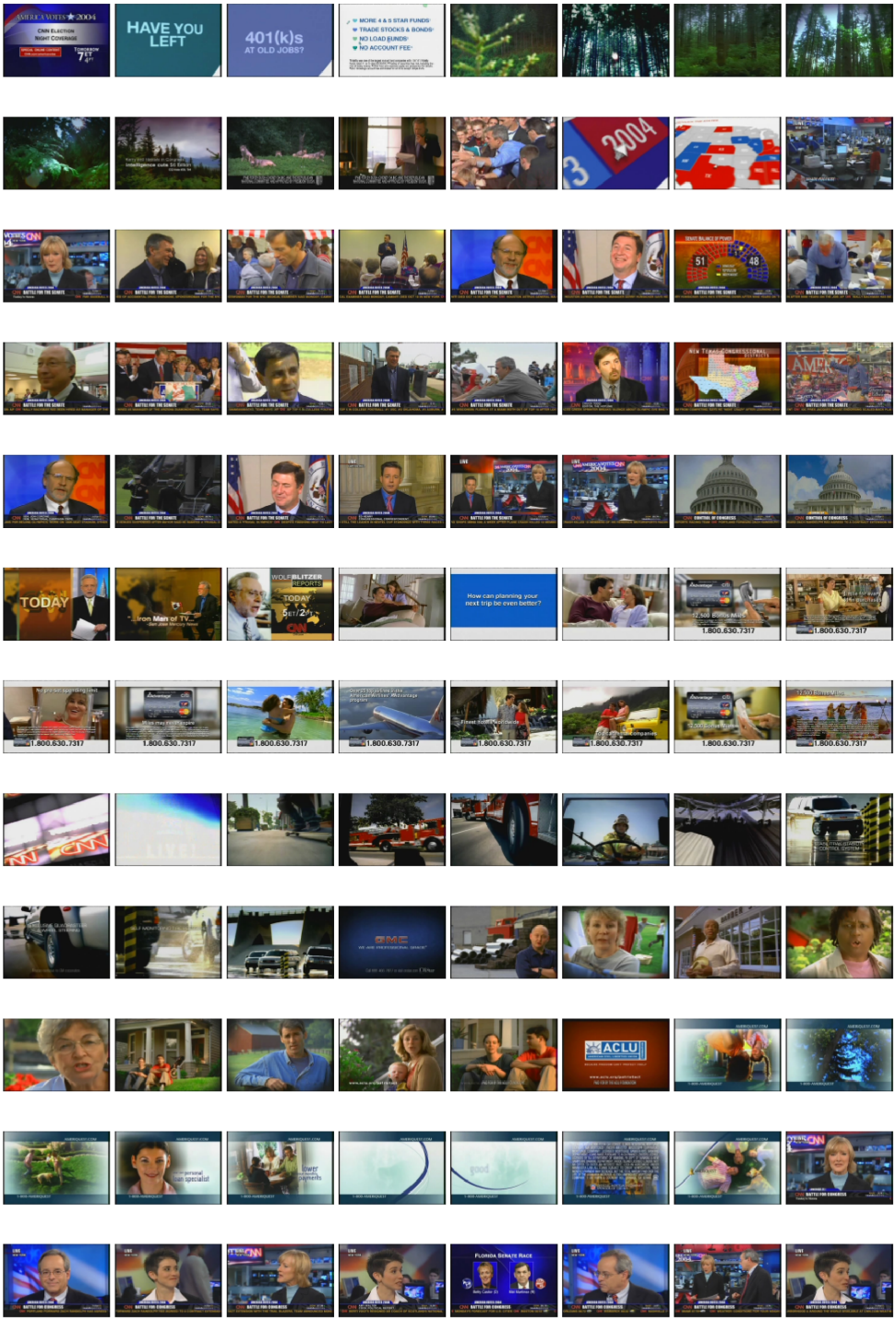
(β′) Περισσότερα χαρακτηριστικά καρέ

Σχήμα 5.5: *Το χαρακτηριστικό καρέ που εξάγεται από τον αλγόριθμο του [168] και παρέχεται από τους οργανωτές του TRECVID και τα χαρακτηριστικά καρέ που εξήχθησαν με τη χρήση του προτεινόμενου αλγορίθμου. Είναι φανερό ότι το βίντεο περιέχει περισσότερες έννοιες από αυτές που απεικονίζονται στο μοναδικό χαρακτηριστικό καρέ.*

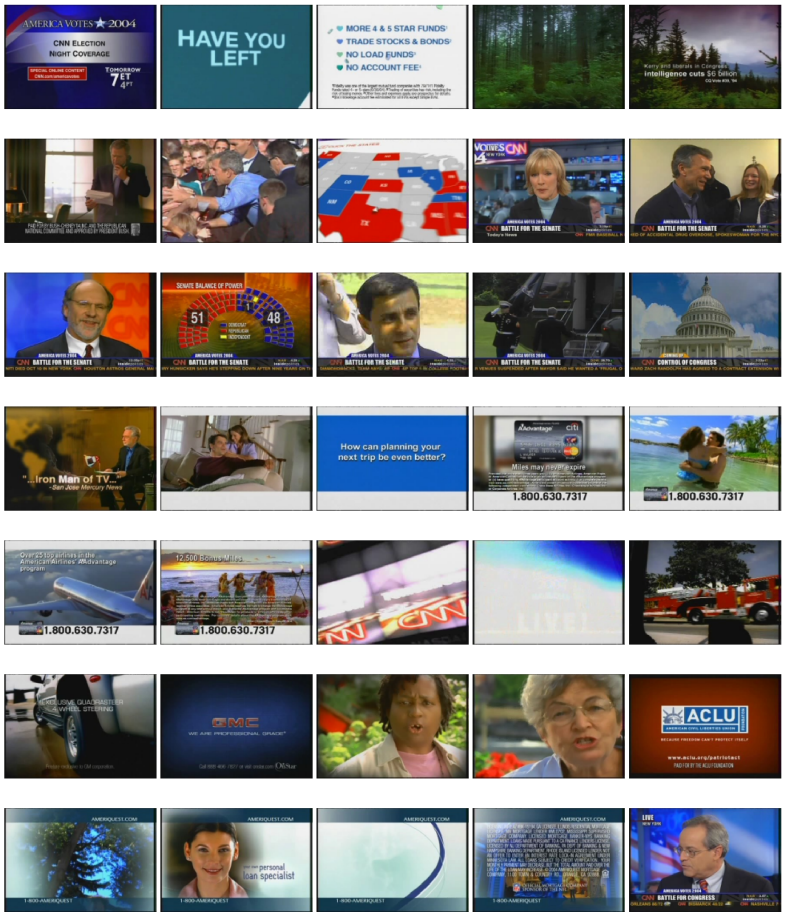
περίπου ίση με 5 δευτερόλεπτα. Οι έννοιες υψηλού επιπέδου που περιέχονται και ανήκουν σε αυτές που έχουν επιλεχθεί προς ανίχνευση είναι προφανώς *εξωτερικός χώρος*, *βλάστηση* και *ουρανός*. Το μοναδικό καρέ που επιλέχθηκε για την αναπαράσταση του περιεχομένου του πλάνου είναι αυτό που φαίνεται στο Σχήμα 5.5(α′). Ο προτεινόμενος αλγόριθμος εξήγαγε τα 4 χαρακτηριστικά καρέ που φαίνονται στο Σχήμα 5.5(β′). Είναι προφανές ότι το 1ο χαρακτηριστικό καρέ στο Σχήμα 5.5(β′) είναι καλύτερο για την ανίχνευση της έννοιας *ουρανός* ενώ το δέυτερο είναι καλύτερο για την ανίχνευση της έννοιας *βλάστηση*, σε σχέση πάντα με το μοναδικό καρέ. Επιπρόσθετα, τα 4 καρέ θα αυξήσουν την πιθανότητα σωστής ανίχνευσης της έννοιας *εξωτερικός χώρος*.

5.5.2 Εξαγωγή Χαρακτηριστικών Καρέ για τη δημιουργία Περιλήψεων

Στην Ενότητα αυτή παρουσιάζονται αποτελέσματα της εφαρμογής του προτεινόμενου αλγορίθμου εξαγωγής χαρακτηριστικών καρέ για τη δημιουργία περιλήψεων ακολουθιών βίντεο. Για λόγους ευκρίνειας της παρουσίασης, χρησιμοποιήθηκε ένα μέρος από βίντεο του TRECVID που περιέχει δελτίο ειδήσεων, μιας και στην περί-



Σχήμα 5.6: *Ένα σχετικά μεγάλο υποσύνολο των καρέ από τα οποία αποτελείται ένα βίντεο.*



Σχήμα 5.7: *Τα χαρακτηριστικά καρέ που εξήχθησαν με τη χρήση της τεχνικής που προτείνεται σε αυτό το Κεφάλαιο, από το βίντεο του Σχήματος 5.6.*

πτωση αυτή, τα καρέ είναι πιο ανομοιογενή και η επιλογή των χαρακτηριστικών καρέ είναι πιο δύσκολη.

Το βίντεο περιέχει τον παρουσιαστή των ειδήσεων, μερικά διαφημιστικά μηνύματα, κάποιες συνεντεύξεις εντός και εκτός στούντιο, χάρτες, διαγράμματα κλπ. Η διάρκειά του είναι περίπου 620 δευτερόλεπτα. Στο Σχήμα 5.6 παρουσιάζεται συνοπτικά το οπτικό του περιεχόμενο. Στη συνέχεια, στο Σχήμα 5.7 παρουσιάζονται τα εξαχθέντα χαρακτηριστικά καρέ. Πρέπει να τονιστεί ότι τα χαρακτηριστικά καρέ είναι όλα αυτά που βρίσκονται κοντύτερα στα κέντρα των συστάδων που σχηματίστηκαν μετά από εφαρμογή του αλγορίθμου αφαιρετικής συσταδοποίησης στα διανύσματα αναπαράστασης.

Για κάθε καρέ, ο μέσος χρόνος επεξεργασίας ήταν 0.75 δευτερόλεπτα για την κατάτμηση και 1.15 δευτερόλεπτα για την εξαγωγή των χαμηλού επιπέδου περιγραφέων. Έτσι, για το συγκεκριμένο βίντεο, αφού χρησιμοποιήθηκαν καρέ που εξήχθησαν με ρυθμό 5 καρέ/δευτερόλεπτο και μαζί με το χρόνο που απαιτήθηκε για την επεξεργασία του (συσταδοποίηση, σχηματισμός διανυσμάτων αναπαράστασης κλπ), κάτι λιγότερο από 2 ώρες χρειάστηκαν για την εξαγωγή των χαρακτηριστικών καρέ. Πρέπει να τονιστεί ότι ο πιο σημαντικός χρονικός περιορισμός έχει να κάνει με την εξαγωγή των περιγραφέων και την συσταδοποίηση, κάτι που μπορεί να βελτιωθεί με τη χρήση τεχνικών που να έχουν μεγαλύτερη ταχύτητα, κάτι που ξεφεύγει από τα ερευνητικά πλαίσια της παρούσας εργασίας.

Για την αξιολόγηση των περιλήψεων που δημιουργήθηκαν, ζητήθηκε η γνώμη από 12 χρήστες. Σε κάθε χρήστη ζητήθηκε να ελέγξει ποια από τα επιλεγμένα καρέ είναι απαραίτητα για να περιγράψουν ικανοποιητικά το περιεχόμενο του βίντεο, καθώς και ποια από αυτά που δεν επιλέχθηκαν θα έπρεπε να επιλεχθούν προκειμένου να επιτευχθεί πιο ολοκληρωμένη περιγραφή. Ο αλγόριθμος που παρουσιάστηκε στην Ενότητα 5.4 συγκρίθηκε με τα μοναδικά χαρακτηριστικά καρέ που εξήχθησαν από κάθε πλάνο με την τεχνική του Petersohn [168]. Για τον προτεινόμενο αλγόριθμο, το 80% των επιλεχθέντων χαρακτηριστικών καρέ κρίθηκε σαν απαραίτητο και μόνο 12% επιπλέον καρέ κρίθηκαν απαραίτητα για τη δημιουργία πληρέστερης περίληψης. Στην περίπτωση του αλγορίθμου του [168], το 82% των επιλεχθέντων καρέ κρίθηκε ως αναγκαίο. Αυτό ήταν αναμενόμενο, μιας και για κάθε πλάνο, η εξαγωγή ενός τουλάχιστον καρέ είναι αναγκαία. Όπως ήταν επίσης αναμενόμενο, ένα επιπλέον 76% από καρέ κρίθηκε αναγκαίο για την πληρέστερη περίληψη, μιας και στην πλειοψηφία των περιπτώσεων, περισσότερα καρέ είναι αναγκαία για να καλύψουν όλα τα οπτικά και σημασιολογικά χαρακτηριστικά ενός πλάνου.

5.6 Συμπεράσματα

Στο Κεφάλαιο αυτό παρουσιάστηκε μια τεχνική εξαγωγής χαρακτηριστικών καρέ από βίντεο. Η τεχνική αυτή βασίστηκε στην κατασκευή ενός τοπικού οπτικού θησαυρού περιοχών. Ο θησαυρός αυτός χαρακτηρίζεται ως τοπικός, καθώς κατασκευάζεται με βάση τις περιοχές των καρέ που απαρτίζουν ένα πλάνο ή γενικότερα μια ακολουθία βίντεο. Με βάση το θησαυρό αυτό κατασκευάζεται ένα διάνυσμα αναπαράστασης, το οποίο και χρησιμοποιείται για την περιγραφή του οπτικού περιεχομένου των καρέ.

Με την τεχνική αυτή κατασκευάστηκαν περιλήψεις, οι οποίες είχαν ως πρώτο στόχο να κάνουν το χρήστη τους να αντιληφθεί το σημασιολογικό περιεχόμενο των βίντεο, παρατηρώντας μόνο ένα σχετικά μικρό από καρέ. Ο δεύτερος στόχος ήταν να

χρησιμοποιηθούν αυτές οι περιλήψεις και να διερευνηθεί αν εφαρμόζοντας τις τεχνικές ανίχνευσης εννοιών του Κεφαλαίου 4 καταστεί δυνατή η βελτίωση της ακρίβειας της ανίχνευσης.

Οι περιλήψεις που κατασκευάστηκαν με τη χρήση του αλγορίθμου, διαφάνηκε μέσα από πειράματα για έναν μικρό αριθμό από βίντεο και με τη συμμετοχή ενός ικανού αριθμού χρηστών, ότι έχουν σαφώς πιο πλούσιο περιεχόμενο από ένα μοναδικό καρέ, αλλά και ότι μπορούν να δώσουν στο χρήστη τη δυνατότητα να αντιληφθεί πλήρως το σημασιολογικό περιεχόμενο ενός βίντεο, όσον αφορά, τουλάχιστον τις έννοιες που μπορούν να γίνουν αντιληπτές από το οπτικό του περιεχόμενο.

Επίσης, η εφαρμογή της τεχνικής ανίχνευσης στις περιλήψεις που αποτελούνται απο πολλά χαρακτηριστικά καρέ, εμφάνισε αυξημένη ακρίβεια. Αυτό ήταν κάτι αναμενόμενο, καθώς απλές παρατηρήσεις στα βίντεο και τα μοναδικά χαρακτηριστικά καρέ που εξάγονται καταδεικνύουν ότι σπάνια περιέχονται όλες οι έννοιες σε ένα καρέ. Το τελευταίο θα μπορούσε να συμβεί μόνο στην περίπτωση που τα βίντεο έχουν πολύ μικρή διάρκεια και αναζητείται ένας πολύ μικρός αριθμός εννοιών. Η περίληψη που βασίστηκε στον τοπικό οπτικό θησαυρό συνδέεται άμεσα με την τεχνική ανίχνευσης με χρήση του οπτικού θησαυρού. Οι θησαυροί κατασκευάζονται με βάση πρακτικά την ίδια τεχνική. Η επιλογή των χαρακτηριστικών καρέ με βάση τους πιο σημαντικούς τύπους περιοχής που αυτά περιέχουν, αυτόματα συνεπάγεται ότι επιλέγονται τα καρέ που περιέχουν ικανό αριθμό από έννοιες.

Διαφάνηκε επίσης μια από τις πιο συνηθισμένες αιτίες που πολλές τεχνικές ανίχνευσης εννοιών αποτυγχάνουν να επιτύχουν ικανοποιητικές τιμές ακρίβειας στη διαδικασία αξιολόγησης του TRECVID. Αυτό συμβαίνει όταν μια τεχνική εφαρμόζεται σε μοναδικό χαρακτηριστικό καρέ των βίντεο. Ωστόσο η αξιολόγηση δεν γίνεται σε επίπεδο καρέ, αλλά σε επίπεδο πλάνου. Έτσι είναι παραπάνω από πιθανό κάποιες έννοιες να μην εμφανίζονται στα καρέ στα οποία εφαρμόζονται οι τεχνικές, αυτές σωστά να μην τις αναγνωρίζουν στα καρέ αυτά, αλλά να αγνοούν την ύπαρξη και τη σημασιολογία των υπόλοιπων καρέ, χάνοντας έτσι σε ακρίβεια.

Το βασικό μειονέκτημα της τεχνικής έχει να κάνει με την αυξημένη της πολυπλοκότητα. Καθώς απαιτείται η εφαρμογή του αλγορίθμου κατάτμησης εικόνων σε μεγάλο αριθμό από καρέ, ο χρόνος που απαιτείται για την εφαρμογή της σε όλα τα καρέ από βίντεο είναι ιδιαίτερα μεγάλος, κάτι που καθιστά την εφαρμογή της σε μεγάλο αριθμό βίντεο ιδιαίτερα δύσκολη.

Κεφάλαιο 6

Ανάκτηση Εικόνων με χρήση Οπτικού Θησαυρού

6.1 Εισαγωγή

Όπως έχει αναφερθεί αρκετές φορές εώς τώρα στην παρούσα Εργασία, οι τεχνολογικές εξελίξεις ιδιαίτερα των δύο τελευταίων δεκαετιών έχουν σταθεί αφορμή για προσανατολισμό της έρευνας προς κατευθύνσεις που σκοπό έχουν να καλύψουν τις νέες ανάγκες που προκύπτουν. Μία από τις ανάγκες που έχει προκύψει εξαιτίας της ραγδαίας αύξησης πολυμεσικού υλικού αποτελεί η ανάκτηση του περιεχομένου αυτού, καθώς όλη αυτή η πληροφορία που έχει συσσωρευτεί, είναι άχρηστη εάν οι χρήστες δεν μπορούν να την ανακτήσουν εύκολα, γρήγορα και αποδοτικά. Το παρόν Κεφάλαιο ασχολείται με μια συγκεκριμένη περιοχή της ανάκτησης πληροφορίας με βάση το περιεχόμενο, αυτή της *ανάκτησης εικόνων* και έμφαση θα δοθεί στην ανάκτηση με βάση τα οπτικά τους χαρακτηριστικά.

Η έρευνα στο χώρο της ανάκτησης εικόνων ξεκίνησε, ωστόσο, από τη δεκαετία του '70. Οι δύο προσεγγίσεις που ακολουθήθηκαν ήταν αφενός αυτές που εμπνεύστηκαν από τις τεχνικές διαχείρισης βάσεων δεδομένων και αφετέρου αυτές από το χώρο της όρασης υπολογιστών. Οι τεχνικές που εντάσσονται στην πρώτη κατηγορία βασίζονται στην ανάκτηση με βάση το κείμενο που συνοδεύει τις εικόνες, ενώ αυτές που εντάσσονται στη δεύτερη κατηγορία χρησιμοποιούν τα οπτικά χαρακτηριστικά των εικόνων. Ωστόσο, όπως σημειώνουν οι Smeulders et al. [193], "οι εικόνες είναι υπεράνω λέξεων", δηλαδή το οπτικό τους περιεχόμενο δεν μπορεί να περιγραφεί πλήρως από ένα σύνολο λέξεων. Συνεπώς, η έμφαση πρέπει να δοθεί στην ανάκτηση με βάση το οπτικό περιεχόμενο των εικόνων και η άλλη κατηγορία ερευνητικών προσπαθειών να έχει επικουρικό ρόλο στη διαδικασία.

Η ανάκτηση εικόνων αποτελεί ένα ερευνητικό πεδίο το οποίο μπορεί να ωφελήσει και να ωφεληθεί άμεσα ή έμμεσα από πολλές εφαρμογές που σχετίζονται με την ανάλυση πολυμεσικού περιεχομένου. Ξεχωρίζουν η οργάνωση ψηφιακών αρχείων, η όραση υπολογιστών, οι τεχνικές μηχανικής μάθησης, οι βάσεις δεδομένων, η ανίχνευση εννοιών υψηλού επιπέδου κ.ο.κ.

Στο Κεφάλαιο αυτό παρουσιάζεται αρχικά η παρούσα κατάσταση σχετικά με τις ερευνητικές προσπάθειες που έχουν επικεντρωθεί στην ανάκτηση εικόνας. Έπειτα, παρουσιάζεται μια τεχνική, η οποία επωφελούμενη από την ιδέα περιγραφής μιας εικόνας με βάση τις περιοχές στις οποίες μπορεί αυτή να χωριστεί και τη βοήθεια ενός

οπτικού θησαυρού, προσπαθεί να συνεισφέρει στο πρόβλημα της ανάκτησης εικόνων. Στοχεύει σε μια σημασιολογική ανάλυση, όπου σημασία έχουν πρωτίστως οι έννοιες που περιέχουν οι εικόνες που ανακτήθηκαν και έπειτα η εμφάνισή τους σε σχέση με τις εικόνες των ερωτημάτων.

6.2 Περιγραφή του Προβλήματος

Το πρόβλημα της ανάκτησης εικόνων είναι περισσότερο σύνθετο από ότι ίσως φαίνεται στον τελικό χρήστη. Προκειμένου να καταστεί κατανοητός ο τρόπος με τον οποίο πρέπει να σχεδιαστεί ένα σύστημα ανάκτησης, προκειμένου να επιλύσει ένα συγκεκριμένο πρόβλημα, πρέπει αρχικά να γίνουν κατανοητές οι επιδιώξεις και οι προθέσεις των υποψήφιων χρηστών του. Έπειτα, είναι πολύ σημαντικό να ληφθούν υπόψη οι ιδιαιτερότητες των δεδομένων τα οποία θα κληθεί να ανακτήσει. Στη συνέχεια πρέπει να αποφασιστεί ο τρόπος με τον οποίο θα τίθενται τα ερωτήματα και οι τεχνικές που θα χρησιμοποιηθούν για την επεξεργασία τους και την παραγωγή των αποτελεσμάτων, τα οποία και θα πρέπει να παρουσιαστούν με βάση τις ανάγκες των χρηστών.

6.2.1 Επιδιώξεις και προθέσεις των χρηστών

Ένα από τα ιδιαίτερα αλλά και βασικότερα χαρακτηριστικά των προβλημάτων ανάκτησης αποτελεί ο σκοπός των πιθανών χρηστών τους. Τις περισσότερες φορές, αυτό που επιδιώκει ο χρήστης που χρησιμοποιεί ένα σύστημα ανάκτησης και ψάχνει για εικόνες δεν είναι ιδιαίτερα σαφές. Έτσι, αν οι επιθυμίες των χρηστών διαφέρουν από αυτό που προσφέρει ένα σύστημα ανάκτησης, τότε αυτοί θα μείνουν ανικανοποίητοι από αυτό και η αξιολόγησή του με βάση υποκειμενικά κριτήρια θα είναι κακή. Σύμφωνα με τους Datta et al. [56], υπάρχουν τρεις κατηγορίες χρηστών, ανάλογα με τις προθέσεις τους, αλλά και τη σαφήνεια του σκοπού τους:

- *Περιηγητής* (Browser): Σε αυτή την κατηγορία ανήκουν οι χρήστες που δεν έχουν κάποιο συγκεκριμένο στόχο και απλά περιηγούνται σε συλλογές εικόνων. Μια συνεδρία ενός τέτοιου χρήστη θα περιλαμβάνει διαδοχικά, χωρίς συνοχή μεταξύ τους ερωτήματα. Μπορεί να ψάχνει αρχικά για *αυτοκίνητα* και έπειτα για *ηλιοβασιλέματα* στη *θάλασσα*. Οι χρήστες που ανήκουν στην κατηγορία αυτή συνήθως δεν έχουν ιδιαίτερες απαιτήσεις από το σύστημα ανάκτησης και δίνουν έμφαση σε χαρακτηριστικά όπως υποδείξεις για το τι ψάχνουν άλλοι χρήστες και πιο δημοφιλείς εικόνες προκειμένου να βρουν αφορμή για να ξεκινήσουν μια συνεδρία.

- *Τυπικός χρήστης* (Surfer): Σε αυτή την κατηγορία ανήκουν οι χρήστες που έχουν μια σχετική σαφήνεια ως προς τον σκοπό της αναζήτησης. Τα αρχικά ερωτήματα μπορεί να είναι διερευνητικά, τα επόμενα όμως δείχνουν με μεγαλύτερη σαφήνεια τον στόχο της αναζήτησης, επηρεαζόμενα ίσως και από τα αρχικά αποτελέσματα. Για παράδειγμα, ένας τέτοιος χρήστης μπορεί αρχικά να ψάχνει για *παραλία* και έπειτα για *θάλασσα*. Οι χρήστες που ανήκουν στην κατηγορία αυτή επιθυμούν πρωτίστως ένα περιβάλλον που να διευκολύνει την αναζήτηση.

- *Ερευνητής* (Searcher): Σε αυτή την κατηγορία ανήκουν οι χρήστες οι οποίοι ξέρουν εκ των προτέρων τι ψάχνουν, ο σκοπός τους είναι δηλαδή ξεκάθαρος από την αρχή. Μια συνεδρία ενός χρήστη που ανήκει σε αυτή την κατηγορία είναι συνήθως μικρή σε διάρκεια, έχει υψηλή συνοχή, δηλαδή τα ενδιάμεσα αποτελέσματα είναι παρόμοια σε οπτικό ή σε εννοιολογικό επίπεδο. Μια τέτοιου τύπου αναζήτηση οδηγεί τις περισσότερες φορές σε κάποια τελικά αποτελέσματα. Οι χρήστες που ανήκουν σε αυτή την κατηγορία επιθυμούν ακρίβεια στα αποτελέσματα της αναζήτησης και καλή παρουσίασή τους.

Οι μελέτες που ασχολούνται με τα συστήματα ανάκτησης από την πλευρά του χρήστη σπανίζουν. Οι Christel και Conescu [48] χώρισαν τους χρήστες σε *αρχάριους* και *προχωρημένους* και μελέτησαν τα μοτίβα αλληλεπίδρασης μέσω ενός συστήματος ανάκτησης βίντεο. Για το σκοπό αυτό χρησιμοποίησαν τη βάση με βίντεο από το TRECVID 2004 [192] και κατέληξαν στο συμπέρασμα ότι οι έμπειροι χρήστες μπορούν να εκμεταλλευτούν καλύτερα τις πιο προχωρημένες δυνατότητες των συστημάτων ανάκτησης και σε συνδυασμό με την εμπειρία τους να επιτύχουν καλύτερα αποτελέσματα στην αναζήτηση. Επιπρόσθετα, οι Armitage και Esner [4] προσπάθησαν να αναλύσουν τις ανάγκες του χρήστη όσον αφορά στην ανάκτηση οπτικής πληροφορίας.

6.2.2 Κατηγορίες Δεδομένων

Όσον αφορά τη συλλογή των εικόνων που περιέχει ένα σύστημα ανάκτησης, θα πρέπει να τονιστεί ιδιαίτερα ότι αυτές παίζουν ιδιαίτερο ρόλο στο σχεδιασμό και τη λειτουργία του. Σύμφωνα με τους Datta et al. [56], οι εικόνες που μπορεί να περιλαμβάνει ένα σύστημα ανάκτησης χωρίζονται στις παρακάτω κατηγορίες:

- *Προσωπική Συλλογή* (Personal Collection): Στην κατηγορία αυτή εντάσσονται συνήθως μικρές σε μέγεθος και αρκετά ετερογενείς συλλογές, οι οποίες ανήκουν όλες στον ίδιο χρήστη και συνήθως μόνο αυτός έχει πρόσβαση σε αυτές. Οι συλλογές αυτές αποθηκεύονται κατά κύριο λόγο σε τοπικά αποθηκευτικά μέσα, ενώ τα τελευταία χρόνια έχουν αρχίσει να κάνουν την εμφάνισή τους ιστοσελίδες για αυτό το σκοπό. Παραδείγμα τέτοιων συλλογών είναι οι φωτογραφίες από τις διακοπές μιας οικογένειας. Στο Σχήμα 6.1 απεικονίζεται ένα παράδειγμα μιας προσωπικής συλλογής στη γνωστή ιστοσελίδα Flickr[1].

- *Συλλογή Θεματικού Πεδίου* (Domain-Specific Collection): Στην κατηγορία αυτή περιλαμβάνονται ομοιογενείς συλλογές, το περιεχόμενο των οποίων ανήκει σε συγκεκριμένο θεματικό πεδίο. Σε τέτοιου τύπου συλλογές δίνεται συνήθως πρόσβαση σε επιλεγμένους χρήστες, καθώς αυτές γενικά δεσμεύονται από πνευματικά δικαιώματα. Παράδειγμα τέτοιων συλλογών είναι φωτογραφίες που ανήκουν στα θεματικά πεδία *Παραλία*, *Πόλη* και λοιπά. Όσο πιο καλά ορισμένο είναι το θεματικό πεδίο, τόσο πιο ομοιογενείς θα είναι και οι εικόνες τέτοιων συλλογών. Για παράδειγμα, οι εικόνες του θεματικου πεδίου *Παραλία* θα είναι πιο ομοιογενείς από τις αντίστοιχες του θεματικού πεδίου *εξωτερικός χώρος*, δηλαδή θα εμφανίζουν παρόμοια χαρακτηριστικά χαμηλού και υψηλού επιπέδου. Τέτοιες συλλογές συναντώνται συνήθως σε ερευνητικά έργα, ή σε διάφορες μεθόδους αξιολόγησης ή συνοδεύουν διάφορα εμπορικά προϊόντα, όπως είναι

[1] `http://www.flickr.com`

για παράδειγμα η συλλογή εικονων του Corel[2], η οποία χρησιμοποιείται κατά κόρον για αξιολόγηση τεχνικών, αλλά και για εμπορικούς σκοπούς. Ένα τέτοιο παράδειγμα αποτελεί και το σύνολο εικόνων που χρησιμοποιήθηκε στα πειράματα που έγιναν στο πλαίσιο του Κεφαλαίου αυτού, ένα μέρος του οποίου απεικονίζεται στο Σχήμα 6.8.

- *Συλλογή Επιχείρησης* (Enterprise Collection): Στην κατηγορία αυτή περιλαμβάνονται εικόνες που περιέχονται στο τοπικό δίκτυο μιας εταιρίας ή επιχείρησης, γενικότερα. Οι εικόνες αυτές είναι συνήθως πολύ ετερογενείς. Για παράδειγμα, στην περίπτωση μιας διαφημιστικής εταιρίας, μπορεί να περιέχονται εικόνες με *λογότυπα* προϊόντων, εικόνες με *πρόσωπα*, με τοπία *εξωτερικού χώρου* και ούτω καθεξής. Οι συλλογές που ανήκουν στην κατηγορία αυτή δεν είναι συνήθως διαθέσιμες σε απλούς χρήστες.

- *Αρχεία* (Archives): Η κατηγορία αυτή τυπικά περιλαμβάνει μεγάλες συλλογές συνήθως ιστορικού περιεχομένου. Οι συλλογές αυτές είναι πιο οργανωμένες από τις αυτές των παραπάνω κατηγοριών, είτε ανά χρονική περίοδο, είτε ανά γεγονός. Πολλές φορές ιστορικά αρχεία είναι διαθέσιμα στον Παγκόσμιο Ιστό για όλες τις χρήσεις. Οι εικόνες αυτές τυπικά αποθηκεύονται σε εξυπηρετητές αρχείων. Στο Σχήμα 6.2 απεικονίζεται ένα μικρό μέρος από το ψηφιοποιημένο αρχείο φωτογραφιών της ΕΡΤ[3] από το Μικρασιατικό Πόλεμο.

- *Ιστός* (Web): Η τελευταία αυτή κατηγορία περιλαμβάνει έναν τεράστιο αριθμό εικόνων, το σύνολο αυτών που περιέχονται στον Παγκόσμιο Ιστό. Οι εικόνες αυτές είναι κάποιες φορές οργανωμένες, σε επίπεδο δικτυακού τόπου, αλλά συνήθως δεν έχουν καμία οργάνωση. Φυσικά, η ετερογένεια των εικόνων αυτής της κατηγορίας είναι πολύ μεγάλη.

Ιδιαίτερη μνεία θα πρέπει να γίνει σε ιστοσελίδες που περιέχουν οργανωμένες συλλογές εικόνων από χρήστες, όπως είναι για παράδειγμα το Flickr, τα Picasa Web Albums[4] και το Panoramio[5], όπου οι χρήστες ανεβάζουν τις φωτογραφίες τους, τις σχολιάζουν είτε με λέξεις-κλειδιά, είτε με ελεύθερο κείμενο και τις μοιράζονται με άλλους χρήστες. Οι συλλογές εικόνων αυτής της κατηγορίας μπορεί να περιέχουν από προσωπικές φωτογραφίες, εώς και εικόνες που οι χρήστες έχουν αντιγράψει από διάφορες ιστοσελίδες. Τέτοιες συλλογές δεν μπορούν πάντα να ενταχθούν σε μία από τις παραπάνω κατηγορίες, ωστόσο αναφέρονται εξαιτίας της μεγάλης άνθησης που έχουν γνωρίσει ιδιαίτερα τα τελευταία χρόνια.

6.2.3 Τρόποι Ερωτημάτων

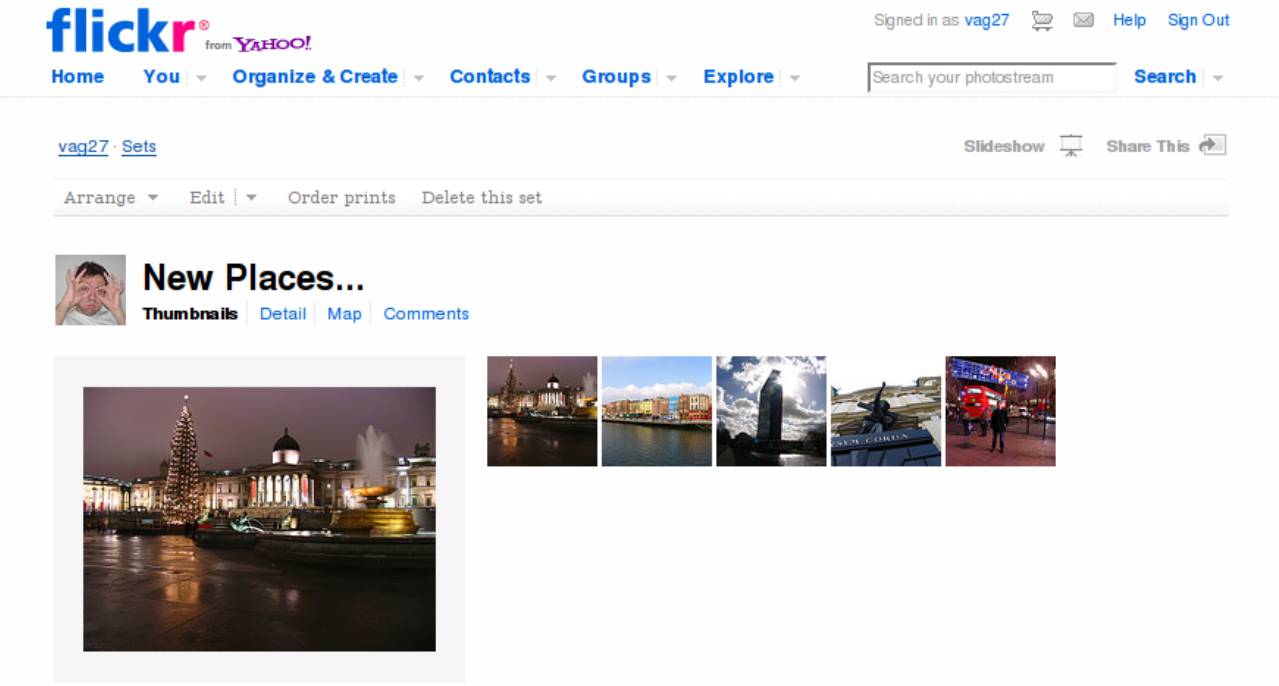
Μία από τις βασικότερες παραμέτρους ενός συστήματος ανάκτησης εικόνων είναι οι δυνατότητες που δίνει στο χρήστη σχετικά με τον τρόπο που αυτός σχηματίζει τα ερωτήματά του. Από την πλευρά του χρήστη αυτό συνεπάγεται τη δυνατότητα να δίνει πολυτροπικά ερωτήματα, συνδυάζοντας περισσότερους από έναν τρόπους αναπαράστασης του ερωτήματος. Οι τρόποι αυτοί είναι οι παρακάτω:

[2] `http://www.corel.com`

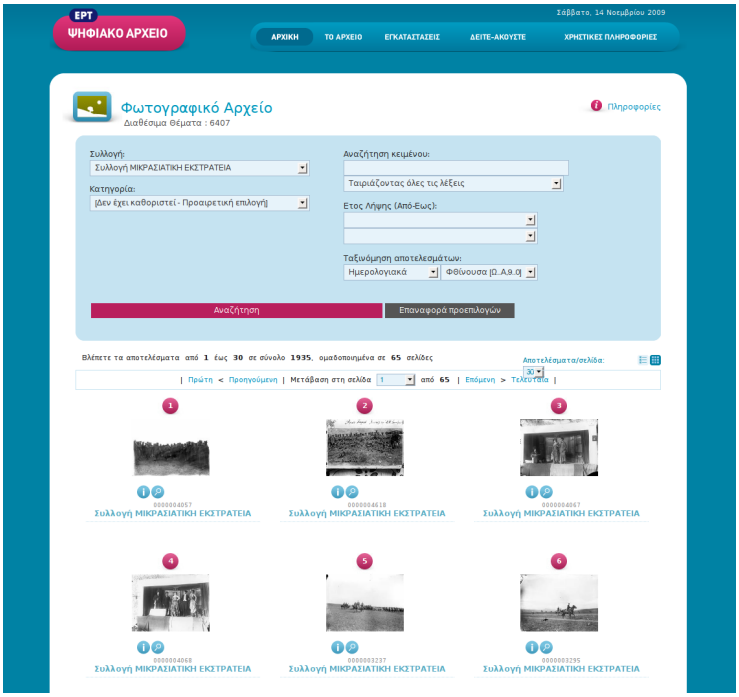
[3] `http://www.ert-archives.gr/V3/public/index.aspx`

[4] `http://picasaweb.google.com`

[5] `http://www.panoramio.com`



Σχήμα 6.1: *Ένα παράδειγμα μιας προσωπικής συλλογής στη γνωστή ιστοσελίδα Flickr.*



Σχήμα 6.2: *Ένα παράδειγμα ενός ψηφιοποιημένου αρχείου φωτογραφιών είναι το αρχείο της ΕΡΤ.*

- *Λέξεις-Κλειδιά* (Keywords): Ο χρήστης θέτει το ερώτημά του χρησιμοποιώντας μία ή και περισσότερες λέξεις. Αυτός είναι ένας από τους πιο συνηθισμένους τρόπους ανάκτησης εικόνων, όντας παράλληλα ο πρώτος τρόπος που χρησιμοποιήθηκε για το πρόβλημα αυτό. Μια πολύ συνηθισμένη μέθοδος ανάκτησης

που βασίζεται σε λέξεις-κλειδιά χρησιμοποιεί το σχολιασμό των διαθέσιμων εικόνων, ο οποίος μπορεί να γίνει είτε από έναν ειδήμονα ή από τον χρήστη που του ανήκει η εικόνα ή συλλογικά από τους επισκέπτες του συστήματος. Η πρώτη περίπτωση συναντάται κυρίως σε συλλογές εικόνων που εντάσσονται στην κατηγορία των *συλλογών θεματικού πεδίου*, ενώ οι υπόλοιπες συναντώνται πλέον πολύ συχνά σε δικτυακές κοινότητες όπως το Flickr. Η δεύτερη και πιο συνηθισμένη περίπτωση έχει να κάνει με το συνοδευτικό κείμενο των εικόνων που βρίσκονται σε ιστοσελίδες. Με αυτόν τον τρόπο γίνεται η ανάκτηση εικόνων σε δημοφιλείς μηχανές αναζήτησης όπως το Google[6] και το Yahoo![7].

- *Ελεύθερο κείμενο* (Free-text): Η περίπτωση αυτή είναι παρόμοια με την προηγούμενη, με τη διαφορά ότι ο χρήστης στην περίπτωση αυτή σχηματίζει το ερώτημα με τη μορφή μιας πρότασης, η οποία περιγράφει τι επιθυμεί ιδανικά να περιέχουν οι εικόνες που θα επιστρέψει το σύστημα.

- *Εικόνα*: Στην περίπτωση αυτή ο χρήστης είτε επιλέγει μια από τις εικόνες που περιέχονται στη συλλογή, είτε χρησιμοποιεί μια δική του εικόνα. Σαν ιδιαίτερες περιπτώσεις της κατηγορίας αυτής μπορούν να θεωρηθούν τα συστήματα που δίνουν τη δυνατότητα στο χρήστη να επιλέξει είτε ένα τμήμα της εικόνας του ερωτήματος, ή να "συνθέσει" μια εικόνα χρησιμοποιώντας π.χ. κομμάτια από τις διαθέσιμες εικόνες της συλλογής.

- *Γραφικά* (Graphics): Στην περίπτωση αυτή ο χρήστης δημιουργεί μια εικόνα (σκίτσο) με τη χρήση κάποιων απλών "εργαλείων", είτε σε κάποια online διαδικτυακή εφαρμογή είτε τέλος με τη χρήση ειδικού εξοπλισμού. Διαφέρει από την προηγούμενη γιατί οι εικόνες που δημιουργούνται δεν περιέχουν πραγματικά κομμάτια, αλλά συνθετικά. Προκειμένου να γίνει κατανοητή αυτή η περίπτωση ερωτήματος, στο Σχήμα 6.3 παρατίθεται ένα τέτοιο παράδειγμα από το σύστημα VideoQ που θα αναφερθεί στην Ενότητα 6.2.6. Ο χρήστης μπορεί να ζωγραφίσει απλά σχήματα, να τους προσθέσει χρώμα ή/και υφή και να σχηματίσει ένα οπτικό ερώτημα. Προαιρετικά μπορεί να προσθέσει και λέξεις κλειδιά, αλλά και βάρη στις διάφορες παραμέτρους της αναζήτησης.

- *Μεταδεδομένα τοποθεσίας* (Geo-tags): Στην περίπτωση αυτή, οι εικόνες της συλλογής συνοδεύονται από τις γεωγραφικές τους συντεταγμένες. Ο χρήστης επιλέγει μια τοποθεσία είτε σε ένα χάρτη, είτε δίνοντας κάποιες γεωγραφικές συντεταγμένες και το σύστημα επιλέγει εικόνες σε μια ακτίνα γύρω από την τοποθεσία αυτή.

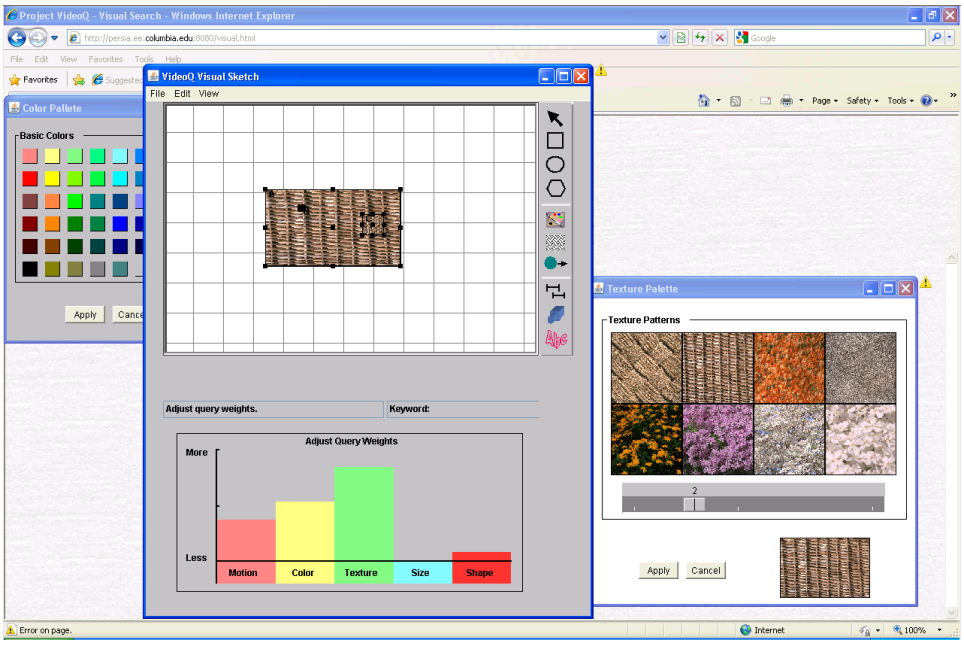
- *Σύνθετα ερωτήματα* (Composite): Η κατηγορία αυτή περιλαμβάνει όλες τις περιπτώσεις που γίνεται με συνδυασμό δύο ή και περισσότερων από τους προαναφερθέντες τρόπους.

6.2.4 Επεξεργασία Ερωτημάτων

Η επεξεργασία των ερωτημάτων που θέτουν οι χρήστες γίνεται με αυτόματο τρόπο από την πλευρά του συστήματος. Οι περιπτώσεις που συναντώνται είναι συνοπτικά

[6] `http://www.google.com`

[7] `http://www.yahoo.com`



Σχήμα 6.3: *Ένα παράδειγμα ενός συστήματος ανάκτησης που παρέχει στο χρήστη τη δυνατότητα να σχηματίζει ερωτήματα "ζωγραφίζοντας" απλά γραφικά.*

οι παρακάτω:

- *Επεξεργασία Κειμένου* (Text-based): Οι εικόνες δεικτοδοτούνται με βάση λέξεις που περιέχονται στην ίδια ιστοσελίδα με αυτές και τα συστήματα ανάκτησης αναζητούν τις λέξεις-κλειδιά του ερωτήματος. Στην περίπτωση που το ερώτημα έχει τη μορφή ελεύθερου κειμένου, η επεξεργασία του ερωτήματος συνήθως έχει να κάνει με την εξαγωγή κάποιων συχνά χρησιμοποιούμενων λέξεων, οι οποίες και τελικά αναζητούνται στις εικόνες της συλλογής. Φυσικά, η συλλογή έχει επίσης δεικτοδοτηθεί με παρόμοιο τρόπο. Πολύ συχνά στον τομέα αυτό βρίσκουν εφαρμογή τεχνικές επεξεργασίας της φυσικής γλώσσας.

- *Επεξεργασία Περιεχομένου* (Content-based): Οπτικά χαρακτηριστικά εξάγονται από την εικόνα ή από περιοχές της και η αναζήτηση γίνεται στις εικόνες της συλλογής με βάση αυτά. Η κατηγορία αυτή έχει ιδιαίτερη σημασία για την ερευνητική κοινότητα και όπως θα φανεί στην Ενότητα 6.3, υπάρχει πλήθος από σχετικές εργασίες. Σημαντικές περιοχές έρευνας στο χώρο αυτό έχουν να κάνουν με την εξαγωγή χαρακτηριστικών χαμηλού επιπέδου, την αναπαράσταση του οπτικού περιεχομένου και το ταίριασμα μεταξύ των εικόνων.

- *Σύνθετη Επεξεργασία* (Composite): Στην κατηγορία αυτή εντάσσονται οι μέθοδοι επεξεργασίας που συνδυάζουν τις δύο προαναφερθέντες κατηγορίες, κειμένου και περιεχομένου.

- *Διαδραστική Επεξεργασία* (Interactive): Στην περίπτωση αυτή γίνεται απλή επεξεργασία και ζητείται η παρέμβαση του χρήστη. Κυριότερη περίπτωση που συναντάται εδώ είναι τα συστήματα ανάδρασης σχετικότητας.

- *Διαδραστική Σύνθετη Επεξεργασία* (Interactive Composite): Η κατηγορία αυτή διαφέρει από την προηγούμενη καθώς ο χρήστης μπορεί να αλληλεπιδράσει με

το σύστημα σε περισσότερους τρόπους επεξεργασίας, π.χ. κείμενο και εικόνα.

6.2.5 Παρουσίαση Αποτελεσμάτων

Αν και το ερευνητικό ενδιαφέρον εστιάζεται συνήθως στην αξιολόγηση των αποτελεσμάτων των συστημάτων ανάκτησης, στην περίπτωση των πραγματικών συστημάτων που είναι διαθέσιμα στο ευρύ κοινό μεγαλύτερη σημασία έχει ο τρόπος με τον οποίο παρουσιάζονται αυτά. Έτσι, οι πλέον συνήθεις τρόποι παρουσίασης των αποτελεσμάτων της ανάκτησης είναι οι ακόλουθοι:

- *Κατάταξη με βάση τη σχετικότητα* (Relevance-Ordered): Στην περίπτωση αυτή που είναι και η πλέον συνηθισμένη, τα αποτελέσματα παρουσιάζονται με βάση τη σχετικότητα που έχουν προς την εικόνα του ερωτήματος. Χρησιμοποιείται ένα μέτρο για τη σύγκριση, το οποίο και σε κάποιες περιπτώσεις επίσης περιλαμβάνεται. Τέτοια παραδείγματα απεικονίζονται στα Σχήματα 6.6 και 6.7.

- *Κατάταξη με βάση το χρόνο* (Time-Ordered): Στην περίπτωση αυτή και εφόσον οι εικόνες που περιλαμβάνονται στο σύστημα ενσωματώνουν μεταδεδομένα στα οποία υπάρχει η πληροφορία για το χρόνο λήψης, η παρουσίαση των αποτελεσμάτων γίνεται με αύξουσα ή φθίνουσα χρονολογική σειρά. Ο τρόπος αυτός συνηθίζεται σε συστήματα που περιέχουν προσωπικές συλλογές.

- *Ομαδοποιημένη παρουσίαση* (Clustered): Στην περίπτωση αυτή γίνεται ομαδοποίηση ή συσταδοποίηση των αποτελεσμάτων της ανάκτησης με βάση τα οπτικά ή/και τα μεταδεδομένα που συνοδεύουν τις εικόνες. Το κέντρο της συστάδας πολλές φορές χρησιμοποιείται για να αναπαραστήσει όλες τις εικόνες που ανήκουν σε αυτήν. Συνήθως αυτός ο τρόπος αναπαράστασης χρησιμοποιείται για να διευκολύνει την περιήγηση του χρήστη στα αποτελέσματα.

- *Ιεραρχική παρουσίαση* (Hierarchical): Στην περίπτωση αυτή τα μεταδεδομένα παρουσιάζονται με τη μορφή δέντρου, χρησιμοποιώντας κάποια γνωστή ιεραρχία, όπως είναι για παράδειγμα το Wordnet [68].

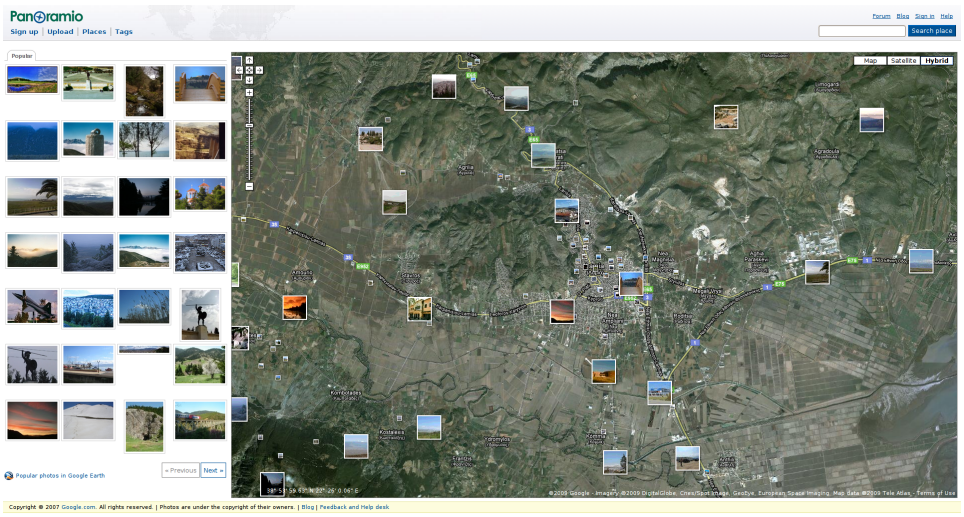
- *Παρουσίαση με βάση την τοποθεσία* (Geo-tag based): Στην περίπτωση αυτή τα αποτελέσματα παρουσιάζονται ταξινομημένα με βάση την απόστασή τους στο χώρο από την εικόνα ή την τοποθεσία του ερωτήματος. Ένα παράδειγμα ενός συστήματος ανάκτησης από την ιστοσελίδα Panoramio[8] που παρουσιάζει τα αποτελέσματα με βάση την απόσταση από την τοποθεσία του ερωτήματος απεικονίζεται στο Σχήμα 6.4. Τα αποτελέσματα της ανάκτησης εμφανίζονται πάνω σε έναν χάρτη.

- *Σύνθετη Παρουσίαση* (Composite): Στην περίπτωση αυτή τα αποτελέσματα παρουσιάζονται με σύνθεση δύο ή και περισσότερων από τους παραπάνω τρόπους.

6.2.6 Συστήματα Ανάκτησης Εικόνων

Προς το τέλος της προηγούμενης δεκαετίας αναπτύχθηκαν πολλά συστήματα ανάκτησης εικόνων με βάση το οπτικό τους περιεχόμενο. Τα περισσότερα από αυτά

[8] `http://www.panoramio.com`



Σχήμα 6.4: *Ένα παράδειγμα από την ιστοσελίδα Panoramio που παρουσιάζει τα αποτελέσματα με βάση την απόσταση από την τοποθεσία του ερωτήματος.*

έχουν πλέον εγκαταλειφθεί. Στη βιβλιογραφία υπάρχουν αρκετές μελέτες που παρουσιάζουν διεξοδικά τα συστήματα αυτά. Μια από τις πιο ολοκληρωμένες μελέτες για συστήματα που αναπτύχθηκαν την προηγούμενη δεκαετία και εως τα τέλη του 2002 είναι αυτή των Veltkamp και Tanase [222]. Σε αυτή περιγράφονται αναλυτικά 46 συστήματα ανάκτησης. Τα περισσότερα από αυτά δημιουργήθηκαν για ερευνητικούς σκοπούς αλλά αρκετά από αυτά είχαν τελικά και εμπορική εφαρμογή. Τα τελευταία χρόνια, όχι μόνο έχει σταματήσει να εξελίσσεται η συντριπτική πλειοψηφία των παλαιότερων συστημάτων, αλλά και δεν εμφανίζεται πλέον μεγάλος αριθμός από νέα συστήματα, όπως γινόταν στα τέλη της δεκαετίας του '90.

Από τα παλαιότερα συστήματα ανάκτησης εικόνας θα αναφερθούν στη συνέχεια κάποιες από τις πιο αξιοπρόσεκτες περιπτώσεις. Το *AMORE*[9] παρουσιάστηκε από τους Mukherjea et al. [143] και χρησιμοποιούσε οπτικά χαρακτηριστικά χρώματος, υφής και σχήματος και τα ερωτήματα πραγματοποιούνταν με λέξεις-κλειδιά ή/και εικόνες. Ένα από τα ιδιαίτερα χαρακτηριστικά του ήταν ότι προσπαθούσε να "μαντέψει" τη σημασιολογία των εικόνων με βάση το περιεχόμενο των ιστοσελίδων στις οποίες αυτές περιέχονται.

Μια ενδιαφέρουσα προσέγγιση για την ανάκτηση εικόνων με βάση το σχήμα των αντικειμένων που αυτές απεικονίζουν ήταν το *SQUID*[10]. Το σύστημα αυτό βασίστηκε στην css αναπαράσταση του σχήματος που πρότειναν οι Mokhtarian et al. [141], όπως ακριβώς και ο Περιγραφέας Σχήματος με βάση το Περίγραμμα του προτύπου MPEG-7. Δυστυχώς και για πρακτικούς λόγους, η ανάκτηση περιορίστηκε σε δυαδικές εικόνες που περιείχαν ένα μοναδικό αντικείμενο.

Το *QBIC*[11] αναπτύχθηκε από την IBM και πιο συγκεκριμένα από τους Niblack et al. [152]. Το σύστημα αυτό χρησιμοποιήθηκε και στο μουσείο Hermitage για αναζήτηση έργων τέχνης. Χρησιμοποιήθηκαν οπτικοί περιγραφής χρώματος, υφής και σχήματος, τόσο καθολικοί, όσο και από περιοχές. Και το σύστημα αυτό επέτρεπε

[9] `http://www.ccrl.com/amore/`

[10] `http://www.ee.surrey.ac.uk/Research/VSSP/imagedb/demo.html`

[11] `http://www.qbic.almaden.ibm.com/`

τη δημιουργία ερωτημάτων με τη χρήση λέξεων-κλειδιών ή φράσεων.

Το *ImageRover*[12] των Sclaroff et al. [181] χρησιμοποίησε πράκτορες για τη συλλογή εικόνων από το διαδίκτυο. Τα χαρακτηριστικά χαμηλού επιπέδου περιορίστηκαν στη χρήση του χρώματος. Το σύστημα αυτό υποστήριζε ερωτήματα με εικόνα, αλλά και με λέξεις-κλειδιά και ήταν από τα πρώτα που υποστήριξαν μεγάλο αριθμό εικόνων. Επιπρόσθετα, το ανάκτησης *WebSEEk*[13] αναπτύχθηκε από τους Smith και Chang [195] και είναι παρόμοιο με το ImageRover αν και προέρχεται από διαφορετική ερευνητική ομάδα. Πρόκειται για ένα παλιό σύστημα αναζήτησης, το οποίο συνδυάζει ανάλυση κειμένου και οπτικού περιεχομένου για εικόνες και βίντεο.

Το *Surfimage*[14] αναπτύχθηκε από την ομάδα της INRIA και συγκεκριμένα από τους Nastar et al.[151]. Το σύστημα αυτό δίνει ιδιαίτερη έμφαση στην ανάδραση σχετικότητας, προσπαθώντας να "μάθει" από τους χρήστες ποια από τα αποτελέσματα της ανάκτησης θεωρούν αυτοί ως σχετικά και ποια όχι. Επίσης, μπορεί κατ'επιλογή του χρήστη να συνδυάζει διάφορα χαρακτηριστικά χαμηλού επιπέδου και να επιλέγονται διάφορες μετρικές για τη σύγκριση των εικόνων.

Ένα από τα πιο γνωστά συστήματα που μάλιστα έχουν εφαρμοστεί στο TRECVID και συγκεκριμένα στη δοκιμασία της αναζήτησης χαρακτηριστικών καρέ σε βίντεο από δελτία ειδήσεων (όπως αναφέρθηκε στην Ενότητα 4.5) είναι το *MARVEL*[15] που αναπτύχθηκε από την ομάδα της IBM και συγκεκριμένα από την ομάδα του Smith [194]. Η μηχανή ανάκτησης του MARVEL ενσωματώνει πολυμεσική αναζήτηση με βάση το σημασιολογικό περιεχόμενο καθώς και τεχνικές αναζήτησης με βάση την ομιλία (στα βίντεο), το κείμενο, τα μεταδεδομένα και τα οπτικοακουστικά χαρακτηριστικά.

Το *VideoQ*[16] των Chang et al. [42] δίνει ιδιαίτερη έμφαση στους τρόπους με τους οποίους μπορεί ο χρήστης να θέτει τα ερωτήματά του. Έτσι, είναι εφικτός ο σχηματισμός οπτικών ερωτημάτων από σκίτσα που φτιάχνει ο χρήστης, στα οποία δύναται να εισάγει χαρακτηριστικά χρώματος, υφής και κίνησης, ερωτημάτων που σχηματίζονται από λέξεις-κλειδιά, σύνθετα ερωτήματα που συνδυάζουν τις δύο αυτές κατηγορίες και τέλος συνδυασμό από πολλά ερωτήματα προκειμένου να διευκολυνθεί η ανάκτηση βίντεο.

Η μηχανή αναζήτησης *Virage*[17] που αναπτύχθηκε από τους Bach et al. [7] παρέχει ένα ανοιχτό πλαίσιο στο οποίο μπορούν να κατασκευαστούν συστήματα ανάκτησης. Η μηχανή αυτή εκφράζει τα οπτικά χαρακτηριστικά των εικόνων ως "πρωταρχικά" χαρακτηριστικά ή "εξειδικευμένα" ανάλογα με το θεματικό πεδίο εφαρμογής τους. Στην πρώτη κατηγορία εντάσσονται χαρακτηριστικά όπως το χρώμα, η υφή και το σχήμα, ενώ στη δεύτερη τεχνικές όπως εντοπισμός προσώπου. Η βασική φιλοσοφία πίσω από την αρχιτεκτονική αυτή επιδιώκει έναν μετασχηματισμό από μια πλούσια αναπαράσταση των εικονοστοιχείων σε μία συμπαγή και πλούσια σημασιολογικά αναπαράσταση των οπτικών χαρακτηριστικών. Φυσικά, ο σχεδιασμός και η επιλογή των χαρακτηριστικών αυτών στην πράξη δεν είναι προφανής και περιορίζεται από ένα σύνολο από περιορισμούς που επιβάλλουν οι εφαρμογές στον πραγματικό κόσμο. Η

[12] `http://www.cs.bu.edu/groups/ivc/ImageRover/`

[13] `http://www.ctr.columbia.edu/webseek/`

[14] `http://www-roc.inria.fr/imedia/Articles/MM98/mm98.html`

[15] `http://mp7.watson.ibm.com/marvel/`

[16] `http://www.ctr.columbia.edu/VideoQ/`

[17] `http://www.virage.com/`

μηχανή του Virage παρέχει την αρχιτεκτονική που επιτρέπει στους ερευνητές να χρησιμοποιήσουν τα δικά τους χαρακτηριστικά για να επιλύσουν το δικό τους πρόβλημα ανάκτησης.

Η παρουσίαση της "πρώτης γενιάς" συστημάτων ανάκτησης εικόνων κλείνει με το *Blobworld*[18] των Carson et al. [38]. Το σύστημα αυτό επιδίωξε την ανάκτηση περιοχών εικόνων και χρησιμοποίησε χαρακτηριστικά χρώματος, υφής, σχήματος των περιοχών, αλλά και τη θέση τους στην εικόνα. Ο χρήστης επέλεγε μια κατηγορία στην οποία ήθελε να ανήκουν οι εικόνες του αποτελέσματος, με σκοπό να περιοριστεί ο αριθμός των εικόνων στον οποίο θα γινόταν η αναζήτηση. Στη συνέχεια και από την εικόνα του ερωτήματος ο χρήστης επέλεγε μια περιοχή (blob) και καθόριζε πόσο σημαντικά ήταν τα χαρακτηριστικά της περιοχής αυτής σε σχέση με ολόκληρη την εικόνα και έπειτα πόσο σημαντικό είναι κάθε χαρακτηριστικό σε σχέση με τα υπόλοιπα. Για το σχηματισμό του τελικού ερωτήματος ήταν δυνατόν να συμπεριληφθούν περισσότερες από μία εικόνες.

Ένα σύστημα ανάκτησης το οποίο έχει αντέξει στο χρόνο και είναι ακόμη διαθέσιμο στο διαδίκτυο είναι το *Ikona*[19], το οποίο έχει υλοποιηθεί από την ομάδα του INRIA και πιο συγκεκριμένα από τους Boujemaa και Nastar [22] αρχικά και κατόπιν δέχτηκε πολλές προσθήκες και βελτιώσεις. Σε αυτό χρησιμοποιούνται για τη σύγκριση οπτικά χαρακτηριστικά χρώματος, σχήματος και υφής. Υπάρχει επίσης η δυνατότητα ανάκτησης με χαρακτηριστικά εξαγόμενα από περιοχές της εικόνας, ανάκτηση με ανάδραση σχετικότητας, ανάκτηση με τοπικά χαρακτηριστικά των εικόνων καθώς και ανάκτηση τρισδιάστατων μοντέλων.

Δύο από τα πιο ενδιαφέροντα συστήματα ανάκτησης που εξελίσσονται σήμερα προέρχονται από την ιδιαίτερα ενεργή στον χώρο ομάδα του πανεπιστημίου της Οξφόρδης. Σε πλήρη εφαρμογή υπάρχουν δύο συστήματα, το ένα από τα οποία είναι το *Video Google*[20] από τους Sivic και Zisserman [190], στο οποίο ο χρήστης διαλέγει ένα αντικείμενο από κάποιο χαρακτηριστικό καρέ μιας ταινιας. Το σύστημα πραγματοποιεί αναζήτηση στα υπόλοιπα καρέ της ταινίας, αναζητώντας το αντικείμενο αυτό. Το έτερο σύστημα της ομάδας αυτής, γνωστό ως Oxford Buildings Dataset[21] από τους Philbin et al. [171] κάνει ανάκτηση εικόνων που περιέχουν κτίρια από την πόλη της Οξφόρδης, μέσα σε μια μεγάλη συλλογή εικόνων. Το σύστημα αυτό βασίζεται εν μέρει στο Video Google, αλλά έχουν γίνει μεγάλες βελτιώσεις στον έλεγχο χωρικής συνεκτικότητας, τη δεικτοδότηση και την ταχύτητα.

Ένα άλλο σύγχρονο σύστημα ανάκτησης είναι το *SIMPLIcity*[22] (Semantics-sensitive Integrated Matching for Picture LIbraries), το οποίο αναπτύχθηκε από τους Wang et al [227] και είναι, όπως λέει και το όνομα του, ευαίσθητο σε σημασιολογικές έννοιες. Το σύστημα αυτό συνδυάζει μεταδεδομένα της εικόνας με τα οπτικά της χαρακτηριστικά, για την εξαγωγή του αποτελέσματος της ανάκτησης. Το σύστημα αυτό συνδέεται άμεσα και με το *Alipr*[23] των Li και Wang [117], το οποίο συνδυάζει επίσης μεταδεδομένα των εικόνων και δίνει στο χρήστη την επιλογή ανάκτησης "σχετικών" (related) ή "παρόμοιων" (similar) εικόνων. Στην πρώτη

[18] `http://elib.cs.berkeley.edu/photos/blobworld/`

[19] `http://www-roc.inria.fr/cgi-bin/imedia/circario.cgi/demos`

[20] `http://www.robots.ox.ac.uk/~vgg/research/vgoogle/index.html`

[21] `http://www.robots.ox.ac.uk/~vgg/research/oxbuildings/index.html`

[22] `http://wang14.ist.psu.edu/cgi-bin/zwang/regionsearch_show.cgi`

[23] `http://alipr.com/`

περίπτωση, η αναζήτηση γίνεται χρησιμοποιώντας ταμπέλες των εικόνων (με την μορφή κειμένου), ενώ στη δεύτερη με οπτικά χαρακτηριστικά. Στοχεύει στην αυτόματη γλωσσική δεικτοδότηση/ταξινόμηση των εικόνων με μεταδεδομένα και επιδιώκει μια στατιστική προσέγγιση προς αυτή την κατεύθυνση, όπως μαρτυρά και το όνομά του (τα αρχικά Alipr σημαίνουν Automatic Linguistic Indexing and Picture Retrieval). Ένα ακόμα σύστημα που συνδυάζει παρόμοιες τεχνολογίες και δανείζεται το ίδιο γραφικό περιβάλλον είναι το *CLUE*[24] των Chen et al. [45]. Τα τρία αυτά συστήματα έχουν αναπτυχθεί από την ίδια ερευνητική ομάδα.

Η κολοσσιαία ανάπτυξη των φωτογραφικών συλλογών που συναντώνται στο διαδίκτυο ενέπνευσε τους Torralba et al. για την ανάπτυξη ενός συστήματος ανάκτησης το οποίο επιθυμούσαν να περιέχει όσο το δυνατόν περισσότερες εικόνες. Το σύστημα *80 Million Tiny Images*[25] [212] περιέχει μια τεράστια συλλογή από πολύ μικρές εικόνες (32×32 εικονοστοιχείων), όπως μαρτυρά και το όνομά του. Στις περιοχές κάθε εικόνας αντιστοιχούνται λέξεις που τις περιγράφουν με όσο το δυνατόν καλύτερο τρόπο.

Ένα ακόμα σύστημα ανάκτησης που μπορεί να βρει κανείς online στο διαδίκτυο είναι το *Accio!*[26] των Rahmani et al. [175]. Σε αυτό, ο χρήστης δίνει στο σύστημα κάποιες εικόνες που περιέχουν το αντικείμενο προς ανάκτηση και κάποιες που δεν το περιέχουν, και με την εφαρμογή μιας τεχνικής κατάτμησης από σημεία ενδιαφέροντος και ενός αλγορίθμου μάθησης από πολλαπλές εικόνες ταξινομούνται οι εικόνες και επιστρέφονται στον χρήστη οι κοντινότερες ως προς το ερώτημα του.

Οι *ιατρικές εικόνες* είναι ένας τομέας στον οποίο η χρήση συστημάτων ανάκτησης με βάση το οπτικό περιεχόμενο είναι ιδιαίτερα διαδεδομένη. Τα αποτελέσματα στον τομέα αυτόν είναι εντυπωσιακά και ιατρικά κέντρα σε όλο τον κόσμο χρησιμοποιούν καθημερινά την τεχνολογία αυτή για αναγνώριση παθήσεων από ακτινογραφίες, μαγνητικές τομογραφίες ή εγκεφαλογραφήματα [207]. Τα συστήματα αυτά δεν διαφέρουν σε πολλά σημεία από τα συστήματα ανάκτησης εικόνων γενικού περιεχομένου, αλλά αποτελούν στις περισσότερες περιπτώσεις επεκτάσεις τους. Στην κατηγορία αυτή εντάσσεται το σύστημα των Lehmann et al. [112], το οποίο είναι μια επέκταση του Blobworld.

Τέλος, έχουν αρχίσει και κάνουν την εμφάνισή τους και συστήματα ανάκτησης που ταυτόχρονα έχουν και άλλες λειτουργίες και δεν περιορίζονται μόνο στην εμφάνιση παρόμοιων αποτελεσμάτων με τις εικόνες των ερωτημάτων. Το *ViRaL*[27] των Kalantidis et al. [96] στοχεύει στον προσδιορισμό της τοποθεσίας στην οποία έχει ληφθεί η εικόνα του ερωτήματος. Αυτό το επιτυγχάνει εκμεταλλευόμενο τις γεωγραφικές συντεταγμένες που περιέχονται στα μετα-δεδομένα των εικόνων που έχει αποθηκευμένες, αλλά και τις λέξεις-κλειδιά με τις οποίες έχει σχολιαστεί. Αρχικά γίνεται σύγκριση των οπτικών χαρακτηριστικών της εικόνας του ερωτήματος και έπειτα, με βάση τα μετα-δεδομένα των πρώτων εικόνων που ανακτήθηκαν γίνεται μια εκτίμηση της τοποθεσίας της εικόνας του ερωτήματος, η απεικόνισή της σε χάρτη, της πόλης στην οποία έχει ληφθεί και ο προσδιορισμός του μνημείου/κτιρίου που τυχόν περιέχει.

[24] `http://wang14.ist.psu.edu/cgi-bin/yixin/spectralsearch_show.cgi`

[25] `http://people.csail.mit.edu/torralba/tinyimages/`

[26] `http://www.cs.wustl.edu/accio/`

[27] `http://www.image.ntua.gr/iva/tools/viral`

6.3 Σχετικές Εργασίες

Το ερευνητικό πεδίο της ανάκτησης εικόνων με βάση το περιεχόμενο έχει τραβήξει σε μεγάλο βαθμό το ερευνητικό ενδιαφέρον, με συνέπεια να υπάρχει πλήθος σχετικών εργασιών. Σαν συνέπεια αυτού, πολλές είναι οι εργασίες που κάνουν μια σύνοψη των ερευνητικών προσπαθειών. Συνοπτικά, αξίζει να αναφερθούν αυτή των Rui et al. [177], που καλύπτει αρκετό κομμάτι της έρευνας εώς το 1997, η εργασία των Smeulders et al. [193], η οποία θεωρεί τα χρόνια εώς το 2000 ως τα "πρώιμα χρόνια" της ανάκτησης με βάση το περιεχόμενο και συνοψίζει ένα πλήθος από εργασίες και τέλος, την εργασία των Datta et al. [56], η οποία είναι ενημερωμένη με πιο σύγχρονες τεχνικές και εφαρμογές. Καθώς η παρούσα εργασία εστιάζει στην αναπαράσταση και το ταίριασμα των εικόνων, ιδιαίτερη βαρύτητα δίνεται στην παρουσίαση της σχετικής βιβλιογραφίας στο χώρο αυτό. Έτσι στη συνέχεια παρουσιάζονται πολλές και σημαντικές εργασίες οι οποίες δίνουν έμφαση στις τεχνικές που χρησιμοποιούν αποκλειστικά τα οπτικά χαρακτηριστικά των εικόνων.

Η απλούστερη αλλά ταυτόχρονα και λιγότερο αποδοτική προσέγγιση είναι η εξαγωγή των οπτικών χαρακτηριστικών που θα χρησιμοποιηθούν για την περιγραφή της εικόνας να γίνει καθολικά. Έτσι το όχι προφανές βήμα της επιλογής περιοχών της εικόνας από τις οποίες θα γίνει η εξαγωγή των περιγραφών παραλείπεται.

Οι Anthoine et al. [2], όρισαν ένα μέτρο για την ομοιότητα μεταξύ των εικόνων του συστήματος. Το μέτρο αυτό βασίζεται στην KL απόκλιση (KL divergence) για τη σύγκριση αραιών πολυκλιμακωτών περιγραφών εικόνων που βασίζονται στα wavelets. Η μέθοδος αυτή επέδειξε αρκετά καλή απόδοση στο πρόβλημα της ανάκτησης και πρέπει να σημειωθεί ότι απεδείχθη εύρωστη σε απλούς γεωμετρικούς μετασχηματισμούς, διατηρώντας χαμηλό υπολογιστικό κόστος και ευκολία στην υλοποίηση. Οι Yang et al. [234], εξάγουν ταυτόχρονα καθολικά και τοπικά χαρακτηριστικά των εικόνων, υπολογίζοντας προσαρμοζόμενα ιεραρχικά γεωμετρικά κέντρα της εικόνας, τα οποία ονομάζουν "γειτονιές". Αποδεικνύουν ότι τα κέντρα αυτά δεν είναι ευαίσθητα σε μεταβολές της φωτεινότητας και της κλίμακας. Τη μέθοδο αυτή την εφαρμόζουν σε πρόβλημα ανάκτησης "σχεδόν ίδιων" εικόνων (near-duplicate), αλλά και σε κλασσικό πρόβλημα ανάκτησης εικόνων.

Μια άλλη δημοφιλής κατηγορία τεχνικών ανάκτησης είναι η εξαγωγή περιγραφών από περιοχές της εικόνας, σε μια προσπάθεια να εξαχθούν τοπικά χαρακτηριστικά της εικόνας και να επιτύχουν καλύτερη απόδοση από τις τεχνικές που χρησιμοποιούν καθολικά χαρακτηριστικά, μιας και είναι πολύ συνηθισμένο φαινόμενο δύο εικόνες να έχουν παρόμοια καθολικά χαρακτηριστικά, αλλά σε τοπικό επίπεδο να είναι πολύ διαφορετικές. Τυπικά, η γνώση για τις ιδιότητες των περιοχών της εικόνας κωδικοποιείται με ένα οπτικό λεξικό και κάθε μία από τις περιοχές αντιστοιχείται σε μια οπτική λέξη. Χρησιμοποιούνται, έτσι, τεχνικές που βασίζονται σε μοντέλα τύπου bag-of-words, παρόμοιες με αυτές που έχουν παρουσιαστεί στην Ενότητα 4.4. Πέρα από αυτές τις τεχνικές που τυπικά βασίζονται σε περιοχές εικόνων που έχουν προκύψει με κάποιο τρόπο όπως ένα πλέγμα ή έναν αλγόριθμο κατάτμησης, πολλές σύγχρονες τεχνικές χρησιμοποιούν τις περιοχές γύρω από σημεία, τα οποία προσδιορίζονται με τρόπο ώστε να είναι αμετάβλητα κάτω από συγκεκριμένους γεωμετρικούς και άλλους μετασχηματισμούς. Αν και τέτοιες τεχνικές εξαγωγής χαρακτηριστικών δεν εμπίπτουν στο ερευνητικό πεδίο της παρούσας εργασίας, οι αλγόριθμοι που χρησιμοποιούνται εμφανίζουν πολλές ομοιότητες με τους αυτούς των μοντέλων bag-of-words. Αν και αυτές οι τεχνικές πρωτοεφαρμόστηκαν σε προβλήματα αναγνώρισης αντικει-

μένων, γρήγορα επεκτάθηκαν και σε προβλήματα ανάκτησης εικόνων. Για το λόγο αυτό θα παρουσιαστούν ορισμένες χαρακτηριστικές τεχνικές της κατηγορίας αυτής.

Οι Philbin et al. [172] διερευνούν τεχνικές για την αντιστοίχηση κάθε περιοχής της εικόνας σε ένα σύνολο με βάρη από οπτικές λέξεις, επιτρέποντας την περίληψη από χαρακτηριστικά που συνήθως χάνονται κατά τη φάση της κβαντοποίησης. Το σύνολο από τις οπτικές λέξεις σχηματίζεται με την επιλογή λέξεων βασισμένων στην γειτνίαση στον χώρο των περιγραφέων. Τα πειραματικά τους αποτελέσματα αποδεικνύουν τη βελτίωση της απόδοσης και ιδιαίτερα στην περίπτωση που το οπτικό λεξικό έχει κατασκευαστεί με διαφορετικό σύνολο δεδομένων από αυτό που χρησιμοποιείται για την αξιολόγηση. Οι Duygulu et al. [62] προτείνουν μια τεχνική όπου οι εικόνες κατατέμνονται σε περιοχές και οι περιοχές ταξινομούνται σε οπτικές λέξεις χρησιμοποιώντας ένα πλήθος από χαρακτηριστικά. Στη συνέχεια κάνουν μια απεικόνιση ανάμεσα σε οπτικές λέξεις και έννοιες χρησιμοποιώντας την μέθοδο Expectation-Maximization. Έτσι δημιουργείται η περιγραφή της εικόνας. Οι Li και Wang [116], προτείνουν μια τεχνική για τη δεικτοδότηση συλλογών εικόνων. Για την εξαγωγή των χαρακτηριστικών από εικόνες χρησιμοποίησαν wavelets. Έπειτα, με τη χρήση κρυφών μαρκοβιανών μοντέλων έγινε η αντιστοίχηση των χαρακτηριστικών με τις λέξεις-κλειδιά που περιγράφουν τις εικόνες. Το βασικό προτέρημα της μεθόδου των αποδείχτηκε η δυνατότητά να εκπαιδευθούν τα μοντέλα για κάθε έννοια ανεξάρτητα και για πολύ μεγάλο αριθμό από έννοιες. Οι Chum και Matas [49] πρότειναν μια μέθοδο εξόρυξης δεδομένων η οποία χρησιμοποιείται για να βρει συστάδες από εικόνες που έχουν μερική χωρική επικάλυψη. Η μέθοδος αυτή δε χρειάζεται επίβλεψη, εφαρμόζεται σε μεγάλες συλλογές εικόνων και πέρα από την ανακάλυψη συστάδων από περιοχές που εμφανίζουν παρόμοιες χωρικές ιδιότητες, χρησιμοποιείται και αυτή για ανάκτηση σχεδόν ίδιων εικόνων. Στηρίζεται στον min-Hash αλγόριθμο, ο οποίος γενικά χρησιμοποιείται για την εύρεση παρόμοιων ζευγών εικόνων. Επιπρόσθετα, η τεχνική που προτείνουν οι Jegou et al. [89] χρησιμοποιεί μια περιγραφή της εικόνας που βασίζεται σε χρήση οπτικού λεξικού. Στη συνέχεια επιδιώκει να δημιουργήσει μια πιο ακριβή περιγραφή με τη χρήση της τεχνικής Hamming embedding και χαλαρών γεωμετρικών περιορισμών. Οι περιορισμοί αυτοί ενσωματώνονται στο inverted file και χρησιμοποιούνται για όλες τις εικόνες της συλλογής. Καταλήγουν να βελτιώσουν αισθητά την απόδοση της ανάκτησης, ιδιαίτερα στην περίπτωση μεγάλων συλλογών εικόνων. Το αξιοσημείωτο είναι ότι τελικά δεν έχουν αρνητική επίπτωση στο χρόνο που απαιτείται για την εξαγωγή των αποτελεσμάτων του ερωτήματος.

Οι Gemert et al. [220] επεκτείνουν το μοντέλο bag-of-words εισάγοντας ασάφεια στη διαδικασία της μοντελοποίησης. Η ασάφεια αυτή επιτεύχθηκε με τη χρήση τεχνικών βασισμένων σε εκτίμηση πυκνότητας πυρήνα και συγκεκριμένα μέσω πυρήνων γκαουσιανών. Η τεχνική αυτή υποφέρει λιγότερο από τις κλασσικές τεχνικές στο φαινόμενο της "κατάρας των μεγάλων διαστάσεων" (curse of dimensionality), επιτυγχάνοντας καλά αποτελέσματα σε χώρους μεγάλης διάστασης. Οι Philbin et al. [171] πρότειναν μια τεχνική που σκοπεύει στην ανάκτηση εικόνων που περιέχουν κάποιο συγκεκριμένο αντικείμενο σε πολύ μεγάλες συλλογές. Το αντικείμενο του ερωτήματος επιλέγεται ως μια περιοχή από μια εικόνα. Η τεχνική αυτή χρησιμοποίησε τοπικά χαρακτηριστικά και διερεύνησε διάφορες μεθόδους για τη βελτίωση της ποιότητας του οπτικού λεξικού, αλλά και της αντιστοιχίας οπτικών λέξεων με περιοχές της εικόνας. Οι Chum και Matas [50] χρησιμοποίησαν τα MSER χαρακτηριστικά [133] για την εξαγωγή των περιοχών και προκειμένου να επιτύχουν ταίριασμα μεταξύ των περιοχών αυτών χρησιμοποίησαν την τεχνική geometric hashing. Η αναπαράστα-

ση των χαρακτηριστικών της εικόνας αποδείχτηκε ότι δεν είναι ευαίσθητη σε αλλαγές της φωτεινότητας και της οπτικής γωνίας. Οι Ke et al. [99] αντιμετώπισαν το πρόβλημα της ανάκτησης σχεδόν ταυτόσημων εικόνων αλλά και της ανάκτησης υποπεριοχών της εικόνας. Χρησιμοποίησαν μια προσέγγιση που βασίστηκε σε αναπαράσταση των μερών μιας εικόνας. Εφάρμοσαν το πρόβλημά τους σε ανίχνευση παραβίασης πνευματικών δικαιωμάτων έργων τέχνης, επιτυγχάνοντας υψηλή ακρίβεια στην ανάκτηση. Ωστόσο αντιμετώπισαν τον περιορισμό ότι λόγω της φύσης των PCA-SIFT χαρακτηριστικών που χρησιμοποίησαν και της ιδιαιτερότητας του προβλήματος, κάποιες εικόνες θεωρήθηκαν λανθασμένα ότι ήταν αντίγραφα της εικόνας του ερωτήματος, απλά και μόνο επειδή περιείχαν το ίδιο ορόσημο. Τέλος, οι Schmidt και Mohr [180] χρησιμοποίησαν τοπικά χαρακτηριστικά σε γκρι εικόνες και προσπάθησαν να τα αξιοποιήσουν στο πρόβλημα της ανάκτησης. Στα διανύσματα αναπαράστασης ενσωματώνει και πληροφορία για τη χωρική διάταξη των χαρακτηριστικών και των γειτονικών τους σημείων. Η τεχνική τους έδειξαν ότι δεν επηρεάζεται από απλούς γεωμετρικούς μετασχηματισμούς.

6.4 Ανάκτηση Εικόνων με χρήση Οπτικού Θησαυρού

Στην παρούσα Ενότητα αρχικά περιγράφεται ο τρόπος με τον οποίο σχηματίζεται το διάνυσμα αναπαράστασης που θα χρησιμοποιηθεί για να εκφράσει το οπτικό, αλλά και σημασιολογικό περιεχόμενο των εικόνων. Ο αλγόριθμος σχηματισμού του διανύσματος αυτού θα βασιστεί στις ιδέες που παρουσιάστηκαν στο Κεφάλαιο 4. Έπειτα, παρουσιάζεται η διαδικασία με την οποία πραγματοποιείται τελικά η ανάκτηση μέσα σε συλλογές εικόνων.

6.4.1 Κατασκευή Διανύσματος Αναπαράστασης

Στο παρόν Κεφάλαιο θα χρησιμοποιηθεί η ιδέα της αναπαράστασης του οπτικού περιεχομένου μιας εικόνας P_i με τη χρήση ενός διανύσματος αναπαράστασης M_i, το οποίο και εκφράζει τη σχέση της με έναν οπτικό θησαυρό περιοχών. Η γενική μορφή του διανύσματος αναπαράστασης περιγράφεται από την (4.9). Στο παρόν πρόβλημα της ανάκτησης εικόνων, ο οπτικός θησαυρός κατασκευάζεται με τη χρήση των εικόνων της διαθέσιμης συλλογής. Έπειτα, για κάθε εικόνα εξάγεται ένα διάνυσμα αναπαράστασης. Για λόγους που θα γίνουν κατανοητοί στη συνέχεια, η διαδικασία με την οποία κατασκευάζεται το διάνυσμα αναπαράστασης τροποποιείται.

Αρχικά και σε κάθε εικόνα γίνεται κατάτμηση, με τη χρήση της μεθόδου των Avrithis et al. [5], όπως ακριβώς έγινε και στα προηγούμενα Κεφάλαια. Έτσι, για μια εικόνα P_i προκύπτει ένα σύνολο από περιοχές R_i. Έστω N_i ο αριθμός των περιοχών της εικόνας αυτής και r_{ij} η j-οστή περιοχή της.

Θεωρώντας ότι έχει κατασκευαστεί ένας οπτικός θησαυρός $T = \{w_i\}$, που αποτελείται από N_T τύπους περιοχής, ακολουθώντας τη διαδικασία που έχει περιγραφεί αναλυτικά στην Ενότητα 4.6.3, η διαδικασία κατασκευής του διανύσματος αναπαράστασης ξεκινά με την επιλογή των K-κοντινότερων τύπων περιοχής για κάθε μία από τις περιοχές της εικόνας. Για το σκοπό αυτό, για κάθε περιοχή r_{ij} ορίζεται πρώτα το διατεταγμένο σύνολο $\mathcal{W}_{ij}$ το οποίο περιέχει όλους τους τύπους περιοχής του οπτικού θησαυρού T, διατεταγμένους ως προς την απόστασή τους d_{ij} από την περιοχή

αυτή, ως

$$\mathcal{W}_i = \{w_{ij} \mid \forall k, l \leq N_T, k \leq l : w_{ik} \leq w_{il}\} \ . \tag{6.1}$$

Στη συνέχεια, για κάθε περιοχή επιλέγονται οι K κοντινότεροι τύποι περιοχής, ως τα K πρώτα στοιχεία του διατεταγμένου σύνολου $\mathcal{W}_i$. Έτσι, για κάθε περιοχή ορίζεται ένα σύνολο από τους κοντινότερους τύπους περιοχής ως

$$\mathcal{W}_i^K = \{w_{ij} : j \leq K\} \ . \tag{6.2}$$

Αφού επιλεχθούν οι κοντινότεροι τύποι περιοχής για κάθε μία από τις περιοχές της εικόνας, το σύνολο των περιοχών που θα χρησιμοποιηθούν για την κατασκευή του διανύσματος αναπαράστασης $\hat{m}_i$ προκύπτει σαν η ένωση W^K των διατεταγμένων συνόλων $\mathcal{W}_i^K$

$$W^K = \bigcup_i \mathcal{W}_i^K \ . \tag{6.3}$$

Ο ορισμός του συνόλου W^K με τη χρήση της πράξης της ένωσης συνόλων εξασφαλίζει ότι το σύνολο αυτό θα περιλαμβάνει κάθε τύπο περιοχής μία φορά, καθώς είναι σύνηθες να μοιάζουν περισσότερες από μία περιοχές μιας εικόνας με τον ίδιο τύπο περιοχής. Έτσι το σύνολο W^K περιέχει τους κοντινότερους τύπους περιοχής του οπτικού θησαυρού ως προς τις περιοχές της εικόνας. Με τη βοήθεια του συνόλου αυτού είναι πλέον δυνατόν να κατασκευαστεί το διάνυσμα αναπαράστασης, το οποίο θα περιέχει τις ελάχιστες αποστάσεις όλων των περιοχών της εικόνας ως προς τους επιλεχθέντες τύπους περιοχής που περιέχονται στο σύνολο W^K. Έτσι, το διάνυσμα αναπαράστασης $\hat{m}_i$ που περιγράφει με μοναδικό τρόπο το οπτικό περιεχόμενο της εικόνας P_i ως προς τον οπτικό θησαυρό T κατασκευάζεται ως

$$\hat{m}_i = \{\hat{m}_i(1)\ \hat{m}_i(2)\ \ldots\ \hat{m}_i(N_T)\} \ , \tag{6.4}$$

όπου N_T είναι ο αριθμός των τύπων περιοχής. Ως $m_i(j)$ ορίζεται η ελάχιστη απόσταση από όλες τις περιοχές της εικόνας P_i και υπολογίζεται ως

$$\hat{m}_i(j) = \begin{cases} min\{d(f(w_{ij}), f(r_{ij}))\} & \text{, αν } w_{ij} \in W^K \\ 0 & \text{, αλλιώς} \end{cases} \ , \tag{6.5}$$

όπου είναι φανερή η ομοιότητά της ως προς την 4.10 που χρησιμοποιήθηκε στο Κεφάλαιο 4. Υπενθυμίζεται ότι ως $f(\bullet)$ ορίζεται το διάνυσμα χαρακτηριστικών μιας περιοχής ή ενός τύπου περιοχής.

Η μέθοδος που ακολουθήθηκε για την κατασκευή του διανύσματος αναπαράστασης m_i περιέχει ένα ενδιάμεσο βήμα σε σχέση με αυτή του Κεφαλαίου 4. Όπως θα φανεί και στα πειραματικά αποτελέσματα της Ενότητας 6.5, στην περίπτωση που το διάνυσμα αναπαράστασης χρησιμοποιεί όλους τους τύπους περιοχής του οπτικού θησαυρού, η ανάκτηση γίνεται πολύ δύσκολη εώς αδύνατη. Στην περίπτωση της ανίχνευσης εννοιών έχει γίνει σαφές ότι η χρήση ταξινομητή βασισμένου σε μηχανική μάθηση αναθέτει στην ουσία βάρη στους διάφορους τύπους περιοχής. Έτσι, αυτοί που δεν έχουν ιδιαίτερη "σημασία" όσον αφορά την ταξινόμηση, τελικά θα λάβουν μικρό βάρος και θα αγνοηθούν. Ωστόσο, στην περίπτωση της ανάκτησης, όλοι οι τύποι περιοχής έχουν ίδιο βάρος και άρα η ύπαρξη όλων των τύπων περιοχής έστω και με μικρή τιμή θα λειτουργούσε ως "θόρυβος" και θα δυσχέραινε τη σωστή ανάκτηση των εικόνων.

Σχήμα 6.5: *Οι 2 ($K = 2$) κοντινότεροι τύποι περιοχής για κάθε μία από τις περιοχές της εικόνας στα αριστερά.*

Προκειμένου να ερμηνευθεί το γιατί επιλέγονται τελικά K τύποι περιοχής για κάθε περιοχή της εικόνας και όχι ένας μοναδικός, μπορεί να δοθεί και μια σημασιολογική ερμηνεία. Έχει παρατηρηθεί ότι πολλές έννοιες υψηλού επιπέδου αντιστοιχούν σε παραπάνω από έναν τύπους περιοχής. Έτσι αν για παράδειγμα δύο τύποι περιοχής μοιάζουν με την έννοια *άμμος* και στο διάνυσμα αναπαράστασης λαμβανόταν υπόψη μονάχα ο κοντινότερος τύπος περιοχής, τότε για δύο εικόνες που περιέχουν την έννοια αυτή και οι αντίστοιχες περιοχές τους αντιστοιχούν η μια στον πρώτο και η άλλη στον δεύτερο τύπο, λανθασμένα δεν θα είχαν καμία ομοιότητα. Μειώνεται με αυτόν τον τρόπο το σφάλμα της κβαντοποίησης. Στο Σχήμα 6.5 απεικονίζεται μια εικόνα και διακρίνονται οι επιμέρους περιοχές της όπως έχουν προκύψει έπειτα από κατάτμηση. Σε κάθε εικόνα έχουν αντιστοιχηθεί οι δύο κοντινότεροι τύποι περιοχής.

Όσον αφορά τώρα κάποιες λεπτομέρειες ως προς την υλοποίηση, θα πρέπει να επισημανθεί οτι για τον υπολογισμό των αποστάσεων μεταξύ δυο διανυσμάτων χαρακτηριστικών, χρησιμοποιούνται βάρη για κάθε περιγραφέα. Το βάρος που χρησιμοποιείται για κάθε περιγραφέα ισούται με την μέγιστη δυνατή απόσταση μεταξύ δύο ίδιων περιγραφέων ανάμεσα σε όλες τις περιοχές των εικόνων της συλλογής. Έτσι, η απόσταση μεταξύ δύο διανυσμάτων χαρακτηριστικών f_i και f_j υπολογίζεται ως

$$d_{ij} = d(f_i, f_j) = \frac{1}{d^k_{max}} d^k(D^k_i, D^k_j) \ , \tag{6.6}$$

όπου d^k_{max} είναι η μέγιστη απόσταση για τον k-οστό περιγραφέα, D^k_i, D^k_j ο k-οστός περιγραφέας για τις δύο περιοχές και $d^k(\bullet)$ η συνάρτηση που υπολογίζει την απόσταση για τον k-οστό περιγραφέα.

Επιλέχθηκε η χρήση της Ευκλείδειας απόστασης γιατί αφενός είναι μία από τις πλέον συνήθεις στην περιοχή της ανάκτησης εικόνων, δίνοντας ικανοποιητικά αποτελέσματα σε πλήθος εφαρμογών και αφετέρου ο υπολογίσμός της είναι απλός και πολύ γρήγορος. Τα διανύσματα αναπαράστασης των εικόνων της συλλογής αποθηκεύονται σε βάση. Σε περίπτωση που κριθεί απαραίτητο, είναι δυνατόν να αποθηκευθούν για κάθε εικόνα μόνο οι τύποι περιοχής και οι αντίστοιχοι βαθμοί βεβαιότητας. Έτσι αντί για ένα N_T-διάστατο διάνυσμα για μια εικόνα, αποθηκεύεται ένα διάνυσμα διάστασης το πολύ $2 * K$, κάτι αρκετά αποδοτικό στην περίπτωση που ο αριθμός K των κοντινότερων τύπων περιοχής για κάθε περιοχή της εικόνας είναι αρκετά μικρότερος από το μέγεθος N_T του οπτικού θησαυρού.

6.4.2 Ανάκτηση Εικόνων

Όπως επισημάνθηκε στην Ενότητα 6.4.1, για τον υπολογισμό της απόστασης μεταξύ δύο εικόνων, υπολογίζεται η Ευκλείδεια απόσταση μεταξύ των διανυσμάτων αναπαράστασής τους. Όταν γίνεται ένα ερώτημα ανάκτησης, εάν η εικόνα είναι εικόνα της συλλογής, φορτώνεται από την βάση το διάνυσμα αναπαράστασης και υπολογίζεται η απόσταση του από όλα τα αντίστοιχα διανύσματα όλων των εικόνων της συλλογής. Οι αποστάσεις ταξινομούνται και στο χρήστη του συστήματος επιστρέφονται ως παρόμοιες οι εικόνες που έχουν την μικρότερη απόσταση από την αρχική εικόνα.

Μερικά παραδείγματα αναζήτησης φαίνονται στα Σχήματα 6.6 και 6.7. Στο πάνω μέρος εμφανίζεται η εικόνα του ερωτήματος ακολουθούμενη από τις εικόνες που επιστρέφονται από το σύστημα ταξινομημένες από τα αριστερά προς τα δεξιά και από πάνω προς τα κάτω. Για πρακτικούς λόγους, στα Σχήματα απεικονίζονται οι κοντινότερες 12 εικόνες. Είναι φανερό ότι οι πρώτες εικόνες που επιστρέφονται από το σύστημα περιέχουν πράγματι τις έννοιες *ηλιοβασίλεμα* και *χιόνι*, που απεικονίζονται στην εικόνα του ερωτήματος. Αν παρατηρήσει κανείς προσεκτικά όλες τις εικόνες του αποτελέσματος, είναι προφανές ότι σε περίπτωση που γινόταν ανάκτηση με χρήση καθολικών χαρακτηριστικών, οι εικόνες που έχουν επιστραφεί στην 7η και 8η θέση στην περίπτωση του *χιονιού* και στην 7η και 10η θέση στην περίπτωση του *ηλιοβασιλέματος* δε θα είχαν ανακτηθεί. Αυτό θα συνέβαινε καθώς τα καθολικά χαρακτηριστικά θα επηρεάζονταν από π.χ. στην περίπτωση της 7ης εικόνας του *ηλιοβασιλέματος* από το γεγονός ότι η *θάλασσα* καταλαμβάνει μεγάλη περιοχή της εικόνας, με χρώμα που σπάνια συναντάται σε εικόνες *ηλιοβασιλέματος*.

Έτσι, η ανάκτηση με την αναπαράσταση της εικόνας με τη χρήση της προτεινόμενης τεχνικής φαίνεται ότι είναι πιο αποδοτική σε έννοιες όπως αυτές που εμπίπτουν στην κατηγορία των "υλικών", όπως και ορίστηκαν στο Κεφάλαιο 1.

6.5 Πειραματικά Αποτελέσματα

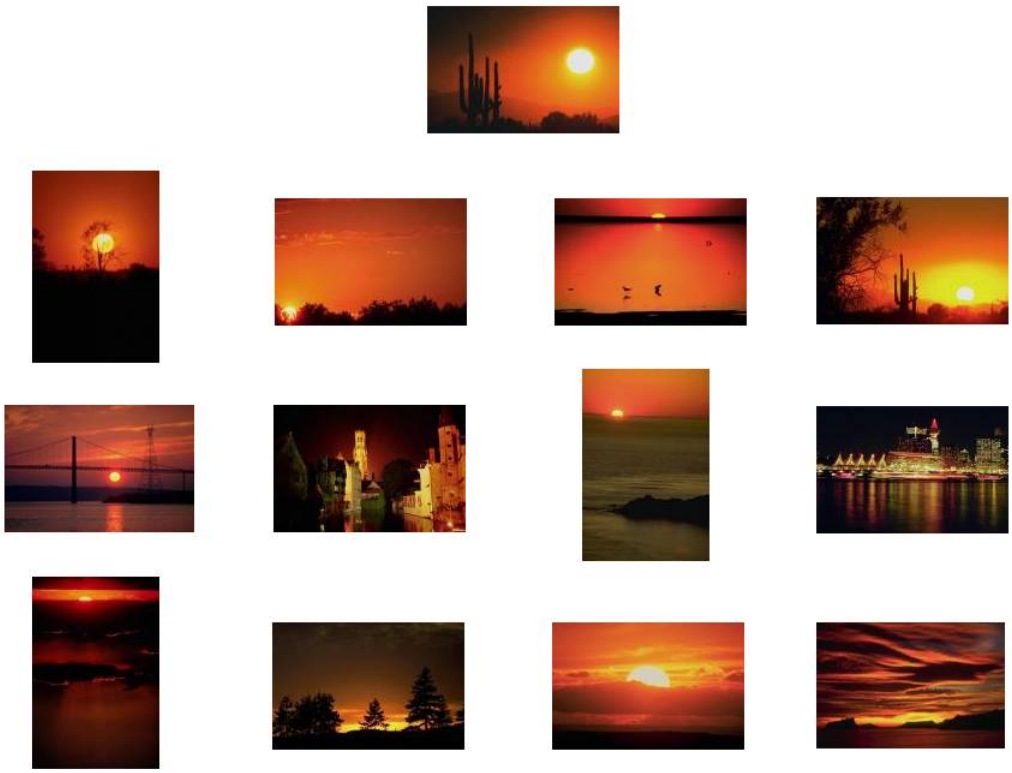
6.5.1 Κατασκευή Συνόλου Δεδομένης Αλήθειας

Έστω p_i μία τυχαία εικόνα, $\mathcal{P}$ το σύνολο των εικόνων της συλλογής και $\mathcal{C}$ το σύνολο όλων των εννοιών προς ανάκτηση. Για την εικόνα p_i, έστω C_i το σύνολο των εννοιών που περιέχονται σε αυτή. Για μια έννοια $c_i \in \mathcal{C}$, το σύνολο $\mathcal{P}$ μπορεί να χωριστεί σε δύο υποσύνολα, με το ένα να αποτελείται από τις εικόνες που απεικονίζουν την έννοια c_i και το άλλο από εκείνες που δεν την απεικονίζουν. Τα δύο αυτά σύνολα θα συμβολίζονται στο εξής με G_i και $\bar{G}_i$ αντίστοιχα:

$$G_i = \{p_i \in \mathcal{P} : \quad c_i \in C(i)\} \tag{6.7}$$

$$\bar{G}_i : \{p_i \in \mathcal{P} : \quad c_i \notin C(i)\} \tag{6.8}$$

Σύμφωνα με τους Smeulders et al. [193], η όλη διαδικασία αξιολόγησης στο πρόβλημα της ανάκτησης εικόνων είναι προβληματική καθώς το σύνολο δεδομένης αλήθειας δεν μπορεί να οριστεί πάντα απόλυτα και με ακρίβεια. Στις περισσότερες περιπτώσεις, το πρόβλημα αυτό ανάγεται στο πρόβλημα του ορισμού της τέλειας ανάκτησης, κάτι που αποδεικνύεται ότι είναι εξαρτώμενο από την περίσταση και σχεδόν ποτέ δεν μπορεί να οριστεί με μονοσήμαντο και αντικειμενικό τρόπο. Τα μέτρα ακρίβειας και ανάκτησης που έχουν ήδη χρησιμοποιηθεί. Ωστόσο, αν και τα μέτρα αυτά είναι πολύ χρήσιμα



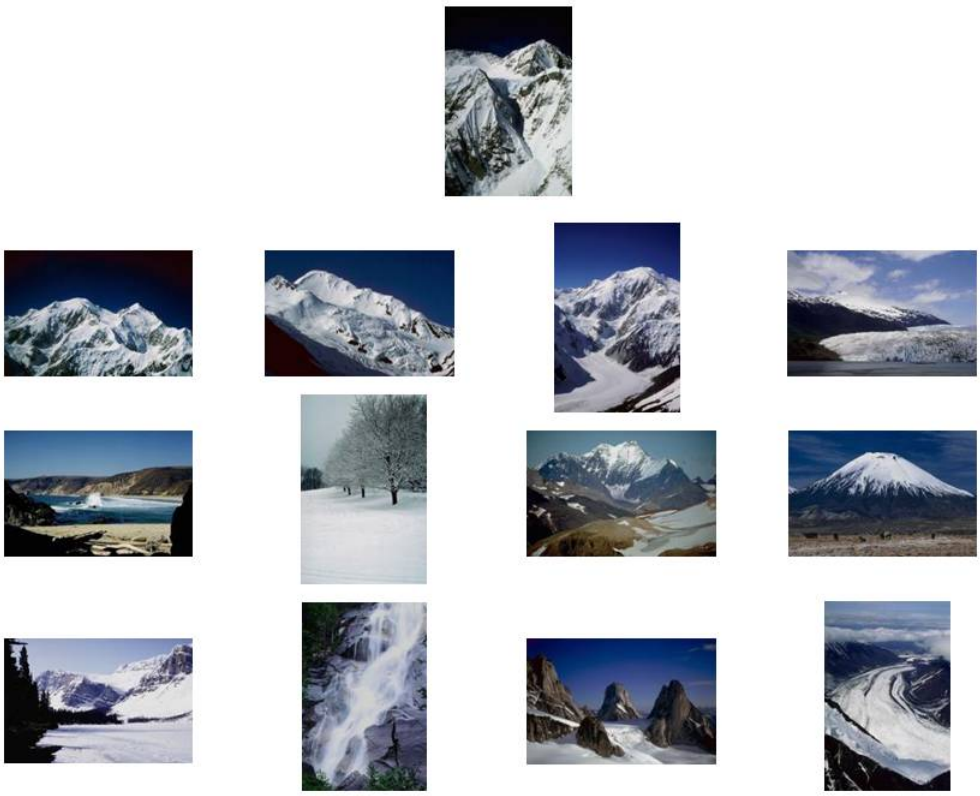
Σχήμα 6.6: *Αποτελέσματα αναζήτησης με χαρακτηριστικά εξαγόμενα από περιοχές σε συλλογή με φυσικές εικόνες από το Corel για την έννοια ηλιοβασίλεμα.*

για την αξιολόγηση τεχνικών ταξινόμησης και ανίχνευσης, στο παρόν πρόβλημα της ανάκτησης ο υπολογισμός τους δεν προσφέρει καμία ένδειξη για την απόδοση των σχετικών τεχνικών.

Για να γίνει κατανοητό αυτό, αν αναλογιστεί κανείς ότι ένα σύστημα ανάκτησης επιστρέφει στο χρήστη ολόκληρη την συλλογή εικόνων, ταξινομημένη βέβαια ως προς την απόσταση από την εικόνα του ερωτήματος, τότε τα προαναφερθέντα μέτρα δεν έχουν ιδιαίτερη σημασία, μιας και δεν εμπεριέχουν καθόλου πληροφορία για την θέση των σωστών και λάθος ανακτώμενων εικόνων. Το μέτρο ανάκτησης σε αυτή τη περίπτωση θα είναι πάντα ίσο με 1, γιατί όλες οι εικόνες που περιέχουν τις έννοιες της εικόνας του ερωτήματος θα έχουν επιστραφεί, μιας και αποτελούν υποσύνολο των εικόνων της συλλογής. Το μέτρο ακρίβειας θα είναι επίσης σταθερό και ίσο με τον λόγο του αριθμού των εικόνων που περιέχουν τις έννοιες της εικόνας του ερωτήματος προς το σύνολο των εικόνων της συλλογής. Απαιτείται λοιπόν ένα μέτρο το οποίο θα λαμβάνει υπόψη και την θέση ανάκτησης των "σωστών" και "λάθος" σύμφωνα με το ερώτημα ανακτήσεων.

Για το λόγο αυτό υιοθετείται η χρήση του μέτρου της *μέσης ακρίβειας*, το οποίο έχει οριστεί στην (4.28). Η σχέση αυτή προσαρμόζεται στο παρόν πρόβλημα και συγκεκριμένα, η ακρίβεια υπολογιζόμενη στις m κοντινότερες στην εικόνα του ερωτήματος εικόνες που επέστρεψε το συστημα επαναορίζεται ως

$$p_m = \frac{1}{m} \sum_{k=1}^{m} x_k \ , \tag{6.9}$$



Σχήμα 6.7: *Αποτελέσματα αναζήτησης με χαρακτηριστικά εξαγόμενα από περιοχές σε συλλογή με φυσικές εικόνες από το Corel για την έννοια χιόνι.*

όπου p_m είναι η μέση τιμή των $x_1, x_2, \ldots, x_m$, και συμβολίζεται με $\bar{x}_m$ και το $x_i \in \{0, 1\}$ είναι ίσο με 0 αν δεν υπάρχει στην εικόνα στη i-οστή θέση η έννοια της εικόνας του ερωτήματος και 1 αν υπάρχει. Η μέση ακρίβεια ορίζεται σαν η μέση τιμή των ακριβειών μετά από κάθε σχετική εικόνα που συναντάται στην λίστα. Μαθηματικά εκφράζεται ως

$$AP_c^N = \frac{1}{|G_c|} \sum_{j=1}^{N} x_j p_j = \frac{1}{|G_c|} \sum_{j=1}^{N} \frac{x_j}{j} \sum_{k=1}^{j} x_k \ , \tag{6.10}$$

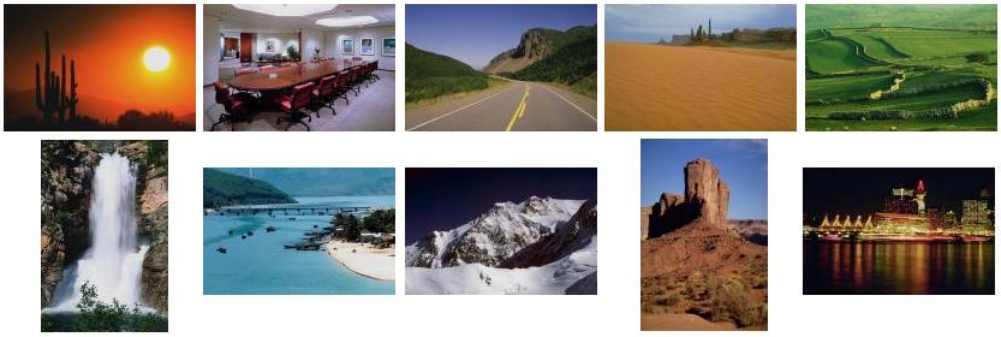
η οποία αντιστοιχεί στη μέση ακρίβεια για την έννοια c με παράθυρο N. Όπως γίνεται εύκολα αντιληπτό, το μέτρο της μέσης ακρίβειας επηρεάζεται πολύ περισσότερο από τις σωστές ανακτήσεις που επιστρέφονται σε υψηλή θέση, παρά από αυτές που επιστρέφονται σε χαμηλότερες. Αν θεωρηθεί ο αριθμητικός μέσος όρος από μέσες ακρίβειες, για μια σειρά από ερωτήματα τότε προκύπτει το μέτρο mAP (mean Average Precision)[28]

Για την κατασκευή των καμπυλών ακρίβειας-ανάκτησης ή mAP, είναι αναγκαίο να καθοριστεί και μια παράμετρος που θα μεταβάλλεται, για τις διαφορετικές τιμές. Συνήθως και σε προβλήματα ανάκτησης εικόνας, αυτή είναι στις περισσότερες περιπτώσεις το *παράθυρο των επιστρεφόμενων αποτελεσμάτων* γνωστό και ως *scope*. Σε άλλες περιπτώσεις, σημαντικότατος παράγοντας για την απόδοση είναι το μέγεθος του οπτικού λεξικού, άρα μεταβάλλεται ο αριθμός των συστάδων κατά τη διαδικασία

[28]Το μέτρο mean Average Precision είναι δύσκολο να μεταφραστεί εύηχα στα ελληνικά, καθώς δεν υπάρχουν δύο λέξεις για τον μέσο όρο, σε αντίθεση με την αγγλική γλώσσα (mean - average).

της συσταδοποίησης, όπως π.χ. προτείνεται από τους Jing et al. [92]. Οι Huijsmans και Sebe προτείνουν να μεταβάλλεται ο αριθμός των εικόνων που θεωρούνται ως "σωστές" ως προς το εικόνα του ερωτήματος, στη συλλογή εικόνων (size of the *embedding*) [87].

6.5.2 Συλλογές εικόνων που χρησιμοποιήθηκαν για την αξιολόγηση



Σχήμα 6.8: *Δείγμα από το υποσύνολο της συλλογής εικόνων Corel που χρησιμοποιήθηκε.*

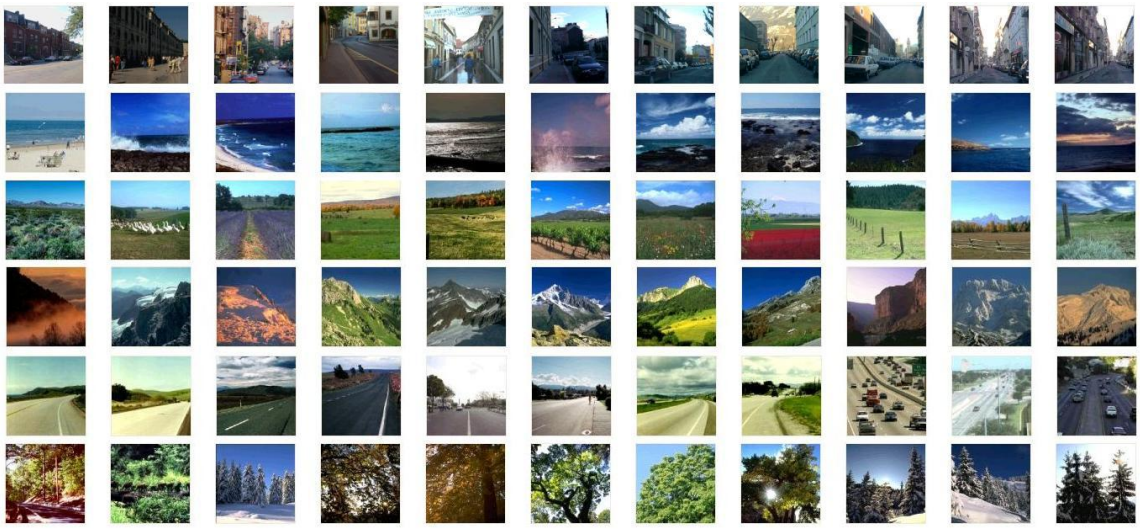
Για την πληρότητα της παρουσίασης και αξιολόγησης των τεχνικών που αναπτύχθηκαν, χρησιμοποιήθηκαν δύο διαφορετικές συλλογές εικόνων.

Η πρώτη συλλογή των εικόνων αποτελείται από ένα υποσύνολο των εικόνων του Corel. Οι εικόνες αυτές περιέχουν τις έννοιες *χιόνι*, *βλάστηση*, *ηλιοβασίλεμα*, *δρόμος*, *έρημος*, *καταρράκτης* και *παραλία*. Χρησιμοποιήθηκαν επίσης και εικόνες από τις έννοιες *νυχτερινές φωτογραφίες πόλης*, *εσωτερικός χώρος*, *υποθαλάσσιες φωτογραφίες* και *γάτες*. Οι εικόνες αυτές χρησιμοποιήθηκαν για να δυσκολέψουν το έργο της ανάκτησης. Η συλλογή εικόνων του Corel ανήκει στην κατηγορία των συλλογών θεματικού πεδίου. Οι εικόνες που περιέχει είναι υψηλής ποιότητας φωτογραφίες και είναι σχεδόν σε όλες προφανές ποια έννοια απεικονίζουν. Έχει χρησιμοποιηθεί κατά κόρον στην ερευνητική κοινότητα για αξιολόγηση τεχνικών σε προβλήματα ανάκτησης, αναζήτησης, ταξινόμησης, αναγνώρισης και ούτω καθεξής. Ένα δείγμα από τη συλλογή αυτή απεικονίζεται στο Σχήμα 6.8.

Η δεύτερη συλλογή που χρησιμοποιήθηκε, δημιουργήθηκε από τους Oliva και Torralba [159] για χρήση σε πρόβλημα αναγνώρισης σκηνής. Η συλλογή αυτή είναι διαθέσιμη[29] στην ερευνητική κοινότητα για αξιολογήσεις και συγκρίσεις τεχνικών. Οι κατηγορίες που επιλέχθηκαν από αυτή τη συλλογή είναι οι εξής: *ακτή*, *δάσος*, *αυτοκινητόδρομος*, και *δρόμος πόλης*. Και η συλλογή αυτή μπορεί να καταταχθεί στην κατηγορία συλλογών θεματικού πεδίου, ωστόσο δεν είναι εμπορική συλλογή σαν αυτή του Corel. Ένα δείγμα από τη συλλογή αυτή απεικονίζεται στο Σχήμα 6.9.

Οι εικόνες της συλλογής των Oliva και Torralba έχουν σχολιαστεί για έναν πολύ μεγάλο αριθμό από έννοιες. Αντίθετα, οι εικόνες του Corel έχουν έναν πιο "χαλαρό" σχολιασμό. Έτσι, οι εικόνες της συλλογής αυτής σχολιάστηκαν και δημιουργήθηκε έτσι κατάλληλο σύνολο δεδομένης αλήθειας. Προσδιορίστηκε δηλαδή σε κάθε εικόνα

[29] `http://people.csail.mit.edu/torralba/code/spatialenvelope/`



Σχήμα 6.9: *Δείγμα από τη συλλογή εικόνων του Torralba που χρησιμοποιήθηκε.*

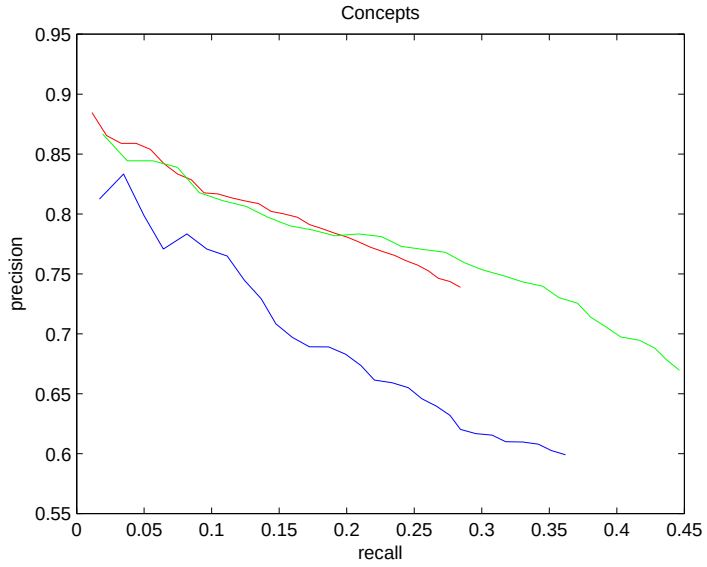
η έννοια που αυτή περιέχει για να μπορεί να αξιολογηθεί η ανάκτηση με τα μέτρα που παρουσιάστηκαν στην Ενότητα 6.5.1.

6.5.3 Αποτελέσματα ανάκτησης

Στην Ενότητα αυτή παρουσιάζονται τα αποτελέσματα των πειραμάτων της αναζήτησης. Πέρα από την τεχνική που προτείνεται σε αυτό το Κεφάλαιο, υλοποιήθηκε και μια μέθοδος που βασίζεται στην εξαγωγή καθολικών χαρακτηριστικών, παρόμοια με αυτές που χρησιμοποιήθηκαν σε παλαιότερα συστήματα ανάκτησης. Αρχικά υλοποιήθηκε μια απλοϊκή μέθοδος αναζήτησης που χρησιμοποιεί καθολικά χαρακτηριστικά από τις εικόνες. Για την περιγραφή των χαρακτηριστικών χαμηλού επιπέδου των εικόνων χρησιμοποιούνται οι περιγραφείς του προτύπου MPEG-7. Επιλέχθηκαν και συγχωνεύτηκαν, σύμφωνα με την τεχνική του Κεφαλαίου 2 ο Περιγραφέας Κλιμακωτού Χρώματος, ο Περιγραφέας Διάταξης Χρώματος και ο Περιγραφέας Ομοιογενούς Υφής. Μια εικόνα περιγράφεται έτσι με ένα διάνυσμα, το οποίο περιέχει την πληροφορία για τα χαρακτηριστικά χρώματος και υφής της. Υπολογίστηκαν η ακρίβεια, η ανάκτηση και το μέτρο mAP και για τις δύο τεχνικές. Τα μέτρα mAP για όλες τις έννοιες και τεχνικές συνοψίζονται στον Πίνακα 6.1.

6.5.3.1 Ανάκτηση με περιγραφείς από όλη την εικόνα

Παρατηρείται ότι οι έννοιες που εντοπίζονται πιο εύκολα από τους περιγραφείς του MPEG-7 εξαγόμενους από ολόκληρη την εικόνα είναι η *βλάστηση*, το *ηλιοβασίλεμα*, η *έρημος*, το *χιόνι* και ο *εσωτερικός χώρος*. Από την άλλη, έννοιες όπως ο *δρόμος* και ο *καταρράκτης* που μπορεί να έχουν χαρακτηριστική υφή και χρώμα, δεν καταλαμβάνουν όμως συνήθως μεγάλο ποσοστό των εικόνων, δεν μπορούν να ανακτηθούν σωστά. Μικρή είναι και η τιμή του μέτρου mAP για την έννοια *παραλία*, καθώς οι εικόνες που ανήκουν σε αυτήν την κατηγορία παρουσιάζουν μεγάλη οπτική ποικιλομορφία. Επιπλέον, καθώς η περιγραφή εξάγεται από ολόκληρη την εικόνα είναι μάλλον αναμενόμενο να ανακτώνται σωστά οι έννοιες που χαρακτηρίζουν μεγάλο μέρος της εικόνας. Στο Σχήμα 6.10 απεικονίζεται το διάγραμμα ανάκτησης-ακρίβειας



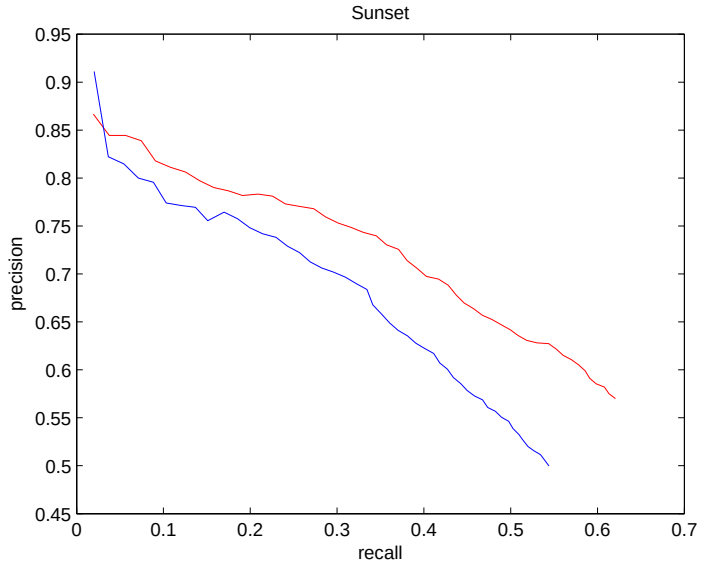
Σχήμα 6.10: *Διαγράμματα ακρίβειας-ανάκτησης για τις έννοιες χιόνι (μπλέ), ηλιοβασίλεμα (πράσινο) και βλάστηση (κόκκινο). Οι περιγραφείς εξάγονται από ολόκληρη την εικόνα.*

ως προς τον αριθμό των εικόνων που ανακτήθηκαν, ενδεικτικά για τις έννοιες *χιόνι*, *ηλιοβασίλεμα* και *βλάστηση* της συλλογής του Corel. Παρατηρείται ότι οι τιμές της ακρίβειας είναι αρκετά ικανοποιητικές, χωρίς ωστόσο να συνοδεύονται από αντίστοιχες τιμές ανάκτησης.

Ιδιαίτερη σημασία στην ανάκτηση έχει η ταυτόχρονη παρουσία των περιγραφέων χρώματος και υφής. Αν και η διερεύνηση του πόσο επηρεάζεται η ανάκτηση με βάση τα καθολικά χαρακτηριστικά χρώματος και υφής δεν αποτελεί σκοπό αυτού του Κεφαλαίου, στο Σχήμα 6.11 απεικονίζεται το διάγραμμα ακρίβειας-ανάκτησης για την έννοια ηλιοβασίλεμα, με όλους τους περιγραφείς και μόνο με τους περιγραφείς χρώματος. Παραητρείται ότι η ταυτόχρονη ύπαρξη και των δύο τύπων περιγραφέων βελτιώνει ουσιαστικά τα μέτρα αυτά.

6.5.4 Ανάκτηση με περιγραφείς από περιοχές της εικόνας

Η εφαρμογή της τεχνικής ανάκτησης στη συλλογή του Corel, έδωσε σε μερικές από τις έννοιες που χρησιμοποιήθηκαν χειρότερα αποτελέσματα, όπως φαίνεται στον Πίνακα 6.1, παρότι είναι πιο σύνθετη από την καθολική εξαγωγή χαρακτηριστικών. Αυτό εξηγείται από το γεγονός ότι οι εικόνες της συγκεκριμένης συλλογής περιγράφουν έννοιες υψηλού επιπέδου οι οποίες έχουν ομοιόμορφη υφή και χρώμα και κατανέμονται συνολικά σε ολόκληρη την εικόνα. Στο συγκεκριμένο πείραμα χρησιμοποιήθηκε θησαυρός 70 τύπων περιοχής, καθώς διαπιστώθηκε πειραματικά ότι εμφανίζει τα καλύτερα αποτελέσματα, χωρίς να μεγαλώνει ιδιαίτερα το μήκος της περιγραφής. Επίσης, το διάνυσμα αναπαράστασης σχηματίστηκε με βάση τους 3 κοντινότερους τύπους περιοχών για κάθε περιοχή της εικόνας. Από τα αποτελέσματα μπορεί να παρατηρηθεί ότι ενώ το μέτρο mAP είναι μικρότερο σε έννοιες όπως *χιόνι*, *έρημος* και *ηλιοβασίλεμα*, είναι μεγαλύτερο σε πιο "σύνθετες" έννοιες, όπως *δρόμος* και *βλάστηση* που συνήθως συνυπάρχουν στις εικόνες με άλλες έννοιες, δεν έχουν, δηλαδή, "κυρίαρχο" ρόλο στις εικόνες.



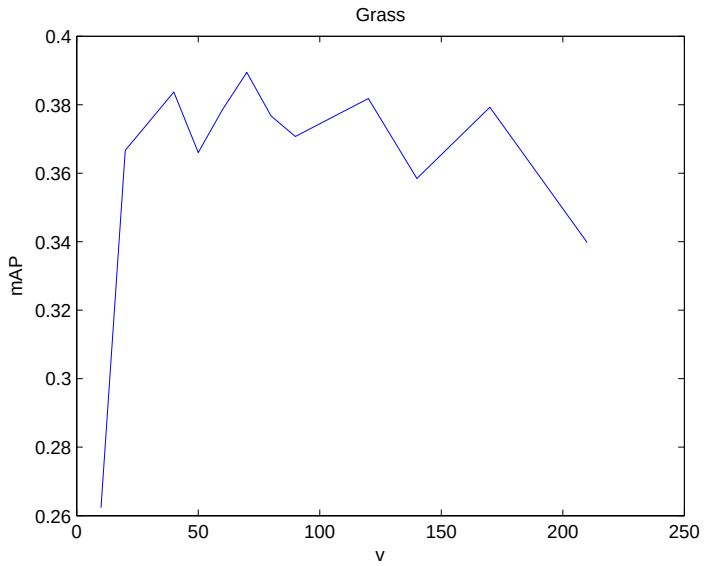
Σχήμα 6.11: *Διαγράμματα ακρίβειας-ανάκτησης για την έννοια ηλιοβασίλεμα με όλους τους περιγραφείς (κόκκινο) και μόνο με τους περιγραφείς χρώματος (μπλέ). Οι περιγραφείς εξάγονται από ολόκληρη την εικόνα.*

Έγιναν επίσης, πειράματα με διαφορετικά μεγέθη οπτικού θησαυρού, για να διερευνηθεί πόσο σημαντικό ρόλο παίζει το μέγεθός του στην ανάκτηση. Στο Σχήμα 6.12 απεικονίζεται ενδεικτικά το μέτρο mAP για μεγέθη οπτικού θησαυρού από 10 εώς και 210 τύπους περιοχής. Η βέλτιστη τιμή επιτυγχάνεται για 70 τύπους περιοχής, αλλά φαίνεται ότι οι διαφορές του mAP είναι μικρες για ένα εύρος τιμών θησαυρού από 30 εώς 160 λέξεις. Για μεγαλύτερο μέγεθος θησαυρού, παρατηρείται ιδιαίτερη πτώση του μέτρου.

Έννοια	**Καθολικά**	**Περιοχές**
χιόνι	**0.52**	0.24
ηλιοβασίλεμα	**0.63**	0.30
βλάστηση	0.60	**0.62**
δρόμος	0.15	**0.17**
έρημος	**0.60**	0.43
καταρράκτης	**0.26**	0.20
παραλία	**0.24**	0.18
εσωτερικός χώρος	**0.78**	**0.78**

Πίνακας 6.1: *Το μέτρο mAP που επιτεύχθηκε στη συλλογή του Corel για ανάκτηση με καθολικά χαρακτηριστικά και ανάκτηση με χρήση οπτικού θησαυρού.*

Έχει διαφανεί εώς τώρα ότι η ανάκτηση των εικόνων χρησιμοποιώντας μια περιγραφή που βασίζεται σε οπτικό θησαυρό δεν παρέχει συνολικά κάποια βελτίωση ως προς την ανάκτηση με βάση τα καθολικά χαρακτηριστικά. Βελτίωση επιτεύχθηκε για μεμονωμένες έννοιες. Εύλογο είναι να διερευνηθεί το φαινόμενο αυτό. Για το λόγο αυτό, επιλέχθηκε η συλλογή εικόνων του Torralba. Όπως φαίνεται στο Σχήμα 6.9, οι εικόνες της συλλογής αυτής είναι ιδιαίτερα ανομοιογενείς και φαίνονται πιο σύνθε-



Σχήμα 6.12: *Η εξέλιξη του μέτρου mAP για την έννοια βλάστηση, για μεγέθη θησαυρού από 10 εώς 210 τύπους περιοχής.*

τες από αυτές της συλλογής του Corel. Τα συνολικά αποτελέσματα παρουσιάζονται στον Πίνακα 6.2. Όπως είναι φανερό, η χρήση καθολικής περιγραφής με περιγραφείς χρώματος και υφής του MPEG-7 παρουσίασε άσχημα αποτελέσματα. Υπολογίστηκαν οι τιμές του mAP για τις έννοιες *ακτή*, *δάσος*, *αυτοκινητόδρομος* και *δρόμος πόλης*. Οι υπόλοιπες έννοιες της συλλογής ήταν οι *εξοχή*, *βουνό* και *πόλη*, οι οποίες θεωρήθηκαν ως αρνητικά παραδείγματα. Τα μέτρα mAP για αυτές τις έννοιες αυτές, χρησιμοποιώντας την τεχνική ανάκτησης που παρουσιάστηκε στο παρόν Κεφάλαιο και για διάφορους συνδυασμούς του μεγέθους του οπτικού θησαυρού και του αριθμού των τύπων περιοχής που θεωρούνται οι κοντινότεροι σε μια περιοχή της εικόνας παρουσιάζονται στον Πίνακα 6.3.

Έννοια	**Καθολικά**	**Περιοχές**
ακτή	0.18	**0.66**
δάσος	0.11	**0.24**
δρόμος πόλης	0.05	**0.09**
αυτοκινητόδρομος	0.03	**0.10**

Πίνακας 6.2: *Το μέτρο mAP που επιτεύχθηκε στη συλλογή του Torralba για ανάκτηση με καθολικά χαρακτηριστικά και ανάκτηση με χρήση οπτικού θησαυρού με N_T=270 και K=2.*

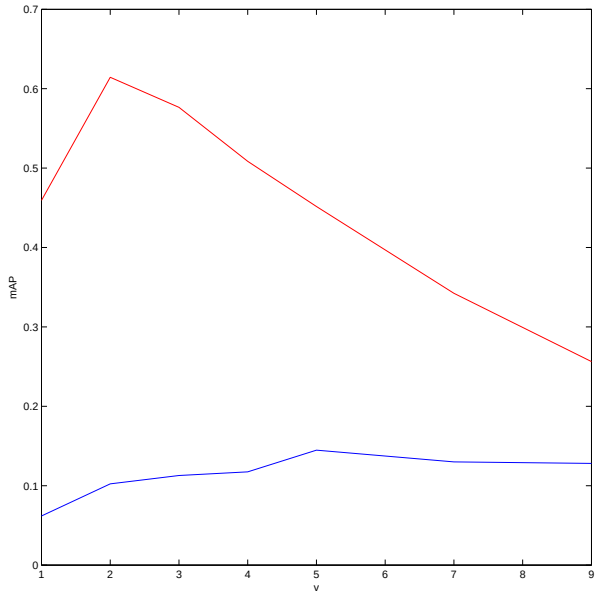
Μπορεί να παρατηρηθεί ότι η τεχνική ανάκτησης που προτείνεται αποδίδει πάρα πολύ καλά στις έννοιες *ακτή* και *δάσος*, ενώ η απόδοσή της είναι αισθητά χαμηλότερη στις έννοιες *αυτοκινητόδρομος* και *δρόμος πόλης*. Το αποτέλεσμα αυτό μπορεί να ερμηνευθεί, αν ληφθεί υπόψη η ιδιαιτερότητα των εννοιών αυτών. Κατά πρώτον, στην περίπτωση της *ακτής* και του *δάσους* η κατάτμηση δημιουργεί περιοχές που μπορούν πιο εύκολα να διαχωρίσουν τις έννοιες αυτές από τις υπόλοιπες. Αντίθετα, στις εικόνες *δρόμου πόλης* και *αυτοκινητόδρομου*, δεν δημιουργεί συνήθως περιοχές χαρακτηριστικές για τις έννοιες αυτές. Κατά δεύτερον, οι δύο έννοιες μοιάζουν αρ-

Έννοια	$N_T = 150$ $K = 4$	$N_T = 270$ $K = 1$	$N_T = 270$ $K = 2$	$N_T = 270$ $K = 5$
ακτή	0.36	0.46	**0.66**	0.45
δάσος	0.17	**0.27**	0.23	0.18
δρόμος πόλης	0.13	0.06	0.09	**0.14**
αυτοκινητόδρομος	0.09	0.06	0.10	**0.14**

Πίνακας 6.3: *Το μέτρο mAP για όλες τις έννοιες και για διάφορες περιπτώσεις μεγέθους οπτικού θησαυρού N_T και αριθμού κοντινότερων τύπων περιοχών K.*

κετά μεταξύ τους, αλλά μοιάζουν και με τις εικόνες πόλης που περιείχε η συλλογή και χρησιμοποιήθηκαν ως αρνητικά παραδείγματα, προκαλώντας επιπλέον σύγχυση στους ταξινομητές.

Μελετήθηκε επίσης και το πώς επηρεάζεται το μέτρο mAP σε σχέση με τον αριθμό K των κοντινότερων τύπων περιοχής που λαμβάνονται υπόψη ως προς τις περιοχές των εικόνων, για τον σχηματισμό των διανυσμάτων αναπαράστασης. Στο Σχήμα 6.13 απεικονίζεται η εξέλιξη του μέτρου mAP σε σχέση με τον αριθμό K για τις έννοιες *ακτή* και *δρόμος πόλης*. Όπως είναι φανερό, όσο αυξάνεται η τιμή του K, αυξάνεται και η τιμή του mAP για την έννοια *δρόμος πόλης*, ενώ το αντίθετο συμβαίνει για την έννοια *ακτή*. Παρόμοια είναι η συμπεριφορά αντίστοιχα για τις έννοιες αυτοκινητόδρομος και δάσος, οδηγώντας στο συμπέρασμα ότι οι πιο "απλές" διαισθητικά έννοιες, δηλαδή αυτές που μπορεί να περιγραφούν από έναν πιο μικρό αριθμό περιοχών, δεν απαιτούν μεγάλες τιμές του K για την ανάκτησή τους.



Σχήμα 6.13: *Η εξέλιξη του μέτρου mAP για τις έννοιες δρόμος πόλης (μπλέ)και ακτή (κόκκινο) όσο αυξάνεται ο αριθμός K των τύπων περιοχής που θεωρούνται οι κοντινότεροι σε μια περιοχή της εικόνας για το σχηματισμό του διανύσματος αναπαράστασης.*

6.6 Συμπεράσματα

Στο Κεφάλαιο αυτό παρουσιάστηκε μια τεχνική ανάκτησης εικόνων. Η τεχνική αυτή βασίστηκε στις ιδέες και τους αλγορίθμους που μελετήθηκαν και αναπτύχθηκαν στο Κεφάλαιο 4 και συγκεκριμένα στη δημιουργία και χρήση ενός οπτικού θησαυρού περιοχών για την αναπαράσταση του οπτικού περιεχομένου μια εικόνας. Καθώς η αναπαράσταση αυτή διαφάνηκε ότι ήταν ανεπαρκής για το πρόβλημα εικόνων, προτάθηκε μια παραλλαγή της, η οποία αποδείχθηκε ότι είναι ικανή να περιγράψει τόσο το οπτικό, όσο και κατά μια έννοια το σημασιολογικό περιεχόμενο των εικόνων, επιτρέποντας έτσι την αποδοτική ανάκτησή τους.

Η προτεινόμενη τεχνική είναι αρκετά γρήγορη καθώς απαιτείται ένας αριθμός συγκρίσεων που είναι ίσος με το μέγεθος του οπτικού θησαυρού που χρησιμοποιήθηκε. Το μόνο μέρος της διαδικασίας που μπορεί να χαρακτηριστεί ως "αργό" είναι η κατάτμηση της εικόνας, καθώς ο αλγόριθμος που χρησιμοποιείται απαιτεί αρκετό χρόνο, κάτι εμφανές σε εικόνες μεγάλης διάστασης. Αντίθετα, οι περισσότεροι περιγραφείς MPEG-7 μπορούν να εξαχθούν σχετικά γρήγορα. Φυσικά, η διαδικασία αυτή γίνεται μόνο για την εικόνα του παραδείγματος και εφόσον αυτή δεν ανήκει ήδη στη συλλογή εικόνων.

Φυσικά, η ανάκτηση εικόνων με βάση το μοντέλο bag-of-words, στηριζόμενο σε περιοχές εικόνων που προκύπτουν από κατάτμηση δεν είναι δυνατόν να αντιμετωπίσει με επιτυχία όλες τις συλλογές εικόνων και όλες τις πιθανές προσδοκίες των χρηστών. Όπως έδειξαν και τα αποτελέσματα του Κεφαλαίου 4, είναι δυνατόν να διαχωριστούν εικόνες που περιέχουν διαφορετικές έννοιες, με την αναπαράσταση που βασίζεται στον οπτικό θησαυρό περιοχών, εφόσον οι έννοιες ανήκουν στην κατηγορία των "υλικών". Το ίδιο δε θα μπορούσε να μη συμβαίνει και στην περίπτωση της ανάκτησης. Ωστόσο, αν οι χρήστες επιδιώκουν την ανάκτηση εικόνων με άλλες έννοιες, όπως π.χ. *κτίρια*, θα πρέπει να στραφούν σε συστήματα που χρησιμοποιούν άλλου τύπου χαρακτηριστικά.

Συμπερασματικά, η προτεινόμενη τεχνική μπορεί εύκολα και αποδοτικά να εφαρμοστεί σε περιπτώσεις που στόχος είναι η ανάκτηση εικόνων με βάση τη σημασιολογία τους, εφόσον οι έννοιες που αυτές περιέχουν ανήκουν στην κατηγορία των υλικών, αλλά και με βάση το οπτικό τους περιεχόμενο, καθώς μπορεί να παρέχει μια ικανοποιητική περιγραφή των καθολικών οπτικών χαρακτηριστικών, η οποία είναι πληρέστερη π.χ. από τα απλά ιστογράμματα χρώματος.

Κεφάλαιο 7

Ανίχνευση Εννοιών σε Εικόνες με χρήση του Οπτικού Εννοιολογικού Πλαισίου

7.1 Εισαγωγή

Έχει διαπιστωθεί ότι τα συστήματα ανάλυσης του περιεχομένου και ανάκτησης εικόνων περιορίζονται πολλές φορές από τις υπάρχουσες τεχνικές και τεχνολογίες του ερευνητικού πεδίου της κατανόησης εικόνας. Αυτό συμβαίνει, καθώς αυτά συνήθως υιοθετούν προσεγγίσεις χαμηλού επιπέδου, χωρίς να λαμβάνουν υπόψη τη γνώση και τα χαρακτηριστικά υψηλού επιπέδου. Τα τελευταία χρόνια, οι τεχνικές που βασίζονται σε γνώση, στην ανθρώπινη αντίληψη και στην κατανόηση του περιεχομένου των σκηνών έχουν αρχίσει να τραβούν το ενδιαφέρον της ερευνητικής κοινότητας, προσπαθώντας να καλύψουν το σημασιολογικό και το εννοιολογικό κενό που χωρίζει ανθρώπους και υπολογιστές.

Το τελευταίο αποτελεί ένα σχετικά καινούριο ερευνητικό πεδίο. Έτσι, πρόσφατα έχει δοθεί ιδιαίτερο βάρος στην έρενα στο χώρο της μοντελοποίησης και της αξιοποίησης της εννοιολογικής πληροφορίας με σκοπό την εξέλιξη και βελτίωση των υπαρχουσών τεχνικών ανάλυσης. Η πληροφορία αυτή δρα σαν μια προσομοίωση του τρόπου με τον οποίο αντιλαμβάνεται ο άνθρωπος τον κόσμο, λαμβάνοντας, δηλαδή, υπόψη όλη την πληροφορία που είναι παρούσα στο οπτικό περιεχόμενο μιας σκηνής. Η πληροφορία αυτή, γνωστή και ως *εννοιολογικό πλαίσιο* μπορεί να ενσωματωθεί σε τεχνικές ανάλυσης πολυμεσικού υλικού με βάση τη γνώση, συστήματα σημασιολογικής δεικτοδότησης, καθώς και ανάκτησης.

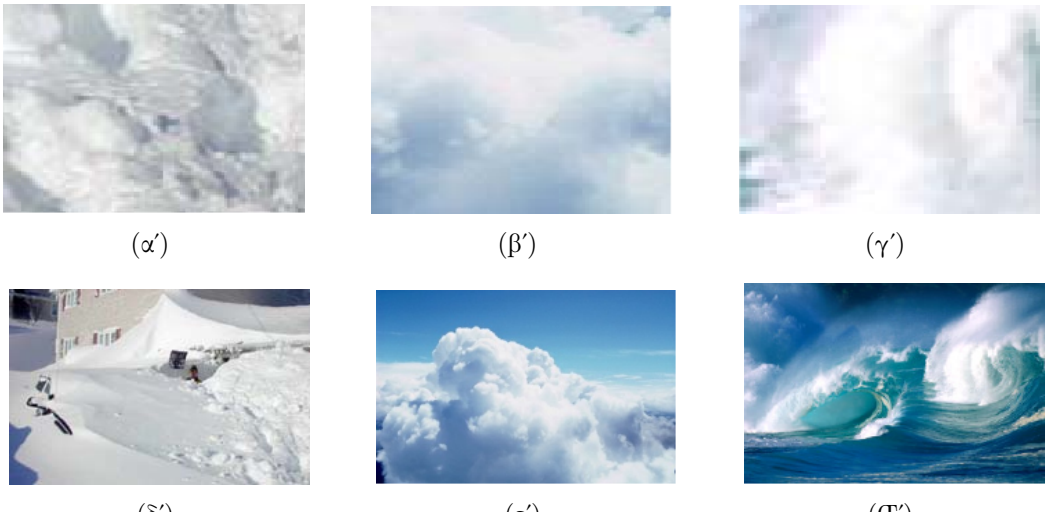
Στο Κεφάλαιο αυτό αρχικά περιγράφεται το εννοιολογικό πλαίσιο ως έννοια και παρουσιάζονται κάποια από τα προβλήματα που συναντούν τόσο η ανθρώπινη αντίληψη, όσο και οι τεχνικές ανιχνευσης εννοιών υψηλού επιπέδου. Γίνεται έτσι κατανοητή η χρησιμότητά του στα προβλήματα ανάλυσης εικόνων. Στη συνέχεια παρουσιάζονται οι σχετικές εργασίες, οι οποίες και χωρίζονται σε τέσσερις μεγάλες κατηγορίες, ανάλογα με το πως έχει μοντελοποιηθεί το εννοιολογικό πλαίσιο. Έπειτα, προτείνονται δύο μέθοδοι οι οποίες μοντελοποιούν το εννοιολογικό πλαίσιο των εικόνων ενός θεματικού πεδίου και των τύπων περιοχής. Στη συνέχεια, αυτές οι μέθοδοι ενοποιούνται, σχηματίζοντας το μεικτό εννοιολογικό πλαίσιο. Έτσι, η τεχνική ανίχνευσης εννοιών που προτάθηκε στο Κεφάλαιο 4 εμπλουτίζεται και διερευνάται αν τελικά το εννοιο-

λογικό πλαίσιο μπορεί να βελτιώσει την ακρίβεια που επιτυγχάνεται.

7.2 Περιγραφή του προβλήματος

Το *εννοιολογικό πλαίσιο* (συμφραζόμενα ή συγκείμενο) μιας οντότητας είναι το σύνολο των εννοιών, συνθηκών, παραμέτρων και καταστάσεων που την περιβάλλουν και την καθορίζουν. Για παράδειγμα, μια λέξη ή μια φράση αποκτά διαφορετικό νόημα, ανάλογα με τις λέξεις ή φράσεις που την περιβάλλουν. Το εννοιολογικό πλαίσιο παίζει ιδιαίτερο ρόλο σε αρκετές επιστήμες, όπως η αρχαιολογία, η γλωσσολογία, τα μαθηματικά και άλλες. Στη σημασιολογική ανάλυση εικόνας, το εννοιολογικό πλαίσιο αποκτά ιδιαίτερη σημασία, γιατί η οπτική πληροφορία που συνυπάρχει σε κάποιο πολυμεσικό έγγραφο καθίσταται ιδιαίτερα χρήσιμη σε εφαρμογές ανίχνευσης εννοιών, ταξινόμησης σκηνών και άλλες.

Ξεκινώντας με την ταξινόμηση εικόνας, πολλές είναι οι έννοιες που εμφανίζουν παρόμοια οπτικά χαρακτηριστικά. Έτσι, τα χαμηλού επιπέδου χαρακτηριστικά που εξάγονται από αυτές αδυνατούν να τις διαχωρίσουν, καθιστώντας αμφίβολη εώς αδύνατη τη σωστή ανίχνευσή τους και το σωστό "ορισμό" μιας τέτοιας έννοιας. Για παράδειγμα, θεωρώντας τις εικόνες της πρώτης σειράς του Σχήματος 7.1, ακόμη και για έναν άνθρωπο θα ήταν δύσκολο να αποφανθεί αν πρόκειται για *σύννεφα*, *χιόνι* ή *κύματα*. Αντίθετα, για τις εικόνες της δεύτερης σειράς η παραπάνω απόφαση είναι απλούστερη. Παρατηρώντας ολόκληρη την εικόνα και όχι το μέρος της, δηλαδή το εννοιολογικό πλαίσιο της έννοιας για την οποία δεν είναι εύκολο να ληφθεί απόφαση, γίνεται σαφές ποιες έννοιες απεικονίζονται.



(α′) (β′) (γ′)

(δ′) (ε′) (Γ′)

Σχήμα 7.1: *Επάνω σειρά: χιόνι, σύννεφα και κύματα εκτός εννοιολογικού πλαισίου· κάτω σειρά: Οι ίδιες έννοιες εντός του εννοιολογικού τους πλαισίου.*

Επιπρόσθετα, η ορθή αναγνώριση απομονωμένων τμημάτων καθαρών υλικών χωρίς τη βοήθεια του εννοιολογικού τους πλαισίου αποτελεί και αυτή μια διαδικασία δύσκολη, εώς αδύνατη, ακόμα και για τους ανθρώπους. Όπως φαίνεται στο Σχήμα 7.2, απομονωμένα τμήματα *χιονιού* εμφανίζονται ανάμεικτα με απομονωμένα τμήματα *σύννεφου* και *υφάσματος*. Έτσι, η αναγνώριση του καθενός είναι εξαιρετικά δύσκολη. Αρκετές είναι οι τεχνικές που έχουν προταθεί για την ανίχνευση συγκεκριμένων υλικών, όπως παρουσιάστηκαν στο Κεφάλαιο 4. Στις περισσότερες από αυτές, όμως, η

πληροφορία του εννοιολογικού πλαισίου που τα περιβάλλει δεν αξιοποιείται, οδηγώντας αρκετές φορές σε λανθασμένες ανιχνεύσεις, ιδιαίτερα για υλικά που παρουσιάζουν παρόμοια οπτικά χαρακτηριστικά.

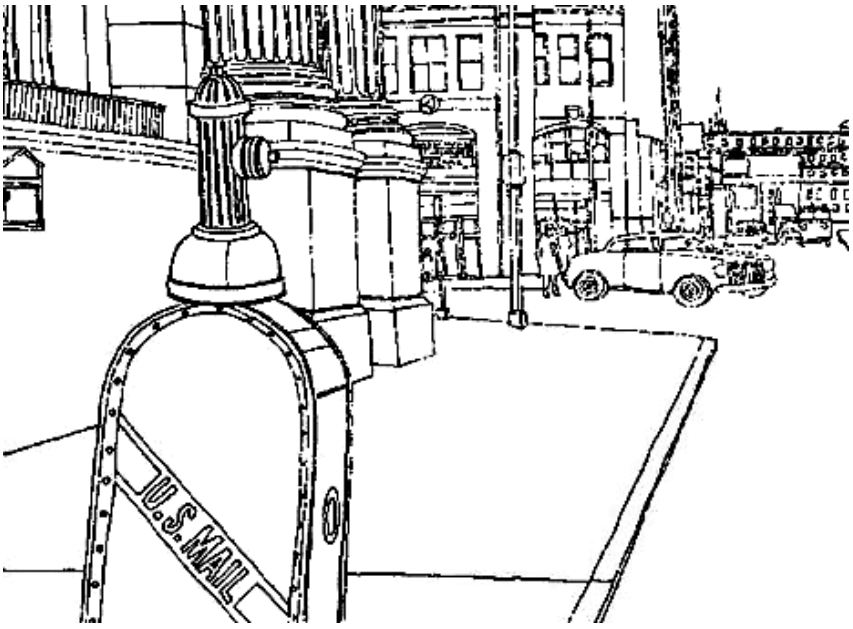
Σχήμα 7.2: *Αναγνώριση παρόμοιων υλικών χωρίς τη βοήθεια του εννοιολογικού πλαισίου· τα επιλεγμένα τμήματα είναι αυτά που απεικονίζουν* χιόνι *στην πραγματικότητα.*

Επίσης, στο ίδιο πλαίσιο και όσον αφορά τα αντικείμενα, είναι προφανές ότι ένα αντικείμενο είναι ορισμένες φορές αρκετά δύσκολο να αναγνωριστεί όταν παρουσιαστεί απομονωμένο. Ωστόσο, η αναγνώρισή του μπορεί να καταστεί ιδιαίτερα εύκολη όταν είναι φυσικά τοποθετημένο στο εννοιολογικό του πλαίσιο. Για να καταστεί αυτό σαφές, στο Σχήμα 7.3 απεικονίζεται ένας *αεραγωγός* εκτός του εννοιολογικού του πλαισίου. Ακόμη κι αν παρατηρηθεί η εικόνα αυτή προσεκτικά από έναν άνθρωπό, είναι ιδιαίτερα δύσκολο να αναγνωρίσει ορθά την έννοια που απεικονίζεται. Στην περίπτωση, όμως, που παρατηρηθεί εντός του εννοιολογικού της πλαισίου, που στο παρόν παράδειγμα είναι το *εσωτερικό ενός αυτοκινήτου*, η αναγνώρισή του δεν αναμένεται ότι θα παρουσιάσει ιδιαίτερη δυσκολία.

Τέλος, έχει αποδειχτεί από ψυχοφυσικές μελέτες όπως αυτή των Biederman et al. [15] ότι η αναγνώριση αντικειμένων εκτός του εννοιλογικού τους πλαισίου ή για την ακρίβεια, όταν αυτά "παραβιάζουν" το εννοιολογικό τους πλαίσιο είναι σημαντικά πιο δύσκολη. Η περίπτωση αυτή απεικονίζεται στο Σχήμα 7.4, όπου ο *πυροσβεστικός κρουνός* έχει τοποθετηθεί επάνω σε ένα *γραμματοκιβώτιο*, κάτι που έχει σαν αποτέλεσμα να δυσχεραίνεται η ορθή του αναγνώριση, όπως πειραματικά αποδείχτηκε, έπειτα από έρευνες σε εθελοντές.

Από όλα τα παραπάνω γίνεται κατανοητό ότι το εννοιολογικό πλαίσιο δύναται να παίξει ιδιαίτερο και πολύ σημαντικό ρόλο στα προβλήματα ανίχνευσης εννοιών υψηλού επιπέδου. Τα αντικείμενα και τα υλικά σπανίως εμφανίζονται στις εικόνες απομονωμένα. Στη συντριπτική πλειοψηφία των περιπτώσεων εμφανίζονται εντός ενός εννοιολογικού πλαισίου, πλούσιου και αυστηρά δομημένου. Αν και θα μπορούσε να πει κανείς ότι το εννοιολογικό πλαίσιο αποτελεί μια από τις αιτίες αφενός του φαινομένου της υπερφόρτωσης πληροφορίας και αφετέρου της αύξησης της πολυπλοκότητας πολλών από τις κλασικές μεθόδους ανίχνευσης εννοιών, μπορεί ταυτόχρονα να υπαγορεύσει και το ποιες έννοιες πρέπει να παρατηρηθούν και τι μπορεί ακίνδυνα να αγνοηθεί, κάτι το οποίο δεν παύει να αποτελεί έναν από τους κύριους στόχους των υπολογιστικών συστημάτων και των μοντέλων οπτικής προσοχής και ανάλυσης εικόνας. Έτσι, η ενσωμάτωσή του στους παραδοσιακούς αλγορίθμους ανάλυσης ει-

Σχήμα 7.3: *Ένας αεραγωγός αρχικά απομονωμένος και έπειτα εντός του εννοιολογικού του πλαισίου. Η αναγνώρισή του είναι ιδιαίτερα δύσκολη εκτός του εννοιολογικού του πλαισίου, αλλά προφανής εντός αυτού.*



Σχήμα 7.4: *Η θέση του πυροσβεστικού κρουνού παραβιάζει το εννοιολογικό του πλαίσιο, καθιστώντας δυσκολότερη την ορθή του αναγνώριση. Το Σχήμα κατασκευάστηκε από τους Biederman et al. [15].*

κόνας μπορεί να αυξήσει σημαντικά την ακρίβεια που αυτοί επιτυγχάνουν. Αυτή η προσέγγιση ακολουθείται και στο παρόν Κεφάλαιο, όπου ορίζονται κατά σειρά το εννοιολογικό πλαίσιο των εικόνων ενός θεματικού πεδίου, το εννοιολογικό πλάισιο των τύπων περιοχής και το μεικτό εννοιολογικό πλαίσιο που αποτελεί το συνδυασμό τους. Στη συνέχεια, ενσωματώνονται στην τεχνική ανίχνευσης εννοιών που προτάθηκε στο Κεφάλαιο 4 και διερευνάται κατά πόσο μπορεί να βελτιώσει τα αποτελέσματα της ανάλυσης.

7.3 Σχετικές Εργασίες

Υπάρχουν αρκετές παραλλαγές στις τεχνικές που χρησιμοποιούν το εννοιολογικό πλαίσιο. Αυτές μπορούν να χωριστούν σε κατηγορίες σύμφωνα με τον τρόπο που αυτό ορίζεται. Έτσι, στη βιβλιογραφία συναντάται το εννοιολογικό πλαίσιο σκηνής, το χωρικό εννοιολογικό πλαίσιο, το χρονικό εννοιολογικό πλαίσιο και το εννοιολογικό πλαίσιο μεταδεδομένων. Φυσικά υπάρχουν και άλλες παραλλαγές, ή άλλες ονομασίες,

ωστόσο στην παρούσα εργασία χρησιμοποιούνται οι πιο συνήθεις όροι.

7.3.1 Εννοιολογικό Πλαίσιο Σκηνής

Το *εννοιολογικό πλαίσιο σκηνής* είναι ίσως ο πιο απλός τρόπος για τη μοντελοποίηση του εννοιολογικού πλαισίου σε ένα πρόβλημα ανάλυσης εικόνας. Σύμφωνα με τον ορισμό των Torralba et al. [213], το εννοιολογικό πλαίσιο σκηνής ορίζεται σαν ένας συνδυασμός από αντικείμενα/έννοιες, που έχουν συσχετιστεί κατά την κοινή ανθρώπινη αντίληψη και έχουν την ιδιότητα να "αλληλοσυμπληρώνονται", προκειμένου να διευκολύνουν την αναγνώριση αντικειμένων και την κατηγοριοποίηση σκηνής. Προκειμένου να μπορέσει κάποιος να αξιοποιήσει την έκφραση αυτή του εννοιολογικού πλαισίου, πρέπει πρώτα να αντιληφθεί με κάποιο τρόπο την ύπαρξη μιας "διαγνωστικής" έννοιας, δηλαδή μιας έννοιας που συσχετίζεται άμεσα είτε με τη σκηνή, είτε με άλλες έννοιες και με βάση την πεποίθησή του αυτή να συμπεράνει την πιθανότητα για ύπαρξη και άλλων εννοιών, την τοποθεσία τους στη σκηνή κ.ο.κ. Για παράδειγμα, σε ένα πρόβλημα ανίχνευσης εννοιών, αν ανιχνευτεί με μεγάλο βαθμό βεβαιότητας η έννοια *θάλασσα*, εύκολα μπορεί να περιμένει κανείς την ύπαρξη της έννοιας *ουρανός* ή το χαρακτηρισμό της σκηνής ως *παραλία*. Την προσέγγιση αυτή για τη μοντελοποίηση του εννοιολογικού πλαισίου ακολούθησαν οι πρώτες ερευνητικές προσπάθειες.

Οι Naphade και Smith [149] προτείνουν ένα υβριδικό σύστημα το οποίο μπορεί να συνδυάσει μοντέλα της δομής των εννοιών που περιέχονται σε μια εικόνα και του εννοιολογικού πλαισίου. Προτείνουν μια κατηγορία εννοιών που την ονομάζουν *Πολυμεσικό Αντικείμενο* (multimedia object - multijet). Η κατηγορία αυτή περιέχει έννοιες που χαρακτηρίζονται ως αντικείμενα (π.χ. *αυτοκίνητο*, *άνθρωπος*, *ελικόπτερο*), σκηνές (π.χ. *εξωτερικός χώρος*, *παραλία*) και γεγονότα (π.χ. *έκρηξη*, *άνθρωπος που περπατάει*). Ορμώμενοι από την παρατήρηση ότι η παρουσία συγκεκριμένων εννοιών υποδεικνύει μεγάλη πιθανότητα για την ανίχνευση άλλων πολυμεσικών αντικειμένων. Αυτό σημαίνει ότι η ανίχνευση μιας έννοιας σημαίνει ότι όχι μόνο αυξάνεται η πιθανότητα για ανίχνευση μιας άλλης έννοιας, αλλά ο συνδυασμός δύο εννοιών που ανιχνεύτηκαν μπορεί να οδηγησει σε αυξημένη πιθανότητα για την ανίχνευση π.χ. ενός γεγονότος. Οι Fan et al. [64] προτείνουν μια πολυεπίπεδη προσέγγιση για το σχολιασμό εικόνων που περιέχουν φυσικές σκηνές χρησιμοποιώντας τόσο τα κυρίαρχα συστατικά της εικόνας, δηλαδή τα μέρη της εικόνας που παρουσιάζουν κάποια εξέχοντα χαρακτηριστικά, όσο και των σχετικών σημασιολογικών εννοιών. Από τα πρώτα εξάγουν χαρακτηριστικά χαμηλού επιπέδου και στη συνέχεια χρησιμοποιούν μια νέα τεχνική ταξινόμησης που αντιστοιχεί τις εικόνες στις πιο σχετικές έννοιες. Η τεχνική αυτή στηρίζεται στον αλγόριθμο EM. Οι Yang et al. [233] περιγράφουν διάφορες τεχνικές για τη μάθηση των σχέσεων ανάμεσα στις έννοιες, μέσω μιας αναπαράστασης ενός πιθανοτικού μοντέλου. Πιο συγκεκριμένα, διερευνούν τη χρήση μπεϋσιανών δικτύων, μηχανών Boltzmann, τυχαίων μαρκοβιανών πεδίων και υπό συνθήκη τυχαίων πεδίων για να εκφράσουν τις σχέσεις ανάμεσα στις έννοιες του TRECVID 2005, στο σύνολο δεδομένων του οποίου αξιολογούν τις προτεινόμενες τεχνικές. Τελος, οι Murphy et al. [145] χρησιμοποιούν το εννοιολογικό πλάισιο σκηνής σαν μια επιπρόσθετη πληροφορία στις καθολικές περιγραφές, με σκοπό αυτό να βοηθήσει να επιλυθούν αμφιβολίες σχετικά με τις έννοιες που περιέχονται τοπικά στις εικόνες. Χρησιμοποιούν ένα υπό συνθήκη τυχαίο πεδίο για να επιλύσουν από κοινού τα προβλήματα της αναγνώρισης αντικειμένων και της ταξινόμησης σκηνής.

7.3.2 Χωρικό Εννοιολογικό Πλαίσιο

Το *χωρικό εννοιολογικό πλαίσιο* αφορά όπως φανερώνει και το όνομά του τις χωρικές σχέσεις ανάμεσα στις έννοιες. Είναι το προφανές επόμενο βήμα στην ενίσχυση των μοντέλων εννοιολογικού πλαισίου σκηνής. Έτσι, ορίζεται σαν ένα συνδυασμός από αντικείμενα/έννοιες που εκτός από την ιδιότητα της συνύπαρξης, διαθέτουν και την ιδιότητα να εμφανίζονται σε συγκεκριμένη χωρική θέση το ένα ως προς το άλλο. Προκειμένου να εκμεταλλευτεί κάποιος το χωρικό εννοιολογικό πλαίσιο, αφού ανιχνεύσει την ύπαρξη μιας διαγνωστικής έννοιας, όπως στην περίπτωση του εννοιολογικού πλαισίου σκηνής, στη συνέχεια και για έννοιες που μοιάζουν μεταξύ τους, θα βγάλει συμπεράσματα για την ύπαρξή τους με βάση τη θέση τους ως προς τη διαγνωστική έννοια. Στην περίπτωση που υπάρχει η βεβαιότητα για τη σκηνή που απεικονίζεται, τότε ανάλογα με τη θέση κάποιων περιοχών μπορούν να εξαχθούν συμπεράσματα για την έννοια που περιέχουν. Για παράδειγμα, στο πρόβλημα της ανάλυσης, σε περίπτωση που αναγνωριστεί μια σκηνή ως *παραλία*, τότε αν για μια περιοχή υπάρχει αμφιβολία για το αν περιέχει την έννοια *ουρανός*, ή την έννοια *θάλασσα*, καθώς τα οπτικά χαρακτηριστικά των δύο αυτών εννοιών είναι παρόμοια, η θέση της θα καθορίσει ποια από τις δύο έννοιες περιέχεται. Η έννοια *ουρανός* βρίσκεται συνήθως πάνω από όλες τις έννοιες, ενώ η έννοια *θάλασσα* συνήθως βρίσκεται στο κέντρο της εικόνας. Επίσης, ψυχοφυσικές μελέτες όπως για παράδειγμα αυτή των Cave και Kosslyn [39] έδειξαν ότι αν ένα αντικείμενο χωριστεί στα μέρη του και αυτά διασκορπιστούν στο χώρο με τυχαία θέση, όσο πιο πολύ απέχει η διάταξή τους από την πραγματική κατά το σχηματισμό του αντικειμένου, τόσο δυσκολεύει η αναγνώρισή τους. Στο πλαίσιο αυτό και με βάσεις παρόμοιες παρατηρήσεις κινήθηκαν αρκετές ερευνητικές προσπάθειες.

Οι Lipson et al. [122] χρησιμοποίησαν ποσοτικές χωρικές και φωτομετρικές σχέσεις ανάμεσα σε περιοχές, σε εικόνας χαμηλής ανάλυσης. Εφάρμοσαν την τεχνική τους σε πρόβλημα αναγνώρισης φυσικών σκηνών και έδειξαν πώς μπορούν οι σχέσεις ανάμεσα στις έννοιες να προκύψουν από παραδείγματα. Η εργασία αυτή ήταν από τις πρώτες που αξιοποίησαν το εννοιολογικό πλαίσιο σε πρόβλημα αναγνώρισης σκηνής και μάλιστα την εποχή που οι ερευνητικές προσπάθειες είχαν επικεντρωθεί σε καθολικά χαρακτηριστικά. Οι Singhal et el. [187] εργάζονται στο πρόβλημα της ανίχνευσης υλικών, όπως *ουρανός*, *γρασίδι*, *νερό* κλπ. Χρησιμοποιούν μοντέλα χωρικού εννοιολογικού πλαισίου, προκειμένου να ξεπεράσουν τις λανθασμένες ταξινομήσεις που προκύπτουν όταν δύο υλικά μοιάζουν οπτικά μεταξύ τους, όπως για παράδειγμα ο *ουρανός* και το *νερό*. Ωστόσο, κάποιες έννοιες συχνά έχουν συγκεκριμένες χωρικές σχέσεις ανάμεσά τους όταν εμφανίζονται στις εικόνες και αυτό ακριβώς μοντελοποιούν με τη χρήση πιθανοτικών μοντέλων. Σε συνέχεια της εργασίας αυτής, Luo et al. [128] εξελίσσουν την τεχνική τους, μοντελοποιώντας τις χωρικές σχέσεις μεταξύ των εννοιών με συναρτήσεις πυκνότητας πιθανότητας, οι οποίες προκύπτουν από μια διαδικασία μάθησης. Οι Carbonetto et al. [36] προτείνουν ένα μοντέλο το οποίο λαμβάνει υπόψην του τις χωρικές σχέσεις ανάμεσα στις έννοιες. Το μοντέλο αυτό χρησιμοποιεί μια πιθανοτική αντιστοίχηση ανάμεσα σε διανύσματα που περιέχουν την περιγραφή χαμηλού επιπέδου των εικόνων και τις λέξεις που περιγράφουν τις πιθανές έννοιες. Έτσι, μέσω μιας διαδικασίας μάθησης, το μοντέλο αυτό κατασκευάζει σχέσεις ανάμεσα σε περιοχές εικόνων και εννοιών, αλλά και χωρικές σχέσεις ανάμεσα σε έννοιες. Μια από τις εφαρμογές του είναι η ένωση περιοχών σε υπερ-κατατμημένες εικόνες. Βασίζεται σε μια προσεγγιστική εκδοχή του EM αλγορίθμου. Τα πειραματικά αποτε-

λέσματα δείχνουν ότι επιτυγχάνεται αξιοσημείωτη βελτίωση της ακρίβειας, ακόμη και σε εικόνες που οι περιοχές τους έχουν προκύψει έπειτα από υπερ-κατάτμηση. Οι Li και Sun [118] χωρίζουν την εικόνα με τη χρήση ενός πλέγματος και χρησιμοποιώντας μια τεχνική που βασίζεται σε υπό συνθήκη τυχαία πεδία και ημι-επιβλεπόμενη μάθηση, κωδικοποιούν τόσο τις σχέσεις ανάμεσα σε περιγραφείς και έννοιες, όσο και χωρικές σχέσεις ανάμεσα σε γειτονικές έννοιες. Οι Li et al. [115] υιοθετούν επίσης τη μέθοδο χωρισμού των περιοχών σε εικόνες με τη χρήση ενός πλέγματος. Από κάθε ένα από τα δομικά στοιχεία της εικόνας που προκύπτουν, εξάγουν χαρακτηριστικά χαμηλού επιπέδου. Στη συνέχεια, για να εκμεταλλευτούν τις χωρικές σχέσεις ανάμεσα στις έννοιες, χρησιμοποιούν ένα 2-Δ κρυφό μαρκοβιανό μοντέλο. Παρόμοια είναι και η προσέγγιση των Jiten et al. [93] οι οποίοι χρησιμοποιώντας επίσης ένα 2-Δ κρυφό μαρκοβιανό μοντέλο, διερευνούν πώς η ισορροπία ανάμεσα στην πληροφορία για τη χωρική δομή των εικόνων και στην περιγραφή χαμηλού επιπέδου των εικόνων επηρεάζουν την ακρίβεια στην ανίχνευση εννοιών. Οι Yuan et al. [238] χρησιμοποιούν διάφορα μοντέλα γράφων, με σκοπό να μοντελοποιήσουν το χωρικό εννοιολογικό πλαίσιο. Χρησιμοποιούν πιθανοτική στατιστική και αποδεικνύουν ότι τα μοντέλα που συμπεριλαμβάνουν χωρικές σχέσεις αποδίδουν καλύτερα από τα απλά. Οι Mylonas και Avrithis [147] προτείνουν ένα σύνολο από χωρικές σχέσεις που μπορούν να χρησιμοποιηθούν για την μοντελοποίηση του χωρικού εννοιολογικού πλαισίου σε πολυμεσικό περιεχόμενο. Τέλος, οι Boutell και Brown [23] διερευνούν τρόπους για τη μάθηση των σχέσεων σε σκηνές ανάμεσα στα υλικά από τα οποία αυτές αποτελούνται. Προτείνουν τη χρήση ενός generative μοντέλου που χρησιμοποιεί ανιχνευτές υλικών, αλλά και γνώση που έχει προκύψει στατιστικά για τη διάταξη που έχουν οι σκηνές. Όπως παραδέχονται και οι ίδιοι, η τεχνική τους εμφανίζει το βασικό μειονέκτημα ότι απαιτεί τη συλλογή σχετικά μεγάλου αριθμού από δεδομένα εκπαίδευσης για να κατασκευαστεί η γνώση. Η τεχνική αυτή βασίζεται σε μεθόδους ταιριάσματος γράφων. Στη συνέχεια, επεκτείνουν το μοντέλο αυτό, ενσωματώνοντας και πληροφορίες του εννοιολογικού πλαισίου σκηνής [28].

7.3.3 Χρονικό Εννοιολογικό Πλαίσιο

Ως *χρονικό εννοιολογικό πλαίσιο* ορίζεται αυτό που αφορά τις σχέσεις μιας εικόνας με εικόνες που έχουν ληφθεί σε κοντινές χρονικές στιγμές και ανήκουν στην ίδια συλλογή εικόνων. Έτσι, ορίζεται σαν ένα σύνολο από εικόνες που έχουν ληφθεί σε ένα στενό χρονικό πλαίσιο και ανήκουν στην ίδια συλλογή, ανεξάρτητα από το αν έχουν το ίδιο οπτικό περιεχόμενο. Η ανάλυση μιας εικόνας ή, ακόμη καλύτερα, ο σχολιασμός της μπορεί να οδηγήσει σε χρήσιμα συμπεράσματα για το σημασιολογικό περιεχόμενο των εικόνων που απαρτίζουν το χρονικό εννοιολογικό της πλαίσιο και αντίστροφα. Έτσι αν μια εικόνα σχολιαστεί με την έννοια *παραλία* ή αναγνωριστεί με μεγάλη βεβαιότητα η απεικόνιση της συγκεκριμένης σκηνής και για μια εικόνα που έχει ληφθεί από τον ίδιο χρήστη μέσα σε μικρό χρονικό διάστημα, υπάρχει αμφιβολία για το αν μια εικόνα απεικονίζει την έννοια *θάλασσα* ή την έννοια *ποτάμι*, τότε με βάση το χρονικό εννοιολογικό της πλαίσιο, η ανάλυση οδηγείται στο σύμπερασμα της ύπαρξης της πρώτης έννοιας. Προκειμένου να διαφανεί η σημασία του χρονικού εννοιολογικού πλαισίου, στο Σχήμα 7.5 απεικονίζεται μια εικόνα και το χρονικό εννοιολογικό της πλαίσιο. Οι εικόνες αυτές αποτελούν τμήμα προσωπικής συλλογής. Στο Σχήμα αυτό μπορεί να παρατηρηθεί ότι για την μεσαία εικόνα που απεικονίζει μια σκηνή *παραλίας*, εξαιτίας του τρόπου με τον οποίο έχει ληφθεί είναι δύσκολο να

εξαχθεί με βεβαιότητα το συμπέρασμα αυτό, ακόμη και από πολύ καλά εκπαιδευμένους ανιχνευτές. Ωστόσο, οι εικόνες που αποτελούν το χρονικό εννοιολογικό τους πλαίσιο και έχουν ληφθεί σε διάστημα 10 λεπτών, μπορούν σχετικά εύκολα να ταξινομηθούν ως σκηνές *παραλίας* και έτσι να οδηγήσουν σε αυτό το συμπέρασμα για τη μεσαία εικόνα.

Σχήμα 7.5: *Το χρονικό εννοιολογικό πλαίσιο της εικόνας που απεικονίζεται στη μέση.*

Η τεχνική των Boutell et al. [27] ξεκινά με την ιδέα ότι σε συλλογές εικόνων, όταν υπάρχει η πληροφορία σχετικά με τη χρονική στιγμή που έχουν αυτές ληφθεί, οι εικόνες που βρίσκονται χρονικά γύρω από αυτή που είναι υπό ταξινόμηση, μπορούν να χρησιμοποιηθούν ως το χρονικό εννοιολογικό πλαίσιό της. Οι συγγραφείς προτείνουν τη χρήση ενός πιθανοτικού μοντέλου, μέσω του οποίου προσπαθούν να συνδυάσουν τα χαρακτηριστικά χαμηλού επιπέδου του περιεχομένου μιας εικόνας με αυτά του χρονικού εννοιολογικού πλαισίου της, δηλαδή με αυτά των γειτονικών χρονικά εικόνων της. Κάνουν την παραδοχή ότι όσο πιο κοντά χρονικά έχουν ληφθεί οι εικόνες, τόσο περισσότερο είναι πιθανό να σχετίζονται οπτικά και άρα και σημασιολογικά. Επίσης, προτείνουν μια παραλλαγή της μεθόδου, η οποία μπορεί να εφαρμοστεί σε συλλογές εικόνων όπου υπάρχει η πληροφορία για την ταξινόμησή τους χρονικά, χωρίς όμως να είναι γνωστές οι ακριβείς χρονικές στιγμές που αυτές έχουν ληφθεί. Παρόμοια είναι και η ιδέα της τεχνικής των Paletta et al. [163], οι οποίοι χρησιμοποιούν πολλές διαδοχικές εικόνες από διαφορετικές γωνίες απλών αντικειμένων και αξιοποιούν το χρονικό εννοιολογικό τους πλαίσιο στο πρόβλημα ανάκτησης 3-Δ εικόνων. Το σύστημα που προτείνουν δίνει έμφαση στο γεγονός ότι η χρονική πληροφορία μπορεί να δώσει λύση σε περιπτώσεις όπου υπάρχει σχετική αμφιβολία και βασίζεται σε ένα μπεϋσιανό μοντέλο. Οι Moldovan et al. [142] εργάζονται στο χώρο της ανάλυσης κειμένου και προσπαθούν να ανιχνεύσουν γεγονότα που σχετίζονται χρονικά σε κείμενα φυσικής γλώσσας. Έπειτα, μετατρέπουν τα γεγονότα σε μια εμπλουτισμένη λογική αναπαράσταση. Η συλλογιστική παρέχεται από μια τεχνική αποδείξεων πρώτης τάξης λογικής, η οποία προσαρμόζεται κατάλληλα ώστε να λειτουργεί σε κείμενο. Οι O'Hare et al. [155] συνδυάζουν το χρονικό εννοιολογικό πλαίσιο με την πληροφορία της τοποθεσίας στην οποία έχει ληφθεί η φωτογραφία, για να διευκολύνουν την ανάκτηση εικόνων. Οι Pauty et al. [166] παρουσιάζουν μια εφαρμογή για την πλοήγηση σε συλλογές φωτογραφιών, η οποία χρησιμοποιεί το εννοιολογικό πλαίσιο. Το χρονικό εννοιολογικό πλαίσιο συνδυάζεται με τη γνώση για την τοποθεσία που έχει ληφθεί μια φωτογραφία και έτσι διευκολύνεται η πλοήγηση σε εικόνες που έχουν ληφθεί στις επιθυμητές χρονικές περιόδους. Τέλος, οι Mulhem και Lim [144] προτείνουν τη χρήση χρονικών γεγονότων για την οργάνωση και την αναπαράσταση οικιακών φωτογραφιών και μια νέα μέθοδο για την ανάκτηση εικόνων χρησιμοποιώντας τόσο το οπτικό περιεχόμενό τους, όσο και το χρονικό εννοιολογικό τους πλαίσιο. Περιγράφουν ένα ιεραρχικό μοντέλο από χρονικά γεγονότα, το οποίο κατασκευάζεται από μια συλλογή εικόνων.

7.3.4 Εννοιολογικό Πλαίσιο Μετα-Πληροφοριών

Το *εννοιολογικό πλαίσιο μετα-πληροφοριών* αφορά τις σχέσεις ανάμεσα στις μετα-πληροφορίες που είναι διαθέσιμες στα ψηφιακά αρχεία φωτογραφιών, όπως είναι για παράδειγμα οι ρυθμίσεις που είχε η φωτογραφική μηχανή τη στιγμή της λήψης. Οι μετα-πληροφορίες ενσωματώνονται στα αρχεία, ακολουθώντας το ευρέως διαδεδομένο πρότυπο EXIF (EXchangeable Image File [52]). Το πρότυπο αυτό αποθηκεύει τις παραμέτρους με τις οποίες ήταν ρυθμισμένη η φωτογραφική μηχανή κατά τη λήψη, όπως ο χρόνος έκθεσης, το διάφραγμα, η εστιακή απόσταση και άλλες. Φυσικά αποθηκεύει και τη χρονική στιγμή που έγινε η λήψη. Έτσι, αν για παράδειγμα για μια εικόνα δεν υπάρχει βεβαιότητα με βάση τα οπτικά της χαρακτηριστικά αν απεικονίζει μια σκηνή *εσωτερικού ή εξωτερικού χώρου*, η γνώση της εστιακής απόστασης μπορεί να οδηγήσει σε συμπέρασμα. Αν αυτή είναι ιδιαίτερα μεγάλη, τότε είναι πιθανότερο να απεικονίζεται μια σκηνή *εξωτερικού χώρου*.

Οι Sinha και Jain [188] αξιοποιούν τις μετα-πληροφορίες που είναι διαθέσιμες στις ψηφικές φωτογραφίες και με βάση αυτές καταλήγουν σε συμπεράσματα για τη σημασιολογική πληροφορία που αυτές περιέχουν. Οι συγγραφείς αξιοποιούν μόνο τη μετα-πληροφορία των ρυθμίσεων και τη συνδυάζουν με οπτικά χαρακτηριστικά χαμηλού επιπέδου, προκειμένου να επιτύχουν τη δημιουργία αυτόματων σχολιασμών σε φωτογραφίες. Τα πειραματικά τους αποτελέσματα δείχνούν ότι η πληροφορία αυτή είναι πραγματικά πολύ σημαντική και προσφέρει αξιοσημείωτη βελτίωση της ακρίβειας στην ταξινόμηση. Οι Liu et al. [125] αντιμετωπίζουν το πρόβλημα ταξινόμησης σκηνής *εσωτερικού/εξωτερικού χώρου*, *πόλης/τοπίου*, καθώς και το πρόβλημα του προσανατολισμού των εικόνων, με την αξιοποίηση των μεταπληροφοριών. Χρησιμοποιώντας έναν αλγόριθμο boosting έδειξαν ότι η ακρίβεια που επιτυγχάνεται με τη χρήση μόνο των μεταπληροφοριών είναι συγκρίσιμη με αυτή που επιτυγχάνεται με τα οπτικά χαρακτηριστικά. Επιπρόσθετα, ο συνδυασμός τους δημιουργεί έναν νέο τρόπο περιγραφής, ο οποίος όταν χρησιμοποιηθεί για την ταξινόμηση επιτυγχάνει μεγαλύτερη ακρίβεια και τελικά παρουσιάζει σταθερή απόδοση για διαφορετικά σύνολα δεδομένων. Καθώς οι μεταπληροφορίες είναι διαθέσιμες στα ψηφιακά αρχεία φωτογραφιών, η τεχνική αυτή είναι ιδιαίτερα γρήγορη. Οι Tuffield et al. [216] χρησιμοποιούν τις διαθέσιμες μετα-πληροφορίες των φωτογραφιών και τις συνδυάζουν με οπτικά χαρακτηριστικά χρώματος, καθώς και με ανιχνευτές *προσώπου* και *φυσικού χώρου*. Το σύστημα που παρουσιάζουν αξιοποιεί όλες αυτές τις πληροφορίες και τελικά προτείνει κάποιους πιθανούς σχολιασμούς, τους οποίους και αποδέχονται ή όχι οι χρήστες. Οι Boutell και Luo [24], [26], [25] αντιμετωπίζουν δύο προβλήματα, αυτό της ταξινόμησης *εσωτερικού/εξωτερικού χώρου* και αυτό της ανίχνευσης *ηλιοβασιλέματος*. Η στατιστική ανάλυση των μετα-πληροφοριών που συνοδεύουν τις φωτογραφίες κάθε κατηγορίας αποδεικνύει ότι κάποια πεδία όπως ο χρόνος έκθεσης, η χρήση ή όχι του φλας και η απόσταση του θέματος παρέχουν ισχυρή διαχωρισιμότητα και στα δύο προβλήματα. Ένα μπεϋσιανό δίκτυο χρησιμοποιείται για να συγχωνεύσει τις πληροφορίες από τα χαρακτηριστικά χαμηλού επιπέδου χρώματος και υφής και αυτά των μετα-πληροφοριών. Η τεχνική που προτείνουν αποδίδει ικανοποιητικά ακόμα και στην περίπτωση που κάποια από τα δεδομένα απουσιάζουν. Οι Viana et al. [223] κατασκεύασαν μια οντολογία εννοιολογικού πλαισίου, στην οποία ιδιαίτερο ρόλο παίζουν οι μετα-πληροφορίες που είναι ενσωματωμένες στις ψηφιακές φωτογραφίες. Οι Pigeau και Gelgon [173] συνδυάζουν τις μετα-πληροφορίες με την πληροφορία σχετικά με το πού έχει τραβηχτεί η φωτογραφία. Τέλος, οι Boll et al. [18] προτείνουν ένα σύστημα

που συνδυάζει τις μεταπληροφορίες και τα οπτικά χαρακτηριστικά των εικόνων και προσπαθούν να εκμεταλλευτούν το εννοιολογικό πλαίσιο για να παρέχουν προχωρημένες ψηφιακές υπηρεσίες, όπως την αυτόματη κατασκευή φωτογραφικών συλλογών.

7.4 Το Εννοιολογικό Πλαίσιο των Εικόνων ενός Θεματικού Πεδίου

Στην Ενότητα αυτή προτείνεται μια τεχνική η οποία αποσκοπεί στο να βελτιώσει την ακρίβεια στα αποτελέσματα της ανίχνευσης εννοιών υψηλού επιπέδου που παρουσιάστηκε στο Κεφάλαιο 4, αξιοποιώντας το εννοιολογικό πλαίσιο του θεματικού πεδίου, στο οποίο και εφαρμόζεται. Το εννοιολογικό πλαίσιο ορίστηκε και περιγράφηκε στην Ενότητα 7.2 και αποτελείται από ένα σύνολο εννοιών και ένα σύνολο από σχέσεις ανάμεσά τους. Αρχικά προτείνεται η χρήση μιας οντολογίας η οποία αποτελείται από τις έννοιες αυτές και τις μεταξύ τους σχέσεις. Οι σχέσεις αυτές επιλέγονται έτσι ώστε να έχουν νόημα σε προβλήματα ανίχνευσης εννοιών και επαναορίζονται προκειμένου να συμπεριλάβουν *ασάφεια* στον ορισμό τους. Τέλος, προτείνεται ένας αλγόριθμος που επιδρώντας στις αρχικές εκτιμήσεις για την ύπαρξη μιας έννοιας και λαμβάνοντας υπόψη τις σχέσεις αυτών των εννοιών στο συγκεκριμένο θεματικό πεδίο, επαναϋπολογίζει και βελτιστοποιεί τις αρχικές εκτιμήσεις.

7.4.1 Οντολογία Εννοιολογικού Πλαισίου

Προκειμένου να αξιοποιηθεί το εννοιολογικό πλαίσιο ενός θεματικού πεδίου, το πρώτο βήμα που πρέπει να γίνει είναι να βρεθεί κατάλληλος τρόπος για την αναπαράσταση των εννοιών και των σημασιολογικών σχέσεων που τις διέπουν σε ένα συγκεκριμένο θεματικό πεδίο. Οι Mylonas et al [146] έδειξαν ότι κάθε σύνολο από σημασιολογικές σχέσεις μεταξύ εννοιών μπορεί να αναπαρασταθεί από μια κατάλληλα ορισμένη οντολογία. Στην περίπτωση που εξετάζεται, καθώς η φύση των σχέσεων ανάμεσα στις έννοιες ενός θεματικού πεδίου στον πραγματικό κόσμο διέπεται από ασάφεια και αμφιβολία, είναι αναγκαίο οι σχέσεις αυτές να αποκτήσουν *ασάφεια*. Έτσι, προτείνεται η κατασκευή μιας *ασαφούς* οντολογίας, η οποία εισάγεται με σκοπό να μοντελοποιήσει τις σχέσεις ανάμεσα στις έννοιες που καθορίζουν ένα θεματικό πεδίο του πραγματικού κόσμου.

Για τον ορισμό μιας οντολογίας αρκεί να οριστεί το σύνολο των Εννοιών της καθώς και το σύνολο των σχέσεων μεταξύ τους. Στην προκειμένη περίπτωση, μια οντολογία O_c που μοντελοποιεί το εννοιολογικό πλαίσιο ενός θεματικού πεδίου αποτελείται από το σύνολο των εννοιών υψηλού επιπέδου του θεματικού πεδίου, καθώς και τις σημασιολογικές σχέσεις που ορίζονται μεταξύ τους. Πιο αυστηρά, έστω

1. $C = \{c_i\}, i = 1, 2, \ldots, n$ το σύνολο των εννοιών υψηλού επιπέδου που περιέχονται στο υπό εξέταση θεματικό πεδίο και

2. $R_c = \{R_{c,ij}\}, i, j = 1, 2, \ldots n$ το σύνολο όλων των σημασιολογικών σχέσεων ανάμεσα στις έννοιες. Το σύνολο $R_{c,ij} = r^{(k)}_{c,ij}, k = 1, 2, \ldots, K$ περιλαμβάνει τις K σχέσεις που ορίζονται ανάμεσα σε δύο έννοιες c_i, c_j. Επιπρόσθετα, για μια σχέση $r^{(k)}_{c,ij}$ ορίζεται και η αντίστροφή της $\bar{r}^{(k)}_{c,ji}$.

Άρα για μια οντολογία O_c και τα σύνολα C και $R_{c,ij}$ μπορούν να διατυπωθούν οι σχέσεις

$$O_c = \{C, R_c\} \tag{7.1}$$

και

$$R_{c,ij} : C \times C \rightarrow \{0, 1\} \ , \tag{7.2}$$

όπου φαίνεται ότι μια οντολογία O_c μπορεί να περιγραφεί σαν σύνολο από έννοιες και τις μεταξύ τους σχέσεις και μια σχέση $R_{c,ij}$ έχει σύνολο τιμών το $\{0, 1\}$, δηλαδή είτε ορίζεται ανάμεσα σε δύο έννοιες της οντολογίας, είτε παραλείπεται.

Όπως είναι φανερό, η μοντελοποίηση ενός θεματικού πεδίου όπως περιγράφεται από την (7.1) δεν περιέχει την επιθυμητή ασάφεια. Για το λόγο αυτό, το μοντέλο στη συνέχεια επεκτείνεται προκειμένου οι σχέσεις μεταξύ των εννοιών να γίνουν ασαφείς, όπως αυτές που συναντώνται στον πραγματικό κόσμο. Έτσι, ορίζεται μια ασαφοποιημένη εκδοχή $\mathcal{O}_c$ της οντολογίας εννοιολογικού πλαισίου O_c ως

$$\mathcal{O}_c = \{C, \mathcal{R}_c\} \ , \tag{7.3}$$

όπου το C είναι το σύνολο όλων των πιθανών εννοιών και δεν επηρεάζεται, ενώ το $\mathcal{R}_c$ είναι το σύνολο των ασαφών σημασιολογικών σχέσεων. Κατά αντιστοιχία με την προηγούμενη περίπτωση, $\mathcal{R}_c = \{\mathcal{R}_{c,ij}\}$. Το σύνολο $\mathcal{R}_{c,ij} = \mathrm{r}_{c,ij}^{(k)}, k = 1, 2, \ldots, K$ περιλαμβάνει τις K σχέσεις που ορίζονται ανάμεσα σε δύο έννοιες c_i, c_j. Επιπρόσθετα, για μια σχέση $\mathrm{r}_{c,ij}^{(k)}$ ορίζεται και η αντίστροφή της $\bar{\mathrm{r}}_{c,ji}^{(k)}$. Στην περίπτωση αυτή, ισχύει ότι

$$\mathrm{r}_{c,ij} : C \times C \rightarrow [0, 1] \ . \tag{7.4}$$

Καθώς είναι πιθανό να υπάρχουν περισσότερες από μία σχέσεις ανάμεσα σε δύο έννοιες, ορίζεται ο συνδυασμός των σχέσεων ως

$$\mathcal{U}_{c,ij} = \bigcup_k [\mathrm{r}_{c,ij}^{(k)}]^p, \ i, j = 1, 2, \ldots, N, \ k = 1, 2, \ldots, K \ . \tag{7.5}$$

Με τον τρόπο αυτό ορίζεται το μοντέλο αναπαράστασης του εννοιολογικού πλαισίου που θα χρησιμοποιηθεί κατά τη φάση της ανάλυσης. Η τιμή του p καθορίζεται από τη σημασιολογία της κάθε σχέσης $\mathrm{r}_{c,ij}$ που χρησιμοποιείται κατά την κατασκευή της $\mathcal{U}_{c,ij}$. Πιο συγκεκριμένα:

- $p = 1$, αν η σημασιολογία της $\mathrm{r}_{c,ij}$ υπονοεί ότι πρέπει να χρησιμοποιηθεί ως έχει,
- $p = -1$, αν η σημασιολογία της $\mathrm{r}_{c,ij}$ υπονοεί τη χρήση της αντίστροφης σχέσης $\bar{\mathrm{r}}_{c,ji}$ και
- $p = 0$, αν η σημασιολογία της $\mathrm{r}_{c,ij}$ δεν επιτρέπει τη συμμετοχή της, αλλά ούτε και τη συμμετοχή της αντίστροφής της $\bar{\mathrm{r}}_{c,ji}$ στο σχηματισμό της συνδυασμένης σχέσης $\mathcal{U}_{c,ij}$.

Προκειμένου να εξασφαλιστεί η περιγραφικότητα της οντολογίας που μοντελοποιεί το εννοιολογικό πλαίσιο ενός θεματικού πεδίου, θα πρέπει αυτή να περιέχει έναν αντιπροσωπευτικό αριθμό από ποικίλες και σωστά ορισμένες σχέσεις ανάμεσα στις έννοιες, με σκοπό να διασκορπίσει ανάμεσά τους την υπάρχουσα γνώση.

7.4.2 Σημασιολογικές Σχέσεις μεταξύ Εννοιών

Στο πλαίσιο του προτύπου MPEG-7, οι Benitez et al. [12] όρισαν έναν μεγάλο αριθμό από σημασιολογικές σχέσεις, οι οποίες μπορούν να εφαρμοστούν σε ανάλυση πολυμεσικών εγγράφων. Από τις σχέσεις αυτές επιλέγονται όσες έχουν νόημα ανάμεσα σε δύο έννοιες υψηλού επιπέδου και μπορούν να εφαρμοστούν στο παρόν πρόβλημα. Οι σχέσεις που επιλέχθηκαν επεκτείνονται και επαναορίζονται, προκειμένου να συμπεριλάβουν την απαιτούμενη ασάφεια. Αυτό επιτυγχάνεται με το να ανατεθεί σε κάθε σχέση ένας *βαθμός βεβαιότητας*. Όλες οι σχέσεις και οι αντίστροφές τους συνοψίζονται στον Πίνακα 7.3.

Σχέση	**Αντίστροφη**	**Σύμβολο**	**Σημασία**
Ειδίκευση	*Γενίκευση*	$Sp(a,b)$	η b είναι μια γενίκευση της ερμηνείας της a
Μέρος	*ΜέροςΤου*	$P(a,b)$	η b είναι μέρος της a
Παράδειγμα	*ΠαράδειγμαΤου*	$Ex(a,b)$	η b είναι ένα παράδειγμα της a
Όργανο	*ΌργανοΤου*	$Ins(a,b)$	η b είναι ένα όργανό της ή εμπλέκεται από την a
Θέση	*ΘέσηΤου*	$Loc(a,b)$	η b είναι η θέση της a
Ασθενής	*ΑσθενήςΤου*	$Pat(a,b)$	η b επηρεάζεται από ή υπομένει τη δράση της a
Ιδιότητα	*ΙδιότηταΤου*	$Pr(a,b)$	η b είναι μια ιδιότητα της a

Πίνακας 7.1: *Οι σημασιολογικές σχέσεις που επιλέχθηκαν για τη μοντελοποίηση του εννοιολογικού πλαισίου.*

Πρέπει να καταστεί ξεκάθαρο ότι οι βαθμοί με τους οποίους ισχύουν οι σχέσεις που επιλέχτηκαν δεν μπορούν να προσδιοριστούν με κάποια υπολογιστική ή στατιστική μέθοδο, λόγω της φύσης τους. Για παράδειγμα, δεν υπάρχει κάποιος ποσοτικός τρόπος για να προσδιοριστεί ο βαθμός βεβαιότητας με τον οποίο π.χ. η έννοια *παραλία* είναι *Θέση* της έννοιας *ομπρέλα*. Έτσι, οι βαθμοί βεβαιότητας καθορίζονται από κάποιον ειδικό του θεματικού πεδίου στο οποίο γίνεται η ανάλυση, που θα αποκαλείται εφεξής "ειδήμων". Στη συνέχεια, ακολουθεί επεξήγηση της σημασιολογίας των σχέσεων αυτών, με κατάλληλα παραδείγματα, προκειμένου να αποσαφηνιστεί η ερμηνεία τους.

- Η σχέση *Ειδίκευση* σημαίνει ότι μια έννοια εξειδικεύει τη σημασία κάποιας άλλης έννοιας. Για παράδειγμα, η έννοια *αρκούδα* αποτελεί μια εξειδίκευση της έννοιας *ζώο*.

- Η σχέση *Μέρος* σημαίνει ότι μια έννοια αποτελεί μέρος ή υποσύνολο μιας άλλης έννοιας. Για παράδειγμα η έννοια *δέντρο* αποτελεί μέρος της έννοιας *δάσος*, ενώ η έννοια *θάλασσα* αποτελεί μέρος της έννοιας *παραλία*.

- Η σχέση *Παράδειγμα* σημαίνει ότι μία έννοια αποτελεί ένα παράδειγμα μιας άλλης έννοιας. Έτσι, η έννοια *χώμα* αποτελεί ένα παράδειγμα της έννοιας *υλικό*.

- Η σχέση *Όργανο* σημαίνει ότι μια έννοια αποτελεί όργανο μιας άλλης ή εμπλέκεται από μια άλλη. Έτσι, η έννοια *ρόδα* αποτελεί ένα όργανο της έννοιας *αυτοκίνητο*.

- Η σχέση *Θέση* δηλώνει ότι μια έννοια αποτελεί τη χωρική θέση μιας άλλης. Έτσι, η έννοια *παραλία* αποτελεί θέση της έννοιας *ομπρέλα*.

- Η σχέση *Ασθενής* σημαίνει ότι σημαίνει ότι μια έννοια επηρεάζεται από μια άλλη, ή υπομένει τη δράση της. Έτσι, η έννοια *όπλο* αποτελεί υπομένει τη δράση της έννοιας *στρατιώτης*.

- Τέλος, η σχέση *Ιδιότητα* σημαίνει ότι μια έννοια αποτελεί ιδιότητα μιας άλλης έννοιας. Έτσι, η έννοια *κυματιστός* αποτελεί ιδιότητα της έννοιας *θάλασσα*.

Μετά το συνδυασμό των σχέσεων ανάμεσα σε δύο έννοιες c_i και c_j και την κατασκευή μοναδικής σχέσης $\mathcal{U}_{c,ij}$ ανάμεσά τους, η οντολογία που μοντελοποιεί το εννοιολογικό πλαίσιο είναι ισοδύναμη με έναν RDF γράφο. Η περιγραφή του βαθμού εμπιστοσύνης χρησιμοποιεί την τεχνική RDF reification [10]. Η τεχνική αυτή χρησιμοποιείται στο χώρο της γνώσης για να αναστήει γεγονότα για τα οποία η γνώση που υπάρχει προέρχεται από κάποια *μαρτυρία*, όπως για παράδειγμα, η σύγκριση ανάμεσα σ λογικές καταθέσεις από διαφορετικούς μάρτυρες, με σκοπό να καθοριστεί τελικά η αξιοπιστία τους. Το μήνυμα "*Ο Γιάννης έχει ύψος 1.63 μέτρα*" είναι μια κατάθεση για την αξιοπιστία της οποίας βαρύνεται αυτός που το αναφέρει. Αντίθετα, για την κατάθεση "*Ο Γιώργος αναφέρει ότι ο Γιάννης έχει ύψος 1.63 μέτρα*", ο Γιώργος βαρύνεται για την αξιοπιστία της. Με αυτό τον τρόπο, οι διάφορες καταθέσεις μπορεί να περιέχουν ασαφή πληροφορία, για παράδειγμα, "*Ο Γιάννης έχει ύψος 1.63 μέτρα, με βαθμό βεβαιότητας ίσο με 0.90*", χωρίς να δημιουργούν ανακολουθίες στη συλλογιστική, μιας και μια νέα κατάθεση κατασκευάζεται για την αρχική και περιέχει την πληροφορία του βαθμού βεβαιότητας. Φυσικά, η κατάθεση που έχει χρησιμοποιήσει την προαναφερθείσα τεχνική δεν πρέπει να γίνει αποδεκτή αυτόματα, ένα γεγονός που αποδεικνύει ότι η χρήση της είναι ουσιαστική. Για παράδειγμα, μια RDF τριπλέτα όπως: "*άμμος μέροςΤου παραλια*" με βαθμό βεβαιότητας "*0.85*" για την κατάθεση, προφανώς και δεν σημαίνει ότι η *άμμος* θα αποτελεί *πάντα* μέρος από μια σκηνή *παραλίας*.

Οι κόμβοι του γράφου αντιστοιχούν προφανώς στις έννοιες υψηλού επιπέδου που περιέχει το επιλεγμένο θεματικό πεδίο και οι ακμές στις συνδυασμένες σχέσεις ανάμεσα στις αντίστοιχες έννοιες. Ο συνδυασμός αυτός αποτελεί μια ασαφή τομή και στην πράξη πραγματοποιείται με τη χρήση κατάλληλου τελεστή της ασαφούς άλγεβρας. Στην παρούσα προσέγγιση υιοθετείται η χρήση κατάλληλης t-νόρμας. Ο βαθμός εμπιστοσύνης που αντιστοιχεί σε κάθε ακμή αντιπροσωπεύει την ασάφεια του μοντέλου. Ανάμεσα σε έννοιες που δεν υπάρχουν ακμές υποννοείται ότι δεν υπάρχουν και σχέσεις (δηλαδή παραλείπονται σχέσεις με μηδενικές τιμές).

Πρέπει να αποσαφηνιστεί ότι δεν εφαρμόζονται όλες οι σχέσεις ανάμεσα σε οποιεσδήποτε έννοιες, αλλά μόνο αυτές που η σημασιολογία τους το επιτρέπει. Έτσι, μια ακμή ανάμεσα σε ένα ζεύγος από έννοιες παράγεται με βάση το σύνολο των σχέσεων που έχουν νόημα για το συγκεκριμένο ζεύγος. Για παράδειγμα, η ακμή ανάμεσα στις έννοιες *βράχος* και *άμμος* παράγεται από το συνδυασμό των σχέσεων *Θέση* και *Ασθενής*, ενώ για τις έννοιες *νερό* και *θάλασσα*, η ακμή παράγεται από τις σχέσεις *Ειδίκευση*, *ΜέροςΤου*, *Παράδειγμα*, *Όργανο*, *Θέση* και *Ασθενής*, για την κατασκευή της. Φυσικά, κάθε έννοια έχει διαφορετική πιθανότητα να εμφανιστεί σε μια σκηνή. Έτσι, ένα επίπεδο μοντέλο εννοιολογικού πλαισίου δε θα ήταν επαρκές. Αντίθετα, οι έννοιες σχετίζονται μεταξύ τους, υποννοώντας ότι οι σχέσεις του γράφου είναι τελικά μεταβατικές, δηλαδή δύο έννοιες μπορούν να σχετιστούν έμμεσα, μέσω τρίτης έννοιας.

7.4.3 Ανίχνευση Εννοιών Υψηλού Επιπέδου

Για την εξαγωγή των χαμηλού επιπέδου χαρακτηριστικών της εικόνας, ακολουθείται η μεθοδολογία που περιγράφηκε στο Κεφάλαιο 4. Συνοπτικά, εξάγονται χαμηλού

επιπέδου περιγραφές χρώματος και υφής από περιοχές της εικόνας που έχουν προκύψει έπειτα από χονδροειδή κατάτμησή της. Στη συνέχεια κατασκευάζεται ένα διάνυσμα χαρακτηριστικών για κάθε περιοχή. Με χρήση μιας τεχνικής συσταδοποίησης κατασκευάζεται ένας οπτικός θησαυρός που περιέχει τους πιο συνηθισμένους τύπους περιοχών του συνόλου εκπαίδευσης. Έπειτα και για την περιγραφή μιας εικόνας κατασκευάζεται ένα διάνυσμα αναπαράστασης. Ένας ανιχνευτής εκπαιδεύεται για κάθε έννοια, ο οποίος με είσοδο το διάνυσμα αναπαράστασης που την περιγράφει δίνει στην έξοδό του έναν βαθμό βεβαιότητας για την παρουσία της αντίστοιχης έννοιας στην εικόνα εισόδου.

7.4.4 Αξιοποίηση του Εννοιολογικού Πλαισίου στην Ανίχνευση Εννοιών

Στη συνέχεια και αφού μοντελοποιηθεί η γνώση με τη μορφή της οντολογίας εννοιολογικού πλαισίου, το επόμενο βήμα είναι να χρησιμοποιηθεί προκειμένου να βελτιώσει τα αρχικά αποτελέσματα της ανάλυσης. Για το σκοπό αυτό προτείνεται η χρήση μιας παραλλαγής του αλγορίθμου που προτάθηκε από τους Mylonas et al. [146]. Ο αλγόριθμος αυτός εφαρμόζεται στους βαθμούς βεβαιότητας που εκτιμούν οι ανιχνευτές για κάθε έννοια υψηλού επιπέδου. Σκοπός είναι από τη στιγμή που είναι γνωστό το θεματικό πεδίο που ανήκουν οι εικόνες, να γίνει εκμετάλλευση της γνώσης που παρέχει το εννοιολογικό πλαίσιο, με τη μορφή των ασαφών σχέσεων ανάμεσα στις έννοιες, προκειμένου να βελτιωθούν τα αποτελέσματα της ανίχνευσης. Μέσω μιας επαναληπτικής διαδικασίας, ο αλγόριθμος αξιοποίησης του εννοιολογικού πλαισίου επαναπροσδιορίζει τις εκτιμήσεις των βαθμών βεβαιότητας για την ύπαρξη των εννοιών.

Όπως έχει ήδη αναφερθεί, το εννοιολογικό πλαίσιο συνδέεται σε μεγάλο βαθμό με τις οντολογίες, μιας και αυτές αποτελούν μια προσπάθεια προς τη μοντελοποίηση των οντοτήτων του πραγματικού (ασαφούς) κόσμου και το εννοιολογικό πλαίσιο καθορίζει την "πραγματική" σημασία κάθε έννοιας, δηλαδή, μια έννοια μπορεί σε διαφορετικά εννοιολογικά πλαίσια να έχει και διαφορετικές σημασίες. Σε αυτή την Ενότητα, τα προβλήματα που αντιμετωπίζονται περιλαμβάνουν το πώς είναι δυνατόν να επαναπροσδιοριστούν οι αρχικοί βαθμοί εμπιστοσύνης για τις έννοιες με τρόπο που να σχετίζεται με το υπάρχον εννοιολογικό πλαίσιο και το πώς είναι δυνατό να αξιοποιηθεί αυτό, προκειμένου να βελτιωθούν τα αποτελέσματα της ανίχνευσης.

Μια εκτίμηση για το βαθμό συμμετοχής κάθε έννοιας προκύπτει από άμεσες και έμμεσες σχέσεις της εν λόγω έννοιας με τις υπόλοιπες, χρησιμοποιώντας έναν κατάλληλο ενδείκτη συμβατότητας ή μια μετρική συνάρτηση απόστασης. Ανάλογα με τη φύση του θεματικού πεδίου στο οποίο ανήκει η οντολογία, ο καλύτερος ενδείκτης μπορεί να επιλεχθεί με τη χρήση του τελεστή *max* ή του τελεστή *min*. Φυσικά, για δύο έννοιες, η ιδεατή μετρική απόστασης είναι αυτή που ποσοστοποιεί τη σημασιολογική τους συσχέτιση. Για το πρόβλημα που εξετάζεται στην παρούσα εργασία, η *max* τιμή επιλέγεται ως ένα κατάλληλο μέτρο της συσχέτισης δύο εννοιών.

Ο αλγόριθμος που επαναϋπολογίζει τους βαθμούς βεβαιότητας των εννοιών αποτελείται από τα παρακάτω βήματα:

1. Για το θεματικό πεδίο στο οποίο ανήκει η εικόνα στην οποία ανιχνεύθηκαν οι έννοιες υψηλού επιπέδου, επιλέγεται κατάλληλη τιμή του μέτρου ομοιότητας (ή ανομοιότητας) $w_s \in [0, 1]$.

2. Για κάθε εικόνα p ορίζεται το ασαφές σύνολο L_p που περιέχει τους βαθμούς βεβαιότητας $\mu_p(c_i)$, για όλες τις πιθανές έννοιες $c_i, i = 1, 2, \ldots, k$ του θεματικού πεδίου.

3. Για κάθε έννοια c_i στο ασαφές σύνολο L_p με βαθμό βεβαιότητας $\mu_p(c_i)$, γίνεται ανάκτηση της αντίστοιχης πληροφορίας από το εννοιολογικό πλαίσιο με τη μορφή του συνόλου $\mathcal{U}_{c,i}$ των σχέσεών της με τις υπόλοιπες έννοιες: $\mathcal{U}_{c,i} = \{\mathcal{U}_{c,ij} : c_i, c_j \in C, \forall i \neq j\}$.

4. Υπολογίζονται οι νέοι βαθμοί βεβαιότητας $\mu'_p(c_i)$, αφού ληφθεί υπόψη το μέτρο ομοιότητας του εν λόγω θεματικού πεδίου. Στην περίπτωση πολλαπλών σχέσεων μεταξύ εννοιών στην οντολογία, όταν η έννοια c_i σχετίζεται με περισσότερες έννοιες πέραν της έννοιας που αποτελεί τη *ρίζα* του θεματικού πεδίου, παρεμβάλλεται ένα ενδιάμεσο στάδιο συνάθροισης για τον υπολογισμό του $\mu'_p(c_i)$ με την εφαρμογή της έννοιας της *σχετικότητας του εννοιολογικού πλαισίου* cr_i (context relevance notion) που έχει οριστεί από τους Mylonas et al. [146], $cr_i = \max_j\{\mathcal{U}_{c,ij}\}$, $j = 1, 2, \ldots c_k$. Ο υπολογισμός του $\mu^l_p(c_i)$ που αντιστοιχεί στο βαθμό βεβαιότητας για την ύπαρξη της έννοιας c_i κατά την l-οστή επανάληψη του αλγορίθμου εκφράζεται με τον αναδρομικό τύπο

$$\mu^l_p(c_i) = \mu^{l-1}_p(c_i) - w_s(\mu^{l-1}_p(c_i) - cr_i) \ . \tag{7.6}$$

Ισοδύναμα, για την επανάληψη l ισχύει

$$\mu^l_p(c_i) = (1 - w_s)^l \cdot \mu^0_p(c_i) + (1 - (1 - w_s)^l) \cdot cr_i \ , \tag{7.7}$$

όπου $\mu^0_p(c_i)$ είναι ο αρχικός βαθμός βεβαιότητας για την ύπαρξη της έννοιας c_i, δηλαδή αυτός που προκύπτει από τον αντίστοιχο ανιχνευτή. Για την περάτωση του αλγορίθμου, ο αριθμός των επαναλήψεων που απαιτείται καθορίζεται εμπειρικά και απαιτούνται συνήθως 3 με 6 επαναλήψεις.

7.5 Το Εννοιολογικό Πλαίσιο των Περιοχών Εικόνων

Είναι προφανές ότι από τη στιγμή που οι τύποι περιοχής μπορεί να χαρακτηριστούν σαν "έννοιες μεσαίου επιπέδου", όπως εξηγήθηκε στο Κεφάλαιο 4, είναι δυνατόν να οριστούν σημασιολογικές σχέσεις μεταξύ τους με τρόπο αντίστοιχο του εννοιολογικού πλαισίου ενός θεματικού πεδίου όπως αυτό ορίστηκε στην Ενότητα 7.4. Έτσι, στην Ενότητα αυτή, ορίζεται το εννοιολογικό πλαίσιο των τύπων περιοχής και στη συνέχεια προτείνεται μια τεχνική που αξιποιώντας το, αποσκοπεί στη βελτίωση των αποτελέσματων της ανίχνευσης των εννοιών υψηλού επιπέδου. Και στην περίπτωση αυτή οι σχέσεις μεταξύ των τύπων περιοχής κωδικοποιούνται σε μια κατάλληλη οντολογία. Η βασική διαφορά στην προσέγγιση αυτή έγκειται στο γεγονός ότι η προτεινόμενη μεθοδολογία αποσκοπεί στο να βελτιώσει την περιγραφή μιας εικόνας με βάση τους τύπους περιοχής που περιέχονται σε αυτή, αντί να επιδρά στους αρχικούς βαθμούς βεβαιότητας που εξήχθησαν για τις έννοιες υψηλού επιπέδου.

7.5.1 Οντολογία Εννοιολογικού Πλαισίου Τύπων Περιοχής

Η προτεινόμενη μεθοδολογία ακολουθεί επακριβώς τα βήματα αυτής που παρουσιάστηκε στην Ενότητα 7.4. Έτσι, για την αναπαράσταση των σημασιολογικών σχέσεων που διέπουν τους τύπους περιοχής σε ένα θεματικό πεδίο ακολουθείται μια προσέγγιση παρόμοια με αυτή της Ενότητας 7.4, δηλαδή ορίζεται μια κατάλληλη ασαφής οντολογία. Η ασάφεια επιλέγεται και στην περίπτωση αυτή προκειμένου να κωδικοποιηθεί καλύτερα η ασάφεια που χαρακτηρίζει τις σχέσεις ανάμεσα στους τύπους περιοχής σε εικόνες που συναντώνται στον πραγματικό κόσμο. Στην περίπτωση σαφών σχέσεων ανάμεσα στους τύπους περιοχής, η οντολογία O_T του εννοιολογικού πλαισίου τους μπορεί να περιγραφεί σαν ένα σύνολο T από m τύπους περιοχής και ένα σύνολο $R_{T,ij}$ από σημασιολογικές σχέσεις ανάμεσά τους. Πιο συγκεριμένα, έστω

1. $T = \{t_i\},\ i = 1, 2, \ldots, m$ το σύνολο όλων των τύπων περιοχής, οι οποίοι και αποτελούν τον οπτικό θησαυρό που χρησιμοποιείται στο υπό εξέταση πρόβλημα και

2. $R_T = \{R_{T,ij}\}, i, j = 1, 2, \ldots, m$ το σύνολο όλων των σημασιολογικών σχέσεων ανάμεσα στους τύπους περιοχής. Το σύνολο $R^{(k)}_{T,ij},\ k = 1, 2, \ldots, K'$ περιλαμβάνει τις K' σχέσεις που ορίζονται ανάμεσα σε δύο τύπους περιοχής t_i, t_j. Επιπρόσθετα, για μια σχέση $r^{(k)}_{T,ij}$ ορίζεται και η αντίστροφή της $\bar{r}^{(k)}_{T,ji}$.

Άρα, για μια οντολογία O_T και τα σύνολα T και $R_{T,ij}$ μπορούν να διατυπωθούν οι σχέσεις

$$O_T = \{C, R_T\} \tag{7.8}$$

και

$$R_{T,ij} : T \times T \rightarrow \{0, 1\}\ , \tag{7.9}$$

όπου φαίνεται ότι μια οντολογία τύπων περιοχής O_T μπορεί να περιγραφεί από το σύνολο των τύπων περιοχής και τις μεταξύ τους σχέσεις και μια σχέση $R_{T,ij}$ έχει σύνολο τιμών το $\{0, 1\}$, δηλαδή είτε ορίζεται ανάμεσα σε δύο τύπους περιοχής της οντολογίας, είτε παραλείπεται.

Η παραπάνω μοντελοποίηση του εννοιολογικού πλαισίου των τύπων περιοχής επεκτείνεται και στην περίπτωση αυτή, προκειμένου οι σχέσεις μεταξύ τους να αποκτήσουν την απαραίτητη ασάφεια που συναντάται στις σχέσεις που τις χαρακτηρίζουν στον πραγματικό κόσμο. Για το λόγο αυτό, ορίζεται μια ασαφοποιημένη εκδοχή $\mathcal{O}_T$ της οντολογίας εννοιολογικού πλαισίου των τύπων περιοχής O ως

$$\mathcal{O}_T = \{T, \mathcal{R}_T\}\ , \tag{7.10}$$

όπου T είναι το σύνολο όλων των τύπων περιοχής που απαρτίζουν τον οπτικό θησαυρό και δεν επηρεάζεται, ενώ $\mathcal{R}_T$ είναι το σύνολο των ασαφών σημασιολογικών σχέσεων. Κατά αντιστοιχία με την περίπτωση της σαφούς οντολογίας, $\mathcal{R}_T = \{\mathcal{R}_{T,ij}\}$. Το σύνολο $\mathcal{R}_{T,ij} = \mathrm{r}^{(k)}_{T,ij},\ k = 1, 2, \ldots, K'$ περιλαμβάνει τις K' σχέσεις που ορίζονται ανάμεσα σε δύο τύπους περιοχής t_i και t_j. Επιπρόσθετα, για μια σχέση $\mathrm{r}^{(k)}_{T,ij}$ ορίζεται και η αντίστροφή της $\bar{\mathrm{r}}^{(k)}_{T,ji}$. Τέλος, στην περίπτωση των ασαφών σχέσεων ισχύει ότι

$$r_{T,ij} : T \times T \rightarrow [0, 1]\ . \tag{7.11}$$

Και στην περίπτωση των τύπων περιοχής είναι πιθανό να υπάρχουν περισσότερες από μία σχέσεις ανάμεσά τους. Έτσι, ορίζεται ο συνδυασμός των σχέσεων ως

$$\mathcal{U}_{T,ij} = \bigcup_k [\mathrm{r}_{T,ij}^{(k)}]^p,\ i,j = 1,2,\ldots,N,\ k = 1,2,\ldots,K' \ . \tag{7.12}$$

Με τον τρόπο αυτό ορίζεται το μοντέλο αναπαράστασης του εννοιολογικού πλαισίου που θα χρησιμοποιηθεί κατά τη φάση της ανάλυσης. Η τιμή του p καθορίζεται από τη σημασιολογία της κάθε σχέσης $\mathrm{r}_{\mathrm{T},ij}$ που χρησιμοποιείται κατά την κατασκευή της $\mathcal{U}_{\mathrm{T},ij}$. Πιο συγκεκριμένα:

- $p = 1$, αν η σημασιολογία της $\mathrm{r}_{T,ij}$ υπονοεί ότι πρέπει να χρησιμοποιηθεί ως έχει,
- $p = -1$, αν η σημασιολογία της $\mathrm{r}_{T,ij}$ υπονοεί τη χρήση της αντίστροφης σχέσης $\bar{\mathrm{r}}_{T,ji}$ και
- $p = 0$, αν η σημασιολογία της $\mathrm{r}_{T,ij}$ δεν επιτρέπει τη συμμετοχή της, αλλά ούτε και τη συμμετοχή της αντίστροφής της $\bar{\mathrm{r}}_{T,ji}$ στο σχηματισμό της συνδυασμένης σχέσης $\mathcal{U}_{T,ij}$.

Και στην περίπτωση αυτή και προκειμένου να εξασφαλιστεί η περιγραφικότητα της οντολογίας που μοντελοποιεί το εννοιολογικό πλαίσιο των τύπων περιοχής, θα πρέπει αυτή να αποκτήσει έναν αντιπροσωπευτικό αριθμό από ποικίλες και σωστά ορισμένες σχέσεις ανάμεσα στους τύπους περιοχής, με σκοπό να διασκορπίσει ανάμεσά τους την υπάρχουσα γνώση.

7.5.2 Σχέσεις μεταξύ των Τύπων Περιοχής

Οι σχέσεις που επιλέγονται για τη μοντελοποίηση του εννοιολογικού πλαισίου των τύπων περιοχής επιλέγονται και στην περίπτωση αυτή από το σύνολο των σχέσεων που ορίζονται στο πρότυπο MPEG-7 [12]. Η επιλογή γίνεται έτσι ώστε οι σχέσεις αυτές να μπορούν να εφαρμοστούν στο παρόν πρόβλημα ανάλυσης εικόνων. Προφανώς και για τους λόγους που αναπτύχθηκαν στην προηγούμενη ενότητα, επαναορίζονται προκειμένου να συμπεριλάβουν την απαραίτητη ασάφεια. Οι σχέσεις που επιλέχθηκαν παρουσιάζονται στον Πίνακα 7.2.

Σχέση	**Αντίστροφη**	**Σύμβολο**	**Σημασία**
Παρόμοιος	Παρόμοιος	$Sim(a,b)$	ο τύπος περιοχής a είναι παρόμοιος με τον τύπο b
Συνυπάρχων	Συνυπάρχων	$Acc(a,b)$	ο τύπος περιοχής a συνυπάρχει με τον τύπο b
Μέρος	ΜέροςΤου	$P(a,b)$	ο τύπος περιοχής a είναι μέρος του τύπου b
Συνδυασμός	-	$Comb(a,b)$	συνδυασμός δύο ή περισσότερων τύπων περιοχής

Πίνακας 7.2: *Οι σημασιολογικές σχέσεις που επιλέχθηκαν για τον καθορισμό του εννοιολογικού πλαισίου των τύπων περιοχής.*

Στην περίπτωση του εννοιολογικού πλαισίου των τύπων περιοχής, οι σχέσεις που επιλέχτηκαν λόγω της φύσης τους μπορούν να προσδιοριστούν υπολογιστικά. Έτσι, οι βαθμοί βεβαιότητας καθορίζονται από μια διαδικασία στατιστικής ανάλυσης στο σύνολο εκπαίδευσης. Στη συνέχεια, ακολουθεί επεξήγηση της σημασιολογίας των σχέσεων αυτών, με κατάλληλα παραδείγματα, προκειμένου να αποσαφηνιστεί η ερμηνεία τους, καθώς και ο τρόπος με τον οποίο υπολογίζονται.

- Η σχέση *Παρόμοιος* δηλώνει ότι ένας τύπος περιοχής είναι παρόμοιος με κάποιον άλλο τύπο περιοχής με συγκεκριμένο βαθμό βεβαιότητας. Ο βαθμός αυτός καθορίζεται με βάση τις χαμηλού επιπέδου περιγραφές των τύπων περιοχής και κατάλληλη απόσταση, που ορίζεται ανάλογα με την περίσταση.

- Η σχέση *Συνυπάρχων* δηλώνει τη συνύπαρξη δύο τύπων περιοχής σε μια εικόνα και υπολογίζεται σαν το ποσοστό των εικόνων στις οποίες συνυπάρχουν οι δύο τύπο, ως προς το ποσοστό των εικόνων στις οποίες υπάρχει είτε ο ένας είτε ο άλλος τύπος.

- Η σχέση *Μέρος* δηλώνει ότι ένας τύπος περιοχής αποτελεί μέρος κάποιου άλλου. Η σχέση αυτή μπορεί να προσδιορισθεί είτε στην περίπτωση που ο οπτικός θησαυρός περιοχών έχει κατασκευαστεί από κάποιο ειδήμονα, οπότε και υπάρχει η γνώση ότι ένας τύπος περιοχής πράγματι αποτελεί μέρος κάποιου άλλου, είτε έπειτα από την παρατήρηση ότι αυτό οφείλεται στην κατάτμηση.

- Η σχέση *Συνδυασμός* δηλώνει ότι δύο τύποι περιοχής συνδυάζονται μεταξύ τους για να σχηματίσουν κάποιον άλλο τύπο περιοχής (όχι αναγκαστικά τον ίδιο). Πρέπει να σημειωθεί ότι λόγω της ιδιαίτερης σημασιολογίας αυτής της σχέσης, είναι προφανές ότι δεν ορίζεται η αντίστροφή της.

Είναι προφανές ότι και στην περίπτωση αυτή η μοντελοποίηση του εννοιολογικού πλαισίου με τη μορφή οντολογίας οδηγεί στην κατασκευή ενός RDF γράφου, με τους κόμβους του να αντιστοιχούν στους τύπους περιοχής και τις ακμές του να αντιστοιχούν στις ασαφείς σχέσεις μεταξύ τους. Σε κάθε ακμή αντιστοιχεί ένας βαθμός βεβαιότητας, ο οποίος και εκφράζει την ασάφεια του μοντέλου. Όταν μεταξύ δύο κόμβων δεν υπάρχει ακμή, αυτό υπονοεί ότι οι αντίστοιχοι τύποι περιοχής δεν σχετίζονται με κανέναν τρόπο και έτσι, για λόγους απλότητας, σχέσεις με μηδενική βεβαιότητα παραλείπονται. Επειδή κάθε τύπος περιοχής έχει διαφορετική πιθανότητα να εμφανιστεί σε μια σκηνή, ένα επίπεδο μοντέλο εννοιολογικού πλαισίου δεν θα ήταν επαρκές ούτε σε αυτή την περίπτωση. Η περιγραφή του συνοδευτικού βαθμού εμπιστοσύνης γίνεται και εδώ με τη χρήση της μεθόδου RDF reification [10]. Δημιουργούνται, δηλαδή, τριπλέτες RDF όπως για παράδειγμα η "*μπλε partOf πράσινο*", με βαθμό εμπιστοσύνης "*0.85*" για το γεγονός αυτό. Η τριπλέτα αυτή προφανώς και δεν υπονοεί ότι ένας *μπλε* τύπος περιοχής θα είναι *πάντα* μέρος ενός *πράσινου* τύπου περιοχής σε μια σκηνή.

7.5.3 Επιλογή Τύπων Περιοχής και Ανίχνευση Εννοιών Υψηλού Επιπέδου

Για την επιλογή των τύπων περιοχής που θα περιέχονται στην οντολογία, χρησιμοποιείται η τεχνική κατασκευής του οπτικού θησαυρού που περιγράφηκε στο Κεφάλαιο 4. Πιο συγκεκριμένα, επιλέγεται ένας αρκετά μεγάλος αριθμός από εικόνες που περέχουν όλες τις υπό ανίχνευση έννοιες και μπορούν σύμφωνα με τη γνώμη ενός ειδήμωνα να θεωρηθούν αντιπροσωπευτικές για την περιγραφή του θεματικού πεδίου. Στη συνέχεια, από τις εικόνες αυτές και με τη βοήθεια ενός εργαλείου κατάτμησης, προκύπτει ένας πολύ μεγάλος αριθμός από χονδροειδείς περιοχές. Μέσω μιας διαδικασίας συσταδοποίησης οι περιοχές χωρίζονται σε συστάδες και από κάθε

μία επιλέγεται αυτή που βρίσκεται πλησιέστερα στο κέντρο της. Το σύνολο των περιοχών αυτών αποτελεί τον οπτικό θησαυρό μέσω του οποίου μια εικόνα περιγράφεται με ένα διάνυσμα αναπαράστασης.

Στη συνέχεια, οι βαθμοί βεβαιότητας για τις σχέσεις μεταξύ των περιοχών υπολογίζονται και τελικά οι σχέσεις συνδυάζονται. Για την ανίχνευση των εννοιών ψηλού επιπέδου ακολουθείται επακριβώς η μεθοδολογία του Κεφαλαίου 4. Οι εικόνες περιγράφονται μέσω του οπτικού θησαυρού και εκπαιδεύονται κατάλληλοι ανιχνευτές. Όταν πρόκειται να γίνει ανίχνευση εννοιών σε μια άγνωστη εικόνα, εφαρμόζεται στο διάνυσμα αναπαράστασής της ο αλγόριθμος που αξιοποιεί το εννοιολογικό πλαίσιο των τύπων περιοχής και το τροποποιεί, όπως παρουσιάζεται στην επόμενη Ενότητα και έπειτα, το τροποποιημένο διάνυσμα αναπαράστασης οδηγείται στον αντίστοιχο ταξινομητή.

7.5.4 Αξιοποίηση του Εννοιολογικού Πλαισίου των Περιοχών στην Ανίχνευση Εννοιών

Μετά την κατασκευή του διανύσματος αναπαράστασης για μια εικόνα, εφαρμόζεται μια τροποποιημένη εκδοχή του αλγορίθμου που παρουσιάστηκε στην Ενότητα 7.4.4, η οποία χρησιμοποιεί το εννοιολογικό πλαίσιο των τύπων περιοχής με σκοπό να επαναπροσδιορίσει το διάνυσμα αναπαράστασης, μεταβάλλοντας τις τιμές του που αντιστοιχούν στο βαθμό βεβαιότητας για την ύπαρξη ενός τύπου περιοχής σε μια εικόνα. Αυτό αποτελεί το τελευταίο στάδιο προεπεξεργασίας και προσφέρει μια βελτιωμένη εκτίμηση του διανύσματος αναπαράστασης, με την αξιοποίηση του εννοιολογικού του πλαισίου, με σκοπό να παρέχει στους ανιχνευτές εννοιών μια καλύτερη εκτίμηση του διανύσματος αναπαράστασης και έτσι να οδηγήσει σε καλύτερη ακρίβεια. Ο σκοπός είναι τελικά να φέρει την αναπαράσταση της εικόνας πιο κοντά σε αυτές με τις οποίες εκπαιδεύτηκαν οι ταξινομητές.

Έτσι, για παράδειγμα ο ταξινομητής που ανιχνεύει την έννοια *θάλασσα* στο θεματικό πεδίο *Παραλία* μπορεί να την έχει συσχετίσει με την ύπαρξη μιας *μπλε*, μιας *γαλάζιας* και μιας *καφέ* περιοχής (καθώς μια τυπική εικόνα *Παραλίας* περιέχει π.χ. *θάλασσα*, *ουρανό* και *άμμο*). Αν σε αυτόν οδηγηθεί μια εικόνα στην οποία λόγω του φωτισμού ο τύπος περιοχής που αντιστοιχεί στην έννοια *θάλασσα* είναι *πράσινος*, ενώ οι υπόλοιποι είναι όπως αναμένεται, η παραπλάνηση του ταξινομητή είναι σχεδόν βέβαια. Έτσι η έννοια *θάλασσα* είτε δεν θα ανιχνευθεί, είτε θα ανιχνευθεί με χαμηλή βεβαιότητα, μειώνοντας τη συνολική επίδοση του συστήματος.

Ο προτεινόμενος αλγόριθμος προσπαθεί να αντιμετωπίσει ακριβώς αυτό το πρόβλημα. Στη συγκεκριμένη περίπτωση, είναι προφανές ότι θα έπρεπε το διάνυσμα αναπαράστασης της εικόνας να έχει μεγάλο βαθμό βεβαιότητας για τον *μπλε* τύπο περιοχής και μικρό για τον *πράσινο* τύπο περιοχής, προκειμένου να πλησιάσει τις αντίστοιχες τιμές των διανυσμάτων αναπαράστασης με τα οποία έχει εκπαιδευθεί ο ανιχνευτής. Αυτό ακριβώς επιτυγχάνεται με την εφαρμογή του αλγορίθμου αξιοποίησης. Ο βαθμός βεβαιότητας για τον *μπλε* τύπο περιοχής θα αυξηθεί (χωρίς φυσικά να γίνει πολύ υψηλός, μιας και σε καμία περίπτωση δεν μπορεί να θεωρηθεί ότι η ύπαρξή του θα βοηθήσει την ανάλυση) και ο βαθμός της *πράσινης* περιοχής θα μειωθεί (χωρίς φυσικά να γίνει πολύ χαμηλή, γιατί δεν μπορεί να παραβλεφθεί το γεγονός ότι ο τύπος αυτός υπάρχει στην εικόνα και μπορεί η ύπαρξή του να είναι αναγκαία για την ανίχνευση κάποιας άλλης έννοιας).

Ο αρχικός προσδιορισμός του διανύσματος αναπαράστασης για μια εικόνα γίνεται

με τη διαδικασία που περιγράφεται στο Κεφάλαιο 4. Η γενική δομή του αλγόριθμου αξιοποίησης του εννοιολογικού πλαισίου των τύπων περιοχής, που επαναπροσδιορίζει τους βαθμούς βεβαιότητάς τους έχει ως εξής:

1. Για το θεματικό πεδίο στο οποίο ανήκει η εικόνα στην οποία ανιχνεύονται οι έννοιες υψηλού επιπέδου επιλέγεται κατάλληλη τιμή του μέτρου ομοιότητας (ή ανομοιότητας): $w_T \in [0, 1]$.

2. Για κάθε εικόνα p ορίζεται το ασαφές σύνολο L_T που περιέχει τους βαθμούς βεβαιότητας $\mu_p(t_i)$, για όλους τους τύπους περιοχής $t_i, i = 1, 2, \ldots, k'$ του οπτικού θησαυρού.

3. Για κάθε τύπο περιοχής t_i στο ασαφές σύνολο L_T με βαθμό βεβαιότητας $\mu_p(T_i)$, γίνεται ανάκτηση της αντίστοιχης πληροφορίας από το εννοιολογικό του πλαίσιο, με τη μορφή του συνόλου $\mathcal{U}_{T,i}$ των σχέσεών του με τους υπόλοιπους τύπους περιοχής: $\mathcal{U}_{T,i} = \{\mathcal{U}_{T,ij} : t_i, t_j \in T, \forall i \neq j\}$.

4. Υπολογίζονται οι νέοι βαθμοί βεβαιότητας $\mu'_p(T_i)$, αφού ληφθεί υπόψη το μέτρο ομοιότητας του εν λόγω θεματικού πεδίου. Στην περίπτωση πολλαπλών σχέσεων μεταξύ τύπων περιοχής στην οντολογία, όταν ο τύπος περιοχής t_i σχετίζεται με περισσότερους τύπους πέραν της ρίζας της οντολογίας, παρεμβάλλεται ένα επιπλέον στάδιο συνάθροισης για τον υπολογισμό του $\mu'_p(T_i)$, με την εφαρμογή της έννοιας της *σχετικότητας του εννοιολογικού πλαισίου* cr_i, που έχει οριστεί από τους Mylonas et al. [146], $cr_i = \max_j\{\mathcal{U}_{T,ij}\}$, $j = 1, 2, \ldots c_k$. Ο υπολογισμός του $\mu^l_p(t_i)$ που αντιστοιχεί στο βαθμό βεβαιότητας για την ύπαρξη του τύπου περιοχής t_i κατά την l-οστή επανάληψη του αλγορίθμου εκφράζεται με τον αναδρομικό τύπο

$$\mu^l_p(t_i) = \mu^{l-1}_p(t_i) - w_t(\mu^{l-1}_p(t_i) - cr_i) \ . \tag{7.13}$$

Ισοδύναμα, για την επανάληψη l ισχύει

$$\mu^l_p(t_i) = (1 - w_t)^l \cdot \mu^0_p(t_i) + (1 - (1 - w_t)^l) \cdot cr_i \ , \tag{7.14}$$

όπου $\mu^0_p(c_i)$ είναι ο αρχικός βαθμός βεβαιότητας για την ύπαρξη του τύπου περιοχής t_i, δηλαδή αυτός που προκύπτει από την αρχική περιγραφή της εικόνας με τη χρήση της πληροφορίας που παρέχει ο οπτικός θησαυρός. Για την περάτωση του αλγορίθμου, ο αριθμός των επαναλήψεων που απαιτείται καθορίζεται εμπειρικά και είναι συνήθως ίσος με 3 εώς 6 επαναλήψεις.

7.6 Μεικτό Εννοιολογικό Πλαίσιο

Όπως έχει γίνει σαφές στις προηγούμενες Ενότητες, αλλά και στο Κεφάλαιο 4 η ύπαρξη των τύπων περιοχής σε μια σκηνή είναι συνυφασμένη με την ύπαρξη των εννοιών αλλά και η ύπαρξη ή η απουσία μιας έννοιας μπορεί να οδηγήσει σε σχετικά ασφαλή συμπεράσματα για την ύπαρξη ή την απουσία άλλων εννοιών ή τύπων περιοχής που ανήκουν στο ίδιο θεματικό πεδίο. Αντίστοιχα συμπεράσματα μπορεί να εξαχθούν και από την ύπαρξη ή απουσία ενός τύπου περιοχής. Στις προηγούμενες δυο Ενότητες παρουσιάστηκαν δύο προσεγγίσεις που αντιμετωπίζουν εν μέρει το πρόβλημα της

αξιοποίησης του εννοιολογικού πλαισίου ενός θεματικού πεδίου, μιας και η μέθοδος της Ενότητας 7.4 εκμεταλλεύεται τις σχέσεις μεταξύ των εννοιών, ενώ αυτή της Ενότητας 7.5 εκμεταλλεύεται τις σχέσεις μεταξύ των τύπων περιοχής. Είναι προφανές ότι μια πιο πλήρης αντιμετώπιση του προβλήματος με αξιοποίηση του εννοιολογικού πλαισίου θα πρέπει να περιέχει και να εκμεταλλεύεται και τις σημασιολογικές σχέσεις μεταξύ εννοιών και τύπων περιοχής. Έτσι, στην ενότητα αυτή ορίζεται το "Μεικτό Εννοιολογικό Πλαίσιο", το οποίο και περιέχει τόσο τις έννοιες ενός θεματικού πεδίου όσο και τους τύπους περιοχής του, καθώς και κατάλληλες σχέσεις μεταξύ τους. Έννοιες υψηλού επιπέδου και τύποι περιοχής θα αποκαλούνται εφεξής "οντότητες".

7.6.1 Οντολογία Μεικτού Εννοιολογικού Πλαισίου

Στην Ενότητα αυτή γίνεται σύνθεση των ιδεών που περιγράφηκαν στις Ενότητες 7.4 και 7.5. Προτείνεται μια πρωτότυπη αναπαράσταση της γνώσης, η οποία έχει τη μορφή μιας διευρυμένης οντολογίας εννοιολογικού πλαισίου. Η διαφορά με τις προηγούμενες Ενότητες είναι ότι η προτεινόμενη σε αυτή την Ενότητα οντολογία περιγράφεται από ένα σύνολο εννοιών υψηλού επιπέδου, ένα σύνολο από τύπους περιοχής και ένα σύνολο από σχέσεις ανάμεσά τους. Οι σχέσεις αυτές μπορεί να συνδέουν είτε έννοιες, είτε τύπους περιοχής, είτε μια έννοια και έναν τύπο περιοχής. Το σύνολο των εννοιών καθορίζεται και πάλι από έναν ειδήμονα του θεματικού πεδίου, ενώ το σύνολο των τύπων περιοχής από μια διαδικασία συσταδοποίησης στο σύνολο εκπαίδευσης που είναι διαθέσιμο.

Γενικά, και σε αντιστοιχία με τις προηγούμενες Ενότητες, μια οντολογία O που μοντελοποιεί το μεικτό εννοιολογικό πλαίσιο ενός θεματικού πεδίου αποτελείται από τα σύνολα

- $C = \{c_i\}, i = 1, 2, \ldots, n$, που είναι το σύνολο όλων των εννοιών υψηλού επιπέδου του υπό εξέταση θεματικού πεδίου,
- $T = \{t_i\}, i = 1, 2, \ldots, m$ που είναι το σύνολο όλων των τύπων περιοχής του οπτικού θησαυρού που έχει επιλεχθεί για την ανάλυση στο συγκεκριμένο θεματικό πεδίο και
- $R = R_{ij}, i, j = 1, 2, \ldots, n + m$ που είναι το σύνολο όλων των σημασιολογικών σχέσεων ανάμεσα σε δύο οντότητες x_i και x_j. Το σύνολο $R_{ij} = r_{ij}^{(k)}, k = 1, 2, \ldots, K+K'$ περιλαμβάνει τις το πολύ $K+K'$ σχέσεις που ορίζονται ανάμεσα σε δύο οντότητες x_i και x_j.

Άρα για μια οντολογία μεικτού εννοιολογικού πλαισίου O και τα συνολα C, R και T μπορούν να διατυπωθούν οι σχέσεις

$$O = \{C, T, R_{ij}\} \tag{7.15}$$

και

$$r_{ij}^{(k)} : (C \cup T) \times (C \cup T) \rightarrow \{0, 1\}, \quad i, j = 1, 2, \ldots m + n, \quad i \neq j \ . \tag{7.16}$$

Όπως φαίνεται στην (7.16), επειδή ακριβώς η προτεινόμενη μοντελοποίηση του μεικτού εννοιολογικού πλαισίου δεν περιορίζει τις σχέσεις μόνο ανάμεσα σε μέλη είτε του C είτε του T, επιβάλλει την ύπαρξη σχέσεων ανάμεσα σε έννοιες και τύπους

περιοχής. Επίσης, είναι προφανές ότι μια οντολογία μεικτού εννοιολογικού πλαισίου O μπορεί να περιγραφεί από τις έννοιες υψηλού επιπέδου, τους τύπους περιοχής και τις μεταξύ τους σχέσεις. Μια τέτοια σχέση έχει σύνολο τιμών το $\{0, 1\}$, δηλαδή είτε ορίζεται είτε παραλείπεται για δύο οντότητες.

Η μοντελοποίηση αυτή επεκτείνεται και στην περίπτωση του μεικτού εννοιολογικού πλαισίου και οι σχέσεις επαναορίζονται προκειμένου να αποκτήσουν την απαραίτητη ασάφεια που συναντάται ανάμεσα στις οντότητες του πραγματικού κόσμου, όπως έγινε και στις Ενότητες 7.4 και 7.5. Έτσι, ορίζεται μια ασαφοποιημένη εκδοχή $\mathcal{O}$ της οντολογίας του μεικτού εννοιολογικού πλαισίου O, ως

$$\mathcal{O} = \{C, T, \mathcal{R}\} \ , \tag{7.17}$$

όπου τα σύνολα C και T που περιέχουν τις έννοιες υψηλού επιπέδου και τους τύπους περιοχής, αντίστοιχα, παραμένουν αμετάβλητα, ενώ το σύνολο $\mathcal{R}$ περιέχει πια τις ασαφοποιημένες σχέσεις ανάμεσα στις οντότητες. Κατά αντιστοιχία με την περίπτωση της σαφούς οντολογίας, $\mathcal{R} = \mathcal{R}_{ij}$. Το σύνολο $\mathcal{R}_{ij} = r_{ij}^{(k)}, k = 1, 2, \ldots, K + K'$ περιλαμβάνει τις $K + K'$ σχέσεις ανάμεσα σε δύο οντότητες x_i και x_j. Επίσης, για μια σχέση $r_{ij}^{(k)}$, ορίζεται και η αντίστροφή της $\bar{r}_{ji}^{(k)}$.

Τέλος, στην περίπτωση των ασαφών σχέσεων ισχύει ότι

$$r_{ij} : (C \cup T) \times (C \cup T) \rightarrow [0, 1] \ . \tag{7.18}$$

Φυσικά, είναι πιθανό να υπάρχουν περισσότερες από μία σχέσεις ανάμεσα σε δύο οντότητες. Έτσι, ορίζεται ο συνδυασμός των σχέσεων ως

$$\mathcal{U}_{ij} = \bigcup_k [\mathrm{r}_{ij}^{(k)}]^p, \ i, j = 1, 2, \ldots, N, \ k = 1, 2, \ldots, K' \ . \tag{7.19}$$

Με τον τρόπο αυτό ορίζεται το μοντέλο αναπαράστασης του εννοιολογικού πλαισίου που θα χρησιμοποιηθεί κατά τη φάση της ανάλυσης. Η τιμή του p καθορίζεται από τη σημασιολογία της κάθε σχέσης r_{ij} που χρησιμοποιείται κατά την κατασκευή της $\mathcal{U}_{ij}$. Πιο συγκεκριμένα:

- $p = 1$, αν η σημασιολογία της r_{ij} υπονοεί ότι πρέπει να χρησιμοποιηθεί ως έχει,
- $p = -1$, αν η σημασιολογία της r_{ij} υπονοεί τη χρήση της αντίστροφης σχέσης $\bar{\mathrm{r}}_{ji}$ και
- $p = 0$, αν η σημασιολογία της r_{ij} δεν επιτρέπει τη συμμετοχή της, αλλά ούτε και τη συμμετοχή της αντίστροφής της $\bar{\mathrm{r}}_{T,ji}$ στο σχηματισμό της συνδυασμένης σχέσης $\mathcal{U}_{ij}$.

Τέλος, και στην περίπτωση αυτή και προκειμένου να εξασφαλιστεί η περιγραφικότητα της οντολογίας που μοντελοποιεί το εννοιολογικό πλαίσιο των τύπων περιοχής, θα πρέπει αυτή να αποκτήσει έναν αντιπροσωπευτικό αριθμό από ποικίλες και σωστά ορισμένες σχέσεις ανάμεσα στους τύπους περιοχής, με σκοπό να διασκορπίσει ανάμεσά τους την υπάρχουσα γνώση. Θα χρησιμοποιηθούν οι σχέσεις που ορίστηκαν στις Ενότητες 7.4.2 και 7.5.2 ανάμεσα σε έννοιες και τύπους περιοχής και ανάμεσα σε αυτές θα επιλεχθεί ένα υποσύνολό τους που θα μπορεί να εφαρμοστεί ανάμεσα σε διαφορετικού τύπου οντότητες.

7.6.2 Σημασιολογικές Σχέσεις ανάμεσα σε δύο Οντότητες

Τόσο οι σχέσεις ανάμεσα σε έννοιες υψηλού επιπέδου, όσο και αυτές ανάμεσα σε τύπους περιοχής, επιλέχθηκαν από αυτές που προτείνονται από το πρότυπο MPEG-7 [12] και επαναορίστηκαν προκειμένου να αποκτήσουν την απαραίτητη ασάφεια. Από τις σχέσεις αυτές επιλέγεται ένα υποσύνολο κατάλληλο να εφαρμοστεί και ανάμεσα σε έννοιες υψηλού επιπέδου και τύπους περιοχής. Οι σχέσεις που τελικά επιλέχθηκαν συνοψίζονται στον Πίνακα 7.3.

Σχέση	**Αντίθετη**	**Συμβολισμός**	**Σημασία**
Παρόμοιος	*Παρόμοιος*	$Sim(a,b)$	ομοιότητα μεταξύ a και b
Συνυπάρχων	*Συνυπάρχων*	$Acc(a,b)$	συνύπαρξη μεταξύ a και b
Μέρος	*ΜέροςΤου*	$P(a,b)$	η a είναι μέρος της b
Συστατικό	*ΣυστατικόΤου*	$Comp(a,b)$	η a είναι συστατικό της b
Ειδίκευση	*Γενίκευση*	$Sp(a,b)$	η b εξειδικεύει τη σημασία της a
Παράδειγμα	*ΠαράδειγμαΤου*	$Ex(a,b)$	η b είναι ένα παράδειγμα της a
Τοποθεσία	*ΤοποθεσίαΤου*	$Loc(a,b)$	η b είναι η τοποθεσία της a
Ιδιότητα	*ΙδιότηταΤου*	$Pr(a,b)$	η b είναι ιδιότητα της a

Πίνακας 7.3: *Οι σημασιολογικές σχέσεις που χρησιμοποιούνται στο μεικτό εννοιολογικό πλαίσιο και η ερμηνεία τους.*

Σχέση	$C \times C$	$T \times T$	$C \times T$
Παρόμοιος	-	•	-
Συνυπάρχων	•	•	•
Μέρος	•	•	•
Συστατικό	•	•	•
Εξειδίκευση	•	-	-
Παράδειγμα	•	-	-
Τοποθεσία	•	-	-
Ιδιότητα	-	•	•

Πίνακας 7.4: *Επιτρεπτές σχέσεις μεταξύ όμοιων και διαφορετικών οντοτήτων.*

Κάθε οντότητα μπορεί να σχετίζεται με κάποια άλλη μέσω μίας ή και περισσοτέρων από τις προαναφερθείσες σχέσεις. Παρολαυτά, θα πρέπει να ξεκαθαριστεί ότι δεν είναι όλες οι σχέσεις κατάλληλες για να εφαρμοστούν μεταξύ οποιωνδήποτε δύο οντοτήτων. Για παράδειγμα, η σχέση *Όμοιος* δεν έχει νόημα μεταξύ δύο εννοιών ή μεταξύ μιας έννοιας και ενός τύπου περιοχής, δηλαδή η έννοια *θάλασσα* δεν μπορεί να σχετιστεί με την έννοια *άμμος* με τη σχέση αυτή, ούτε και με έναν *μπλε* τύπο περιοχής. Ωστόσο, η ομοιότητα αποτελεί ένα μέτρο που έχει νόημα για τη σχέση μεταξύ δύο τύπων περιοχής και μπορεί να υπολογιστεί με τη σύγκριση των χαμηλού επιπέδου χαρακτηριστικών τους, ως ένας βαθμός ομοιότητας ή μια απόσταση. Οι πιθανές σχέσεις για κάθε ζεύγος από οντότητες φαίνονται στον Πίνακα 7.4.

Όσον αφορά τις σχέσεις που χρησιμοποιούνται, ο βαθμός βεβαιότητας για αυτές προσδιορίζεται είτε μέσω της γνώμης κάποιου ειδήμονα σχετικά με το θεματικό πεδίο, είτε υπολογίζεται στατιστικά, μέσω κατάλληλης επεξεργασίας. Στη συνέχεια ακολουθούν παραδείγματα για κάθε σχέση, καθώς και ο τρόπος με τον οποίο γίνεται ο υπολογισμός του αντίστοιχου βαθμού βεβαιότητας.

- η σχέση *Παρόμοιος* μπορεί να οριστεί μόνο μεταξύ δύο τύπων περιοχής και ο βαθμός βεβαιότητάς αντιστοιχεί στη μεταξύ τους απόσταση.

- η σχέση *Συνυπάρχων* σημαίνει ότι δύο οντότητες συνυπάρχουν στο ίδιο πολυμεσικό έγγραφο. Είναι προφανές ότι η σχέση αυτή μπορεί να οριστεί μεταξύ οποιωνδήποτε οντοτήτων. Πρέπει να αποσαφηνιστεί ότι η συνύπαρξη ενός τύπου περιοχής με μια έννοια δεν σημαίνει απαραίτητα ότι ο τύπος περιοχής απεικονίζει την έννοια αυτή. Επιπρόσθετα, ακόμη και στην περίπτωση που ένας τύπος περιοχής απεικονίζει μια έννοια, θα θεωρείται ότι συνυπάρχει με αυτήν, μιας και ο σχολιασμός γίνεται σε καθολικό επίπεδο και όχι τοπικά. Ο υπολογισμός του βαθμού βεβαιότητας γίνεται στατιστικά.

- η σχέση *Μέρος* ορίζεται για οποιεσδήποτε οντότητες και σημαίνει ότι η μία αποτελεί μέρος της άλλης. Στην περίπτωση των εννοιών, για παράδειγμα, η *θάλασσα* αποτελεί *Μέρος* της *παραλίας*. Στην περίπτωση των τύπων περιοχής, ένας *πράσινος με τραχειά υφή* τύπος περιοχής μπορεί να αποτελεί *Μέρος* ενός *πράσινου* τύπου περιοχής. Τέλος, στην περίπτωση μιας έννοιας και ενός τύπου περιοχής, ένας *πράσινος* τύπος περιοχής μπορεί να αποτελεί *Μέρος* ενός *δέντρου*.

- η σχέση *Συστατικό* μπορεί να οριστεί μεταξύ δύο οποιωνδήποτε οντοτήτων. Έτσι, στην περίπτωση δύο εννοιών, για παράδειγμα το *δέντρο* είναι συστατικό του *δάσους*, στην περίπτωση δύο τύπων περιοχής, ένας *σκούρος πράσινος* τύπος μπορεί να είναι ένα συστατικό ενός *πράσινου* τύπου και τέλος, ένας *πορτοκαλί* τύπος περιοχής μπορεί να είναι συστατικό της έννοιας *ηλιοβασίλεμα*. Υπάρχει και η περίπτωση μια έννοια να αποτελεί συστατικό ενός τύπου περιοχής, κάτι πολύ συνηθισμένο εξαιτίας της χονδροειδούς κατάτμησης, αλλά η περίπτωση αυτή δεν χρησιμοποιείται, λόγω του καθολικού σχολιασμού που καθιστά αδύνατο τον προσδιορισμό της. Πρέπει να αποσαφηνιστεί ότι ο συνδυασμός δύο οντοτήτων μπορεί να οδηγήσει είτε σε έννοια είτε σε τύπο περιοχής και δεν υπάρχει κάποιος γενικός κανόνας για το τι θα προκύψει.

- η σχέση *Ειδίκευση* ορίζεται μεταξύ εννοιών, όπως αναφέρθηκε στην Ενότητα 7.4.

- η σχέση *Παράδειγμα* ορίζεται μεταξύ εννοιών, όπως αναφέρθηκε στην Ενότητα 7.4.

- η σχέση *Τοποθεσία* ορίζεται μεταξύ εννοιών, όπως αναφέρθηκε στην Ενότητα 7.4.

- η σχέση *Ιδιότητα* μπορεί να οριστεί ανάμεσα σε μια έννοια και ένα τύπο περιοχής, ή ανάμεσα σε δύο έννοιες. Στην πρώτη περίπτωση, ένας τύπος περιοχής να αποτελεί ιδιότητα μιας έννοιας. Για παράδειγμα, ένας "*πράσινος*" τύπος περιοχής μπορεί να αποτελεί ιδιότητα της έννοιας *βλάστηση*. Στη δεύτερη περίπτωση, μια έννοια αποτελεί ιδιότητα μιας άλλης έννοιας. Έτσι, η έννοια *κυματιστός* αποτελεί ιδιότητα της έννοιας *θάλασσα*.

Όλες οι παραπάνω σχέσεις μοντελοποιούν το εννοιολογικό πλαίσιο ανάμεσα στις έννοιες και τους τύπους περιοχής ενός θεματικού πεδίου. Ανάμεσα σε δύο οντότητες x_i και x_j δημιουργείται μοναδική σχέση $\mathcal{U}_{ij}$ και η οντολογία $\mathcal{O}$ που προκύπτει αποτελεί και στην περίπτωση αυτή έναν RDF γράφο, με την περιγραφή του βαθμού

βεβαιότητας ανάμεσα σε δύο κόμβους να γίνεται με την τεχνική RDF reification [10]. Πρέπει να τονιστεί και πάλι ότι η ακμή ανάμεσα σε κάθε ζευγάρι από έννοιες παράγεται με βάση το σύνολο των σχέσεων που έχουν νόημα για το συγκεκριμένο ζευγάρι. Για παράδειγμα, η ακμή ανάμεσα στις έννοιες *βράχος* και *άμμος* παράγεται με το συνδυασμό των σχέσεων *Τοποθεσία* και *Συνοδός*, ενώ η ακμή ανάμεσα στις έννοιες *νερό* και *θάλασσα* χρησιμοποιεί τις σχέσεις *Εξειδίκευση*, *ΜέροςΤου*, *Παράδειγμα* και *Τοποθεσία*. Κατά τον ίδιο τρόπο, ένας *πράσινος* και ένας *μπλε* τύπος περιοχής χρησιμοποιούν τις σχέσεις *Παρόμοιος*, *Συνοδός* και *Συστατικό*. Τέλος, ένας *μπλε* τύπος περιοχής και η έννοια *θάλασσα* συνδέονται με τις σχέσεις *Συνυπάρχων*, *Μέρος* και *Συστατικό*.

7.6.3 Αξιοποίηση του Μεικτού Εννοιολογικού Πλαισίου

Στην Ενότητα αυτή παρουσιάζεται ο προτεινόμενος αλγόριθμος αξιοποίησης του μεικτού εννοιολογικού πλαισίου. Η βασική ιδέα του αλγορίθμου είναι ίδια με αυτή του αλγορίθμου αξιοποίησης του εννοιολογικού πλαισίου ενός θεματικού πεδίου που παρουσιάστηκε στην Ενότητα 7.4.4 αλλά και μ'αυτή του αλγορίθμου αξιοποίησης του εννοιολογικού πλαισίου των τύπων περιοχής ενός θεματικού πεδίου που παρουσιάστηκε στην Ενότητα 7.5.4. Ο σκοπός του αλγορίθμου στην περίπτωση αυτή είναι να επαναπροσδιορισθούν οι βαθμοί βεβαιότητας των οντοτήτων που απαρτίζουν την οντολογία που μοντελοποιεί το μεικτό εννοιολογικό πλαίσιο, μέσω μιας επαναληπτικής διαδικασίας.

Η γενική δομή του προτεινόμενου αλγορίθμου επαναπροσδιορισμού των βαθμών βεβαιότητας των διαφόρων οντοτήτων μιας εικόνας έχει ως εξής:

1. Για το θεματικό πεδίο στο οποίο ανήκει η εικόνα στην οποία ανιχνεύονται οι έννοιες υψηλού επιπέδου επιλέγεται κατάλληλη τιμή του μέτρου ομοιότητας (ή ανομοιότητας): $w_m \in [0, 1]$.

2. Για κάθε εικόνα p, ορίζεται το ασαφές σύνολο L_p που περιέχει τους βαθμούς βεβαιότητας $\mu_p(c_i)$, για όλες τις πιθανές έννοιες $c_i, i = 1, 2, \ldots, k$ του θεματικού πεδίου.

3. Για κάθε εικόνα p ορίζεται το ασαφές σύνολο L_T που περιέχει τους βαθμούς βεβαιότητας $\mu_p(t_i)$, για όλους τους τύπους περιοχής $t_i, i = 1, 2, \ldots, k'$ του οπτικού θησαυρού.

4. Για κάθε έννοια c_i στο ασαφές σύνολο L_p με βαθμό βεβαιότητας $\mu_p(c_i)$, γίνεται ανάκτηση της αντίστοιχης πληροφορίας από το εννοιολογικό πλαίσιο με τη μορφή του συνόλου $\mathcal{U}_{c,i}$ των σχέσεών της με τις υπόλοιπες έννοιες: $\mathcal{U}_{c,i} = \{\mathcal{U}_{c,ij} : c_i, c_j \in C, \forall i \neq j\}$.

5. Για κάθε τύπο περιοχής t_i στο ασαφές σύνολο L_T με βαθμό βεβαιότητας $\mu_p(T_i)$, γίνεται ανάκτηση της αντίστοιχης πληροφορίας από το εννοιολογικό του πλαίσιο, με τη μορφή του συνόλου $\mathcal{U}_{T,i}$ των σχέσεών του με τους υπόλοιπους τύπους περιοχής: $\mathcal{U}_{T,i} = \{\mathcal{U}_{T,ij} : t_i, t_j \in T, \forall i \neq j\}$.

6. Υπολογίζονται οι νέοι βαθμοί βεβαιότητας $\mu'_p(c_i)$ και $\mu'_p(T_i)$, αφού ληφθεί υπόψη το μέτρο ομοιότητας του εν λόγω θεματικού πεδίου. Στην περίπτωση πολλαπλών σχέσεων μεταξύ τύπων περιοχής στην οντολογία, όταν μια οντότητα x_i σχετίζεται με περισσότερες οντότητες πέραν της ρίζας της οντολογίας, παρεμβάλλεται ένα επιπλέον στάδιο συνάθροισης για τον υπολογισμό των $\mu'_p(c_i)$ και $\mu'_p(T_i)$, με την εφαρμογή της έννοιας της *σχετικότητας του εννοιολογικού πλαισίου* cr_i, που έχει οριστεί από τους Mylonas et al. [146], $cr_i = \max_j\{\mathcal{U}_{ij}\}$, $j = 1, 2, \ldots c_k$. Ο υπολογισμός των $\mu^l_p(c_i)$ και $\mu^l_p(t_i)$ που αντιστοιχούν στους βαθμούς βεβαιότητας για την ύπαρξη της έννοιας c_i και του τύπου περιοχής t_i, αντίστοιχα, κατά την l-οστή επανάληψη του αλγορίθμου εκφράζεται με τον αναδρομικό τύπο

$$\mu^l_p(x_i) = \mu^{l-1}_p(x_i) - w_m(\mu^{l-1}_p(x_i) - cr_i) \ . \tag{7.20}$$

Ισοδύναμα, για την επανάληψη l ισχύει

$$\mu^l_p(x_i) = (1 - w_m)^l \cdot \mu^0_p(x_i) + (1 - (1 - w_m)^l) \cdot cr_i \ , \tag{7.21}$$

όπου $\mu^0_p(x_i)$ είναι ο αρχικός βαθμός βεβαιότητας για την ύπαρξη της οντότητας x_i, δηλαδή αυτός που προκύπτει από την αρχική περιγραφή της εικόνας με τη χρήση της πληροφορίας που παρέχει ο οπτικός θησαυρός στην περίπτωση του τύπου περιοχής ή αυτός που προκύπτει από την αρχική εφαρμογή του αλγορίθμου ανίχνευσης, στην περίπτωση μιας έννοιας υψηλού επιπέδου. Για την περάτωση του αλγορίθμου, ο αριθμός των επαναλήψεων που απαιτείται καθορίζεται εμπειρικά και είναι συνήθως ίσος με 3 εώς 6 επαναλήψεις.

7.7 Πειραματικά Αποτελέσματα

Στην Ενότητα αυτή παρουσιάζονται πειραματικά αποτελέσματα από την εφαρμογή των αλγορίθμων αξιοποίησης του εννοιολογικού πλαισίου που προτάθηκαν σε αυτό το Κεφάλαιο σε ένα πραγματικό πρόβλημα ανίχνευσης εννοιών υψηλού επιπέδου σε εικόνες. Για κάθε μία από τις τεχνικές που προτάθηκαν παρουσιάζεται ένα απλό παράδειγμα εφαρμογής τους, προκειμένου να γίνει κατανοητός ο τρόπος με τον οποίο επιδρούν στα αρχικά αποτελέσματα της ανίχνευσης εννοιών και δημιουργούν τους νέους βαθμούς βεβαιότητας. Επίσης οι τεχνικές αυτές συγκρίνονται με αυτές του Κεφαλαίου 4 καθώς και με δύο τεχνικές που έχουν προταθεί στη βιβλιογραφία και ασχολούνται με το ίδιο πρόβλημα, χρησιμοποιώντας παρόμοια μοντέλα.

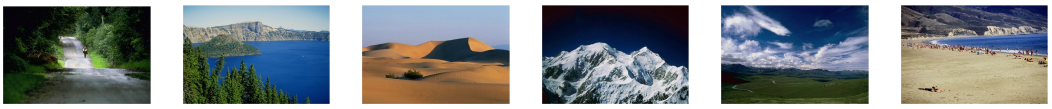
7.7.1 Σύνολα Ελέγχου

Για το σκοπό της αξιλόγησης των προτεινόμενων τεχνικών επιλέχτηκε ένα σύνολο εικόνων από τη συλλογή του Corel[1], καθώς και ένα σύνολο εικόνων από το TRECVID 2007 [192]. Επιλέχτηκε ένα σύνολο από έννοιες που ανήκουν στην κατηγορία των υλικών. Στην περίπτωση του Corel, οι έννοιες αυτές ήταν οι *δρόμος*, *άμμος*, *θάλασσα*, *ουρανός*, *χιόνι* και *βλάστηση*, ενώ στην περίπτωση του TRECVID ήταν οι *βλάστηση*, *δρόμος*, *άμμος*, *νερό*, *ουρανός*, *χιόνι* και *φωτιά*. Το σύνολο του Corel σε

[1] `http://www.corel.com`

πρώτη ματιά είναι πιο απλό και πιο ομοιογενές από το αντίστοιχο του TRECVID. Οι εικόνες έχουν υψηλή ποιότητα και απέχουν από αυτές που βγάζει ο μέσος χρήστης. Αντίθετα, το σύνολο του TRECVID χαρακτηρίζεται από μεγάλη ανομοιογένεια. Οι εικόνες στην περίπτωση αυτή είναι χαρακτηριστικά καρέ από βίντεο πολιτιστικού περιεχομένου. Η ποιότητά τους μπορεί να χαρακτηριστεί απλά ως ικανοποιητική.

Στην περίπτωση του Corel, το σύνολο αποτελείται από 750 εικόνες. Χρησιμοποιήθηκαν 525 εικόνες (το 80%) για την κατασκευή του οπτικού θησαυρού και την εκπαίδευση των ανιχνευτών εννοιών και 275 (το 20%) για την αξιολόγηση της τεχνικής. Ο οπτικός θησαυρός που κατασκευάστηκε αποτελείται από 40 τύπους περιοχής. Το σύνολο δεδομένης αλήθειας κατασκευάστηκε χειρωνακτικά και οι εικόνες σχολιάστηκαν καθολικά. Η παράμετρος κανονικοποίησης του αλγορίθμου τέθηκε ίση με $\mu = 0.12$ για το υπό εξέταση πρόβλημα[2]. Χρησιμοποιήθηκαν 525 εικόνες για την εκπαίδευση των ανιχνευτών, καθώς και για την κατασκευή του οπτικού θησαυρού και 225 σαν σύνολο ελέγχου. Μερικές χαρακτηριστικές εικόνες από τη συλλογή του Corel απεικονίζονται στο Σχήμα 7.6. Είναι φανερό ότι σε μια εικόνα είναι δυνατόν να απεικονίζονται περισσότερες από μια έννοια.



Σχήμα 7.6: *Ενδεικτικές εικόνες από τη συλλογή του Corel, στις οποίες απεικονίζονται οι έννοιες προς ανίχνευση.*

Όσον αφορά τη συλλογή εικόνων του TRECVID, από το σύνολο των χαρακτηριστικών καρέ του 2007 επιλέχτηκε ένα κατάλληλο υποσύνολό, αποτελούμενο απο 4000 εικόνες. Χρησιμοποιήθηκαν 250 εικόνες (το 6.25%) για την κατασκευή του οπτικού θησαυρού και την εκπαίδευση των ανιχνευτών εννοιών και 3750 (το 93.75%) για την αξιολόγηση της τεχνικής. Ο οπτικός θησαυρός που κατασκευάστηκε είχε 100 τύπους περιοχής. Ο σχολιασμός των εικόνων προέκυψε από τη συλλογική προσπάθεια που οργανώθηκε από τους Ayache και Quenot [6], η οποία αναφέρθηκε στην Ενότητα 4.2. Η παράμετρος κανονικοποίησης του αλγορίθμου τέθηκε ίση με $\mu = 0.15$ για το υπό εξέταση πρόβλημα[3]. Μερικές χαρακτηριστικές εικόνες από τη συλλογή του TRECVID απεικονίζονται στο Σχήμα 7.7. Και στην περίπτωση αυτή, είναι φανερό ότι σε μια εικόνα είναι δυνατόν να απεικονίζονται περισσότερες από μια έννοια.

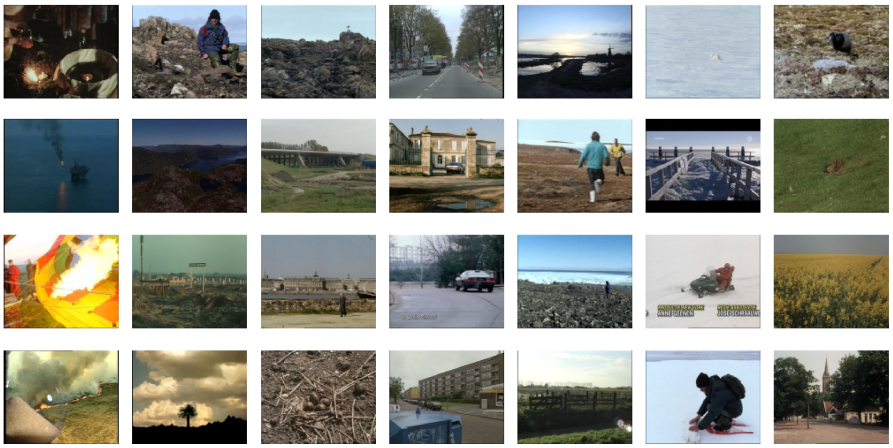
7.7.2 Μέτρα Αξιολόγησης

Για την αξιολόγηση των τεχνικών επιλέχθηκαν και στο Κεφάλαιο αυτό τα μέτρα της ακρίβειας, της ανάκτησης καθώς και το F-μέτρο, τα οποία παρατίθενται ακολούθως για ευκολία. Το μέτρο της ακρίβειας υπολογίζεται ως

$$P_i = \frac{|D_i \bigcap G_i|}{|D_i|}, \quad i = 1, 2, \ldots N_C \ , \tag{7.22}$$

[2]Στο συγκεκριμένο πρόβλημα δεν μπορεί να καθοριστεί το θεματικό πεδίο με την αυστηρή σημασία του, μιας και οι εικόνες προέρχονται από διαφορά θεματικά πεδία. Ωστόσο, η περίπτωση αντιμετωπίζεται σαν να πρόκειται για το θεματικό πεδίο *Corel*.

[3]Ομοίως με πριν θεωρήθηκε ότι οι έννοιες που επιλέχθηκαν αποτελούν το θεματικό πεδίο *TRECVID*.



Σχήμα 7.7: *Ενδεικτικές εικόνες από τη συλλογή του TRECVID, στις οποίες απεικονίζονται οι έννοιες προς ανίχνευση.*

ενώ το μέτρο της ανάκτησης υπολογίζεται ως

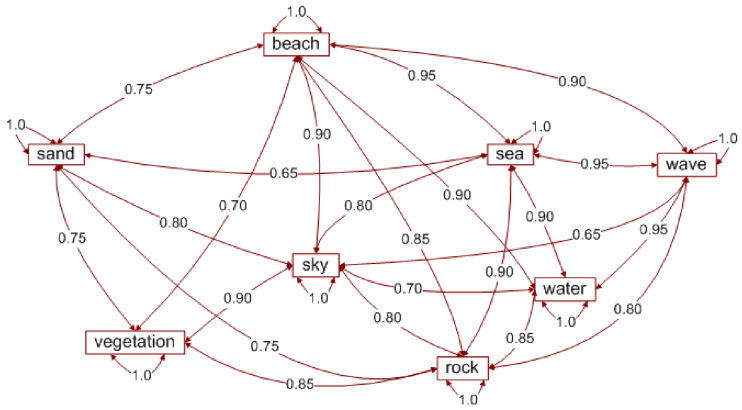
$$R_i = \frac{|D_i \bigcap G_i|}{|G_i|}, \quad i = 1, 2, \ldots N_C \ , \tag{7.23}$$

όπου υπενθυμίζεται ότι D_i είναι το σύνολο των εικόνων για τις οποίες ο ανιχνευτής της έννοιας C_i αποφάσισε ότι την απεικονίζουν, ενώ G_i είναι το σύνολο των εικόνων που πραγματικά απεικονίζουν την έννοια, σύμφωνα με το σύνολο δεδομένης αλήθειας. Τέλος, το F-μέτρο υπολογίζεται ως

$$F_i = \frac{2P_iR_i}{P_i + R_i}, \quad i = 1, 2, \ldots N_C \ . \tag{7.24}$$

7.7.3 Εφαρμογή του Εννοιολογικού Πλαισίου των Εικόνων ενός Θεματικού Πεδίου

Προκειμένου να γίνει κατανοητός ο τρόπος με τον οποίο επιδρά ο αλγόριθμος στους αρχικούς βαθμούς βεβαιότητας, παρατίθεται ένα απλό παράδειγμα για την περίπτωση του συνόλου εικόνων του Corel. Με βάση τις σχέσεις ανάμεσα στο σύνολο των εννοιών, κατασκευάστηκε η οντολογία του εννοιολογικού πλαισίου που απεικονίζεται στο Σχήμα 7.8. Στο Σχήμα αυτό οι σχέσεις ανάμεσα στις έννοιες έχουν προκύψει με εφαρμογή της (7.5). Επιπρόσθετα, στο Σχήμα 7.9 απεικονίζονται τρεις από τις εικόνες του συνόλου ελέγχου, στις οποίες και εφαρμόστηκε ο αλγόριθμος. Στον Πίνακα 7.5 παρατίθενται οι βαθμοί βεβαιότητας για την ύπαρξη των εννοιών της εικόνας, *πριν* και *μετά* την εφαρμογή του αλγορίθμου. Μπορεί να παρατηρηθεί ότι οι βαθμοί βεβαιότητας για τις έννοιες για τις οποίες η αρχική ανίχνευση έδωσε υψηλή βεβαιότητα, μεγαλώσαν ακόμη περισσότερο, σε αντίθεση με όσες είχαν μικρό αρχικό βαθμό βεβαιότητας και αδύναμες σχέσεις με τις πρώτες.



Σχήμα 7.8: *Οντολογία εννοιολογικού πλαισίου για το θεματικό πεδίο Παραλία, που αποτελείται από τις 7 προς ανίχνευση έννοιες.*

Σχήμα 7.9: *3 παραδείγματα εικόνων από το θεματικό πεδίο Corel. Άνω σειρά: αρχικές εικόνες. Κάτω σειρά: Χάρτες Κατάτμησης.*

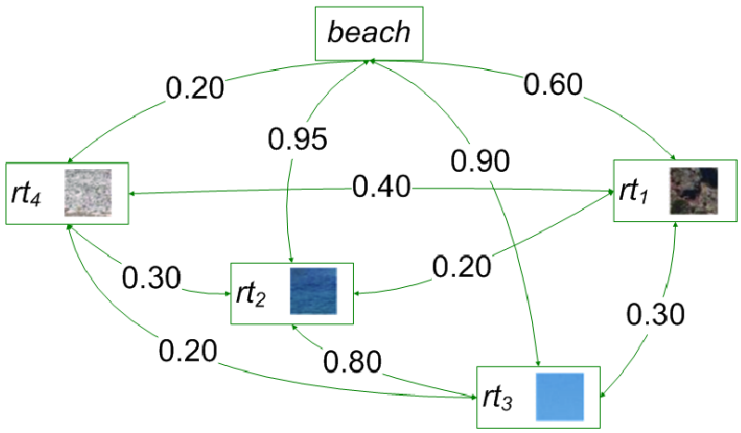
7.7.4 Εφαρμογή του Εννοιολογικού Πλαισίου των Τύπων Περιοχής

Με βάση τις σχέσεις ανάμεσα στους τύπους περιοχής των εικόνων του συνόλου εκπαίδευσης κατασκευάστηκε η οντολογία του εννοιολογικού πλαισίου των τύπων περιοχής. Προκειμένου να γίνει κατανοητός ο τρόπος με τον οποίο επιδρά ο αλγόριθμος αξιοποίησης του εννοιολογικού πλαισίου των τύπων περιοχής, ακολουθεί ένα απλό παράδειγμα. Στο Σχήμα 7.10. απεικονίζεται ένα τμήμα της οντολογίας αυτής, αποτελούμενο από 4 τύπους περιοχής και τις μεταξύ τους σχέσεις, οι οποίες παρατίθενται στον Πίνακα 7.2 και υπολογίστηκαν για αυτές οι αντίστοιχοι βαθμοί βεβαιότητας.

Στο Σχήμα 7.11 απεικονίζεται μια από τις εικόνες του συνόλου ελέγχου και ο αντίστοιχος χάρτης κατάτμησης. Για την εικόνα αυτή, η εφαρμογή του αλγορίθμου αξιοποίησης του εννοιολογικού πλαισίου των τύπων περιοχής με χρήση της απλοϊκής οντολογίας του Σχήματος 7.10 επιφέρει ουσιαστική βελτίωση στο διάνυσμα αναπαράστασης και για το λόγο αυτό επιλέχθηκε για τις ανάγκες του παραδείγματος. Με

Έννοιες	1η		2η		3η	
	Πριν	Μετά	Πριν	Μετά	Πριν	Μετά
θάλασσα	0.77	0.85	0.65	0.75	0.62	0.72
νερό	0.63	0.70	0.60	0.69	0.58	0.67
βλάστηση	0.35	0.43	0.35	0.40	0.62	0.72
ουρανός	0.45	0.57	0.55	0.60	0.53	0.61
άμμος	0.69	0.75	0.45	0.56	0.52	0.60
βράχος	0.25	0.35	0.63	0.68	0.65	0.75
κύμα	0.00	0.00	0.25	0.34	0.20	0.27

Πίνακας 7.5: *Βαθμοί βεβαιότητας για τις εικόνες του Σχήματος 7.9.*



Σχήμα 7.10: *Μια απλή οντολογία εννοιολογικού πλαισίου τύπων περιοχής για μέγεθος οπτικού θησαυρού ίσο με 4 και για το θεματικό πεδίο Παραλία.*

βάση τον οπτικό θησαυρό, το αρχικό διάνυσμα αναπαράστασης είναι:

$$\mathbf{MV}_{\text{πριν}} = \begin{bmatrix} 0.723 & 0.220 & 0.753 & 0.364 \end{bmatrix} . \tag{7.25}$$

Παρατηρώντας ότι η εικόνα του παραδείγματος αποτελείται από *ουρανό* και *θάλασσα*, διαισθητικά θα περίμενε κανείς ότι οι τύποι περιοχής που μοιάζουν οπτικά με τις έννοιες αυτές θα έχουν μεγαλύτερη βεβαιότητα στο διάνυσμα αναπαράστασης. Στην υπό εξέταση περίπτωση, όμως, η *θάλασσα* έχει αρκετά διαφορετικό χρώμα από τον τύπο περιοχής του λεξικού που οπτικά μοιάζει αυτήν. Θα μπορούσε να πει κανείς ότι η περιοχή που αντιστοιχεί στη *θάλασσα* μοιάζει π.χ. με *βράχο*. Το επιθυμητό μετά την αξιοποίηση του εννοιολογικού πλαισίου των τύπων περιοχης θα ήταν να αυξηθεί ο βαθμός βεβαιότητας του τύπου περιοχής που μοιάζει με τη *θάλασσα* και να μειωθεί ο βαθμός του τύπου περιοχής που μοιάζει με το *βράχο* (2η και 4η αντίστοιχα συνιστώσες του διανύσματος αναπαράστασης). Μετά την εφαρμογή του αλγορίθμου αξιοποίησης του εννοιολογικού πλαισίου, το διάνυσμα αναπαράστασης γίνεται:

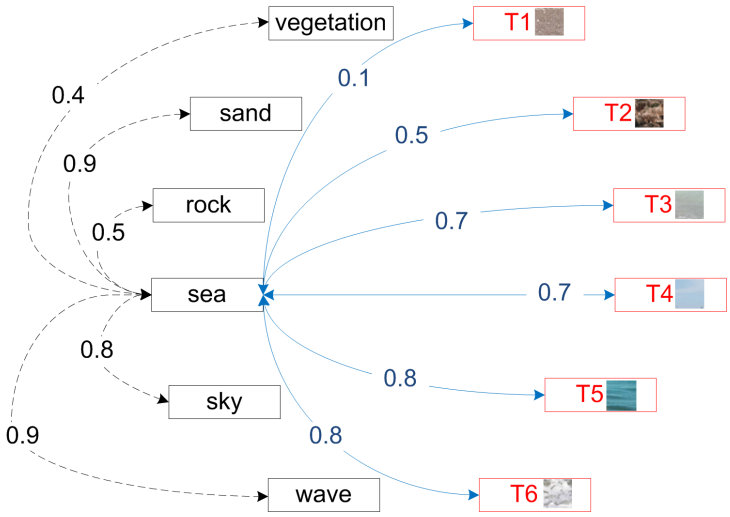
$$\mathbf{MV}_{\text{μετά}} = \begin{bmatrix} 0.778 & 0.452 & 0.800 & 0.338 \end{bmatrix} . \tag{7.26}$$

7.7.5 Εφαρμογή του Μεικτού Εννοιολογικού Πλαισίου

Στην περίπτωση του μεικτού εννοιολογικού πλαισίου, οι σχέσεις ανάμεσα στις έννοιες, ανάμεσα στους τύπους περιοχής και ανάμεσα σε διαφορετικές οντότητες

(α′) Εικόνα Εισόδου (β′) Κατάτμηση

Σχήμα 7.11: *Παράδειγμα εικόνας από το θεματικό πεδίο Παραλία, όπου το διάνυσμα αναπαράστασης είναι διαφορετικό από μια τυπική εικόνα Παραλίας.*

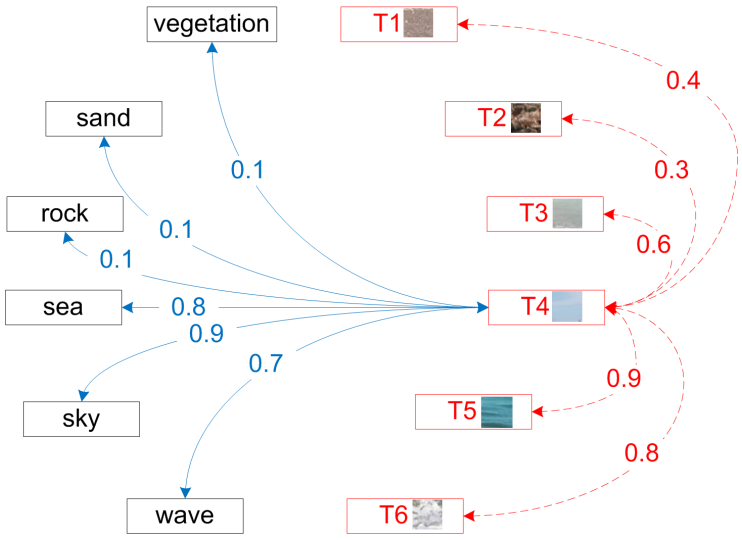


Σχήμα 7.12: *Ένα κομμάτι της οντολογίας εννοιολογικού πλαισίου. Απεικονίζονται οι σχέσεις ανάμεσα στην έννοια θάλασσα και τις υπόλοιπες οντότητες που την απαρτίζουν.*

σχημάτισαν την αντίστοιχη οντολογία. Στα Σχήματα 7.12 και 7.13 παρουσιάζονται δύο τμήματα του γράφου που κατασκευάστηκε με τη διαδικασία που περιγράφηκε στην Ενότητα 7.6 και χρησιμοποιείται για την αναπαράσταση της οντολογίας για το θεματικό πεδίο *Corel*. Πιο συγκεκριμένα, στο Σχήμα 7.12 απεικονίζονται σχέσεις ανάμεσα στην έννοια *θάλασσα* και τις υπόλοιπες οντότητες, ενώ στο Σχήμα 7.13 απεικονίζονται σχέσεις ανάμεσα στον τύπο περιοχής T_4 και τις υπόλοιπες οντότητες. Απεικονίζονται για λόγους ευκρίνειας μόνο οι σχέσεις μεταξύ των εννοιών και ενός μικρού αριθμού τύπων περιοχής. Θα πρέπει επίσης να διευκρινιστεί ότι παραλείπονται οι σχέσεις μεταξύ των οντοτήτων της οντολογίας και της έννοιας *Corel* που αποτελεί το θεματικό πεδίο και άρα και τη ρίζα του γράφου και με την οποία συνδέονται όλες οι έννοιες. Φυσικά, οι γράφοι αυτοί αντιπροσωπεύουν μόνο μικρά τμήματα της συνολικής οντολογίας εννοιολογικού πλαισίου. Η παρουσίαση σε ένα σχήμα της οντολογίας του θεματικού πλαισίου *Corel*, που αποτελείται από 7 έννοιες και 40 τύπους περιοχής (47 οντότητες συνολικά) δεν είναι εύκολη, καθώς απαιτεί την αναπαράσταση 188 σχέσεων το μέγιστο. Παρότι δεν εφαρμόζονται τελικά όλες οι σχέσεις, αριθμός τους παραμένει ιδιαίτερα μεγάλος και φυσικά καθιστά αδύνατη την πλήρη απεικόνισή τους σε ένα σχήμα.

Τέλος, στην Ενότητα αυτή παρουσιάζεται ένα παράδειγμα για τον τρόπο με τον οποίο επιδρά ο αλγόριθμος του μεικτού εννοιολογικού πλαισίου. Για τους σκοπούς του παραδείγματος αυτού χρησιμοποιήθηκε η απλή οντολογία μεικτού εννοιολογικού πλαισίου, μέρη της οποίας απεικονίζονται στα Σχήματα 7.12 και 7.13.

Υπενθυμίζεται ότι η σχέση ανάμεσα σε δύο οντότητες μπορεί να είναι μοναδική



Σχήμα 7.13: *Ένα κομμάτι της οντολογίας εννοιολογικού πλαισίου. Απεικονίζονται οι σχέσεις ανάμεσα στον τύπο περιοχής T_4 και τις υπόλοιπες οντότητες που την απαρτίζουν.*

ή ένας συνδυασμός από περισσότερες από μία σχέσεις, όπως φαίνεται στην Εξίσωση (7.19). Για να διευκολυνθεί η οπτικοποίηση της οντολογίας, ο Πίνακας 7.6, παρουσιάζει όλους τους πιθανούς συνδυασμός ανάμεσα σε μια έννοια υψηλού επιπέδου που αναφέρεται ως c_i και σε ένα υποσύνολο των υπόλοιπων οντοτήτων της οντολογίας. Τιμή βαθμού βεβαιότητας ίση με 0 υπονοεί την απουσία της σχέσης αυτής από τον αντίστοιχο συνδυασμό. Επιπρόσθετα, στον Πίνακα 7.7 παρουσιάζονται οι αντίστοιχες ασαφείς τιμές για κάθε ζεύγος από οντότητες για την σχέση *Συνοδός* και για ένα υποσύνολο των οντοτήτων. Υπενθυμίζεται ότι η σχέση αυτή είναι εφαρμόσιμη για οποιοδήποτε ζεύγος από οντότητες.

	C_1	C_2	...	C_N	T_1	T_2	...	T_M
Sim	0	0	...	0	0	0	...	0
Acc	1	0.5	...	0.9	0.7	0.8	...	0
P	1	0	...	0.3	0.7	0	...	0
$Comp$	1	0.2	...	0.9	0	0.5	...	0
Sp	0	0.8	...	0	0	0	...	0
Ex	0	0.7	...	0	0	0	...	0
Loc	0	0.9	...	0.8	0	0	...	0
Pr	0	0	...	0	0.5	0	...	0.7

Πίνακας 7.6: *Ασαφείς σχέσεις ανάμεσα στην έννοια υψηλού επιπέδου C_1 και σε όλες τις υπόλοιπες οντότητες. Οι αριθμοί δείχνουν τον βαθμό εμπιστοσύνης για κάθε σχέση.*

Ένα απλό παράδειγμα εφαρμογής του αλγορίθμου αξιοποίησης του μεικτού εννοιολογικού πλαισίου παρουσιάζεται προκειμένου να γίνει ευκολότερα κατανοητός ο τρόπος με τον οποίο επιδρά ο αλγόριθμος της Ενότητας 7.6. Για ευκολία στην παρουσίαση, χρησιμοποιείται ο οπτικός θησαυρός των 6 τύπων περιοχής, τμήματα του οποίου απεικονίζονται στα Σχήματα 7.12 και 7.13, στα οποία είναι και ορατοί οι τύποι περιοχής από τους οποίους αποτελείται. Η εικόνα στην οποία εφαρμόζεται απεικονίζεται στο Σχήμα 7.11. Το διάνυσμα αναπαράστασης T που περιέχει τους βαθμούς

	C_1	C_2	...	C_N	T_1	T_2	...	T_M
C_1	1	0.7	...	0	0.7	0.2	...	0.4
C_2	0.7	1	...	0.8	0.6	0.7	...	0.5
.	.	.		.	.	.		.
.	.	.		.	.	.		.
.	.	.		.	.	.		.
C_N	0	0.8	...	1	0.6	0.7	...	0.8
T_1	0.7	0.2	...	0.4	1	0.3	...	0.5
T_2	0.6	0.7	...	0.5	0.3	1	...	0.1
.	.	.		.	.	.		.
.	.	.		.	.	.		.
.	.	.		.	.	.		.
T_M	0.6	0.7	...	0.8	0.5	0.1	...	1

Πίνακας 7.7: *Οι βαθμοί εμπιστοσύνης της σχέσης* Συνοδός *για όλα τα ζεύγη από οντότητες. Οι αριθμοί δείχνουν τον βαθμό εμπιστοσύνης για κάθε σχέση.*

βεβαιότητας για την παρουσία των τύπων περιοχής T_i είναι

$$\mathbf{T} = \{T_i\} = \left[\begin{array}{ccccc} 0.89 & 0.62 & 0.21 & 0.68 & 0.670.31 \end{array} \right] , \tag{7.27}$$

ενώ οι βαθμοί βεβαιότητας c_i για τις υπό ανίχνευση έννοιες που προέκυψαν από την αρχική εφαρμογή των εκπαιδευμένων ανιχνευτών περιέχονται στο διάνυσμα C ως

$$\mathbf{C} = \{c_i\} = \left[\begin{array}{ccccc} 0.32 & 0.91 & 0.12 & 0.87 & 0.35 \end{array} \right] . \tag{7.28}$$

Όπως είναι προφανές, η εικόνα εισόδου περιέχει τις έννοιες *θάλασσα*, *ουρανό* και *κύμα*. Ωστόσο, η αρχική εκτίμηση των ανιχνευτών δεν έδωσε υψηλό βαθμό βεβαιότητας για την έννοια *θάλασσα*. Αυτό οφείλεται στο ότι το σύνολο ελέγχου δεν περιείχε εικόνες στις οποίες η έννοια *θάλασσα* να είχε παρόμοια περιγραφή χαμηλού επιπέδου με την εικόνα του παραδείγματος. Παρολαυτά, έπειτα από την εφαρμογή του αλγορίθμου το ανανεωμένο διάνυσμα αναπαράστασης είναι

$$\mathbf{T}' = \{T_i'\} = \left[\begin{array}{ccccc} 0.89 & 0.62 & 0.21 & 0.68 & 0.670.31 \end{array} \right] , \tag{7.29}$$

και το διάνυσμα που περιέχει τους ανανεωμένους βαθμούς βεβαιότητας είναι

$$\mathbf{C}' = \{c_i'\} = \left[\begin{array}{ccccc} 0.62 & 0.95 & 0.18 & 0.90 & 0.29 \end{array} \right] . \tag{7.30}$$

Ο αλγόριθμος αξιοποίησης του μεικτού εννοιολογικού πλαισίου αξιοποίησε τις παρακάτω πληροφορίες που ήταν διαθέσιμες στην οντολογία:

- Αυτή είναι μια εικόνα από το θεματικό πεδίο *Παραλία*, άρα χρησιμοποιείται η κατάλληλη οντολογία.
- Η έννοια *ουρανός* είχε υψηλό βαθμό βεβαιότητας, έπειτα από την αρχική ανίχνευση.

- Η έννοια *κύμα* είχε υψηλό βαθμό βεβαιότητας, έπειτα από την αρχική ανίχνευση.
- Στην εικόνα υπάρχει ένας *μπλε* τύπος περιοχής.
- Στην εικόνα υπάρχει ένας *άσπρος* τύπος περιοχής.
- Ο *ουρανός* και το *κύμα* έχουν "μεγάλη" σχέση με τη *θάλασσα*.
- Ο *μπλε* τύπος περιοχής και ο *άσπρος* τύπος περιοχής έχουν "μεγάλη" σχέση με τη *θάλασσα*.

Έτσι, τροποποίησε το διάνυσμα αναπαράστασης της εικόνας και τους βαθμούς βεβαιότητας ως εξής:

- Αύξησε το βαθμό βεβαιότητας για την παρουσία του *μπλε* τύπου περιοχής.
- Κράτησε σχεδόν αμετάβλητους τους βαθμούς βεβαιότητας των υπόλοιπων τύπων περιοχής.
- Αύξησε το βαθμό βεβαιότητας για την παρουσία της *θάλασσας*.
- Κράτησε σχεδόν αμετάβλητους τους βαθμούς βεβαιότητας των υπόλοιπων εννοιών.

Σε αυτό το απλοϊκό παράδειγμα γίνεται σαφής η δυνατότητα του αλγορίθμου αξιοποίησης του μεικτού εννοιολογικού πλαισίου σε προβλήματα ανίχνευσης εννοιών, αφού τελικά ανιχνεύθηκαν όλες οι παρούσες έννοιες, αν και η εικόνα θα μπορούσε να θεωρηθεί ως "παραπλανητική" ως προς την έννοια *θάλασσα*, σύμφωνα, πάντα με το σύνολο εκπαίδευσης.

7.7.6 Πειράματα στις συλλογές του TRECVID και του COREL

Στη συνέχεια παρουσιάζεται εκτενής αξιολόγηση των μεθόδων αξιοποίησης του εννοιολογικού πλαισίου που παρουσιάστηκαν στις Ενότητες 7.4, 7.5 και 7.6. Για το σκοπό αυτό χρησιμοποιήθηκαν τα σύνολα εικόνων από τις συλλογές του TRECVID και του Corel, που περιγράφηκαν στην Ενότητα 7.7.1. Για την ορθότερη αξιολόγηση των προτεινόμενων αλγορίθμων, αυτοί συγκρίθηκαν αρχικά με τις τεχνικές που παρουσιάστηκαν στο Κεφάλαιο 4. Τα αποτελέσματα της εφαρμογής τους στα σύνολα του Corel και του TRECVID παρατίθενται στους Πίνακες 7.8 και 7.9, αντίστοιχα.

Έννοιες	**RT**			**RT+LSA**			**C1**			**C2**			**C3**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
δρόμος	0.22	0.40	0.28	0.27	0.38	0.32	0.30	0.35	0.32	0.41	0.34	0.37	0.43	0.35	0.39
άμμος	0.38	0.50	0.43	0.42	0.48	0.45	0.50	0.48	0.49	0.52	0.47	0.49	0.55	0.44	0.49
θάλασσα	0.72	0.85	0.78	0.75	0.84	0.79	0.80	0.83	0.81	0.85	0.79	0.82	0.89	0.80	0.84
ουρανός	0.81	0.88	0.84	0.83	0.86	0.84	0.81	0.86	0.83	0.87	0.82	0.84	0.88	0.82	0.85
χιόνι	0.48	0.68	0.56	0.53	0.64	0.58	0.60	0.60	0.60	0.65	0.58	0.61	0.72	0.57	0.64
βλάστηση	0.67	0.81	0.73	0.70	0.78	0.74	0.74	0.76	0.75	0.72	0.78	0.75	0.81	0.74	0.77

Πίνακας 7.8: *Αποτελέσματα ανίχνευσης εννοιών στο σύνολο του Corel για τις τεχνικές των Κεφαλαίων 4 και 7. RT: ανίχνευση με οπτικό θησαυρό, RT+LSA: ανίχνευση με οπτικό θησαυρό και LSA, C1: εννοιολογικό πλαίσιο, C2: εννοιολογικό πλαίσιο τύπων περιοχής, C3: μεικτό εννοιολογικό πλαίσιο. P: ακρίβεια, R: ανάκτηση, F: F-μέτρο.*

Έννοιες	RT			RT+LSA			C1			C2			C3		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
βλάστηση	0.50	0.64	0.56	0.52	0.62	0.57	0.60	0.55	0.57	0.68	0.49	0.57	0.78	0.45	0.57
δρόμος	0.22	0.31	0.26	0.28	0.31	0.29	0.31	0.30	0.30	0.41	0.27	0.33	0.43	0.24	0.31
άμμος	0.83	0.82	0.81	0.93	0.80	0.86	0.92	0.76	0.83	0.93	0.69	0.79	1.00	0.76	0.86
νερό	0.60	0.67	0.63	0.63	0.68	0.65	0.66	0.65	0.65	0.70	0.58	0.63	0.81	0.57	0.67
ουρανός	0.60	0.79	0.68	0.62	0.80	0.70	0.66	0.74	0.70	0.74	0.68	0.71	0.90	0.57	0.70
χιόνι	0.43	0.50	0.46	0.49	0.48	0.44	0.49	0.45	0.47	0.51	0.38	0.44	0.57	0.37	0.44
φωτιά	0.30	0.47	0.37	0.35	0.47	0.41	0.42	0.45	0.43	0.46	0.37	0.41	0.55	0.36	0.43

Πίνακας 7.9: *Αποτελέσματα ανίχνευσης εννοιών στο σύνολο του TRECVID για τις τεχνικές των Κεφαλαίων 4 και 7. RT: ανίχνευση με οπτικό θησαυρό, RT+LSA: ανίχνευση με οπτικό θησαυρό και LSA, C1: εννοιολογικό πλαίσιο, C2: εννοιολογικό πλαίσιο τύπων περιοχής, C3: μεικτό εννοιολογικό πλαίσιο. P: ακρίβεια, R: ανάκτηση, F: F-μέτρο.*

Για την περαιτέρω αξιολόγηση των αλγορίθμων επιλέχτηκαν δύο τεχνικές που προσπαθούν να αντιμετωπίσουν το ίδιο πρόβλημα και έχουν χρησιμοποιηθεί στο TRECVID.

Η πρώτη που θα αναφέρεται ως "*R.LSA*" προτάθηκε από τους Souvannavong et al. [198] και εφαρμόστηκε στο TREVID από τους Jiten et al. [94]. Η τεχνική αυτή προσθέτει ευθέως δομικούς περιορισμούς στις οπτικές λέξεις του θησαυρού. Η βασική διαφορά της τεχνικής αυτής από τον παραδοσιακό αλγόριθμο LSA έγκειται στο ότι κάθε ζευγάρι λέξεων (χωρίς να έχει σημασία η σειρά) θεωρείται επίσης σαν οπτική λέξη. Έτσι δημιουργείται ένας οπτικός θησαυρός που περιέχει έναν μεγάλο αριθμό λέξεων. Τα οπτικά χαρακτηριστικά χαμηλού επιπέδου που εξάγονται στην περίπτωση αυτή είναι αρκετά πιο απλά από τους MPEG-7 περιγραφείς και περιορίζονται σε ένα απλό ιστόγραμμα χρώματος με 64 κορυφές, στον HSV χώρο και τις ενέργειες από 24 φίλτρα Gabor. Και στην περίπτωση αυτή, ο αριθμός των λέξεων που απαρτίζουν τον οπτικό θησαυρό καθορίζεται εμπειρικά. Η δεύτερη τεχνική που αξιολογείται προτάθηκε από τους Jiang et al. [91]. Η τεχνική αυτή εξάγει τοπικά σημεία ενδιαφέροντος και θα αναφέρεται ως "*LIPs*". Τα σημεία ενδιαφέροντος τείνουν να έχουν αρκετά διαφορετικές ιδιότητες από τα υπόλοιπα εικονοστοιχεία στη γειτονιά τους. Για την εξαγωγή τους, αρχικά εφαρμόζεται αρχικά η μέθοδος *Difference-of-Gaussian (DoG)*. Στη συνέχεια, από κάθε σημείο εξάγεται ένας SIFT περιγραφέας [126]. Έπειτα δημιουργείται ένας οπτικός θησαυρός, μέσα απο μια διαδικασία κβαντισμού των περιγραφέων των σημείων ενδιαφέροντος. Με τη βοήθεια του θησαυρού αυτού κάθε εικόνα περιγράφεται με ένα διάνυσμα αναπαράστασης. Τέλος, για κάθε έννοια εκπαιδεύεται ένας ανιχνευτής. Οι τεχνικές αυτές επιλέχτηκαν γιατί προσπαθούν να επιλύσουν ακριβώς το ίδιο πρόβλημα, έχοντας και περίπου το ίδιο κίνητρο με τις μεθόδους που παρουσιάστηκαν στα Κεφάλαια 4 και 7. Πιο συγκεκριμένα, η πρώτη τεχνική προσπαθεί να εκμεταλλευτεί τη συνύπαρξη των τύπων περιοχής και να ενσωματώσει στη διαδικασία κατασκευής του οπτικού θησαυρό γνώση για τη δομή των τύπων περιοχής. Αντίστοιχα, η δεύτερη ορίζει περιοχές γύρω από τα σημεία ενδιαφέροντος και εξάγει από αυτά κατάλληλους περιγραφείς, υλοποιώντας και αυτή μια παραλλαγή του μοντέλου bag-of-words. Ένας επιπλέον λόγος ήταν και η συμμετοχή των δύο τεχνικών στο TRECVID, με σχετική επιτυχία. Τα αποτελέσματα της σύγκρισης των τεχνικών αυτών με το μεικτό εννοιολογικό πλαίσιο φαίνονται στους Πίνακες 7.10 και 7.11. Μπορεί να παρατηρηθεί ότι το μεικτό εννοιολογικό πλαίσιο επιτυγχάνει την καλύτερη ακρίβεια ανάμεσα σε όλες τις τεχνικές που συγκρίθηκαν. Ωστόσο, σε κάποιες περιπτώσεις χάνει λίγο σε ανάκτηση.

Έννοιες	C3			R.LSA			LIPs		
	P	R	F	P	R	F	P	R	F
δρόμος	0.43	0.35	0.39	0.34	0.37	0.35	0.42	0.35	0.38
άμμος	0.55	0.44	0.49	0.47	0.46	0.46	0.52	0.45	0.48
θάλασσα	0.89	0.80	0.84	0.77	0.83	0.80	0.80	0.82	0.81
ουρανός	0.88	0.82	0.85	0.86	0.85	0.85	0.88	0.83	0.85
χιόνι	0.72	0.57	0.64	0.58	0.61	0.59	0.64	0.57	0.60
βλάστηση	0.81	0.74	0.77	0.73	0.76	0.74	0.76	0.73	0.74

Πίνακας 7.10: *Αποτελέσματα ανίχνευσης εννοιών στο σύνολο του Corel. C3: μεικτό εννοιολογικό πλαίσιο, R.LSA: η τεχνική του [198], LIPs: η τεχνική του [91]. P: ακρίβεια, R: ανάκτηση, F: F-μέτρο.*

Έννοιες	C3			R.LSA			LIPs		
	P	R	F	P	R	F	P	R	F
βλάστηση	0.78	0.45	0.57	0.50	0.59	0.54	0.52	0.55	0.54
δρόμος	0.43	0.24	0.31	0.30	0.30	0.30	0.37	0.27	0.31
άμμος	1.00	0.76	0.86	0.93	0.76	0.84	0.94	0.71	0.81
νερό	0.81	0.57	0.67	0.60	0.66	0.63	0.61	0.64	0.63
ουρανός	0.90	0.57	0.70	0.59	0.79	0.67	0.60	0.76	0.67
χιόνι	0.57	0.37	0.44	0.50	0.44	0.47	0.56	0.40	0.47
φωτιά	0.55	0.36	0.43	0.38	0.45	0.41	0.45	0.43	0.44

Πίνακας 7.11: *Αποτελέσματα ανίχνευσης εννοιών στο σύνολο του TRECVID. C3: μεικτό εννοιολογικό πλαίσιο, R.LSA: η τεχνική του [198], LIPs: η τεχνική του [91]. P: ακρίβεια, R: ανάκτηση, F: F-μέτρο.*

7.8 Συμπεράσματα

Στο Κεφάλαιο αυτό παρουσιάστηκαν τεχνικές που έχουν να κάνουν με την αξιοποίηση του εννοιολογικού πλαισίου των εικόνων ενός θεματικού πεδίου, αλλά και των τύπων περιοχής που περιέχονται σε αυτές. Έγινε σαφές ότι η αξιοποίηση του εννοιολογικού πλαισίου δύναται να παίξει καθοριστικό ρόλο στην ανίχνευση εννοιών υψηλού επιπέδου σε εικόνες, εκμεταλλευόμενη αφενός τη γνώση που υπάρχει για τις σχέσεις μεταξύ των εννοιών ενός θεματικού πεδίου και αφετέρου τη γνώση που υπάρχει για τους τύπους περιοχής που σχηματίζουν τις εικόνες ενός θεματικού πεδίου. Επίσης διαφάνηκε ότι η μείξη των δύο αυτών τεχνικών σε ένα ενιαίο πλαίσιο επέφερε ουσιαστική βελτίωση στα αποτελέσματα της αναλυσης.

Με βάση τα πειραματικά αποτελέσματα, έγινε σαφές προφανές ότι οι σχέσεις που απαρτίζουν το εννοιολογικό πλαίσιο ανάμεσα στις έννοιες και τους τύπους περιοχής είναι ικανές να βοηθήσουν στην βελτίωση της ακρίβειας που επιτυγχάνεται με το απλό μοντέλο bag-of-words. Οι προτεινόμενοι αλγόριθμοι χρησιμοποίησαν και εκμεταλλεύτηκαν τις σχέσεις αυτές. Η προσεκτική παρατήρηση των πειραματικών αποτελεσμάτων, αλλά και των απλών παραδειγμάτων της ανάλυσης που παρουσιάστηκαν στην Ενότητα 7.7, καταδεικνύουν ότι οι προτεινόμενοι αλγόριθμοι ευνοούν τις έννοιες που αρχικά ανιχνεύτηκαν με μεγάλο βαθμό βεβαιότητας Οι έννοιες που αρχικά είχαν μικρό βαθμό βεβαιότητας και σχετίζονται με αδύναμες σχέσεις με αυτές που είχαν μεγάλο βαθμό βεβαιότητας δεν ωφελούνται από την εφαρμογή των αλγορίθμων. Ο αλγόριθμος αξιοποίησης του εννοιολογικού πλαισίου των τύπων περιοχής εφαρμόστηκε με επιτυχία και κατάφερε να παρέχει βελτίωση στην περιγραφή των εικόνων και άρα να οδηγήσει σε υψηλότερη ακρίβεια στην ταξινόμηση.

Στα πλεονεκτήματα των τεχνικών που παρουσιάστηκαν συγκαταλέγονται η ευκολία στην υλοποίηση των προτεινόμενων οντολογιών, η ευκολία στον προσδιορισμό των βαθμών βεβαιότητας για τις σχέσεις ανάμεσα σε οντότητες, καθώς και η ευκολία

αναπαράστασης των οπτικών χαρακτηριστικών με τη βοήθεια του οπτικού θησαυρού. Ωστόσο, υπάρχουν και κάποια μειονεκτήματα, τα οποία και θα πρέπει να επισημανθούν. Καταρχάς, αρκετές από τις σχέσεις που ορίστηκαν δεν προσδιορίζονται με καθαρά υπολογιστικό τρόπο, αλλά με τη γνώμη ενός ειδήμονα. Οπότε, είναι δυνατόν μια οντολογία εννοιολογικού πλαισίου που έχει εφαρμοστεί με επιτυχία σε ένα σύνολο εικόνων, σε κάποιο άλλο που περιέχει εικόνες του ίδιου θεματικού πεδίου να μην δύναται να προσφέρει την ίδια βελτίωση, εξαιτίας των βαθμών αυτών.

Συμπερασματικά, η το εννοιολογικό πλαίσιο αξιοποιεί πληροφορία που υπάρχει στο σύνολο εικόνων που χρησιμοποιείται για την εκπαίδευση, ωστόσο μένει αναξιοποίητη από τις περισσότερες τεχνικές που βασίζονται στο μοντέλο bag-of-words. Η διαφορά στην ακρίβεια που προσφέρει μπορεί να μη φαντάζει ιδιαίτερα εντυπωσιακή, ωστόσο διαφαίνεται ότι υπάρχουν περιθώρια για περαιτέρω βελτιώσεις, οι οποίες για παράδειγμα θα μπορούσαν να επιτευχθούν με την περαιτέρω ενίσχυση του μοντέλου με νέες σχέσεις, όπως για παράδειγμα τοπολογικές.

Κεφάλαιο 8

Συμπεράσματα και Μελλοντικές Επεκτάσεις

8.1 Συμπεράσματα

Όσον αφορά το πρόβλημα της ταξινόμησης σκηνής, η διερεύνηση των τεχνικών που βασίστηκαν σε αλγορίθμους μηχανικής μάθησης κατέστησε σαφές ότι το πρόβλημα μπορεί να αντιμετωπιστεί με σχετική επιτυχία, χρησιμοποιώντας συγχώνευση περιγραφέων MPEG-7. Ανάμεσα στις τεχνικές που παρουσίαστηκαν, η τεχνική ταξινόμησης με απλή χρήση συγχωνευμένης περιγραφής που βασίστηκε σε SVM, παρουσίασε τα πλεονεκτήματα της ευκολίας στην υλοποίηση και της αρκετά ικανοποιητικής ακρίβειας και το μειονέκτημα της αδυναμίας να παρέχει ένα βαθμό βεβαιότητας. Οι τεχνικές πρώιμης και όψιμης συγχώνευσης που βασίστηκαν σε νευρωνικά δίκτυα, αφενός βελτιώνουν την ακρίβεια της ταξινόμησης και αφετέρου παρέχουν επιπλέον πληροφορίες για αυτήν. Η μέθοδος με την οποία γίνεται η ταξινόμηση υπολογίζει σε κάποιο στάδιο τις αποστάσεις μεταξύ δύο εικόνων. Αυτό μπορεί να φανεί χρήσιμο σε περιπτώσεις όπου είναι επιθυμητή η ανάκτηση εικόνων, αντί για την ταξινόμηση. Οι τεχνικές που χρησιμοποίησαν νευροασαφή δίκτυα παρείχαν έναν τρόπο εξαγωγής της γνώσης που απέκτησαν οι ταξινομητές με την εκπαίδευση. Δημιούργησαν ασαφείς κανόνες, μέσω των οποίων μπορεί να εξαχθεί ο μηχανισμός με τον οποίο γίνεται η ταξινόμηση. Γενικά, στο πρόβλημα της ταξινόμησης σκηνής η συγχώνευση των περιγραφέων λειτούργησε αποτελεσματικά και διαφάνηκε ότι όσο μεγαλύτερος είναι ο αριθμός των περιγραφέων που συγχωνεύονται, τόσο βελτιώνεται η ακρίβεια, καθώς διαφορετικοί περιγραφείς συλλαμβάνουν διαφορετικά χαρακτηριστικά των εικόνων.

Η αντιμετώπιση του προβλήματος της ταξινόμησης περιοχών εικόνων με χρήση γνώσης που αποθηκεύτηκε σε μια δομή οντολογιών αξιοποίησε την τεχνική συγχώνευσης των περιγραφέων του MPEG-7 και την εφάρμοσε στον υπολογισμό των αποστάσεων ανάμεσα σε μια περιοχή και ένα πρωτότυπο μιας έννοιας. Το βασικό πλεονέκτημα της μεθοδολογίας αυτής διαφάνηκε ότι ήταν η ευκολία εμπλουτισμού της γνώσης με νέα πρωτότυπα, αλλά και την εισαγωγή νέων περιγραφέων στα ήδη υπάρχοντα πρωτότυπα. Επίσης, η εισαγωγή ενός νέου θεματικού πεδίου στο ήδη υπάρχον σύστημα απαιτεί μόνο την κατασκευή μιας νέας οντολογίας θεματικού πεδίου που να περιέχει τα πρωτότυπα για τις έννοιες που περιλαμβάνει αυτό. Φυσικά, μεγάλο πλεονέκτημα αποτελεί το γεγονός ότι προσδιορίζεται η περιοχή της εικόνας στην οποία απεικονίζεται η έννοια. Το βασικό μειονέκτημα που διαφάνηκε είναι ότι η μεθοδολο-

γία αυτή δεν είναι εύκολο να εφαρμοστεί σε μεγάλα σύνολα εικόνων, αλλά και ότι εξαρτάται σε μεγάλο βαθμό από τα αποτελέσματα της κατάτμησης της εικόνας. Τέλος στα μειονεκτήματά της θα μπορούσε να θεωρηθεί το γεγονός ότι η κατασκευή της γνώσης γίνεται από τους χρήστες που αν δεν είναι ειδήμονες του εκάστοτε θεματικού πεδίου, η γνώση που θα κατασκευαστεί ενδέχεται να δυσκολέψει την ταξινόμηση.

Για το σκοπό της ανίχνευσης εννοιών υψηλού επιπέδου σε εικόνες αναπτύχθηκε μια τεχνική που βασίζεται στο μοντέλο bag-of-words. Η τεχνική που παρουσιάστηκε εμφανίζει το πλεονέκτημα ότι παρέχει μια προσέγγιση για την ανίχνευση εννοιών υψηλού επιπέδου σε εικόνες η οποία μπορεί να εφαρμοστεί για την ανίχνευση πολλών εννοιών, αρκεί αυτές να χαρακτηρίζονται ως υλικά ή σκηνές, χωρίς να εξαρτάται από το μέγεθος της περιοχής της εικόνας στην οποία περιέχεται η έννοια υπό ανίχνευση. Ένα μειονέκτημα της τεχνικής αυτής ήταν η μικρή εξάρτησή της από τον αλγόριθμο κατάτμησης που χρησιμοποιείται στο αρχικό της στάδιο. Επίσης, η περιγραφή των ιδιοτήτων της εικόνας με βάση τις περιοχές της δεν μπορεί σε καμία περίπτωση να οδηγήσει στην αναγνώριση αντικειμένων, καθώς κάτι τέτοιο θα απαιτούσε τέλεια κατάτμηση, κατ'αρχήν, κάτι που δεν είναι εφικτό. Ωστόσο, η εφαρμογή της τεχνικής στο σύνολο εικόνων που προέρχεται από το TRECVID οδήγησε σε ορισμένα χρήσιμα συμπεράσματα αφενός για την απόδοσή της και αφετέρου για τον τρόπο με τον οποίο πρέπει να αντιμετωπίζονται μεγάλα σύνολα δεδομένων. Τα πειραματικά αποτελέσματα απέδειξαν ότι η ανίχνευση εννοιών υψηλού επιπέδου σε εικόνες μπορεί να επιτευχθεί σε ικανοποιητικό βαθμό, όταν το περιεχόμενο μιας εικόνας περιγράφεται με ένα διάνυσμα αναπαράστασης το οποίο βασίζεται σε έναν οπτικό θησαυρό. Η τεχνική LSA λειτούργησε συμπληρωματικά και απέδειξε πειραματικά ότι βελτίωσε την ακρίβεια σε κάποιες έννοιες. Αυτό συμβαίνει καθώς η τεχνική αυτή ελαβε υπόψη τις λανθάνουσες σχέσεις μεταξύ των περιοχών και αυτές οι συσχετίσεις μπορεί να είναι πιο μεγάλες για μερικές έννοιες από κάποιες άλλες και διαφάνηκε ότι για έννοιες με υψηλές συσχετίσεις μεταξύ των περιοχών η τεχνική αποδίδει καλύτερα.

Οι περιλήψεις που κατασκευάστηκαν με τη χρήση του προτεινόμενου αλγορίθμου, φάνηκε ότι έχουν σαφώς πιο πλούσιο περιεχόμενο από ένα μοναδικό καρέ, αλλά και ότι μπορούν να δώσουν στο χρήστη τη δυνατότητα να αντιληφθεί πλήρως το σημασιολογικό περιεχόμενο ενός βίντεο, όσον αφορά, τουλάχιστον τις έννοιες που μπορούν να γίνουν αντιληπτές από το οπτικό του περιεχόμενο. Επίσης, η εφαρμογή της τεχνικής ανίχνευσης στις περιλήψεις που αποτελούνται απο πολλά χαρακτηριστικά καρέ, εμφάνισε αυξημένη ακρίβεια. Αυτό ήταν κάτι αναμενόμενο, καθώς απλές παρατηρήσεις στα βίντεο και τα μοναδικά χαρακτηριστικά καρέ που εξάγονται καταδεικνύουν ότι σπάνια περιέχονται όλες οι έννοιες σε ένα καρέ. Το τελευταίο θα μπορούσε να συμβεί μόνο στην περίπτωση που τα βίντεο έχουν πολύ μικρή διάρκεια και αναζητείται ένας πολύ μικρός αριθμός εννοιών. Η περίληψη που βασίστηκε στον τοπικό οπτικό θησαυρό συνδέεται άμεσα με την τεχνική ανίχνευσης με χρήση του οπτικού θησαυρού. Οι θησαυροί κατασκευάζονται με βάση πρακτικά την ίδια τεχνική. Η επιλογή των χαρακτηριστικών καρέ με βάση τους πιο σημαντικούς τύπους περιοχής που αυτά περιέχουν, αυτόματα συνεπάγεται ότι επιλέγονται τα καρέ που περιέχουν ικανό αριθμό από έννοιες.

Η ανάκτηση εικόνων που βασίστηκε στο μοντέλο bag-of-words δεν κατάφερε αντιμετωπίσει με επιτυχία όλες τις συλλογές εικόνων και όλες τις πιθανές προσδοκίες των χρηστών. Όπως έδειξαν και τα αποτελέσματα, είναι δυνατόν να διαχωριστούν εικόνες που περιέχουν διαφορετικές έννοιες, με την αναπαράσταση που βασίζεται στον οπτικό θησαυρό περιοχών, εφόσον οι έννοιες ανήκουν στην κατηγορία των

"υλικών". Το ίδιο δε θα μπορούσε να μη συμβαίνει και στην περίπτωση της ανάκτησης. Η προτεινόμενη τεχνική μπορεί εύκολα και αποδοτικά να εφαρμοστεί σε περιπτώσεις που στόχος είναι η ανάκτηση εικόνων με βάση τη σημασιολογία τους, εφόσον οι έννοιες που αυτές περιέχουν ανήκουν στην κατηγορία των υλικών, αλλά και με βάση το οπτικό τους περιεχόμενο, καθώς μπορεί να παρέχει μια ικανοποιητική περιγραφή των καθολικών οπτικών χαρακτηριστικών, η οποία είναι πληρέστερη π.χ από τα απλά ιστογράμματα χρώματος.

Τέλος, οι σχέσεις που απαρτίζουν το εννοιολογικό πλαίσιο ανάμεσα στις έννοιες και τους τύπους περιοχής είναι ικανές να βοηθήσουν στην βελτίωση της ακρίβειας που επιτυγχάνεται με το απλό μοντέλο bag-of-words. Οι προτεινόμενοι αλγόριθμοι χρησιμοποίησαν και εκμεταλλεύτηκαν τις σχέσεις αυτές. Τα πειραματικά αποτελέσματα που παρουσιάστηκαν καταδεικνύουν ότι οι προτεινόμενοι αλγόριθμοι ευνοούν τις έννοιες που αρχικά ανιχνεύτηκαν με μεγάλο βαθμό βεβαιότητας. Οι έννοιες που αρχικά είχαν μικρό βαθμό βεβαιότητας και σχετίζονται με αδύναμες σχέσεις με αυτές που είχαν μεγάλο βαθμό βεβαιότητας δεν ωφελούνται από την εφαρμογή των αλγορίθμων. Ο αλγόριθμος αξιοποίησης του εννοιολογικού πλαισίου των τύπων περιοχής εφαρμόστηκε με επιτυχία και κατάφερε να παρέχει βελτίωση στην περιγραφή των εικόνων και άρα να οδηγήσει σε υψηλότερη ακρίβεια στην ταξινόμηση. Έτσι διαφάνηκε ότι το εννοιολογικό πλαίσιο αξιοποιεί πληροφορία που υπάρχει στο σύνολο εικόνων που χρησιμοποιείται για την εκπαίδευση, ωστόσο μένει αναξιοποίητη από τις περισσότερες τεχνικές που βασίζονται στο μοντέλο bag-of-words. Η διαφορά στην ακρίβεια που προσφέρει μπορεί να μη φαντάζει ιδιαίτερα εντυπωσιακή, ωστόσο διαφαίνεται ότι υπάρχουν περιθώρια για περαιτέρω βελτιώσεις, οι οποίες για παράδειγμα θα μπορούσαν να επιτευχθούν με την περαιτέρω ενίσχυση του μοντέλου με νέες σχέσεις, όπως για παράδειγμα τοπολογικές.

8.2 Συνεισφορά της Διατριβής

Η τεχνική ταξινόμησης εικόνων που παρουσιάστηκε στο Κεφάλαιο 2 προτείνει νέες τεχνικές για τη συγχώνευση περιγραφέων MPEG-7, με χρήση τεχνικών μηχανικής μάθησης. Φυσικά οι τεχνικές αυτές μπορούν να επεκταθούν και με τη χρήση άλλων περιγραφέων, αρκεί αυτοί να έχουν τη μορφή διανύσματος. Επίσης προτείνεται η εξαγωγή ασαφών κανόνων για την ταξινόμηση εικόνας, κάτι που επιτεύχθηκε με χρήση νευροασαφών δικτύων. Τα συμπεράσματα αυτά χρησιμοποιήθηκαν και για την ταξινόμηση περιοχών στο Κεφάλαιο 3, όπου προτάθηκε μια τεχνική που χρησιμοποιεί νευρωνικά δίκτυα για τη σύγκριση πρωτοτύπων των εννοιών με περιοχές εικόνων. Η τεχνική ταξινόμησης του Κεφαλαίου 4 προτείνει μια νέα μεθοδολογία τόσο για την κατασκευή ενός οπτικού λεξικού, όσο και για την περιγραφή εικόνων με βάση τους τύπους περιοχής που εμφανίζονται σε αυτό. Επίσης ενσωματώνει στο μοντέλο bag-of-words την τεχνική της λανθάνουσας σημασιολογικής ανάλυσης. Στο Κεφάλαιο 5 προτείνεται ένας αλγόριθμος εξαγωγής χαρακτηριστικών καρέ από βίντεο. Η πρωτοτυπία του αλγορίθμου έγκειται στην περιγραφή του οπτικού περιεχομένου των χαρακτηριστικών καρέ που γίνεται με τυπους περιοχής, αλλά και στην επιλογή τους που πραγματοποιείται αξιοποιώντας τη γνώση για τον οπτικό θησαυρό. Η τεχνική της ανάκτησης εικόνων που παρουσιάστηκε στο Κεφάλαιο 6 προτείνει μια νέα παραλλαγή του μοντέλου bag-of-words, κατάλληλη για ανάκτηση εικόνων. Τέλος, στο Κεφάλαιο 7 προτείνονται νέοι τρόποι αναπαράστασης και αξιοποίησης του εννοιολο-

γικού πλαισίου των εικόνων και των περιοχών τους και εφαρμόζονται στο πρόβλημα της ανίχνευσης εννοιών. Ιδιαίτερη μνεία πρέπει να γίνει στο εννοιολογικό πλαίσιο των τύπων περιοχής, αλλά και στο μεικτό εννοιολογικό πλαίσιο, που εισάγουν σχέσεις ανάμεσα σε τύπους περιοχής και σε έννοιες και τύπους περιοχής, αντίστοιχα.

8.3 Μελλοντικές Επεκτάσεις

Όπως διαφάνηκε από την εώς τώρα ερευνητική ενασχόληση του συγγραφέα με τα προβλήματα που αντιμετωπίστηκαν στην παρούσα Εργασία, ιδιαίτερο ενδιαφέρον παρουσιάζει το ερευνητικό πεδίο που ασχολείται με την περιγραφή εικόνων με βάση το μοντέλο bag-of-words. Παρότι αυτό έχει χρησιμοποιηθεί κατά κόρον, υπάρχουν ακόμη ορισμένες πλευρές του που χρήζουν περαιτέρω έρευνας. Έτσι, ιδιαίτερο ενδιαφέρον παρουσιάζει ο τρόπος με τον οποίο κατασκευάζεται το οπτικό λεξικό. Όλες οι τεχνικές στη βιβλιογραφία υιοθετούν έναν αλγόριθμο συσταδοποίησης. Ωστόσο, σπάνιες είναι οι περιπτώσεις που διερευνάται ποιος θα πρέπει να είναι ο αριθμός των συστάδων από τις οποίες και θα προκύψουν οι οπτικές λέξεις. Επιπρόσθετα, ενδιαφέρον έχει και η μελέτη των χωρικών σχέσεων ανάμεσα στους διάφορους τύπους περιοχής. Οι τεχνικές ενδέχεται να προκύψουν από αυτή τη μελέτη θα μπορούν να χρησιμοποιηθούν για την κατασκευή οπτικού λεξικού ανεξαρτήτως των περιγραφέων που χρησιμοποιούνται.

Πέρα από αυτό, η ανίχνευση των εννοιών με βάση το εννοιολογικό τους πλαίσιο αποτελεί ένα ακόμη πεδίο στο οποίο δύναται να στραφεί μελλοντική έρευνα. Τα μοντέλα που παρουσιάστηκαν στο Κεφάλαιο 7 περιλαμβάνουν μόνο σημασιολογικές σχέσεις ανάμεσα στις έννοιες και τους τύπους περιοχής. Μπορούν να εμπλουτιστούν και ως προς την κατεύθυνση του χωρικού εννοιολογικού πλαισίου, αλλα και προς αυτή του χρονικού εννοιολογικού πλαισίου και του εννοιολογικού πλαισίου μεταπληροφορίας. Στην περίπτωση που εφαρμοστούν σε μεγάλες συλλογές εικόνων που περιέχουν πληροφορία ως προς τον τόπο που έχει ληφθεί μια φωτογραφία, σαν αυτές που συναντώνται σε δημοφιλείς ιστοτόπους όπως για παράδειγμα τη συλλογή του Flickr, είναι δυνατόν να εμπλουτιστεί και προς αυτή την κατεύθυνση το εννοιολογικό πλαίσιο. Φυσικά για να αξιοποιηθεί ένα τέτοιο πολυσύνθετο εννοιολογικό πλαίσιο, απαιτείται και η ανάπτυξη νέων και εξίσου σύνθετων αλγορίθμων.

Τέλος, όπως φάνηκε στην παρούσα Εργασία, οι τεχνικές που χρησιμοποιούνται για την περιγραφή εικόνων μπορούν εύκολα να χρησιμοποιηθούν σε προβλήματα όπως της ανάκτησης πολυμεσικού περιεχομένου και της δημιουργίας περιλήψεων. Έτσι, τεχνικές που τυχόν προκύψουν, όπως αυτές που προαναφέρθηκαν, είναι δυνατόν να εφαρμοστούν και στα προβλήματα αυτά, κάτι που έγινε και στην παρούσα Εργασία.

Βιβλιογραφία

[1] Allen, J. Maintaining knowledge about temporal intervals. *Communications of the ACM 26*, 1 (1983), 832--843.

[2] Anthoine, S., Debreuve, E., Piro, P., and Barlaud, M. Using neighborhood distributions of wavelet coefficients for on-the-fly, multiscale-based image retrieval. In *Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)* (2008).

[3] Antunes, A., Lingnau, C., and Centeno, J. Object oriented analysis and semantic network for high resolution image classification. *Bol. Ciênc. Geod., sec. Artigos 9*, 2 (2003), 233--242.

[4] Armitage, L., and Enser, P. Analysis of user need in image archives. *Journal of information science 23*, 4 (1997).

[5] Avrithis, Y., Doulamis, A., Doulamis, N., and Kollias, S. A Stochastic Framework for Optimal Key Frame Extraction from MPEG Video Databases. *Computer Vision and Image Understanding 75*, 1 (1999), 3--24.

[6] Ayache, S., and Quenot, G. TRECVID 2007 Collaborative Annotation using Active Learning. In *Proceedings of the TRECVID 2007 Workshop* (2007).

[7] Bach, J., Fuller, C., Gupta, A., Hampapur, A., Horowitz, B., Humphrey, R., Jain, R., and Shu, C. Virage image search engine: an open framework for image management. In *Proceedings of SPIE* (1996).

[8] Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D., and Jordan, M. Matching Words and Pictures. *The Journal of Machine Learning Research 3* (2003), 1107--1135.

[9] Bay, H., Tuytelaars, T., and Van Gool, L. SURF: Speeded Up Robust Features. *Lecture Notes in Computer Science 3951* (2006).

[10] Beckett, D., and McBride, B. RDF/XML syntax specification (revised). *W3C Recommendation 10* (2004).

[11] Benitez, A., and Chang, S. Image classification using multimedia knowledge networks. In *IEEE International Conference on Image Processing (ICIP)* (2003).

[12] Benitez, A., Zhong, D., Chang, S., and Smith, J. MPEG-7 MDS content description tools and applications. *Lecture Notes in Computer Science* (2001).

[13] Biederman, I. On the semantics of a glance at a scene. *Perceptual organization* (1981).

[14] Biederman, I. Recognition-by-components: A theoryof human image understanding. *Psychological review 94*, 2 (1987), 115--147.

[15] Biederman, I., Mezzanotte, R., and Rabinowitz, J. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology 14*, 2 (1982), 143--177.

[16] Bishop, C., et al. *Pattern recognition and machine learning*. Springer New York:, 2006.

[17] Bober, M. MPEG-7 visual shape descriptors. *IEEE Transactions on Circuits and Systems for Video Technology 11*, 6 (2001), 716--719.

[18] Boll, S., Sandhaus, P., Scherp, A., and Thieme, S. MetaXa—Context- and content-driven metadata enhancement for personal photo books. In *International Multi-Media Modeling conference (MMM), Singapore* (2007).

[19] Bosch, A., Zisserman, A., and Munoz, X. Scene classification using a hybrid generative/discriminative approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence 30*, 4 (2008).

[20] Bose, B., and Grimson, E. Learning to use scene context for object classification in surveillance. In *Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance* (2003).

[21] Bosi, M., Brandenburg, K., Quackenbush, S., Fielder, L., Akagiri, K., Fuchs, H., Dietz, M., Herre, J., Davidson, G., and Oikawa, Y. ISO/IEC MPEG-2 advanced audio coding. *Journal of the Audio engineering society 45*, 10 (1997), 789--814.

[22] Boujemaa, N., and Nastar, C. Content-based image retrieval at the imedia group of the inria. In *10th DELOS Workshop Audio-Visual Digital Libraries* (1999).

[23] Boutell, M., Brown, C., and Luo, J. Learning spatial configuration models using modified Dirichlet priors. In *ICML 2004 Workshop on Statistical Relational Learning and its Connections to Other Fields*.

[24] Boutell, M., and Luo, J. Bayesian fusion of camera metadata cues in semantic scene classification. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2004).

[25] Boutell, M., and Luo, J. Photo classification by integrating image content and camera metadata. In *International Conference on Pattern Recognition (ICPR)* (2004).

[26] Boutell, M., and Luo, J. Beyond pixels: Exploiting camera metadata for photo classification. *Pattern recognition 38*, 6 (2005), 935--946.

[27] Boutell, M., Luo, J., and Brown, C. A generalized temporal context model for classifying image collections. *Multimedia Systems 11*, 1 (2005), 82--92.

[28] Boutell, M., Luo, J., and Brown, C. Improved semantic region labeling based on scene context. In *IEEE International Conference on Multimedia and Expo (ICME)* (2005).

[29] Boutell, M., Luo, J., and Gray, R. Sunset scene classification using simulated image recomposition. In *International Conference on Multimedia and Expo (ICME)* (2003).

[30] Brandenburg, K., Stoll, G., et al. The ISO-MPEG-1 audio: A generic standard for coding of high-quality digital audio. *Journal of the Audio Engineering Society 42*, 10 (1994), 780--792.

[31] Breen, C., Khan, L., and Ponnusamy, A. Image classification using neural networks and ontologies. In *International Workshop on Database and Expert Systems Applications* (2002).

[32] Brickley, D., and Guha, R. RDF vocabulary description language 1.0: RDF schema. *W3C recommendation 10* (2004).

[33] Burghardt, T., Calic, J., and Thomas, B. Tracking animals in wildlife videos using face detection. In *European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology* (2004).

[34] Calic, J., Kramer, P., Naci, U., Vrochidis, S., Aksoy, S., Zhang, Q., Benois-Pineau, J., Saracoglu, A., Doulaverakis, C., Jarina, R., et al. Cost292 experimental framework for trecvid 2006. In *TRECVID Workshop* (2006).

[35] CandÁes, E., and Donoho, D. Ridgelets: the key to high dimensional intermittency? *Philosophical Transactions of the Royal Society of London 357* (1999), 2495--2509.

[36] Carbonetto, P., de Freitas, N., and Barnard, K. A statistical model for general contextual object recognition. *Lecture Notes in Computer Science* (2004).

[37] Carpenter, G., Grossberg, S., and Rosen, D. Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural networks 4*, 6 (1991), 759--771.

[38] Carson, C., Thomas, M., Belongie, S., Hellerstein, J., and Malik, J. Blobworld: A system for region-based image indexing and retrieval. *Lecture Notes in Computer Science* (1999).

[39] Cave, C., and Kosslyn, S. The role of parts and spatial relations in object identification. *Perception 22* (1993), 229--229.

[40] Chang, C.-C., and Lin, C.-J. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[41] Chang, H., Sull, S., and Lee, S. Efficient video indexing scheme for content-based retrieval. *IEEE Transactions on Circuits and Systems for Video Technology 9*, 8 (1999), 1269--1279.

[42] Chang, S., Chen, W., Meng, H., and Sundaram, H. VideoQ: an automated content based video search system using visual cues. In *ACM International Conference on Multimedia* (1997).

[43] Chang, S., Sikora, T., and Purl, A. Overview of the MPEG-7 Standard. *IEEE Transactions on Circuits and Systems for Video Technology 11*, 6 (2001), 688--695.

[44] Chapelle, O., Haffner, P., and Vapnik, V. SVMs for histogram-based image classification. *IEEE transactions on Neural Networks 10*, 5 (1999).

[45] Chen, Y., Wang, J., and Krovetz, R. Clue: Cluster-based retrieval of images by unsupervised learning. *IEEE Transactions on Image Processing 14*, 8 (2005), 1187--1201.

[46] Chiu, S. *Extracting Fuzzy Rules from Data for Function Approximation and Pattern Classification*. John Wiley and Sons, 1997.

[47] Chopra, R., and Srihari, R. Control structures for incorporating picture-specific context in image interpretation. In *International Joint Conference on Artificial Intelligence* (1995).

[48] Christel, M., and Conescu, R. Addressing the challenge of visual information access from digital image and video libraries. In *ACM/IEEE-CS joint conference on Digital libraries* (2005).

[49] Chum, O., and Matas, J. Web scale image clustering: Large scale discovery of spatially related images. Tech. rep., Technical Report CTU-CMP-2008-15, Czech Technical University in Prague, 2008.

[50] Chum, O., and Matas, J. Geometric hashing with local affine frames. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2006).

[51] Cohn, A., Bennett, B., Gooday, J. M., and Gotts., N. M. *Representing and Reasoning with Qualitative Spatial Relations about Regions*. Kluwer Academic Publishers, 1997.

[52] CP-3451, J. 3451, Exchangeable image file format for digital still cameras: Exif Version 2.2, 2002.

[53] Csurka, G., Dance, C., Fan, L., Willamowski, J., and Bray, C. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV* (2004), vol. 1.

[54] Csurka, G., Dance, C., Fan, L., Willamowski, J., and Bray, C. Visual Categorization with Bags of Keypoints. In *Workshop on Statistical Learning in Computer Vision (ECCV)* (2004).

[55] Cula, O., and Dana, K. Compact representation of bidirectional texture functions. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2001), IEEE Computer Society; 1999.

[56] Datta, R., Joshi, D., Li, J., and Wang, J. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv. 40*, 2 (2008), 1--60.

[57] Deerwester, S., Dumais, S., Furnas, G., Landauer, T., and Harshman, R. Indexing by latent semantic analysis. *Journal of the American society for information science 41*, 6 (1990), 391--407.

[58] Divakaran, A., Radhakrishnan, R., and Peker, K. Motion activity-based extraction of key-frames from video shots. In *IEEE International Conference on Image Processing (ICIP)* (2002).

[59] Dorado, A., Djordjevic, D., Izquierdo, E., and Pedrycz, W. Supervised semantic scene classification based on low-level clustering and relevance feedback. In *European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies (EWIMT)* (2004).

[60] Dorado, A., and Izquierdo, E. Exploiting problem domain knowledge for accurate building image classification. *Lecture notes in computer science* (2004).

[61] Dorairaj, R., and Namuduri, K. Compact combination of MPEG-7 color and texture descriptors for image retrieval. In *38th Asilomar Conference on Signals, Systems and Computers* (2004).

[62] Duygulu, P., Barnard, K., De Freitas, J., and Forsyth, D. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. *Lecture Notes in Computer science* (2002).

[63] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results. http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html.

[64] Fan, J., Gao, Y., and Luo, H. Multi-level annotation of natural scenes using dominant image components and semantic concepts. In *ACM international conference on Multimedia* (2004).

[65] Fauvet, B., Bouthemy, P., Gros, P., and Spindler, F. A geometrical key-frame selection method exploiting dominant motion estimation in video. *Lecture notes in computer science* (2004).

[66] Fei-Fei, L. Bag-of-words models, 2007. `http://vision.cs.princeton.edu/documents/CVPR2007\_tutorial\_bag\_of\_words.ppt`.

[67] Fei-Fei, L., and Perona, P. A Bayesian Hierarchical Model for Learning Natural Scene Categories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2005).

[68] Fellbaum, C. *WordNet: An electronic lexical database*. MIT press Cambridge, MA, 1998.

[69] Fitzpatrick, P. Indoor/outdoor scene classification project.

[70] Fukunage, K., and Narendra, P. A branch and bound algorithm for computing k-nearest neighbors. *IEEE transactions on computers 100*, 24 (1975), 750--753.

[71] Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., and Schneider, L. Sweetening Ontologies with DOLCE. In *13th International Conference on Knowledge Acquisition, Modeling and Management (EKAW)* (2002).

[72] Gibson, D., Campbell, N., and Thomas, B. Visual abstraction of wildlife footage using Gaussian mixture models and the minimum description length criterion. In *International Conference on Pattern Recognition* (2002).

[73] Girgensohn, A., and Boreczky, J. Time-constrained keyframe selection technique. *Multimedia Tools and Applications 11*, 3 (2000), 347--358.

[74] Giro, X., and Marques, F. Detection of semantic objects using description graphs. In *IEEE International Conference on Image Processing (ICIP)* (2005).

[75] Gokalp, D., and Aksoy, S. Scene classification using bag-of-regions representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2007).

[76] Gonzalez, R., and Woods, R. *Digital Image Processing*. Prentice Hall, 2002.

[77] Gorkani, M., and Picard, R. Texture orientation for sorting photos at a glance. In *12th IAPR International Conference on Pattern Recognition* (1994).

[78] Guérin-Dugué, A., and Oliva, A. Classification of scene photographs from local orientations features. *Pattern Recognition Letters 21*, 13-14 (2000), 1135--1140.

[79] Guillemot, M., Wellner, P., Gatica-Pérez, D., and Odobez, J. A hierarchical keyframe user interface for browsing video over the Internet. In *IFIP TC13 International Conference on Human-Computer Interaction (Interact)* (2003).

[80] Hammoud, R., and Mohr, R. A probabilistic framework of selecting effective key frames for video browsing and indexing. In *International workshop on Real-Time Image Sequence Analysis* (2000).

[81] Hanjalic, A., Lagendijk, R., and Biemond, J. A new method for key frame based video content representation. *Image Databases and Multi Media Search.*

[82] Hartigan, J., and Wong, M. A k-means Clustering Algorithm. *Applied Statistics* (1979), 100--108.

[83] Haykin, S. *Neural networks: a comprehensive foundation.* Prentice Hall, 2008.

[84] Hecht-Nielsen, R. Theory of the backpropagation neural network. *Neural Networks 1* (1988), 445.

[85] Hofmann, T. Probabilistic latent semantic indexing. In *International ACM SIGIR conference on Research and development in information retrieval* (1999).

[86] Hudelot, C., and Thonnat, M. A cognitive vision platform for automatic recognition of natural complex objects. In *IEEE International Conference on Tools with Artificial Intelligence* (2003).

[87] Huijsmans, D., and Sebe, N. How to complete performance graphs in content-based image retrieval: Add generality and normalize scope. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2005).

[88] Hunter, J., Drennan, J., and Little, S. Realizing the hydrogen economy through semantic web technologies. *IEEE Intelligent Systems 19*, 1 (2004), 40--47.

[89] Jegou, H., Douze, M., and Schmid, C. Hamming embedding and weak geometric consistency for large scale image search. *European Conference on Computer Vision (ECCV)* (2008).

[90] Jiang, Y., Ngo, C., and Yang, J. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *ACM international conference on Image and video retrieval* (2007).

[91] Jiang, Y., Zhao, W., and Ngo, C. Exploring semantic concept using local invariant features. In *Asia-Pacific Workshop on Visual Information Processing* (2006).

[92] Jing, F., Li, M., Zhang, H., and Zhang, B. An effective region-based image retrieval framework. In *ACM international conference on Multimedia* (2002).

[93] Jiten, J., Merialdo, B., and Huet, B. Semantic feature extraction with multidimensional hidden Markov model. In *Proceedings of SPIE* (2006), vol. 6073, pp. 211--221.

[94] Jiten, J., Souvannavong, F., Merialdo, B., and Huet, B. Eurecom at TRECVid 2005: extraction of high-level features. *TRECVid 2005* (2005).

[95] Jurie, F., and Triggs, B. Creating efficient codebooks for visual recognition. In *IEEE International Conference on Computer Vision (ICCV)* (2005).

[96] Kalantidis, Y., Tolias, G., Spyrou, E., Mylonas, P., and Avrithis, Y. Visual image retrieval and localization. International Workshop on Content-Based Multimedia Indexing (CBMI).

[97] Kang, H. Video abstraction techniques for a digital library. *Distributed multimedia databases: Techniques and applications* (2002).

[98] Kato, Z., Zerubia, J., and Berthod, M. Unsupervised parallel image classification using a hierarchical markovian model. *International Conference on Computer Vision (ICCV)* (1995).

[99] Ke, Y., Sukthankar, R., and Huston, L. Efficient near-duplicate detection and sub-image retrieval. In *ACM Multimedia* (2004).

[100] Kennedy, L., Hauptmann, A., Naphade, M., Smith, A., and Chang, S. LSCOM lexicon definitions and annotations version 1.0. In *DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia* (2006).

[101] Kim, C., and Hwang, J. An integrated scheme for object-based video abstraction. In *ACM international conference on Multimedia* (2000).

[102] Kim, S., and Kweon, I. Simultaneous Classification and Visual Word Selection using Entropy-based Minimum Description. In *18th International Conference on Pattern Recognition (ICPR)* (2006).

[103] Klir, G., Clair, U., and Yuan, B. *Fuzzy set theory: foundations and applications*. Prentice-Hall, 1997.

[104] Kouzani, A. Locating human faces within images. *Computer Vision and Image Understanding 91*, 3 (2003), 247--279.

[105] Laaksonen, J., Koskela, M., Laakso, S., and Oja, E. PicSOM--content-based image retrieval with self-organizing maps. *Pattern Recognition Letters 21*, 13-14 (2000), 1199--1207.

[106] Laaksonen, J., Koskela, M., and Oja, E. PicSOM-self-organizing image retrieval with MPEG-7 content descriptors. *IEEE Transactions on Neural Networks 13*, 4 (2002), 841--853.

[107] Lagendijk, R., Hanjalic, A., Ceccarelli, M., Soletic, M., and Persoon, E. Visual search in a SMASH system. In *International Conference on Image Processing (ICIP)* (1997).

[108] Lavrenko, V., Manmatha, R., and Jeon, J. A model for learning the semantics of pictures.

[109] Lazebnik, S., Schmid, C., and Ponce, J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories.

[110] Le Borgne, H., and O'Connor, N. Natural scene classification and retrieval using ridgelet-based image signatures. In *Advanced Concepts for Intelligent Vision Systems* (2005).

[111] Le Saux, B., and Amato, G. Image recognition for digital libraries. In *ACM SIGMM international workshop on Multimedia information retrieval* (2004).

[112] Lehmann, T., Gold, M., Thies, C., Fischer, B., Spitzer, K., Keysers, D., Ney, H., Kohnen, M., Schubert, H., and Wein, B. Content-based image retrieval in medical applications. *Methods of Information in Medicine 43*, 4 (2004), 354--361.

[113] Leibe, B., and Schiele, B. *Interleaved object categorization and segmentation.* 2005.

[114] Leung, T., and Malik, J. Representing and recognizing the visual appearance of materials using three-dimensional textons. *International Journal of Computer Vision 43*, 1 (2001), 29--44.

[115] Li, J., Najmi, A., and Gray, R. Image classification by a two-dimensional hidden Markov model. *IEEE Transactions on Signal Processing 48*, 2 (2000), 517--533.

[116] Li, J., and Wang, J. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence 2* (2003), 1075--1088.

[117] Li, J., and Wang, J. ALIPR-automatic image tagging and visual image search, 2007.

[118] Li, W., and Sun, M. Semi-supervised learning for image annotation based on conditional random fields. *Lecture Notes in Computer Science 4071* (2006).

[119] Li, X., Wang, L., and Sung, E. Multi-label svm active learning for image classification. In *IEEE International Conference on Image Processing (ICIP)* (2004).

[120] Li, Y., Lee, S., Yeh, C., and Kuo, C. Techniques for movie content analysis and skimming. *IEEE Signal Processing Magazine 23*, 2 (2006), 79.

[121] Li, Y., Zhang, T., and Tretter, D. An overview of video abstraction techniques. *HP Laboratories Palo Alto* (2001).

[122] Lipson, P., Grimson, E., and Sinha, P. Configuration based scene classification and image indexing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (1997).

[123] Little, S., and Hunter, J. Rules-by-example-a novel approach to semantic indexing and querying of images. *Lecture notes in computer science* (2004).

[124] Liu, T., Zhang, H., and Qi, F. A novel video key-frame-extraction algorithm based on perceived motion energy model. *IEEE transactions on circuits and systems for video technology 13*, 10 (2003), 1006--1013.

[125] Liu, X., Zhang, L., Li, M., Zhang, H., and Wang, D. Boosting image classification with LDA-based feature combination for digital photograph management. *Pattern Recognition 38*, 6 (2005), 887--901.

[126] Lowe, D. Object Recognition from Local Scale-Invariant Features. In *Proceedings of the 7th IEEE International Conference on Computer Vision (ICCV)* (1999).

[127] Luo, J., and Savakis, A. Indoor vs outdoor classification of consumer photographs usinglow-level and semantic features. In *International Conference on Image Processing (ICIP)* (2001).

[128] Luo, J., Singhal, A., and Zhu, W. Natural object detection in outdoor scenes based on probabilistic spatial context models. *IEEE International Conference on Multimedia and Expo (ICME)* (2003).

[129] Lynch, R., and Reis, J. Haar transform image coding. In *Proc. National Telecommum. Conf., Dallas, TX* (1976), pp. 44--3.

[130] Makris, A., and Mailis, T. Νέες τεχνικές επιβλεπόμενης μάθησης για δομημένα νευρο-ασαφή δίκτυα. Diploma Thesis, School of Electrical and Computer Engineering, National Technical University of Athens.

[131] Manjunath, B., Ohm, J., Vasudevan, V., Yamada, A., et al. Color and texture descriptors. *IEEE Transactions on circuits and systems for video technology 11*, 6 (2001), 703--715.

[132] Marques, O., and Barman, N. Semi-automatic semantic annotation of images using machine learning techniques. *Lecture Notes in Computer Science* (2003).

[133] Matas, J., Chum, O., Urban, M., and Pajdla, T. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing 22*, 10 (2004), 761--767.

[134] McGuinness, D., Van Harmelen, F., et al. OWL web ontology language overview. *W3C recommendation 10* (2004).

[135] Meghini, C., Sebastiani, F., and Straccia, U. Reasoning about the Form and Content of Multimedia Objects (Extended Abstract). In *AAAI Spring Symposium on Intelligent integration and Use of Text, Image, Video and Audio* (1997).

[136] Meine, A., Hermes, T., Ioannidis, G., Fathi, R., and Herzog, O. Automatic shot boundary detection and classification of indoor and outdoor scenes. In *Information Technology: The 11th Text Retrieval Conference* (2003).

[137] Mikolajczyk, K., and Schmid, C. An Affine Invariant Interest Point Detector. In *European Conference on Computer Vision (ECCV)* (2002).

[138] Mikolajczyk, K., and Schmid, C. A Performance Evaluation of Local Descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence 27*, 10 (2005), 1615--1630.

[139] Mills, M., Cohen, J., and Wong, Y. A magnifier tool for video data. In *SIGCHI conference on Human factors in computing systems* (1992), ACM.

[140] Mitchell, M. *An introduction to genetic algorithms*. The MIT press, 1998.

[141] Mokhtarian, F., Abbasi, S., and Kittler, J. Efficient and robust retrieval by shape content through curvature scale space. *Image Databases and Multi-Media Search* (1997).

[142] Moldovan, D., Clark, C., and Harabagiu, S. Temporal context representation and reasoning. In *International Joint Conference on Artificial Intelligence* (2005).

[143] Mukherjea, S., Hirata, K., and Hara, Y. AMORE: A World Wide Web image retrieval engine. *World Wide Web 2*, 3 (1999), 115--132.

[144] Mulhem, P., and Lim, J. Home photo retrieval: Time matters. *Lecture Notes in Computer Science* (2003).

[145] Murphy, K., Torralba, A., and Freeman, W. Using the forest to see the trees: a graphical model relating features, objects and scenes. *Advances in neural information processing systems 16* (2003).

[146] Mylonas, P., Athanasiadis, T., and Avrithis, Y. Image analysis using domain knowledge and visual context. In *13th International Conference on Systems, Signals and Image Processing (IWSSIP)* (2006).

[147] Mylonas, P., and Avrithis, Y. Context modeling for multimedia analysis. In *International and Interdisciplinary Conference on Modeling and Using Context* (2005).

[148] Naphade, M., Kennedy, L., Kender, J., Chang, S., Smith, J., Over, P., and Hauptmann, A. A Light Scale Concept Ontology for Multimedia Understanding for TRECVID 2005 (LSCOM-Lite). *Research report, IBM* (2005).

[149] Naphade, M., and Smith, J. A hybrid framework for detecting the semantics of concepts and context. *Lecture notes in computer science* (2003).

[150] Naphade, M. R., Kozintsev, I. V., and Huang, T. S. A factor graph framework for semantic video indexing. *IEEE Transactions on Circuits and Systems for Video Technology 12*, 1 (2002), 40--52.

[151] Nastar, C., Mitschke, M., Meilhac, C., and Boujemaa, N. Surfimage: a flexible content-based image retrieval system. In *ACM International Conference on Multimedia* (1998).

[152] Niblack, C., Barber, R., Equitz, W., Flickner, M., Glasman, E., Petkovic, D., Yanker, P., Faloutsos, C., and Taubin, G. QBIC project: querying images by content, using color, texture, and shape. In *Proceedings of SPIE* (1993), vol. 173.

[153] Niebles, J., and Fei-Fei, L. A hierarchical model of shape and appearance for human action classification. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2007).

[154] Nister, D., and Stewenius, H. Scalable Recognition with a Vocabulary Tree. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2006).

[155] O'Hare, N., Gurrin, C., Lee, H., Murphy, N., Smeaton, A., and Jones, G. My digital photos: where and when? In *Proceedings of the 13th annual ACM international conference on Multimedia* (2005).

[156] Ojala, T., Aittola, M., and Matinmikko, E. Empirical Evaluation of MPEG-7 XM Color Descriptors in Content-Based Retrieval of Semantic Image Categories. In *International Conference on Pattern Recognition (ICPR)* (2002).

[157] Ojala, T., Mäenpää, T., Viertola, J., Kyllönen, J., and Pietikäinen, M. Empirical Evaluation of MPEG-7 Texture Descriptors with A Large-Scale Experiment. In *2nd International Workshop on Texture Analysis and Synthesis, Copenhagen, Denmark* (2002).

[158] Oliva, A., and Schyns, P. Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology 34* (1997), 72--107.

[159] Oliva, A., and Torralba, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision 42*, 3 (2001), 145--175.

[160] Oliva, A., and Torralba, A. Building the gist of a scene: The role of global image features in recognition. *Visual Perception* (2006).

[161] Opelt, A., Pinz, A., and Zisserman, A. Incremental learning of object detectors using a visual shape alphabet. In *Conference on Computer Vision and Pattern Recognition(CVPR)* (2006).

[162] Paek, S., Sable, C., Hatzivassiloglou, V., Jaimes, A., Schiman, B., Chang, S., and McKeown, K. Integration of visual and text-based approaches for the content labeling and classification of photographs. In *ACM SIGIR'99 Workshop on Multimedia Indexing and Retrieval* (1999).

[163] Paletta, L., Prantl, M., and Pinz, A. Learning temporal context in active object recognition using bayesian analysis. In *International Conference on Pattern Recognition* (2000).

[164] Panagiotakis, C., Doulamis, A., and Tziritas, G. Equivalent key frames selection based on iso-content principles. *IEEE Transactions on Circuits and Systems for Video Technology* (2008).

[165] Papadias, D., and Theodoridis, Y. Spatial relations, minimum bounding rectangles, and spatial data structures. *International Journal of Geographical Information Science 11* (1997), 111--138.

[166] Pauty, J., Couderc, P., and Banâtre, M. Using context to navigate through a photo collection. In *ACM International conference on Human computer interaction with mobile devices & services* (2005).

[167] Payne, A., and Singh, S. Indoor vs. outdoor scene classification in digital photographs. *Pattern Recognition 38*, 10 (2005), 1533--1545.

[168] Petersohn, C. Fraunhofer hhi at trecvid 2004: Shot boundary detection system. In *TREC Video Retrieval Evaluation Online Proceedings* (2004).

[169] Petkovic, M., and Jonker, W. Content-based video retrieval by integrating spatio-temporal and stochastic recognition of events. In *IEEE Workshop on Detection and Recognition of Events in Video* (2001).

[170] Pfeiffer, S., Lienhart, R., Fischer, S., and Effelsberg, W. Abstracting digital movies automatically. *Journal of Visual Communication and Image Representation 7*, 4 (1996), 345--353.

[171] Philbin, J., Chum, O., Isard, M., Sivic, J., and Zisserman, A. Object retrieval with large vocabularies and fast spatial matching. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2007).

[172] Philbin, J., Chum, O., Isard, M., Sivic, J., and Zisserman, A. Lost in quantization: Improving particular object retrieval in large scale image databases. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2008).

[173] Pigeau, A., and Gelgon, M. Organizing a personal image collection with statistical model-based ICL clustering on spatio-temporal camera phone meta-data. *Journal of Visual Communication and Image Representation 15*, 3 (2004), 425--445.

[174] Qiu, G., Feng, X., and Fang, J. Compressing histogram representations for automatic colour photo categorization. *Pattern Recognition 37*, 11 (2004), 2177--2193.

[175] Rahmani, R., Goldman, S., Zhang, H., and Fritts, J. Accio: A localized content-based image retrieval system. Tech. rep., Technical report, Washington University in St Louis, 2005.

[176] Ratakonda, K. Method for hierarchical summarization and browsing of digital video, Sept. 21 1999. US Patent 5,956,026.

[177] Rui, Y., Huang, T., and Chang, S. Image Retrieval: Past, Present, And Future. *Journal of Visual Communication and Image Representation* (1997), 1--23.

[178] Russell, B., Torralba, A., Murphy, K., and Freeman, W. LabelMe: a Database and Web-based Tool for Image Annotation. *International Journal of Computer Vision 77*, 1 (2008), 157--173.

[179] Savarese, S., Winn, J., and Criminisi, A. Discriminative object class models of appearance and shape by correlatons. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2006).

[180] Schmid, C., and Mohr, R. Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence 19*, 5 (1997), 530--535.

[181] Sclaroff, S., Taycher, L., and La Cascia, M. Imagerover: A content-based image browser for the world wide web. In *IEEE Workshop on Content-based Access of Image and Video Libraries* (1997).

[182] Serrano, N., Savakis, A., and Luo, J. A Computationally Efficient Approach to Indoor/Outdoor Scene Classification. In *International Conference on Pattern Recognition (ICPR)* (2002).

[183] Serrano, N., Savakis, A., and Luo, J. Improved scene classification using efficient low-level features and semantic cues. *Pattern Recognition 37*, 9 (2004), 1773--1784.

[184] Shahraray, B., and Gibbon, D. Automated authoring of hypermedia documents of video programs. In *ACM international conference on Multimedia* (1995).

[185] Sikora, T. The MPEG-4 video standard verification model. *IEEE Transactions on Circuits and Systems for Video Technology 7*, 1 (1997), 19--31.

[186] Simou, N., Tzouvaras, V., Avrithis, Y., Stamou, G., and Kollias, S. A visual descriptor ontology for multimedia reasoning. In *Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)* (2005).

[187] Singhal, A., Luo, J., and Zhu, W. Probabilistic spatial context models for scene content understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2003).

[188] Sinha, P., and Jain, R. Classification and annotation of digital photos using optical context data. In *International Conference on Content-based image and video retrieval (CIVR)* (2008).

[189] Sivic, J., Russell, B., Efros, A., Zisserman, A., and Freeman, W. Discovering Objects and their Location in Images. In *10th IEEE International Conference on Computer Vision (ICCV)* (2005).

[190] Sivic, J., and Zisserman, A. Video Google: A text retrieval approach to object matching in videos. In *International Conference on Computer Vision (ICCV)* (2003).

[191] Skiadopoulos, S., and Koubarakis, M. Composing cardinal direction relations. *Artificial Intelligence 152* (2004), 143--171.

[192] Smeaton, A. F., Over, P., and Kraaij, W. Evaluation campaigns and trecvid. In *MIR '06: 8th ACM International Workshop on Multimedia Information Retrieval* (2006).

[193] Smeulders, A., Worring, M., Santini, S., Gupta, A., and Jain, R. Content-based image retrieval at the end of the early years. *IEEE Transactions on pattern analysis and machine intelligence 22*, 12 (2000), 1349--1380.

[194] Smith, J. MARVEL: Multimedia analysis and retrieval system. *Whitepaper, Intelligent Information Management Dept. IBM TJ Watson Research Center* (2005).

[195] Smith, J., and Chang, S. Searching for images and videos on the world-wide web. *IEEE MultiMedia* (1997).

[196] Smith, J., Naphade, M., and Natsev, A. Multimedia semantic indexing using model vectors. In *International Conference on Multimedia and Expo (ICME)* (2003).

[197] Snoek, C., Worring, M., and Smeulders, A. Early versus late fusion in semantic video analysis. In *13th annual ACM International Conference on Multimedia* (2005).

[198] Souvannavong, F., Merialdo, B., and Huet, B. Region-based video content indexing and retrieval. In *International Workshop on Content-Based Multimedia Indexing (CBMI)* (2005).

[199] Sprague, N., and Luo, J. Clothed people detection in still images. In *International Conference on Pattern Recognition* (2002), vol. 16, pp. 585--589.

[200] Spyrou, E. Μηχανές Διανυσμάτων Στήριξης με χρήση Πυρήνα Ασαφών Βασικών Συναρτήσεων. Diploma Thesis, School of Electrical and Computer Engineering, National Technical University of Athens.

[201] Srihari, R., and Zhang, Z. Show&Tell: A semi-automated image annotation system. *IEEE transactions on multimedia 7*, 3 (2000), 61--71.

[202] S.Staab, and R.Studer. *Handbook on Ontologies*. Springer Verlag, 2004.

[203] Sudderth, E., Torralba, A., Freeman, W., and Willsky, A. Learning hierarchical models of scenes, objects, and parts. In *IEEE International Conference on Computer Vision (ICCV)* (2005).

[204] Sun, X., and Kankanhalli, M. Video summarization using R-sequences. *Real-time imaging 6*, 6 (2000), 449--459.

[205] Sundaram, H., and Chang, S. Video Analysis and Summarization at Structural and Semantic Levels. *Multimedia information retrieval and management: Technological fundamentals and applications* (2003).

[206] Szummer, M., and Picard, R. Indoor-outdoor image classification. In *IEEE International Workshop on Content-Based Access of Image and Video Database* (1998).

[207] Tagare, H., Jaffe, C., and Duncan, J. Medical image databases: A content-based retrieval approach. *Journal of the American Medical Informatics Association 4*, 3 (1997).

[208] Taniguchi, Y., Akutsu, A., and Tonomura, Y. PanoramaExcerpts: extracting and packing panoramas for video browsing. In *Proceedings of the fifth ACM international conference on Multimedia* (1997), ACM New York, NY, USA, pp. 427--436.

[209] Taniguchi, Y., Akutsu, A., Tonomura, Y., and Hamada, H. An intuitive and efficient access interface to real-time incoming video based on automatic indexing. In *ACM international conference on Multimedia* (1995).

[210] Tansley, R., Bird, C., Hall, W., Lewis, P., and Weal, M. Automating the linking of content and concept. In *ACM international conference on Multimedia* (2000).

[211] Tolias, G. *VDE: Visual Descriptor Extraction*, 2008. Software available at `http://image.ntua.gr/smag/tools/vde`.

[212] Torralba, A., Fergus, R., and Freeman, W. 80 million tiny images: a large dataset for non-parametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence 30*, 11 (2008), 1958--1970.

[213] Torralba, A., Oliva, A., Castelhano, M., and Henderson, J. Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review 113*, 4 (2006), 766--786.

[214] Traherne, M., and Singh, S. An Integrated Approach to Automatic Indoor Outdoor Scene Classification in Digital Images. *Lecture notes in computer science* (2004).

[215] Truong, B., and Venkatesh, S. Video abstraction: A systematic review and classification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP) 3*, 1 (2007), 3.

[216] Tuffield, M., Harris, S., Dupplaw, D., Chakravarthy, A., Brewster, C., Gibbins, N., O'Hara, K., Ciravegna, F., Sleeman, D., Shadbolt, N., et al. Image annotation with photocopain. In *Semantic Web Annotation of Multimedia (SWAMM) Workshop at the World Wide Web Conference* (2006).

[217] Uchihashi, S., Foote, J., Girgensohn, A., and Boreczky, J. Video Manga: generating semantically meaningful video summaries. In *ACM international conference on Multimedia* (1999).

[218] Ullman, S., and Vidal-Naquet, M. Visual Features of Intermediate Complexity and their use in Classification. *Nature Neuroscience 5*, 7 (2002), 682--687.

[219] Vailaya, A., Jain, A., and Zhang, H. On Image Classification: City Images vs. Landscapes. *Pattern Recognition 31*, 12 (1998), 1921--1935.

[220] van Gemert, J., Geusebroek, J., Veenman, C., and Smeulders, A. Kernel codebooks for scene categorization. In *European Conference on Computer Vision (ECCV)* (2008), Springer.

[221] Vapnik, V. *The nature of statistical learning theory.* Springer Verlag, 2000.

[222] Veltkamp, R., and Tanase, M. A survey of content-based image retrieval systems. *Content-based image and video retrieval.*

[223] Viana, W., Filho, J., Gensel, J., Oliver, M., and Martin, H. PhotoMap-Automatic Spatiotemporal Annotation for Mobile Photos. *Lecture Notes in Computer Science 4857* (2007).

[224] Vogel, J., and Schiele, B. A Semantic Typicality Measure for Natural Scene Categorization. In *Proceedings of DAGM Pattern Recognition Symposium* (2004).

[225] Vogel, J., and Schiele, B. Natural scene retrieval based on a semantic modeling step. In *Conference on Image and Video Retrieval (CIVR)* (2004).

[226] Wallace, M., Mylonas, P., Akrivas, G., Avrithis, Y., and Kollias, S. Automatic thematic categorization of multimedia documents using ontological information and fuzzy algebra. *Studies in Fuzziness and Soft Computing 204* (2006).

[227] Wang, J., Li, J., and Wiederhold, G. SIMPLIcity: Semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on pattern analysis and machine intelligence* (2001).

[228] Wang, L., Khan, L., and Breen, C. Object boundary detection for ontology-based image classification. In *Multimedia Data Mining - Mining Integrated Media and Complex Data (MDM/KDD'02)* (2002).

[229] Wang, L., and Manjunath, B. A semantic representation for image retrieval. In *IEEE International Conference on Image Processing (ICIP)* (2003).

[230] Wilkins, P., Adamek, T., Byrne, D., Jones, G., Lee, H., Keenan, G., McGuinness, K., O'Connor, N., Smeaton, A., Amin, A., et al. K-Space at TRECVid 2007. In *Proceedings of TRECVID* (2007).

[231] Wolf, W. Key frame selection by motion analysis. In *IEEE International Conference on Acoustics Speech and Signal Processing* (1996).

[232] Yamada, A., Pickering, M., Jeannin, S., Cieplinski, L., Ohm, J., and Kim, M. MPEG-7 visual part of eXperimentation Model Version 9.0 ISO/IEC JTC1/SC29/WG11/N3914. *International Organisation for Standardisation ISO* (2001), 1--83.

[233] Yan, R., Chen, M., and Hauptmann, A. Mining relationship between video concepts using probabilistic graphical model. In *IEEE International Conference on Multimedia and Expo (ICME)* (2006).

[234] Yang, M., Qiu, G., Huang, J., and Elliman, D. Near-duplicate image recognition and content-based image retrieval using adaptive hierarchical geometric centroids. In *International Conference on Pattern Recognition (ICPR)* (2006).

[235] Yeung, M., and Liu, B. Efficient matching and clustering of video shots. In *IEEE International Conference on Image Processing* (1995), vol. 1, pp. 338--341.

[236] Yilmaz, E., and Aslam, J. Estimating average precision with incomplete and imperfect judgments. In *ACM international conference on Information and knowledge management* (2006).

[237] Yu, X., Wang, L., Tian, Q., and Xue, P. Multi-level video representation with application to keyframe extraction. In *IEEE International Multimedia Modelling Conference* (2004).

[238] Yuan, J., Li, J., and Zhang, B. Exploiting spatial context constraints for automatic image region annotation. In *ACM International conference on Multimedia* (2007).

[239] Zadeh, L. Fuzzy sets. *Information and control 8*, 3 (1965), 338--353.

[240] Zhang, H., Wu, J., Zhong, D., and Smoliar, S. An integrated system for content-based video retrieval and browsing. *Pattern recognition 30*, 4 (1997), 643--658.

[241] Zhao, W., Jiang, Y., and Ngo, C. Keyframe retrieval by keypoints: Can point-to-point matching help? *Lecture Notes in Computer Science* (2006).

[242] Zhu, L., and Zhang, A. Theory of keyblock-based image retrieval. *ACM Transactions on Information Systems (TOIS) 20*, 2 (2002), 224--257.

[243] Zhuang, Y., Rui, Y., Huang, T., and Mehrotra, S. Adaptive key frame extraction using unsupervised clustering. In *International Conference on Image Processing (ICIP)* (1998).

Κατάλογος δημοσιεύσεων του συγγραφέα

Άρθρα σε Περιοδικά

- E. Spyrou, G. Tolias, Ph. Mylonas and Y. Avrithis
Concept Detection and Keyframe Extraction Using a Visual Thesaurus
Multimedia Tools and Applications, Springer, Volume 41, Issue 3, pp. 337-373, February 2009

- Ph. Mylonas, E. Spyrou, Y. Avrithis and S. Kollias *Using Visual Context and Region Semantics for High-Level Concept Detection*
IEEE Transactions on Multimedia, Volume 11, Issue 2, pp. 229-243, February 2009

Κεφάλαια σε Βιβλία

- E. Spyrou and Y. Avrithis
Detection of High-Level Concepts in Multimedia
Encyclopedia of Multimedia, 2nd Edition, Springer 2008

- E.Spyrou and Y.Avrithis
High-Level Concept Detection in Video Using a Region Thesaurus
Emerging Artificial Intelligence Applications in Computer Engineering, Series in Frontiers in Artificial Intelligence and Applications, IOS Press, Amsterdam, 2007

- S. Dasiopoulou, E. Spyrou, Y. Avrithis, Y. Kompatsiaris and M.G. Strintzis
Semantic Processing of Color Images
Rastislav Lukac and Konstantinos N. Plataniotis (editors) - Color Image Processing: Emerging Applications - CRC Press - 2006

Άρθρα σε Συνέδρια[1]

- Y. Kalantidis, G. Tolias, E. Spyrou, Ph. Mylonas and Y. Avrithis
  Visual Image Retrieval and Localization
  7th International Workshop on Content-Based Multimedia Indexing, Greece. 2009

- Y. Kalantidis, G. Tolias, E. Spyrou, Ph. Mylonas, Y. Avrithis and S. Kollias
  Visual Image Retrieval and Localization
  3ο Πανελλήνιο Συνέδριο Φοιτητών Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Θεσσαλονίκη, 2009

- E. Spyrou, G. Tolias and Y. Avrithis
  Large Scale Concept Detection in Video Using a Region Thesaurus
  The 15th International MultiMedia Modeling Conference (MMM2009), Sophia Antipolis, France, 2009

- E. Spyrou, G. Tolias and Ph. Mylonas
  A relation-based contextual approach for efficient multimedia analysis
  3rd International Workshop on Semantic Media Adaptation and Personalization (SMAP 2008)

- E. Spyrou, G. Tolias, Ph. Mylonas and Y. Avrithis
  A Semantic Multimedia Analysis Approach Utilizing a Region Thesaurus and LSA
  9th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2008)

- E. Spyrou, Ph. Mylonas and Y. Avrithis
  A Visual Context Ontology for Multimedia High-Level Concept Detection
  5th International Workshop in Modeling and Reasoning in Context (MRC), Held at HCP 08, Delft, The Netherlands, 9-12 June 2008

- E. Spyrou, Ph. Mylonas and Y. Avrithis
  Using Region Semantics And Visual Context For Scene Classification
  1st ICIP Workshop on Multimedia Information Retrieval: New Trends and Challenges October 12, 2008, San Diego, California, U.S.A.

- E. Spyrou and Y. Avrithis
  Keyframe Extraction using Local Visual Semantics in the form of a Region Thesaurus
  2nd International Workshop on Semantic Media Adaptation and Personalization (SMAP 2007), London, United Kingdom, 17-18 December 2007

- E. Spyrou, Ph. Mylonas and Y. Avrithis
  Semantic Multimedia Analysis based on Region Types and Visual Context

[1] Ο συγγραφέας έχει συμμετάσχει ενεργά στη διαδικασίας αξιολόγησης TRECVID, για τις χρονιές 2006-2008, στα πλαίσια ευρύτερων ερευνητικών ομάδων. Έχουν, έτσι προκύψει άλλες 7 δημοσιεύσεις στο Workshop του TRECVID, οι οποίες παραλείπονται από τον κατάλογο αυτό, καθώς δεν έχουν περάσει από διαδικασία κρίσης, καθώς η συμμετοχή στο TRECVID είναι αναγκαία και αρκετή για να γίνουν αποδεκτές.

4th IFIP Conference on Artificial Intelligence Applications Innovations (AIAI), Athens, Greece, 19-21 September 2007

- E. Spyrou, Y. Avrithis
  A Region Thesaurus Approach for High-Level Concept Detection in the Natural Disaster Domain
  2nd international conference on Semantics And digital Media Technologies (SAMT), Italy, Genova, 2007

- J. Molina, E. Spyrou, N. Sofou and J. M. Martínez
  On the selection of MPEG-7 Visual Descriptors and their Level of Detail for Nature Disaster Video Sequences Classification
  2nd international conference on Semantics And digital Media Technologies (SAMT), Italy, Genova, 2007

- Ph. Mylonas, E. Spyrou and Y. Avrithis
  High-Level Concept Detection based on Mid-level Semantic Information and Contextual Adaptation
  2nd International Workshop on Semantic Media Adaptation and Personalization (SMAP 2007), London, United Kingdom, 17-18 December 2007

- Ph. Mylonas, E. Spyrou, and Y. Avrithis
  Enriching a context ontology with mid-level features for semantic multimedia analysis
  1st Workshop on Multimedia Annotation and Retrieval enabled by Shared Ontologies, co-located with SAMT 2007

- E. Spyrou, G. Koumoulos, Y. Avrithis and S. Kollias
  Using Local Region Semantics for Concept Detection in Video 1st International Conference on Semantics And digital Media Technology (SAMT 2006), Athens, Greece

- E.Spyrou, G.Stamou, Y.Avrithis and S.Kollias
  Fuzzy Support Vector Machines for Image Classification fusing MPEG-7 Visual Descriptors
  2nd European Workshop on the Integration of Knowledge,Semantic, and Digital Media Techniques, EWIMT05, London, UK, November 2005

- E. Spyrou, H. Le Borgne, T. Mailis, E. Cooke, Y. Avrithis, and N. O'Connor
  Fusing MPEG-7 visual descriptors for image classification Proc. of International Conference on Artificial Neural Networks (ICANN '05), Warsaw, Poland, September 11-15, 2005

- N.Simou, C.Saathoff, S.Dasiopoulou, E.Spyrou, N.Voisine, V.Tzouvaras, I.Kompatsiaris, Y.Avrithis and S.Staab
  An Ontology Infrastructure for Multimedia Reasoning
  International Workshop VLBV05, Sardinia, Italy, 15-16 September 2005

- N. Voisine, S. Dasiopoulou, V. Mezaris, E. Spyrou, Th. Athanasiadis, I. Kompatsiaris, Y. Avrithis, M. G. Strintzis
  Knowledge-Assisted Video Analysis Using A Genetic Algorithm 6th International

Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2005), Montreux, Switzerland, April 13-15, 2005

□

Βιογραφικό Σημείωμα

Ο υποψήφιος διδάκτορας Ευάγγελος Σπύρου γεννήθηκε στην Αθήνα στις 18 Ιουλίου 1979. Το 1998 έγινε δεκτός στη Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του ΕΜΠ. Ακολούθησε την κατεύθυνση Τηλεπικοινωνιών. Εκπόνησε τη διπλωματική του εργασία με τίτλο "Μηχανές Διανυσμάτων Στήριξης με χρήση Πυρήνα Ασαφών Βασικών Συναρτήσεων" υπό την επίβλεψη του Καθηγητή κ. Στ. Κόλλια. Αποφοίτησε το 2003 με βαθμό 7.62. Τον Οκτώβρη του 2004 έγινε δεκτός από τη Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του ΕΜΠ για εκπόνηση διδακτορικής διατριβής στον τομέα της σημασιολογικής ανάλυσης εικόνας και βίντεο, υπό την επίβλεψη του Καθηγητή κ. Στέφανου Κόλλια.

Ως υποψήφιος διδάκτορας ο Ευάγγελος Σπύρου εργάστηκε ως ερευνητής σε Ελληνικά και Ευρωπαϊκά ερευνητικά και αναπτυξιακά έργα και παρείχε επικουρικό διδακτικό έργο σε 4 μαθήματα του τομέα Τεχνολογίας Πληροφορικής και Υπολογιστών, υπό την επίβλεψη του Καθηγητή κ. Στέφανου Κόλλια.

Συμμετοχή σε ερευνητικά-αναπτυξιακά προγράμματα

- WeKnowIt - Emerging, Collective Intelligence for personal, organisational and social use
- Ontomedia - Σημασιολογική Ανάλυση Πολυμεσικού Υλικού με Χρήση Τεχνολογιών Γνώσης
- MESH -Multimedia sEmantic Syndication for enHanced news services
- X-Media - Large Scale Knowledge Sharing and Reuse Across Media
- K-Space-Knowledge Space of Semantic inference for automatic annotation and retrieval of multimedia content
- aceMedia-Integrating Knowledge, Semantics and Content for User-Centred Intelligent Media Services.

Συμμετοχή σε έργα αξιολόγησης

- TRECVID - TREC Video Retrieval Evaluation, συμμετοχή στη δοκιμασία High-level feature extraction [2006-2008]

Επικουρικό Διδακτικό Έργο

- Νευρωνικά Δίκτυα και Ευφυή Υπολογιστικά συστήματα [2004-2007]
- Τεχνολογία και Ανάλυση Εικόνας και Βίντεο [2004-2006,2008]
- Γραφικά με Υπολογιστές [2004]

□