

Εθνικό Μετσόβιο Πολυτεχνείο

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΥΠΟΛΟΓΙΣΤΩΝ

Τεχνικές ομαδοποίησης και κοντινότερου γείτονα για οπτική αναζήτηση εικόνων

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

του

ΙΩΑΝΝΗ Δ. ΚΑΛΑΝΤΙΔΗ

Διπλωματούχου Ηλεκτρολόγου Μηχανικών &
Μηχανικού Υπολογιστών Ε.Μ.Π. (2009)

Αθήνα, Νοέμβριος 2014



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
& ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΥΠΟΛΟΓΙΣΤΩΝ

Τεχνικές ομαδοποίησης και κοντινότερου γείτονα για οπτική αναζήτηση εικόνων

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

του

ΙΩΑΝΝΗ Δ. ΚΑΛΑΝΤΙΔΗ

Διπλωματούχου Ηλεκτρολόγου Μηχανικών &
Μηχανικού Υπολογιστών Ε.Μ.Π. (2009)

Συμβουλευτική Επιτροπή: Στέφανος Κόλλιας
Ανδρέας Σταφυλοπάτης
Γεώργιος Στάμου

Εγκρίθηκε από την επταμελή επιτροπή την 7η Νοεμβρίου 2014.

. . .	. . .	. . .
Στέφανος Κόλλιας	Ανδρέας Σταφυλοπάτης	Γεώργιος Στάμου
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Επ. Καθηγητής Ε.Μ.Π.
. . .	. . .	. . .
Πέτρος Μαραγκός	Ιωάννης Εμίρης	Παναγιώτης Τσανάκας
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Κ.Π.Α.	Καθηγητής Ε.Μ.Π.
. . .		
Ιωάννης Βασιλείου		
Καθηγητής Ε.Μ.Π.		

Αθήνα, 7 Νοεμβρίου 2014

. . .

ΙΩΑΝΝΗΣ Δ. ΚΑΛΑΝΤΙΔΗΣ
Διδάκτωρ Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

© 2014 - All rights reserved
Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

Περιεχόμενα

1 Εισαγωγή – Βασικές τεχνικές οπτικής αναζήτησης 3

1.1 Εισαγωγή 3
1.2 Σύνοψη συνεισφοράς 5
1.3 Βασικές τεχνικές αναζήτησης εικόνων 7
1.3.1 Εξαγωγή οπτικών χαρακτηριστικών 9
1.3.2 Δεικτοδότηση και στάδιο φιλτραρίσματος 12
1.3.3 Γεωμετρικό ταίριασμα 13

2 Ομαδοποίηση για κατασκευή οπτικών λεξικών 15

2.1 Εισαγωγή 15
2.2 Σχετική βιβλιογραφία 18
2.3 Επεκτατικά μίγματα κανονικών κατανομών 20
2.3.1 Εκμάθηση παραμέτρων 21
2.3.2 Διαγραφή επικαλυπτώμενων κέντρων 23
2.3.3 Επέκταση συνιστωσών 27
2.3.4 Αρχικοποίηση και τερματισμός 28
2.4 Προσεγγιστικά μείγματα κανονικών κατανομών 29
2.5 Πειράματα και συγκρίσεις 31
2.5.1 Πρωτόκολλο πειραμάτων 31
2.5.2 Βελτιστοποίηση παραμέτρων 33
2.5.3 Συγκρίσεις 34

3 Τοπικά Βελτιστοποιημένος Παραγοντικός Κβαντισμός 37

3.1 Εισαγωγή 37
3.2 Σχετική βιβλιογραφία και συνεισφορά 40
3.3 Υπόβαθρο 41
3.3.1 Διανυσματικός Κβαντισμός 41
3.3.2 Παραγοντικός Κβαντισμός 42
3.3.3 Βελτιστοποιημένος Παραγοντικός Κβαντισμός 43
3.3.4 Εξαντλητική αναζήτηση 43
3.3.5 Δεικτοδότηση 44
3.3.6 Ανακατάταξη δευτέρου επιπέδου 44

3.3.7 Πολυ-δεικτοδότηση – multi-index 44
3.4 Τοπικά βελτιστοποιημένος παραγοντικός κβαντισμός 45
3.4.1 Αναζήτηση σε ανεστραμμένο αρχείο 45
3.4.2 Τοπική Βελτιστοποίηση 46
3.4.3 Παράδειγμα 48
3.4.4 Αναζήτηση με multi-index 49
3.4.5 Αναδιάταξη δευτέρου επιπέδου 51
3.5 Πειράματα 52
3.5.1 Πρωτόκολλο πειραμάτων 52
3.5.2 Αποτελέσματα στα σύνολα MNIST, SIFT1M, GIST1M 54
3.5.3 Αποτελέσματα στο σύνολο SIFT1B 58
3.5.4 Ανάλυση της ταχύτητας αναζήτησης και της μνήμης 59

4 Γεωγραφική και οπτική ομαδοποίηση 61
4.1 Εισαγωγή 61
4.2 Σχετική βιβλιογραφία 62
4.2.1 Αναγνώριση Τοποθεσίας 62
4.2.2 Αναγνώριση Ορόσημων 63
4.2.3 Ανακατασκευή τρισδιάστατων σκηνών 64
4.2.4 Δομές Δεικτοδότησης για βάσης μεγάλης κλίμακας 65
4.3 Ομαδοποίηση όψεων 66
4.3.1 Kernel Vector Quantization 67
4.3.2 Γεωγραφική ομαδοποίηση 71
4.3.3 Οπτική ομαδοποίηση 72
4.3.4 Συζήτηση 74
4.4 Χάρτες Σκηνής 75
4.4.1 Κατασκευή χαρτών σκηνής 76
4.4.2 Συζήτηση 78
4.5 Πειράματα 79
4.5.1 Η συλλογή εικόνων European Cities 1M 79
4.5.2 Πρωτόκολλο πειραμάτων 81
4.5.3 Αξιολόγηση της οπτικής αναζήτησης 83

5 Αναγνώριση τοποθεσίας και σκηνής 89
5.1 Εισαγωγή 89
5.2 Οπτικό ταίριασμα και δεικτοδότηση 90
5.2.1 Βασικές διαδικασίες 90
5.2.2 Ευθυγράμμιση όψεων 92
5.2.3 Δεικτοδότηση χαρτών σκηνής 92
5.2.4 Συζήτηση 93
5.3 Αναγνώριση τοποθεσίας και ορόσημων 93

5.3.1 Αναγνώριση τοποθεσίας 94
5.3.2 Συνήθη tags χρηστών 95
5.3.3 Αναγνώριση ορόσημων 96
5.3.4 Συζήτηση 96
5.4 Η εφαρμογή VIRaL 97
5.4.1 Αναγνώριση τοποθεσίας μέ το VIRaL 98
5.4.2 Οι εφαρμογές Viral Explore και Routes 99
5.5 Πειράματα 104
5.5.1 Αξιολόγηση της εκτίμησης τοποθεσίας 104
5.5.2 Αξιολόγηση της αναγνώρισης ορόσημων και σημείων ενδιαφέροντος 107

6 Συμπληρωματικές εργασίες **111**
6.1 Χάρτες Χαρακτηριστικών 111
6.2 Ανίχνευση λογότυπων 113
6.3 Αναγνώριση, κατάτμηση και μεγάλης κλίμακας αναζήτηση προιόντων ρουχισμού 117

7 Συμπεράσματα **119**

Βιβλιογραφία **123**

Κατάλογος Σχημάτων

1.1 Ο όγκος των φωτογραφιών που ανεβαίνουν και μοιράζονται σε μερικές από τις πλέον δημοφιλείς ιστοσελίδες κοινωνικής δικτύωσης τα τελευταία χρόνια. Οι αριθμοί αναφέρονται σε *εκατομμύρια* φωτογραφίες ανά ημέρα. (Πηγή: KPCB) 4

1.2 Εικόνες με διαφορετική πυκνότητα οπτικού περιεχομένου. Η πάνω φωτογραφία, ο οποία τραβήχτηκε στην έρημο της Αριζόνα, μπορεί να περιγραφεί επαρκώς οπτικά με κάποιον ολικό (global) περιγραφέα υφής και χρώματος. Η κάτω εικόνα όμως, τραβηγμένη στην πολύχρωμη Beale Street του Memphis, απαιτεί πιο σύνθετες δομές για την περιγραφή του οπτικού της περιεχομένου. 11

2.1 Εκτίμηση του αριθμού, του πληθυσμού, της θέσης και έκτασης των κέντρων *σε τρεις μόνο επαναλήψεις*, για ένα σύνολο 800 δισδιάστατων δεδομένων τα οποία προέρχονται από ένα μείγμα 8 Κανονικών κατανομών. Κόκκινοι κύκλοι: θέση κέντρων, μπλε: δύο τυπικές αποκλίσεις. Τα κέντρα αρχικοποιούνται σε 50 από τα αρχικά δεδομένα με τυχαίο τρόπο, με αρχικό $\sigma = 0.02$. Είναι εμφανής η «επεκτατική» συμπεριφορά των δύο κέντρων στα αριστερά. 17

2.2 «Δειγματοληψία» μιας απλωμένης συνιστώσας $p_1(x) = \pi_1 \mathcal{N}(x|\mu_1, \sigma_1^2)$ μέσω μιας μικρότερης, $p_2(x) = \pi_2 \mathcal{N}(x|\mu_2, \sigma_2^2)$, σε μία διάσταση. Όταν η μικρή συνιστώσα συρρικνώνεται σε ένα σημείο, η $p_2(x)$ καταρρέει στη συνάρτηση Dirac $\delta(x - \mu_2)$, και το εσωτερικό γινόμενο $\langle p_1, p_2 \rangle$ σε $p_1(\mu_2)$. 25

2.3 Επέκταση συνιστωσών για τις επαναλήψεις 2 και 3 για το παράδειγμα του Σχήματος 2.1. Μπλε κύκλοι: δύο τυπικές αποκλίσεις με επέκταση που δίνεται από τη σχέση (2.26) και $\lambda = 0.25$, όπως και στο Σχήμα 2.1. Ματζέντα: χωρίς επέκταση (2.8); διακεκομμένοι πράσινοι: συνεισφορές εσωτερικών και εξωτερικών αθροισμάτων. 28

2.4 Απόδοση (mAP) για εξειδικευμένα λεξικά στη βάση *Barcelona* (αριστερά) και μεγέθη λεξικών C (δεξιά) ως προς τον αριθμό των επαναλήψεων κατά την εκμάθηση, για διαφορετικές τιμές του *παράγοντα επέκτασης* λ και σταθερό $\tau = 0.5$. 33

2.5 Απόδοση (mAP) για εξειδικευμένα λεξικά στη βάση *Barcelona* ως προς το *κατώφλι επικάλυψης* τ για διάφορετικό αριθμό επαναλήψεων (αριστερά) και τα διαφορετικά μεγέθη λεξικών C ως προς τις επαναλήψεις, για διαφορετικές τιμές του τ (αριστερά), διατηρώντας σταθερό το $\lambda = 0.2$. 34

2.6 (Αριστερά) Απόδοση (mAP) για εξειδικευμένα λεξικά στη βάση *Barcelona* ως προς τον χρόνο εκπαίδευσης, μετρημένο ως αριθμό *διανυσματικών πράξεων* (vop) ανα διάνυσμα εκπαίδευσης, για τις τεχνικές AGM, AKM και RAKM για διαφορετικά επίπεδα ακρίβειας, μετρημένα σε αριθμό *checks* της βιβλιοθήκης FLANN. Κάθε καμπύλη αποτελείται από 5 μετρήσεις, οι οποίες αντιστοιχούν, από τα αριστερά προς τα δεξιά, στις εμαναλήψεις 5, 10, 20, 30 και 40. (Δεξιά) Απόδοση (mAP) για τις τεχνικές AGM και RAKM στη βάση *Oxford buildings*, χρησιμοποιώντας *γενικά* λεξικά και με τη βάση να περιέχει μέχρι και 1 εκατομμύριο εικόνες *περίσπασης*, χρησιμοποιώντας επίσης *πολλαπλή ανάθεση* με 1, 3 και 5 κοντινότερους γείτονες για τους περιγραφείς των εικόνων αναζήτησης. 35

3.1 Τέσσερις κβαντιστές με 64 κέντρα (:centroidkm) ο καθένας, εκπαιδευμένοι σε ένα τυχαίο σύνολο δισδιάστατων σημείων (:datakm), τα οποία ακολουθούν μια κατανομή μείγματος με 100 σημεία ανα συνιστώσα. Οι κβαντιστές των σχημάτων (γ) και (δ) εκτός της περιστροφής του χώρου, πραγματοποιούν και αναδιάρθρωση των διαστάσεων του χώρου, κάτι που δεν μπορεί να παρουσιαστεί σε αυτό το απλό παράδειγμα με τα δισδιάστατα σημεία. 38

3.2 Ανάκληση στα πρώτα R δείγματα (recall@R) για τη συλλογή SYNTH1M—το μέτρο recall@R ορίζεται στην υποενότητα 3.5.1. Για όλες τις μεθόδους χρησιμοποιούμε τις τιμές $K = 1024$ και $w = 8$. Επίσης για όλους τους παραγοντικούς κβαντιστές τις τιμές $m = 8$ και $k = 256$. Οι καμπύλες για τις μεθόδους IVFADC, I-OPQ και I-PCA+RP συμπίπτουν σχεδόν παντού. 49

3.3 Ανάκληση στα πρώτα R δείγματα (recall@R) για τη συλλογή MNIST με $K = 64$, το οποίο επιλέχθηκε ως βέλτιστο και $w = 8$. $\bar{E} = E/n$: μέση παραμόρφωση ανά σημείο. 55

3.4 Ανάκληση στα πρώτα R δείγματα (recall@R) για τη συλλογή SIFT1M με παραμέτρους $K = 1024$, $w = 8$. 56
3.5 Ανάκληση στα πρώτα R δείγματα (recall@R) για τη συλλογή GIST1M με παραμέτρους $K = 1024$, $w = 16$. 56
3.6 Ανάκληση στα πρώτα R δείγματα (recall@R) για τη συλλογή SIFT1M ως πρός το μέγεθος των κωδικών ανα σημείο (bit allocation per point), με παραμέτρους $K = 1024$ και $w = 8$. Για $16, 32, 64$ και 128 bits, η παράμετρος m παίρνει τιμές $2, 4, 8$ και 16 αντιστοίχως. 57
3.7 Ανάκληση στα πρώτα R δείγματα (recall@R) για τη συλλογή SIFT1M ως προς την παράμετρο w, με $K = 1024$ και $m = 8$. 57
3.8 Ανάκληση στο σύνολο SIFT1B με κωδικούς μεγέθους 128-bit και $T = 100$K, όπως παρουσιάζεται και στον πίνακα 3.2. 60

4.1 Παράδειγμα του αλγορίθμου ομαδοποίησης KVQ σε ένα σύνολό από $n = 500$ τυχαία δισδιάστατα δεδομένα, παρμένα από ομοιόμορφη κατανομή στο $[0, 1]^2$, με ακτίνα $r = 0.2$. Ο αλγόριθμος καταλήγει σε 11 από τα αρχικά δεδομένα για κέντρα, σημεία τα οποία απεικονίζονται με κόκκινους κύκλους. 67
4.2 Χάρτης της Αθήνας με τις γεωγραφικές ομάδες για διάφορα επίπεδα μεγέθυνσης. 70
4.3 Φωτογραφίες που ανήκουν στα κέντρα των πλέον πολυπληθών ομάδων από το Πάνθεον της Ρώμης. 72
4.4 Φωτογραφίες σε μερικές από τις οπτικές ομάδες για το Πάνθεον. Η πρώτη φωτογραφία (από τα αριστερά) σε κάθε γραμμής/ομάδες αντιστοιχεί στο κέντρο της ομάδας. 73
4.5 Κατασκευή χάρτη σκηνής από 10 εικόνες του Palau Nacional, Montjuic, Barcelona. 76
4.6 Λεπτομέρεια του σύννεφου σημείων για τον χάρτη σκηνής του Montjuic, που αντιστοιχεί στην επισημειωμένη περιοχή του Σχήματος 4.5, (α) *πριν* και (β) *μετά* τον κβαντισμό διανυσμάτων. Τα διάφορα χρώματα εκφράζουν διαφορετικε΄ς οπτικές λέξεις, modulo 9. 77
4.7 Εικόνες αναζήτησης από τις 17 ομάδες του επισημειωμένου υποσυνόλου που αντιστοιχούν σε ορόσημα. 80
4.8 Εικόνες αναζήτησης από τις 18 ομάδες του επισημειωμένου υποσυνόλου που αντιστοιχούν σε σκηνές. 80
4.9 Παραδείγματα χαρτών σκηνών. 42 εικόνες χρησιμοποιήθηκαν για να κατασκευαστούν οι 6 χάρτες σκηνής του παραδείγματος. 82

4.10 Η διαδικασία κατασκευής ενός χάρτη σκηνής. Οι εικόνες τις αντίστοιχης οπτικής ομάδας ευθυγραμμίζονται πάνω στην εικόνα αναφοράς ακολουθιακά. 82
4.11 Σύγκριση του μέτρου Mean Average Precision για τις τέσσερις μεθόδους στη συλλογή *European Cities 1M* για διαφορετικό αριθμό εικόνων περίσπασης. 83
4.12 Μέση ακρίβεια για κάθε αναζήτηση ως προς το μέγεθος της αντίστοιχης ομάδας. 84
4.13 Παράδειγμα αναζήτησης, με τις παρόμοιες εικόνες που επιστρέφονται από τις τέσσερις μεθόδους, στην περίπτωση μιας μη δημοφιλούς τοποθεσίας. Η εικόνα αναζήτησης φαίνεται στα αριστερά και οι παρόμοιες εικόνες στα δεξιά. Κάθε γραμμή αντιστοιχεί σε κάθε μια από τις μεθόδους. 86
4.14 Παράδειγμα αναζήτησης, με τις παρόμοιες εικόνες που επιστρέφονται από τις τέσσερις μεθόδους, στην περίπτωση μιας δημοφιλούς τοποθεσίας. Η εικόνα αναζήτησης φαίνεται στα αριστερά και οι παρόμοιες εικόνες στα δεξιά. Κάθε γραμμή αντιστοιχεί σε κάθε μια από τις μεθόδους. 86

5.1 Η αρχική σελίδα της διαδικτυακής εφαρμογής VIRaL παρουσιάζει ένα τυχαίο υποσύνολο από εικόνες της συλλογής. 98
5.2 Αποτελέσματα μιας επιτυχημένης αναζήτησης. Πάνω αριστερά: ο χάρτης με μπλε markers για κάθε παρόμοια εικόνα και έναν κόκκινο marker για την εκτιμώμενη τοποθεσία. Πάνω δεξιά: η εικόνα αναζήτησης μαζί με τα συνήθη και προτεινόμενα tags. Κάτω γραμμές: οι οπτικά παρόμοιες εικόνες με φθίνουσα σειρά ομοιότητας. 100
5.3 Το άρθρο της Wikipedia που προτείνεται για την εικόνα αναζήτησης του Σχήματος 5.2. 101
5.4 Άλλη μια σελίδα αποτελεσμάτων με επιτυχημένη αναγνώριση τοποθεσίας και αντικειμένων της εικόνας αναζήτησης που φαίνεται πάνω δεξιά. 101
5.5 Το σύνολο των Similar of Similar εικόνων για την εικόνα αναζήτησης του Σχήματος 5.2. 102
5.6 Αντιστοιχίες μεταξύ της εικόνας αναζήτησης (αριστερά) και μίας παρόμοιας εικόνας (δεξιά). Τα τοπικά χαρακτηριστικά που είναι inliers απεικονίζονται σαν κίτρινοι κύκλοι με κλίμακα και προσανατολισμό και οι αντιστοιχίες ως κόκκινες γραμμές. Τα μπλε ορθογώνια δείχνουν την κοινή περιοχή των δύο εικόνων. 102
5.7 Εκτίμηση τοποθεσίας και αναγνώριση για μια εικόνα αναζήτησης που δεν περιέχει κάποιο ορόσημο. 103

5.8 Η εφαρμογή VIRaL Explore. 104
5.9 Η εφαρμογή VIRaL Routes. 105
5.10 Δείγματα από εικόνες αναζήτησης και οι εκτιμώμενες τοποθεσίες τους στο χάρτη. Για κάθε ζεύγος φαίνεται ο χάρτης στα αριστερά και η εικόνα αναζήτησης δεξιά. Μπλε marker: Παρόμοια εικόνα. Κόκκινος marker: εκτίμηση τοποθεσίας. 106
5.11 Δείγματα εικόνων αναζήτησης μαζί με τα συνήθη και τα προτεινόμενα tags. Τα ορόσημα αναγνωρίζονται επιτυχώς και σε κάθε περίπτωση παρέχεται και σύνδεσμος για το αντίστοιχο άρθρο της Wikipedia. 109

6.1 Πάνω αριστερά: Ένα τυχαίο σύνολο περιοχών. Κάτω αριστερά: Το ίδιο σύνολο μετασχηματισμένο αφινικά, όπου οι θέσεις των περιοχών και το σχήμα τους έχουν διαστρεβλωθεί, και έχουν εισαχθεί νέες περιοχές. Δεξιά: Τα αντίστοιχα αναμορφωμένα σύνολα. Οι πηγές είναι οι δύο μαύρες περιοχές στα αριστερά. 112
6.2 Inliers μεταξύ δύο συνόλων τοπικών χαρακτηριστικών. Ο κάθε ένας αντιστοιχεί σε ένα μη μηδενικό όρο του εσωτερικού γινομένου των αντίστοιχων χαρτών χαρακτηριστικών. Οι μαύρες γραμμές ενώνουν τους inliers. Οι κόκκινες γραμμές ενώνουν τις πηγές. Οι γκρίζες γραμμές ενώνουν τις πηγές με τους inliers. 113
6.3 Τριγωνοποίηση Delaunay στο σύνολο όλων των τοπικών χαρακτηριστικών που εξήχθησαν από ένα λογότυπο της εταιρίας FedEx. 114
6.4 (a) Ένα τρίγωνο με όλα τα σημεία των γωνιών του να έχουν εξαχθεί από σε παρόμοια κλίμακα, έστω s_1. (b) Έστω ότι προστίθεται ένα επιπλέον σημείο εξαγμένο σε κλίμακα s_2. Το σημείο αυτό θα επιρρεάσει την τριγωνοποίηση μόνο αν $|s_1 - s_2| < w$, όπου το w είναι μια παράμετρος κλίμακας. (c) Ένα τρίγωνο με όλα τα σημεία των γωνιών του να έχουν εξαχθεί από σε παρόμοια κλίμακα, έστω s_3. Το τρίγωνο αυτό θα ταιριάξει με το τρίγωνο του Σχήματος (a) αν όλες οι τρείς οπτικές λέξεις των σημείων τους είναι ίδιες. 115
6.5 Τριγωνοποιήσεις τοπικών χαρακτηριστικών σε διάφορες κλίμακες. . 115
6.6 Τρίγωνα που έχουν ταιριάξει μεταξύ δύο παρόμοιων λογότυπων. . . . 116
6.7 Αριστερά: Η εικόνα αναζήτησης, μια καθημερινή εικόνα που παρουσιάζει το στύλ ρουχισμού του ατόμου που θέλουμε να “αντιγράψουμε”. Δεξιά: Προτάσεις προϊόντων ρουχισμού, βασισμένες σε οπτική ανάλυση και αναγνώριση κλάσεων ρουχισμού. 117

6.8 Χωρικοί χάρτες πιθανότητας των κλάσεων, κβαντισμένοι και κανονικοποιημένοι σε μια γενική πόζα. Για την εκμάθησή τους χρησιμοποιήθηκε ένα σχετικό επισημειωμένο σύνολο εικόνων. Αριστερά: Χάρτες για τις κλάσεις ζώνη, μπότες, μπλούζα και φούστα (από αριστερά προς τα δεξιά και από πάνω προς τα κάτω). Δεξιά: Ο συνολικός χάρτης *πιθανότητας εμφάνισης ρουχισμού* όπως προέκυψε μετά την ένωση όλων των επιμέρους χαρτών. 118
6.9 Αριστερά: Η εικόνα αναζήτησης μετά την εξαγωγή της πόζας. Μέση: Η υπερ-κατάτμηση της πόζας. Δεξιά: Ομαδοποίηση με οπτικά κριτήρια. Οι περιοχές αυτές είναι οι υποψήφιες για την ανίχνευση κλάσεων ρουχισμού. 118

Κατάλογος Πινάκων

2.1 Συγκρίσεις στην απόδοση αναζήτησης (mAP) για γενικά (global) λεξικά διαφόρων μεγεθών στη βάση *Oxford Buildings* χωρίς ή με 20 χιλιάδες εικόνες περίσπασης, χρησιμοποιώντας $100/200/200$ checks στη βιβλιοθήκη FLANN για τις τεχνικές AGM/AKM/RAKM αντίστοιχα, 40 επαναλλήψεις για τα AKM/RAKM, και μόλις 15 για το AGM. 36

3.1 Ανάκληση στα πρώτα $\{1, 10, 100\}$ δείγματα, για το σύνολο SIFT1B με κωδικούς μεγέθους 64-bit, $K = 2^{13} = 8192$ και $w = 64$. Για τη μέθοδο Multi-D-ADC, το $K = 2^{14}$ και $T = 100$K. 58

3.2 Ανάκληση στα πρώτα $\{1, 10, 100\}$ δείγματα, για το σύνολο SIFT1B με κωδικούς μεγέθους 128-bit, $K = 2^{13} = 8192$ ($K = 2^{14}$) για απλό ανεστραμμένο αρχείο (multi-index). Για τις μεθόδους IVFADC+R και LOPQ+R, $m' = 8$, $w = 64$. Τα αποτελέσματα των μεθόδων Joint-ADC και KLSH-ADC τα πήραμε από τη δημοσίευση [114]. Οι γραμμές με αναφορά αναφέρονται στα ακριβή δημοσιευμένα νούμερα των αντίστοιχων δημοσιεύσεων. 59

4.1 Ονόματα και μεγέθη των ομάδων που αντιστοιχούν σε ορόσημα (17 ομάδες) και σκηνές (18 ομάδες) του επισημειωμένου υποσυνόλου από τη Βαρκελώνη. 81

4.2 Μέσος χρόνος αναζήτησης και το μέτρο απόδοσης μέσης ακρίβειας (mAP) για τις τέσσερις μεθόδους, στη βάση *European Cities 1M* μαζί με όλες τις εικόνες περίσπασης. 84

4.3 Μέση ακρίβεια (mAP) ανά ορόσημο για τις τέσσερις μεθόδους. Για κάθε ορόσημο χρησιμοποιούνται 5 εικόνες αναζήτησης. 87

4.4 Μέση ακρίβεια (mAP) ανά σκηνή για τις τέσσερις μεθόδους. Για κάθε μία χρησιμοποιούνται 5 εικόνες αναζήτησης ή όλες αν οι εικόνες τις σκηνής είναι λιγότερες από 5. 88

5.1 Ποσοστά σωστού εντοπισμού για διάφορα κατώφλια απόστασης για τις τέσσερις μεθόδους. 105

5.2 Ποσοστό σωστών προτάσεων άρθρων της Wikipedia για κάθε ορόσημο και συνολικός μέσος όρος για τις τέσσερις μεθόδους. 108

ΠΡΟΛΟΓΟΣ

Στην παρούσα διατριβή παρουσιάζονται τα σημαντικότερα αποτελέσματα που προέκυψαν από την έρευνα μου τα τελευταία πέντε χρόνια, έρευνα που τοποθετείται στο ευρύτερο επιστημονικό πεδίο της όρασης υπολογιστών και της μηχανικής μάθησης.

Η παρούσα διατριβή περιέχει νέους αλγορίθμους που είτε ορίζουν είτε στέκονται δίπλα στο state-of-the-art των αντίστοιχων επιστημονικών υποπεριοχών. Μερικοί από αυτούς αποτελούν γενικά εργαλεία για μεγάλης κλίμακας αναζήτηση κοντινότερου γείτονα ή ομαδοποίηση και άλλοι αναφέρονται σε πιο συγκεκριμένες εφαρμογές της οπτικής αναζήτησης. Συνολικά, η διατριβή καλύπτει ένα ευρύ φάσμα της περιοχής αυτής και, μιας και εστιάζει στην αναζήτηση μεγάλης κλίμακας, καθίσταται ιδιαίτερα επίκαιρη σε μια χρονική περίοδο κατά την οποία ο όγκος της διαθέσιμης οπτικής πληροφορίας πολλαπλασιάζεται.

Πρώτα από όλους θα ήθελα να ευχαριστήσω ιδιαίτερα τον υπεύθυνο καθηγητή μου Στεφανο Κόλλια, ο οποίος με καθοδηγούσε και στήριζε όλα αυτά τα χρόνια. Ευχαριστώ επίσης όλα τα παιδιά του εργαστηρίου για τη συνεργασία και την υποστήριξη τους, καθώς και τους συγγενείς και φίλους που, παρότι δεν καταλάβαιναν τι ακριβώς κάνω, δεν σταμάτησαν να με στηρίζουν όπως μπορούσαν.

Τέλος πρέπει να ευχαριστήσω δύο συναδέλφους και φίλους που όχι μόνο με συντρόφευαν καθ' όλη τη διάρκεια της ερευνητικής μου ζωής, αλλά εν τέλει την καθόρισαν σε μεγάλο βαθμό. Ο πρώτος είναι ό Δρ. Γιώργος Τόλιας, ο άνθρωπος με τον οποίο συνεργάστηκα ιδανικά όλα αυτά τα χρόνια σε ένα μεγάλο ποσοστό της έρευνας μου. Ο Γιώργος είναι επίσης και ένας στενός φίλος, του οποίου τα ερευνητικά, κοινωνικά και καλλιτεχνικά ενδιαφέροντα δεν σταματάνε να με εμπνέουν και να με παροτρύνουν να γίνομαι όλο και καλύτερος ερευνητής και άνθρωπος. Ο δεύτερος είναι ο Δρ. Γιάννης Αβρίθης, ο άνθρωπος που με το πάθος του, τις γνώσεις του, την ακεραιότητα του, την αναλυτική σκέψη του, τις αστείρευτες ιδέες του και με αμέτρητες συζητήσεις των "δέκα λεπτών" που κατέληγαν τρίωρες με έκανε να αγαπήσω την έρευνα και να θέλω να ασχοληθώ με αυτήν για την υπόλοιπη μου ζωή. Όσα ευχαριστώ και να γράψω για τον Γιάννη είναι πολύ λιγότερα από όσα του οφείλω, άρα θα του γράψω μόνο ένα: Ευχαριστώ.

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία προτείνονται βελτιώσεις στην οπτική αναζήτηση εικόνων, με τεχνικές που βασίζονται κυρίως σε *ομαδοποίηση*. Η ομαδοποίηση εκτελείται είτε στο χώρο των χαρακτηριστικών είτε στο χώρο των εικόνων, σε *πολυδιάστατους διανυσματικούς* ή *μετρικούς* χώρους, αντίστοιχα.

Αρχικά προτείνουμε μια νέα, γενικότερη μέθοδο ομαδοποίησης, η οποία συνδυάζει την περιγραφική δύναμη των μοντέλων μείγματος κανονικών κατανομών με τις ιδιότητες που απαιτούνται κατά την κατασκευή μεγάλης κλίμακας οπτικών λεξικών για αναζήτηση εικόνων. Είναι μια παραλλαγή του αλγορίθμου *expectation-maximization* που μπορεί να συγκλίνει γρήγορα, ενώ παράλληλα μπορεί να εκτιμήσει δυναμικά τον τελικό αριθμό των συνιστωσών. Επιστρατεύουμε τεχνικές προσεγγιστικών κοντινότερων γειτόνων για την επιτάχυνση του E-step του αλγορίθμου EM και εκμεταλλευόμαστε την επαναληπτική του φύση για να κάνουμε την αναζήτηση *αυξητική*, βελτιώνοντας την ταχύτητα αλλά και την ακρίβεια. Καταλήγουμε να έχουμε απόδοση υψηλότερη από το state of the art της αναζήτησης σε μεγάλες βάσεις εικόνων, ενώ είμαστε ταυτόχρονα το ίδιο γρήγοροι με τις πλέον γρήγορες γνωστές τεχνικές κατασκευής οπτικών λεξικών.

Έπειτα, παρουσιάζουμε μια νέα μέθοδο για αναζήτηση κοντινότερου γείτονα, μια μέθοδο που βελτιστοποιεί παραγοντικούς κβαντιστές *τοπικά* και έτσι μειώνει σημαντικά την παραμόρφωση κατά τον κβαντισμό. Αν συνδυαστεί με τη μέθοδο δεικτοδότησης multi-index, καταφέρνει να ξεπεράσει τα μέχρι τώρα καλύτερα δημοσιευμένα αποτελέσματα στην αναζήτηση κοντινότερου γείτονα σε ένα σύνολο με ένα δισεκατομμύριο πολυδιάστατα σημεία. Παράλληλα απολαμβάνει ταχύτητες αναζήτησης της τάξεως των λίγων millisecond, γεγονός που την καθιστά ανταγωνιστική ως προς το χρόνο ακόμα και σε σχέση με μεθόδους κατακερματισμού (hashing).

Προτείνουμε επίσης τους *χάρτες σκηνών* και θα δείξουμε ότι μια εκ των προτέρων ομαδοποίηση των εικόνων της συλλογής μπορεί να βελτιώσει την απόδοση της οπτικής αναζήτησης, ενώ παράλληλα ένα κριτήριο παραμόρφωσης μπορεί να εγγυηθεί την ανάκτηση ακόμα και απομονωμένων εικόνων από μη δημοφιλής τοποθεσίες όπως σε ένα γενικό σύστημα αναζήτησης εικόνων. Προτείνουμε μια λύση που παρότι μπορεί να δουλέψει σε συλλογές εκατομμυρίων εικόνων, μπορεί

να ανακτήσει ακόμα και τις μη δημοφιλής εικόνες απαιτώντας μονάχα ένα ποσοστό της αρχικής μνήμης.

Παρουσιάσουμε τέλος ένα ολοκληρωμένο σύστημα αναζήτησης εικόνων, το οποίο μπορεί να χρησιμοποιηθεί για αυτόματο γεωγραφικό εντοπισμό καθώς και για αναγνώριση οροσήμων ή σημείων ενδιαφέροντος, όπου αυτό είναι εφικτό. Το VIRaL (Visual Image Retrieval and Localization) παρέχει δημόσια πρόσβαση στις προαναφερθείσες τεχνολογίες μέσω ενός ενοποιημένου γραφικού διαδικτυακού περιβάλλοντος. Η διατριβή καταλήγει με τη συνοπτική περιγραφή μερικών ακόμα δημοσιεύσεων που εστιάζουν σε εφαρμογές της οπτικής αναζήτησης καθώς και τα συμπεράσματα της έρευνας.

ABSTRACT

New applications that exploit the huge data volume in community photo collections are emerging every day and visual image search is therefore becoming increasingly important. In this thesis we propose clustering- and nearest neighbor-based improvements for visual image search. Clustering is either performed on feature space or on image space, *i.e*. on high-dimensional vector spaces or metric spaces, respectively.

We first introduce a clustering method that combines the flexibility of Gaussian mixtures with the scaling properties needed to construct visual vocabularies for image retrieval. It is a variant of expectation-maximization that can converge rapidly while dynamically estimating the number of components. We employ approximate nearest neighbor search to speed-up the E-step and exploit its iterative nature to make search incremental, boosting both speed and precision. We achieve superior performance in large scale retrieval, being as fast as the best known approximate k-means algorithm.

We then present our locally optimized product quantization scheme, an approximate nearest neighbor search method that locally optimizes product quantizers per cell, after clustering the data in the original space. When combined with a multi-index, its performance is unprecedented and sets the new state-of-the-art in a billion scale dataset. At the same time, our approach enjoys query times in the order of a few milliseconds, deeming it comparable in terms of speed even to hashing approaches.

We next focus on large community photo collections. Most applications for such collections focus on popular subsets, *e.g*. images containing landmarks or associated to Wikipedia articles. In this thesis we are concerned with the problem of accurately finding the location where a photo is taken without needing any metadata, that is, solely by its visual content. We also recognize landmarks where applicable, automatically linking them to Wikipedia. We show that the time is right for automating the geo-tagging process, and we show how this can work at large scale. In doing so, we do exploit redundancy of content in popular locations—but unlike most existing solutions, we do not restrict to landmarks. In other words, we can compactly represent the visual content of all thousands of images depicting

e.g. the Parthenon and still retrieve any single, isolated, non-landmark image like a house or a graffiti on a wall. Starting from an existing, geo-tagged dataset, we cluster images into sets of different views of the the same scene. This is a very efficient, scalable, and fully automated mining process. We then align all views in a set to one reference image and construct a 2D *scene map*. Our indexing scheme operates directly on scene maps. We evaluate our solution on a challenging one million urban image dataset and provide public access to our service through our online application, VIRaL.

The thesis concludes with two chapters. The first is a summary of other approaches for visual search and applications, like geometry indexing, logo detection and clothing recognition, while the second presents conclusions and possible future directions.

Κεφάλαιο 1

Εισαγωγή – Βασικές τεχνικές οπτικής αναζήτησης

1.1 Εισαγωγή

Η μηχανή αναζήτησης της Google δεχόταν το 2011 κατά μέσο όρο 4.7 δισεκατομμύρια ερωτήματα *ανά μέρα* [1], 1.2 δισεκατομμύρια περισσότερα από αυτά που δεχόταν το 2010. Στατιστικά όπως αυτό δείχνουν ότι η αναζήτηση πληροφοριών είναι πλέον ένα βασικό εργαλείο του διαδικτύου, του οποίου ο ρόλος γίνεται όλο και πιο σημαντικός. Ακόμα και πιο σύνθετες μορφές αναζήτησης, όπως εκείνες με βάση την ομιλία, τον ήχο ή τις εικόνες έχουν αρχίσει να μπαίνουν στη ζωή του απλού χρήστη. Συστήματα όπως το Siri της Apple ή το αντίστοιχο Google Now της Google επιτρέπουν στους χρήστες φορητών και σταθερών συσκευών να αναζητούν με έναν πιο διαισθητικό τρόπο, «μιλώντας» στη συσκευή, υπαγορεύοντας της τα ερωτήματα με την μορφή ολοκληρωμένων προτάσεων. Η εφαρμογή Shazam[2] μπορεί να αναγνωρίσει ένα μουσικό κομμάτι σε δευτερόλεπτα, ταιριάζοντας ένα μικρό ζωντανό ηχητικό κλιπ με μια βάση πολλών εκατομμυρίων κομματιών.

Η *οπτική* αναζήτηση έχει επίσης αρχίσει να μπαίνει στην καθημερινότητα του χρήστη. Η μηχανή αναζήτησης εικόνων της Google[3] δίνει πλέον τη δυνατότητα αναζήτησης αρχίζοντας από μια εικόνα, διαδικτυακή ή τοπική. Η εφαρμογή Google Goggles[4] προσφέρει την ίδια δυνατότητα σε σχεδόν πραγματικό χρόνο για τους χρήστες φορητών συσκευών, αναγνωρίζοντας ορόσημα, εξώφυλλα βιβλίων, CD, καθώς και κείμενο. Είναι πλέον προφανές ότι το μέλλον της αναζήτησης, διαδικτυακή ή μη, περιλαμβάνει την ανάλυση πολυμεσικού περιεχομένου ως εγγενή διαδικασία.

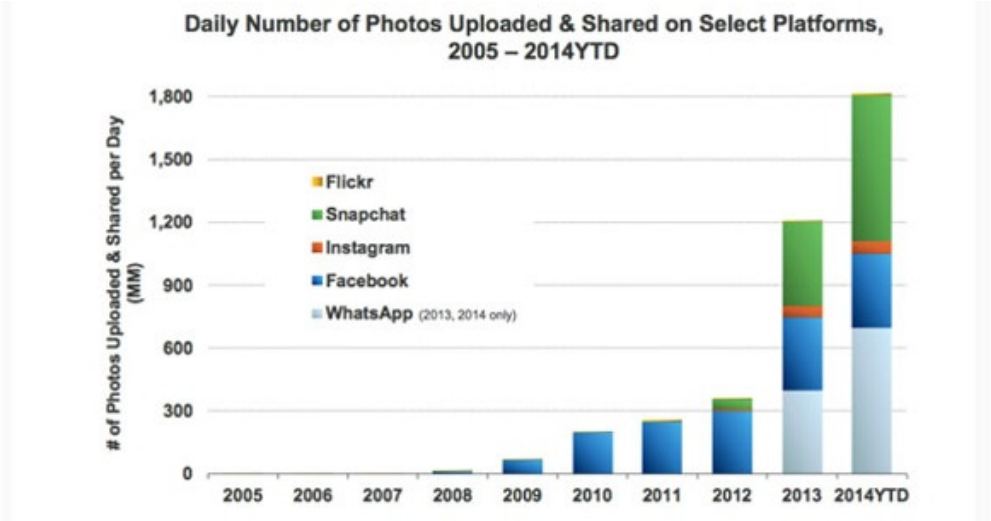
Ο συνολικός αριθμός εικόνων στις ιστοσελίδες κοινωνικής δικτύωσης (Face-

[1] http://www.statisticbrain.com/google-searches/

[2] http://www.shazam.com/

[3] http://images.google.com/

[4] http://www.google.com/mobile/goggles/



Σχήμα 1.1: *Ο όγκος των φωτογραφιών που ανεβαίνουν και μοιράζονται σε μερικές από τις πλέον δημοφιλείς ιστοσελίδες κοινωνικής δικτύωσης τα τελευταία χρόνια. Οι αριθμοί αναφέρονται σε* εκατομμύρια *φωτογραφίες ανά ημέρα. (Πηγή: KPCB)*

book, Instagram, Flickr, Picasa, WhatsApp, Panoramio) ανέρχεται σε πάρα πολλά δισεκατομμύρια, έχοντας παράλληλα όλο και μεγαλύτερους ρυθμούς αύξησης. Στο Σχήμα 1.1 παρουσιάζονται στατιστικά για τον όγκο των φωτογραφιών που ανεβαίνουν και μοιράζονται οι χρήστες σε δημοφιλή ιστοσελίδες κοινωνικής δικτύωσης *κάθε μέρα*, τα τελευταία χρόνια. Συγκεκριμένα, για το τρέχον έτος, παρατηρούμε ότι ο αριθμός των εικόνων που μοιράζονται στο διαδίκτυο, φτάνει τα *1,8 δισεκατομμύρια την ημέρα*!

Η αναζήτηση εικόνων σε τόσο μεγάλες συλλογές βασίζεται παραδοσιακά στην αναζήτηση με κείμενο και άλλα μεταδεδομένα που προέρχονται συνήθως από τον άνθρωπο. Παρά την μεγάλη ανάπτυξη της οπτικής αναζήτησης τα τελευταία χρόνια, οι τεχνολογίες δεν επαρκούν για το χειρισμό δεδομένων τέτοιας κλίμακας. Παράλληλα, έχουν αναπτυχθεί τελευταία διάφορες τεχνικές ομαδοποίησης και εξόρυξης δεδομένων οι οποίες εκμεταλλεύονται συναφή μεταδεδομένα των εικόνων, όπως την τοποθεσία, τον χρόνο, τον χρήστη και τα επισημειωμένα μεταδεδομένα κειμένου ή tags. Οι τεχνικές αυτές εστιάζουν συνήθως σε δημοφιλή υποσύνολα των φωτογραφιών, όπως για παράδειγμα εικόνες που περιέχουν σημεία ενδιαφέροντος ή μέρη που σχετίζονται με σελίδες της Wikipedia[5], περιοχή όπου η αναπαράσταση του οπτικού περιεχομένου των εικόνων μπορεί επίσης να βοηθήσει.

Ενδιαφέρουσα επίσης εξέλιξη είναι η δημοτικότητα που κερδίζουν νέες εφαρμογές, όπως για παράδειγμα εφαρμογές εκτίμησης τοποθεσίας [36], εικονικού τουρισμού [98] και αναγνώρισης ορόσημων (landmark) [117]. Τέτοιου είδους εφαρμογές έχουν αρχίσει να γίνονται μέρος μεγαλύτερων συστημάτων γεωγραφικού

[5] `www.wikipedia.org`

προσανατολισμού και ενδιαφέροντος, δημιουργώντας μια νέου είδους εμπειρία στον χρήστη.

Επιτυχημένο παράδειγμα είναι η εφαρμογή με την οποία ο χρήστης μπορεί να δει εικόνες από τα Flickr[6] και Panoramio[7] πλήρως ευθυγραμμισμένες και ενταγμένες οπτικά στα πανοράματα φωτογραφιών των υπηρεσιών χαρτών Bing Maps και Google Street View αντίστοιχα. Παρότι το ταίριασμα και η ευθυγράμμιση μπορούν να πραγματοποιηθούν αυτόματα με τις υπάρχουσες τεχνολογίες της όρασης υπολογιστών, απαιτείται να είναι ήδη γνωστή η γεωγραφική θέση των φωτογραφιών (geo-tag). Μια άλλη σχετική εφαρμογή είναι το History Pin[8] το οποίο προχωράει ένα βήμα παραπέρα και προσπαθεί να συλλέξει αρχεία που αντιστοιχούν γεωγραφικά ιστορικά γεγονότα και τις αντίστοιχες φωτογραφίες τους. Η όλη διαδικασία, βέβαια, είναι χειροκίνητη και απαιτείται από τους χρήστες της εφαρμογής να «καρφιτσώσουν» τις εικόνες στις κατάλληλες θέσεις του Street View χωρίς άλλη βοήθεια.

Αν και πολλές φωτογραφικές μηχανές και κινητά σήμερα περιέχουν ενσωματωμένο σύστημα εντοπισμού θέσης GPS και μπορούν να παρέχουν τα geo-tags των φωτογραφιών αυτόματα, ο μεγαλύτερος όγκος φωτογραφιών που ανεβαίνει στο διαδίκτυο είναι χωρίς γεωγραφική πληροφορία, πόσο μάλλον σχετικό επεξηγηματικό υλικό. Η αυτοματοποίηση του γεωγραφικού εντοπισμού των εικόνων θα ήταν ένα μεγάλο βήμα μπροστά για εφαρμογές όπως τις παραπάνω.

1.2 Σύνοψη συνεισφοράς

Στην παρούσα εργασία θα προτείνουμε βελτιώσεις στην οπτική αναζήτηση εικόνων, με τεχνικές που βασίζονται κυρίως σε *ομαδοποίηση*.[9] Η ομαδοποίηση εκτελείται είτε στο χώρο των χαρακτηριστικών είτε στο χώρο των εικόνων, σε *πολυδιάστατους διανυσματικούς* ή *μετρικούς* χώρους, αντίστοιχα. Θα προτείνουμε επίσης έναν state-of-the-art αλγόριθμο για προσεγγιστική αναζήτηση κοντινότερου γείτονα, ο οποίος καταφέρνει να δώσει τα καλύτερα αποτελέσματα ακρίβειας σε μια δημόσια συλλογή δισεκατομμυρίων σημείων.

Στις επόμενες ενότητες του τρέχοντος κεφαλαίου θα παρουσιαστούν οι βασικές τεχνικές για αναζήτηση εικόνων. Είναι οι τεχνικές στις οποίες βασιζόμαστε και προσπαθούμε να επεκτείνουμε και να βελτιώσουμε στα επόμενα κεφάλαια, μετρώντας την απόδοση τους με βάση την ακρίβεια σε πειράματα αναζήτησης μεγάλης κλίμακας.

[6] `www.flickr.com`

[7] `www.panoramio.com`

[8] `www.historypin.com/`

[9] Με τον όρο ομαδοποίηση μεταφράζουμε τον αγγλικό όρο *clustering*, του οποίου μια άλλη συνήθης μετάφραση στην ελληνική βιβλιογραφία είναι ο όρος *συσταδοποίηση*.

Στο κεφάλαιο 2 προτείνουμε μια νέα, γενικότερη μέθοδο ομαδοποίησης, η οποία συνδυάζει την περιγραφική δύναμη των μοντέλων μείγματος κανονικών κατανομών (Gaussian mixture model) με τις ιδιότητες που απαιτούνται κατά την κατασκευή μεγάλης κλίμακας οπτικών λεξικών για αναζήτηση εικόνων. Είναι μια παραλλαγή του αλγορίθμου *expectation-maximization* (EM) που μπορεί να συγκλίνει γρήγορα, ενώ παράλληλα μπορεί να εκτιμήσει δυναμικά τον τελικό αριθμό των συνιστωσών. Επιστρατεύουμε τεχνικές προσεγγιστικών κοντινότερων γειτόνων για την επιτάχυνση του E-step του αλγορίθμου EM και εκμεταλλευόμαστε την επαναληπτική του φύση για να κάνουμε την αναζήτηση *αυξητική*, βελτιώνοντας την ταχύτητα αλλά και την ακρίβεια. Καταλήγουμε να έχουμε απόδοση υψηλότερη από το state of the art της αναζήτησης σε μεγάλες βάσεις εικόνων, ενώ είμαστε ταυτόχρονα το ίδιο γρήγοροι με τις πλέον γρήγορες γνωστές τεχνικές κατασκευής οπτικών λεξικών.

Στο κεφάλαιο 3 θα παρουσιάζουμε μια νέα μέθοδο για αναζήτηση κοντινότερου γείτονα, μια μέθοδο που βελτιστοποιεί παραγοντικούς κβαντιστές *τοπικά* και έτσι μειώνει σημαντικά την παραμόρφωση κατά τον κβαντισμό. Αν συνδιαστεί με τη μέθοδο δεικτοδότησης multi-index, καταφέρνει να ξεπεράσει τα μέχρι τώρα καλύτερα δημοσιευμένα αποτελέσματα στην αναζήτηση κοντινότερου γείτονα σε ένα σύνολο με ένα δισεκατομμύριο πολυδιάστατα σημεία, ενώ απολαμβάνει ταχύτητες αναζήτησης της τάξεως των λίγων millisecond.

Στο κεφάλαιο 4 θα προτείνουμε τους *χάρτες σκηνών* και θα δείξουμε ότι μια εκ των προτέρων ομαδοποίηση των εικόνων της συλλογής μπορεί να βελτιώσει την απόδοση της οπτικής αναζήτησης, ενώ παράλληλα ένα κριτήριο παραμόρφωσης μπορεί να εγγυηθεί την ανάκτηση ακόμα και απομονωμένων εικόνων από μη δημοφιλής τοποθεσίες όπως σε ένα γενικό σύστημα αναζήτησης εικόνων. Προτείνουμε μια λύση που παρότι μπορεί να δουλέψει σε συλλογές εκατομμυρίων εικόνων, μπορεί να ανακτήσει ακόμα και τις μη δημοφιλής εικόνες.

Στο κεφάλαιο 5 θα παρουσιάσουμε ένα ολοκληρωμένο σύστημα αναζήτησης εικόνων, το οποίο μπορεί να χρησιμοποιηθεί για αυτόματο γεωγραφικό εντοπισμό καθώς και για αναγνώριση οροσήμων ή σημείων ενδιαφέροντος, όπου αυτό είναι εφικτό. Τέλος θα παρουσιαστεί η διαδικτυακή εφαρμογή μας, VIRaL (Visual Image Retrieval and Localization[10]) η οποία παρέχει δημόσια πρόσβαση στις προαναφερθείσες τεχνολογίες μέσω ενός ενοποιημένου γραφικού περιβάλλοντος. Μέρος του κειμένου των κεφαλαίων 4 και 5 δημοσιεύτηκε πρώτα στις δημοσιεύσεις [5] και [52] αντίστοιχα.

Η διατριβή καταλήγει με τα κεφάλαια 6 και 7. Στο πρώτο παρουσιάζονται συνοπτικά άλλες εργασίες και εφαρμογές που αναπτύχθηκαν κατά τη διάρκεια εκπόνησης της διατριβής και στο δεύτερο θα παρουσιαστούν τα συμπεράσματα της έρευνας και θα προταθούν μελλοντικές κατευθύνσεις.

[10] `viral.image.ntua.gr`

1.3 Βασικές τεχνικές αναζήτησης εικόνων

Η αναζήτηση εικόνων είναι μια ενεργή περιοχή έρευνας από την δεκαετία του ’70, οπότε και γεννήθηκε με μίξη ερευνητών από δύο κυρίως ερευνητικές κοινότητες. Την κοινότητα οργάνωσης βάσεων δεδομένων και την κοινότητα της όρασης υπολογιστών. Οι δύο κλάδοι εξετάζουν την αναζήτηση εικόνων από διαφορετικές σκοπιές, η πρώτη με τεχνικές βασισμένες σε κείμενο και η δεύτερη με βάση τα οπτικά χαρακτηριστικά και το περιεχόμενο των εικόνων.

Οι τεχνικές βασισμένες σε κείμενο αρχίζουν να αναπτύσσονται προς τα τέλη της δεκαετίας του ’70. Ένα τότε δημοφιλές πλαίσιο ανάκτησης προϋπέθετε τον σχολιασμό του περιεχομένου κάθε εικόνας από τον άνθρωπο, και έπειτα η ανάκτηση εκτελούνταν από ένα σύστημα βάσης δεδομένων. Για καιρό υπήρξε έρευνα σε αυτόν τον τομέα, δύο όμως βασικές δυσκολίες κατέστησαν τις μεθόδους αυτές αναποτελεσματικές. Πρώτον, οι συλλογές με εικόνες γίνονταν όλο και μεγαλύτερες και ήταν πλέον αδύνατον να σχολιαστούν όλες οι εικόνες χειροκίνητα. Το δεύτερο και αρκετά ουσιώδες πρόβλημα που παρατηρήθηκε, ήταν η υποκειμενικότητα της αντίληψης του περιεχομένου των εικόνων από άνθρωπο σε άνθρωπο.

Έτσι, στις αρχές της δεκαετίας του ’90, προτάθηκε η αυτόματη εξαγωγή χαρακτηριστικών των εικόνων, με βάση το οπτικό περιεχόμενό τους. Από τότε έχουν αναπτυχθεί πάρα πολλές τεχνολογίες και τεχνικές εξαγωγής χαρακτηριστικών που οδήγησαν σε πολλά, ερευνητικά και εμπορικά, συστήματα αναζήτησης. Η οπτική αναζήτηση εικόνων έχει εξελιχθεί ιδιαίτερα την τελευταία δεκαετία [97]. Στην σύγχρονη/παρούσα φάση της, το πλέον δημοφιλές μοντέλο είναι το *bag-of-words* (BoW)[96], το οποίο με την απλότητα και την αποτελεσματικότητά του έχει καταφέρει να κυριαρχήσει στην περιοχή. Το μοντέλο αυτό είναι ευρέως διαδεδομένο σε μια πληθώρα προβλημάτων της όρασης υπολογιστών, όπως η κατηγοριοποίηση, η ανίχνευση, η αναγνώριση καθώς και η αναζήτηση εικόνων και αντικειμένων.

Με εφαρμογή την αναζήτηση εικόνων, έχουν προταθεί τα τελευταία χρόνια πολλές επεκτάσεις στο μοντέλο *bag-of-words*. Άλλες ασχολούνται με τη γεωμετρική ανακατάταξη των εικόνων [85, 17, 102, 93], άλλες με την επέκταση ερωτήματος [19, 21], άλλες με τη εξαγωγή συνωνύμων για τις οπτικές λέξεις [29, 73, 30], άλλες με την εισαγωγή γεωμετρικής πληροφορίας στη δομή δεικτοδότησης [51, 7, 116, 39, 40] ενώ άλλες προσπαθούν να μειώσουν τα φαινόμενα κβαντισμού μέσω πολλαπλής ανάθεσης λέξεων [108, 86] ή την απαιτούμενη μνήμη[45, 5, 106, 104]

Συνοπτικά, ένα σύγχρονο σύστημα οπτικής αναζήτησης εικόνων βασισμένο στο μοντέλο *bag-of-words* αποτελείται από τις παρακάτω βασικές διαδικασίες:

- **Εξαγωγή οπτικών χαρακτηριστικών**. Αρχικά πρέπει να περιγραφεί το οπτικό περιεχόμενο των εικόνων με έναν τρόπο πιο εκφραστικό από τις αρχικές τιμές φωτεινότητας RGB. Οι εικόνες της συλλογής μπορεί αρχικά να υποστούν κάποιου είδους προ-επεξεργασία, ώστε να μπορέσουμε στη συνέχεια

να εξάγουμε σωστότερα οπτικά χαρακτηριστικά. Σε αυτό το στάδιο μπορεί να περιλαμβάνεται κάποιο φιλτράρισμα ή εξομάλυνση με σκοπό την αποθορυβοποίηση ή απλοποίηση των εικόνων. Σε αυτό το στάδιο μπορεί επίσης να γίνει κατάτμηση των εικόνων με σκοπό την εξαγωγή χαρακτηριστικών κατά περιοχές. Επόμενο στάδιο είναι η εξαγωγή των οπτικών χαρακτηριστικών είτε από ολόκληρη την εικόνα είτε από περιοχές της είτε από κάποια σημεία ενδιαφέροντος. Τα χαρακτηριστικά που χρησιμοποιούμε παρουσιάζονται στην ενότητα 1.3.1.

- **Κατασκευή οπτικού λεξικού**. Ο αλγόριθμος k-*means* είναι η πλέον κοινή επιλογή για τη δημιουργία *οπτικών λεξικών* (*visual vocabularies* ή *codebooks*) κυρίως λόγω της απλότητας και της ταχύτητας του. Η αναζήτηση εναλλακτικών τεχνικών έχει εξελιχθεί σε ενεργό πεδίο έρευνας στις περιπτώσεις λεξικών μικρού και μεσαίου μεγέθους, δηλαδή λεξικών έως 10^4 οπτικές λέξεις ή κέντρα. Για τις περιοχές της αναζήτησης εικόνων και εντοπισμού διπλότυπων σε μεγάλες συλλογές όπου χρησιμοποιούνται τοπικά χαρακτηριστικά, απαιτούνται οπτικά λεξικά πολύ μεγαλύτερου μεγέθους, λεξικά που αποτελούνται από 10^6 οπτικές λέξεις ή ίσως και περισσότερες. Οι τεχνικές ομαδοποίησης για την δημιουργία λεξικών σε αυτή την κλίμακα είναι πολύ περιορισμένες. Μέχρι σήμερα ιδιαίτερα διαδεδομένες παραμένουν απλές παραλλαγές του αλγορίθμου k-means, όπως για παράδειγμα ο *προσεγγιστικός αλγόρθμος* k-*means* (*approximate* k-*means* ή AKM [85]) και ο *ιεραρχικός αλγόριθμος* k-*means* (*hierarchical* k-*means* ή HKM [76]). Η σχετική βιβλιογραφία μαζί με λεπτομέρειες για την κατασκευή οπτικών λεξικών παρουσιάζονται αναλυτικά στο κεφάλαιο 2.

- **Αναπαράσταση της εικόνας**. Τα τοπικά χαρακτηριστικά της κάθε εικόνας αντιστοιχούνται σε οπτικές λέξεις μέσω τεχνικών κοντινότερου γείτονα. Δημιουργείται έπειτα ένα ιστόγραμμα εμφάνισης λέξεων για ολόκληρη την εικόνα, το ιστόγραμμα bag-of-words. Πιο συγκεκριμένα, το διάνυσμα ή αναπαράσταση bag-of-words μιας εικόνας p είναι το ιστόγραμμα που μετρά το πλήθος εμφανίσεων κάθε οπτικής λέξης (*tf* ή *term frequency*) σαν κοντινότερη στα χαρακτηριστικά της εικόνας p. Φυσικά, όταν τα λεξικά είναι της τάξεως του 10^6 δεν φυλάσσεται ποτέ ολόκληρο το διάνυσμα, αλλά μια αραιή αναπαράσταση. Συνήθως, επίσης το όλο διάνυσμα με τις τιμές tf κανονικοποιείται με ℓ_1 ή Ευκλείδεια νόρμα.

- **Δεικτοδότηση και στάδιο φιλτραρίσματος**. Τα διανύσματα bag-of-words από όλες τις εικόνες της βάσης δεικτοδοτούνται με την μορφή ανεστραμμένου αρχείου, για να μπορούμε να έχουμε άμεση πρόσβαση στις εικόνες τις βάσεις που περιέχουν μια οπτική λέξη. Από την εικόνα αναζήτησης εξάγονται οπτικά χαρακτηριστικά, τα οποία αντιστοιχούνται και σε οπτικές λέξεις.

Για κάθε ένα από αυτά, εκτελείται ένα ερώτημα στο ανεστραμμένο αρχείο και αυξάνεται το σκορ στις εικόνες τις βάσεις που ανήκουν στις λίστες των αντίστοιχων οπτικών λέξεων της εικόνας αναζήτησης, με τα κατάλληλα βάρη. Η διαδικασία αυτή, που αναφέρεται και ως *στάδιο φιλτραρίσματος*, περιγράφεται αναλυτικότερα στην ενότητα 1.3.2.

- **Γεωμετρική ανακατάταξη**. Το μοντέλο BoW, όπως παρουσιάστηκε μέχρι τώρα, αγνοεί παντελώς την γεωμετρία των τοπικών χαρακτηριστικών στο επίπεδο της εικόνας. Σε βάσεις μεγάλης κλίμακας, αναζητήσεις κατά τις οποίες δεν έχει ληφθεί υπόψη η γεωμετρική πληροφορία επιστρέφουν πολλά λάθος αποτελέσματα. Για να αποφευχθούν αυτά, οι εικόνες της βάσης που επιστρέφονται με το μεγαλύτερο σκορ από το στάδιο φιλτραρίσματος ανακατατάσσονται μέσω γεωμετρικού ταιριάσματος με την εικόνα αναζήτησης. Η βασική μέθοδος γεωμετρικού ταιριάσματος που χρησιμοποιούμε, και η οποία αποτελεί μια παραλλαγή του αλγορίθμου RANSAC περιγράφεται στην ενότητα 1.3.3.

1.3.1 Εξαγωγή οπτικών χαρακτηριστικών

Τα οπτικά χαρακτηριστικά αποτελούν τη βάση της αναπαράστασης του οπτικού περιεχομένου των εικόνων. Τα χαρακτηριστικά μπορούν να εξαχθούν είτε από ολόκληρη την εικόνα (global features) είτε από ομάδες pixel—περιοχές της εικόνας (region features). Τα πλέον χρησιμοποιούμενα χαρακτηριστικά είναι αυτά που κωδικοποιούν *το χρώμα, την υφή, το σχήμα* ή την περιοχή γύρω από κάποια *σημεία ενδιαφέροντος* της εικόνας. Τα τελευταία, τα οποία θα ονομάζουμε στο εξής και ως *τοπικά χαρακτηριστικά* (*local features*) θα είναι αυτά που θα χρησιμοποιούμε σε όλα τα παρακάτω κεφάλαια.

Στην περίπτωση που τα χαρακτηριστικά εξάγονται από ολόκληρη την εικόνα, επιδιώκεται να κωδικοποιηθούν συνολικά χαρακτηριστικά της. Με αυτόν τον τρόπο η εικόνα αναπαρίσταται ολόκληρη με ένα πολυδιάστατο διάνυσμα, που συνήθως έχει διάσταση της τάξης του 10^3. Τα χαρακτηριστικά που εξάγονται από ολόκληρη την εικόνα χρησιμοποιούνται όλο και λιγότερο τα τελευταία χρόνια για αναζήτηση καθώς δεν είναι αποτελεσματικά σε εικόνες με πυκνό οπτικό περιεχόμενο. Η χρήση τους όμως έχει επεκταθεί σε προβλήματα κατηγοριοποίησης, με νέες ολικές περιγραφές όπως τα VLAD [43] ή τα GIST [81]. Με γενικά χαρακτηριστικά μπορεί να κωδικοποιηθεί αποτελεσματικά μια εικόνα ηλιοβασιλέματος, αλλά όχι μια εικόνα από τον κεντρικό δρόμο του Μέμφις (Σχήμα 1.2). Η πάνω εικόνα αναπαρίσταται αποδοτικά με περιγραφείς εξαγμένους από ολόκληρη την εικόνα (για παράδειγμα με κάποιον περιγραφέα χρώματος) καθώς ολόκληρο το σημασιολογικό της περιεχόμενο μπορεί να περιγραφεί αποδοτικά μονάχα με την έννοια *ηλιοβασίλεμα*. Για την εικόνα στο κάτω μέρος θα χρειαστούν πιο εξελιγμένες τεχνικές

εξαγωγής για να κωδικοποιηθεί αξιοποιήσιμα το περιεχόμενο της.

Πολλές φορές σε μια εικόνα δεν μας ενδιαφέρει το γενικό της περιβάλλον, αλλά ο εντοπισμός συγκεκριμένων αντικειμένων μέσα σε αυτή. Πρέπει λοιπόν να εστιάσουμε την προσοχή μας σε σημεία της εικόνας που μπορεί να περιγράφουν μέρη από κάποια αντικείμενα. Ως σημεία ενδιαφέροντος θεωρούμε τα σημεία, στα οποία η γύρω περιοχή έχει πλούσιο οπτικό περιεχόμενο. Αυτά μπορεί να είναι σημεία στην περιφέρεια, στις γωνίες ή και στο εσωτερικό ενός αντικειμένου. Συνήθως σε ομογενείς επιφάνειες δεν έχουμε σημεία ενδιαφέροντος. Η σπουδαιότητα αυτών των σημείων έγκειται στο ότι περιγράφουν αποτελεσματικά σημαντικές περιοχές τις εικόνας με λίγες σχετικά τιμές, με αποτέλεσμα να είναι ιδιαίτερα διαδεδομένα στην ανάκτηση αντικειμένων.

Για να μην υπάρχει εξάρτηση από την κλίμακα (μέγεθος) η διαδικασία ανίχνευσης σημείων ενδιαφέροντος εκτελείται σε πολλαπλές κλίμακες (multi-scale analysis). Δημιουργείται για αυτό τον σκοπό ένας χώρος κλίμακας με διαδοχικές συνελίξεις της εικόνας με γκαουσιανούς πυρήνες αυξανόμενης τυπικής απόκλισης και σε κάθε επίπεδο εκτελείται η ανίχνευση των σημείων ενδιαφέροντος. Ανάλυση χώρων κλίμακας και εφαρμογές τους στην όραση υπολογιστών δίνει ο Lindeberg [64, 65]. Η περιοχή της ανίχνευσης σημείων ενδιαφέροντος είναι ενεργή μέχρι σήμερα, με νέες προσεγγίσεις να προτείνονται από τους Βαρυτιμίδη *et al*. [109] καθώς και από τους Αβρίθη και Ραπαντζίκο [6, 88].

Αφού έχουν βρεθεί τα σημεία ενδιαφέροντος, επόμενο βήμα είναι να βρεθεί μια εύρωστη και αποτελεσματική αναπαράσταση της γύρω περιοχής. Ο πιο απλός περιγραφέας θα ήταν ένα διάνυσμα με τις εντάσεις της φωτεινότητας των τριγύρω σημείων, με την ετεροσυσχέτιση σαν μέτρο ομοιότητας δύο τέτοιων περιοχών. Μια τέτοια αναπαράσταση όμως δεν είναι ιδιαίτερα αποτελεσματική καθώς απαιτεί αρκετό χώρο αποθήκευσης, και δεν έχει ανοχή σε φαινόμενα όπως περιστροφή ή αλλαγή κλίμακας.

Σημαντικό βήμα στην περιγραφή τοπικών χαρακτηριστικών έγινε το 2004 από τον Lowe, ο οποίος πρότεινε τους περιγραφείς *SIFT* (*Scale Invariant Feature Transform*) [68]. Ο Lowe προσεγγίζει τη λαπλασιανή της Κανονικής κατανομής (Laplacian-of-Gausian – LoG) με *διαφορές* Κανονικών κατανομών (Difference-of-Gaussian – DoG) για ταχύτητα και εκτελεί και αυτός τη διαδικασία σε αυξανόμενες κλίμακες για να επιτύχει ανεξαρτησία από την κλίμακα. Σαν περιγραφέα της περιοχής γύρω από τα σημεία, προτείνει τη χρήση ιστογραμμάτων των παραγώγων, κρατώντας παράλληλα τη θέση και την κατεύθυνσή τους. Με έναν κβαντισμό των κατευθύνσεων και θέσεων των παραγώγων που εκτελείται, επιτυγχάνεται μια ανοχή σε μικρές περιστροφές και γεωμετρικές παραμορφώσεις. Εκτεταμένη αξιολόγηση και σύγκριση τοπικών περιγραφέων γίνεται από τους Mikolajczyk και Schmid [72].

Επειδή η εξαγωγή περιγραφέων όπως τα *SIFT* είναι ιδιαίτερα χρονοβόρα δια-



Σχήμα 1.2: *Εικόνες με διαφορετική πυκνότητα οπτικού περιεχομένου. Η πάνω φωτογραφία, ο οποία τραβήχτηκε στην έρημο της Αριζόνα, μπορεί να περιγραφεί επαρκώς οπτικά με κάποιον ολικό (global) περιγραφέα υφής και χρώματος. Η κάτω εικόνα όμως, τραβηγμένη στην πολύχρωμη Beale Street του Memphis, απαιτεί πιο σύνθετες δομές για την περιγραφή του οπτικού της περιεχομένου.*

δικασία, η ενσωμάτωση τέτοιου είδους περιγραφέων σε συστήματα αναζήτησης εικόνων μεγάλης κλίμακας θα τα επιβάρυνε αρκετά. Έτσι, έχουν γίνει προσπάθειες τα τελευταία χρόνια να αναπτυχθούν περιγραφείς που μοιράζονται την ίδια φιλοσοφία με τα SIFT, είναι όμως πιο γρήγορα υπολογίσιμοι. Σε αυτή την κατηγορία ανήκουν οι περιγραφείς *SURF* (*Speeded-Up Robust Features*) [10], οι οποίοι είναι εκείνοι που θα χρησιμοποιούμε σε όλα μας τα πειράματα στα παρακάτω κεφάλαια.

1.3.2 Δεικτοδότηση και στάδιο φιλτραρίσματος

Έχοντας εξάγει τα διανύσματα bag-of-words για όλες τις εικόνες της βάσης στην οποία θέλουμε να ψάξουμε, κατασκευάζουμε μια δομή δεικτοδότησης *ανεστραμμένου αρχείου*. Αυτή είναι μια τεχνική δεικτοδότησης κατά την οποία δημιουργούνται ανεστραμμένες λίστες για κάθε (οπτική) λέξη. Ακολουθιακά, το id κάθε εικόνας της βάσης εισάγεται στις λίστες των περιεχόμενων οπτικών λέξεων, μαζί με την αντίστοιχη τιμή *tf* της λέξης για τη συγκεκριμένη εικόνα. Απαιτούνται συνολικά 8 bytes για κάθε καταχώριση (4 για το id της εικόνας και 4 για την τιμή *tf*), αλλά μπορούν να μειωθούν εύκολα στα 4 με χρήση απλών τεχνικών συμπίεσης (2 κωδικοποιώντας μόνο τις διαφορές των id και 2 για την τιμή *tf* κβαντισμένη σε 64K τιμές). Υποθέτοντας 1000 χαρακτηριστικά ανά εικόνα, η δομή δεικτοδότησης για ένα εκατομμύριο φωτογραφίες απαιτεί συνολικά περίπου ένα δισεκατομμύριο καταχωρίσεις, άρα περίπου 4GB μνήμης.

Μετά την εισαγωγή όλων των εικόνων της βάσης, το μέγεθος κάθε λίστας οπτικής λέξης μας δίνει μια εκτίμηση για το πόσο «δημοφιλής» ή συχνά εμφανιζόμενη είναι η συγκεκριμένη λέξη σε ολόκληρη τη βάση. Μπορούμε να ορίσουμε λοιπόν, σε αναλογία με την αναζήτηση κειμένου, το μέγεθος *inverse document frequency* ή idf για κάθε οπτική λέξη i, ως

$$\text{idf}_i = \log \frac{N}{n_i} \tag{1.1}$$

όπου N είναι ο αριθμός των εικόνων της βάσης και n_i είναι ο αριθμός εμφανίσεων της οπτικής λέξης i σε ολόκληρη τη βάση, το μέγεθος δηλαδή της αντίστοιχης λίστας στο ανεστραμμένο αρχείο. Έχοντας τις τιμές *idf* για όλες τις λέξεις, μπορούμε να αντικαταστήσουμε τις αποθηκευμένες στη δομή δεικτοδότησης τιμές *tf* με το σταθμισμένο βάρος, γνωστό και ως *TF-IDF*:

$$\text{tf}_{p,i} \cdot \text{idf}_i = \text{tf}_{p,i} \cdot \log \frac{N}{n_i} \tag{1.2}$$

το γινόμενο δηλαδή της κανονικοποιημένης συχνότητας εμφάνισης $\text{tf}_{p,i}$ της λέξης i στην εικόνα p επί το βάρος idf_i.

Μπορούμε επίσης, πάλι σε αναλογία με την αναζήτηση κειμένου, να ορίσουμε μία *stoplist*, μια λίστα οπτικών λέξεων που αγνοούνται. Αυτή η λίστα περιλαμβάνει έναν αριθμό από τις πιο δημοφιλείς και τις πιο σπάνιες λέξεις, οι οποίες στην

πρώτη περίπτωση δεν προσφέρουν διακριτική ικανότητα και στη δεύτερη συνήθως αντιστοιχούν σε θόρυβο.

Κατά τη διαδικασία της αναζήτησης, εξάγονται πρώτα οπτικά χαρακτηριστικά από την εικόνα αναζήτησης, τα οποία αντιστοιχούνται σε οπτικές λέξεις και τέλος παράγεται και από αυτήν το διάνυσμα bag-of-words, παρόμοια με τις εικόνες της βάσης. Για κάθε μη μηδενικό στοιχείο του διανύσματος BoW της εικόνας αναζήτησης, εκτελείται ένα πέρασμα στην αντίστοιχη λίστα της δομής δεικτοδότησης και αθροίζονται στις αντίστοιχες εικόνες τις βάσεις τα αποθηκευμένα βάρη *TF-IDF*. Η διαδικασία αυτή προσπέλασης του αρχείου δεικτοδότησης είναι ιδιαίτερα γρήγορη και αναφερόμαστε συνήθως σε αυτό το στάδιο της αναζήτησης ως *στάδιο φιλτραρίσματος*, καθώς με αυτό μπορούμε πολύ γρήγορα να πάρουμε μια τιμή ομοιότητας της εικόνας αναζήτησης με όλες τις εικόνες της βάσης (απαιτούνται συνήθως κάτω από 100ms για αυτό το στάδιο, για αναζήτηση σε βάση ενός εκατομμυρίου εικόνων).

1.3.3 Γεωμετρικό ταίριασμα

Μετά το στάδιο φιλτραρίσματος μπορούμε να κατατάξουμε τις εικόνες της βάσης ως προς την ομοιότητά τους με την εικόνα αναζήτησης. Όμως, η όλη διαδικασία ταιριάσματος μέχρι τώρα ασχολείται μόνο με την εμφάνιση των τοπικών χαρακτηριστικών (ταίριασμα μέσω οπτικών λέξεων) και αγνοεί παντελώς την γεωμετρία τους στο επίπεδο της εικόνας. Σε βάσεις μεγάλης κλίμακας, αναζητήσεις κατά τις οποίες δεν έχει ληφθεί υπόψη η γεωμετρική πληροφορία επιστρέφουν πολλά λάθος αποτελέσματα. Για να αποφευχθούν αυτά, οι εικόνες της βάσης που επιστρέφονται με το μεγαλύτερο σκορ από το στάδιο φιλτραρίσματος ανακατατάσσονται μέσω γεωμετρικού ταιριάσματος με την εικόνα αναζήτησης.

Για το γεωμετρικό ταίριασμα απαιτούνται κάποιες *πιθανές αντιστοιχίες* (*tentative correspondences*) χαρακτηριστικών μεταξύ των δύο εικόνων, αντιστοιχίες που μπορούμε να πάρουμε εύκολα από τα τοπικά χαρακτηριστικά που αντιστοιχούν σε ίδιες οπτικές λέξεις. Έπειτα μπορούμε να εφαρμόσουμε κάποιον αλγόριθμο τύπου *RANSAC*[26], δηλαδή κάποιον αλγόριθμο που προσπαθεί να βρει το μέγιστο υποσύνολο των αντιστοιχιών που «υπακούουν» σε έναν γεωμετρικό μετασχηματισμό.

Για το γεωμετρικό ταίριασμα στην βασική αναζήτηση χρησιμοποιούμε μια παραλλαγή του fast spatial matching [85] και ένα μοντέλο μετασχηματισμού ομοιότητας (similarity transformation), τεσσάρων βαθμών ελευθερίας. Το μοντέλο αυτό μπορεί να κάνει υποθέσεις από μια αντιστοιχία (*single correspondence* assumption) χρησιμοποιώντας τις σχετικές θέσεις (2 βαθμοί), την κλίμακα και τον προσανατολισμό των χαρακτηριστικών της αντιστοιχίας.

Για κάθε μία από τις πιθανές αντιστοιχίες, χρησιμοποιούμε τη θέση, την κλί-

μακα και τον προσανατολισμό των δύο χαρακτηριστικών της αντιστοιχίας για να υπολογίσουμε τους μετασχηματισμούς ομοιότητας T_1, T_2 που μετασχηματίζουν τα δύο χαρακτηριστικά στο μοναδιαίο κύκλο με κέντρο το κέντρο των αξόνων. Μπορούμε λοιπόν να κατασκευάζουμε μια αρχική υπόθεση μετασχηματισμού ως $T_2^{-1}T_1$. Έπειτα μετράμε τον αριθμό των γεωμετρικά επιβεβαιωμένων αντιστοιχιών ή inliers και επαναλαμβάνουμε για την επόμενη υπόθεση μετασχηματισμού χρησιμοποιώντας την επόμενη πιθανή αντιστοιχία. Όταν βρεθεί ένας νέος μέγιστος αριθμός inliers, μπορούμε μέσω ελαχίστων τετραγώνων να υπολογίσουμε έναν αφινικό μετασχηματισμό απο αυτούς και να αποθηκεύσουμε το μέχρι τότε καλύτερο μοντέλο—μια λογική που μοιάζει με το απλό μοντέλο του Locally Optimized RANSAC (LO-RANSAC) [17]. Παρατηρήσαμε ότι εικόνες που έχουν τουλάχιστον $\tau = 10$ inliers με την εικόνα αναζήτησης συνήθως απεικονίζουν το ίδιο αντικείμενο ή σκηνή. Πιο σύγχρονες και ταχύτερες μέθοδοι για γεωμετρικό ταίριασμα έχουν πρόσφατα εφαρμοστεί και στην αναζήτηση εικόνων με επιτυχία [103].

Κεφάλαιο 2

Ομαδοποίηση για κατασκευή οπτικών λεξικών

2.1 Εισαγωγή

Όπως αναφέρθηκε και στο προηγούμενο κεφάλαιο, το μοντέλο *bag-of-words* (BoW) απαιτεί την κατασκευή *οπτικών λεξικών* (*visual vocabularies* ή *codebooks*) διαδικασία για την οποία η πλέον κοινή επιλογή είναι ο αλγόριθμος k-*means* [69], κυρίως λόγω της απλότητας και της ταχύτητας του. Η αναζήτηση εναλλακτικών τεχνικών έχει εξελιχθεί σε ενεργό πεδίο έρευνας στις περιπτώσεις λεξικών μικρού και μεσαίου μεγέθους, δηλαδή λεξικών έως 10^4 οπτικές λέξεις ή κέντρα. Για τις περιοχές της αναζήτησης εικόνων και εντοπισμού διπλότυπων σε μεγάλες συλλογές όπου χρησιμοποιούνται τοπικά χαρακτηριστικά, απαιτούνται οπτικά λεξικά πολύ μεγαλύτερου μεγέθους, λεξικά που αποτελούνται από 10^6 οπτικές λέξεις ή ίσως και περισσότερες. Οι τεχνικές ομαδοποίησης για την δημιουργία λεξικών σε αυτή την κλίμακα είναι πολύ περιορισμένες. Μέχρι σήμερα ιδιαίτερα διαδεδομένες παραμένουν απλές παραλλαγές του αλγορίθμου k-means, όπως για παράδειγμα ο *προσεγγιστικός αλγόρθμος* k-*means* (*approximate* k-*means* ή AKM [85]) και ο *ιεραρχικός αλγόριθμος* k-*means* (*hierarchical* k-*means* ή HKM [76]).

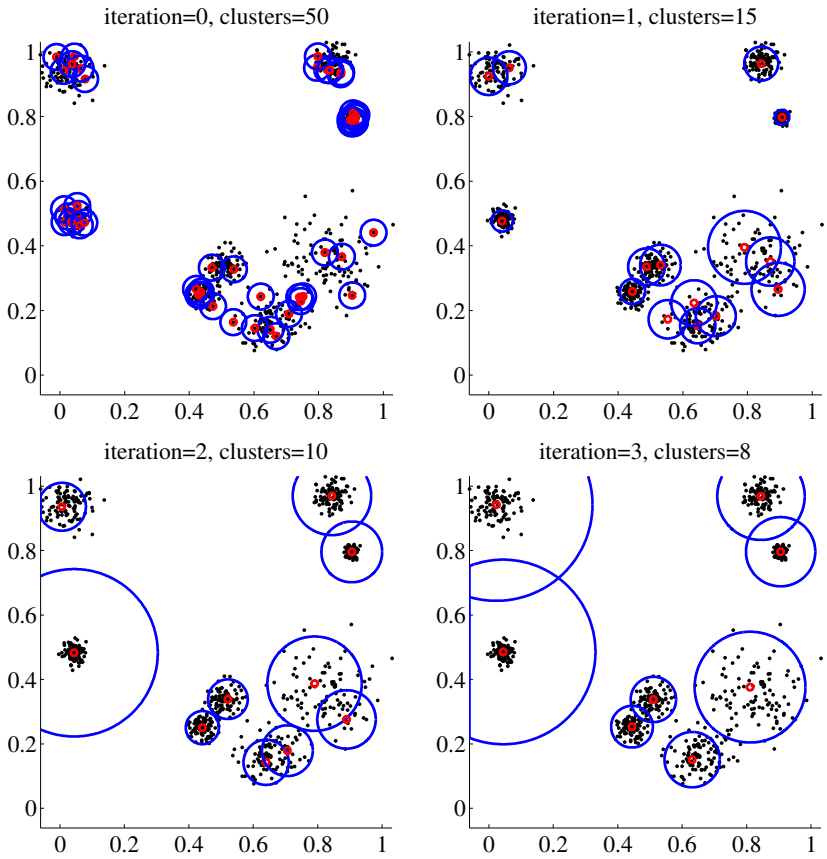
Το *μοντέλο μίγματος κανονικών κατανομών* (*Gaussian mixture model* ή *GMM*) με εκμάθηση *expectation-maximization* (EM) [11], είναι μια επέκταση του αλγορίθμου k-means η οποία έχει με επιτυχία εφαρμοστεί στην κατασκευή οπτικών λεξικών, για αναγνώριση σε επίπεδο κατηγορίας [84]. Εκτός από τη θέση, ο αλγόριθμος αυτός μπορεί να μοντελοποιήσει το *σχήμα* και τον *πληθυσμό* του κάθε κέντρου, δημιουργώντας έτσι ένα πιό περιγραφικό οπτικό λεξικό. Δυστυχώς, ο αλγόριθμος είναι πιό σύνθετος, υποθέτει αλληλεπίδραση μεταξύ όλων των διανυσμάτων δεδομένων με όλα τα κέντρα και συγκλίνει πιο αργά. Η πολυπλοκότητα της κάθε επανάληψης είναι $O(NK)$ όπου N και K είναι τα μεγέθη του συνόλου των δεδομένων και των κέντρων αντίστοιχα, συνεπώς δεν αποτελεί πρακτική επιλογή

για μεγάλο αριθμό κέντρων K. Αν βέβαια αναθέσουμε το κάθε διάνυσμα δεδομένων στο κοντινότερο μονάχα κέντρο μέσω προσεγγιστικών τεχνικών αναζήτησης κοντινότερων γειτόνων (*approximate nearest neighbor* (ANN) search) όπως συμβαίνει στον προσεγγιστικό αλγόριθμο k-means [85], τότε η πολυπλοκότητα πέφτει σε $O(N \log K)$, σε αυτή την περίπτωση όμως κρατείται μοναχά ένα γείτονας ανά διάνυσμα δεδομένων.

Ο *εύρωστος προσεγγιστικός αλγόριθμος k-means* (*Robust approximate k-means* ή RAKM [61]) είναι μια επέκταση του προσεγγιστικού αλγορίθμου k-means, όπου οι κοντινότεροι γείτονες της κάθε επανάληψης επαναχρησιμοποιούνται στην επόμενη για να διευκολύνουν την νέα αναζήτηση γειτόνων. Με αυτό τον τρόπο η διαδικασία αυτή επιταχύνεται, καθώς σε χώρους μεγάλης διάστασης D, για παράδειγμα 64 ή 128 διαστάσεων, βασική επίδραση στο κόστος κάθε επανάληψης κατέχει ο αριθμός των διανυσματικών πράξεων που ξοδεύονται για υπολογισμούς αποστάσεων προς πιθανούς γείτονες. Η παραπάνω προσέγγιση μας παρακίνησε να κρατήσουμε έναν μεγαλύτερο, σταθερό αριθμό m κοντινότερων γειτόνων ανάμεσα στις επαναλήψεις. Με αυτόν τον τρόπο, όχι μόνο βελτιώνεται η αναζήτηση κοντινότερων γειτόνων όπως στο [61], αλλά μας δίνεται επαρκής πληροφορία για να ορίσουμε ένα *προσεγγιστικό μοντέλο μίγματος Κανονικών κατανομών* (*approximate Gaussian mixture* (AGM) model), στο οποίο το κάθε διάνυσμα δεδομένων «αλληλεπιδρά» μοναχά με τα m κοντινότερα σε αυτό κέντρα.

Κατά τη φάση ανάπτυξης της παρούσας τεχνικής προέκυψαν ζητήματα με κέντρα τα οποία επικαλύπτονται στον χώρο, παράγοντας που επιδρά αρνητικά στην «δύναμη» διάκρισης του οπτικού μας λεξικού. Καταλήξαμε, λοιπόν, σε δύο μετατροπές του αλγορίθμου EM, μετατροπές οι οποίες αλλάξουν εμφανώς την συμπεριφορά του αλγορίθμου, οδηγώντας τελικά σε έναν τελείως νέο αλγόριθμο. Ένα παράδειγμα φαίνεται στο Σχήμα 2.1. Αρχίζοντας εφαρμόζοντας τον αλγόριθμο EM με μια σχετικά μεγάλη τιμή για το πλήθος των κέντρων K, η πρώτη μετατροπή είναι να υπολογίσουμε την επικάλυψη των γειτονικών κέντρων και να *διαγράψουμε* τα κέντρα που δείχνουν περιττά μετά από κάθε επανάληψη του EM. Επίσης, τα κέντρα που βρίσκονται κοντά σε κέντρα που διαγράφηκαν, θα πρέπει να γεμίσουν τον χώρο που είναι πλέον «κενός» από κέντρο. Οδηγούμαστε έτσι στη δεύτερη μετατροπή, κατά την οποία τα κέντρα*επεκτείνονται* όσο είναι δυνατό στον κενό χώρο.

Ο αλγόριθμος που προτείνουμε έχει ομοιότητες με άλλες τεχνικές όπως συγχωνευτικές τεχνικές (agglomerative methods) ή τεχνικές εκτίμησης πυκνότητας (density-based mode seeking). Συγκεκριμένα, μπορεί αυτόματα να εκτιμήσει τον τελικό αριθμό των κέντρων, ξεκινώντας με μεγάλο K και διαγράφοντας σε κάθε επανάληψη. Ακόμη, η ιδιότητα επέκτασης του εύρους των κέντρων προς τον κενό χώρο *βελτιώνει τον ρυθμό σύγκλισης*—στο δισδιάστατο παράδειγμα του Σχήματος 2.1 η σύγκλιση επιτυγχάνεται μετά από μόνο 3 επαναλήψεις. Ο αλγόριθμος



Σχήμα 2.1: *Εκτίμηση του αριθμού, του πληθυσμού, της θέσης και έκτασης των κέντρων* σε τρεις μόνο επαναλήψεις*, για ένα σύνολο* 800 *δισδιάστατων δεδομένων τα οποία προέρχονται από ένα μείγμα* 8 *Κανονικών κατανομών. Κόκκινοι κύκλοι: θέση κέντρων, μπλε: δύο τυπικές αποκλίσεις. Τα κέντρα αρχικοποιούνται σε* 50 *από τα αρχικά δεδομένα με τυχαίο τρόπο, με αρχικό* $\sigma = 0.02$*. Είναι εμφανής η «επεκτατική» συμπεριφορά των δύο κέντρων στα αριστερά.*

που προτείνουμε, ο αλγόριθμος *επεκτατικών μιγμάτων κανονικών κατανομών* (*expanding Gaussian mixtures* ή EGM), είναι ιδιαίτερα γενικός και μπορεί να χρησιμοποιηθεί σε πληθώρα εφαρμογών. Στην παρούσα εργασία εστιάζουμε σε ισοτροπικές/σφαιρικές κανονικές κατανομές και εφαρμόζουμε την προσεγγιστική του εκδοχή, AGM, στην κατασκευή οπτικών λεξικών μεγάλου μεγέθους για αναζήτηση εικόνων. Αυτή είναι η πρώτη υλοποίηση μιγμάτων κανονικών κατανομών σε τόσο μεγάλη κλίμακα, μια προσέγγισή που δείχνει να βελτιώνει την απόδοση της αναζήτησης εικόνων χωρίς περαιτέρω υπολογιστικό κόστος σε σχέση με τον προσεγγιστικό αλγόριθμο k-means.

2.2 Σχετική βιβλιογραφία

Όπως έχει ήδη αναφερθεί στο Κεφάλαιο 1, το μοντέλο Bag-of-Words προϋποθέτει την κατασκευή ενός *οπτικού λεξικού*, το οποίο είναι συνήθως προϊόν ομαδοποίησης (clustering) σε ένα σύνολο δεδομένων εκπαίδευσης. Με δεδομένο το λεξικό κάθε νέο διάνυσμα δεδομένων μπορεί να αντιστοιχηθεί σε μία ή περισσότερες *οπτικές λέξεις* ή *κέντρα* μέσω τεχνικών κβαντοποίησης διανυσμάτων ή τεχνικών κοντινότερου γείτονα. Για την διαδικασία ομαδοποίησης η πλέον συνήθης επιλογή είναι ο αλγόριθμος k-*means* [69], του οποίου μια γνωστή αρνητική ιδιότητα είναι το ότι τα κέντρα τείνουν να συγκεντρώνονται στις λίγες πυκνές περιοχές του διανυσματικού χώρου, περιοχές που δεν είναι ιδιαίτερα διακριτικές, ειδικά στην περίπτωση οπού τα τοπικά χαρακτηριστικά είναι εξαγμένα σε πυκνό πλέγμα θέσεων [12].

Για την αποφυγή της δυσάρεστης αυτής κατάστασης έχουν προταθεί αρκετές λύσεις στην βιβλιογραφία. Οι Jurie και Triggs [48] χρησιμοποιούν ακτινική ομαδοποίηση (radius-based clustering) και διατρέχουν τα δεδομένα εφαρμόζοντας ένα κριτήριο ελάχιστης παραμόρφωσης (minimum distortion criterion). Από την άλλη μεριά, οι Leibe *et al*. [59] και Fulkerson *et al*. [27] συνδυάζουν τις διαχωριστικές ιδιότητες του αλγορίθμου k-means με *συγχωνευτική ομαδοποίηση* (*agglomerative clustering*). Στη δημοσίευσή [59], ο διανυσματικός χώρος διαχωρίζεται πρώτα με χρήση του αλγορίθμου k-means και έπειτα εφαρμόζονται τεχνικές αμοιβαίων κοντινότερων γειτόνων (reciprocal nearest neighbors ή RNN) σε κάθε κέντρο ανεξάρτητα. Παρομοίως, στη δημοσίευση [27] χρησιμοποιείται ο ιεραρχικός αλγόριθμος k-means αρχικά, ακολουθούμενος από την συγχωνευτική τεχνική agglomerative information bottleneck (AIB), έτσι ώστε να κατασκευαστούν μικρού μεγέθους οπτικά λεξικά τα οποία έχουν αρκετή διακριτική ικανότητα για κατηγοριοποίηση. Οι Vedaldi and Soatto [110] επεκτείνουν τους αλγορίθμους mean shift και medoid shift σε μη Ευκλείδειους χώρους και παρουσιάζουν τον γρήγορο αλγόριθμο *quick shift*. Τον εφαρμόζουν μεταξύ άλλων σε ομαδοποίηση ιστογραμμάτων bag-of-words, αλλά όχι για την κατασκευή οπτικών λεξικών στο χώρο των χαρακτηριστικών.

Η προσεγγιστική μέθοδος που προτείνεται στο κεφάλαιο αυτό είναι τόσο αποτελεσματική ώστε μπορεί να αρχικοποιηθεί έχοντας ως κέντρα *όλα* τα δεδομένα εκπαίδευσης. Μπορεί, έτσι, αφενός να περιγράψει σωστά ακόμα και τις αραιές περιοχές του χώρου χαρακτηριστικών και από την άλλη να διαγράψει δυναμικά τα επικαλυπτόμενα κέντρα στις πυκνές περιοχές του χώρου, όπως πρέπει.

Τα *μοντέλα μιγμάτων κανονικών κατανομών* (*Gaussian mixture models* ή GMM) μπορούν να αναπαραστήσουν το ακριβές σχήμα των περιοχών των κέντρων και έχουν χρησιμοποιηθεί για καθολικά (universal) οπτικά λεξικά [112], για λεξικά προσαρμοσμένα ανά κλάση [84], ακόμα και για ταξινομητές *SVM μιας κλάσης* [113]. Ο Perronnin [84] εκπαιδεύει ένα καθολικό λεξικό βελτιστοποιόντας με τεχνικές μέ-

γιστης πιθανοφάνειας τις παραμέτρους του GMM μοντέλου και έπειτα εξειδικεύει τα λεξικά ανά κλάση με τεχνικές μεγιστοποίησης της εκ των υστέρων πιθανότητας (maximum posterior ή MAP), ενώ στη δημοσίευση [112] οι συγγραφείς μαθαίνουν ένα βέλτιστου μεγέθους λεξικό μέσω συγχώνευσης κέντρων ενός αρχικά μεγαλύτερου λεξικού. Οι Aggarwal και Triggs [2] επεκτείνουν το μοντέλο μιγμάτων κανονικών κατανομών προσθέτοντας χωρική πληροφορία με σκοπό να αναπαραστήσουν τις εικόνες με υπερ-χαρακτηριστικά (*hyperfeatures*).

Όλες οι παραπάνω μέθοδοι εστιάζουν σε μικρού μεγέθους λεξικά, τα οποία βρίσκουν χρήση σε προβλήματα αναγνώρισης και δεν μπορούν να εφαρμοστούν σε μεγαλύτερη κλίμακα, για παράδειγμα σε προβλήματα ομαδοποίηση με 10^6 κέντρα, ώστε να έχουν την απαιτούμενη διακριτική ικανότητα για να χρησιμοποιηθούν σε αναζήτηση εικόνων μεγάλης κλίμακας. Από τη άλλη, τεχνικές χωρίς στάδιο εκμάθησης όπως για παράδειγμα η σταθερή κβαντοποίηση σε κανονικό πλέγμα [107] ή ο *κατακερματισμός* (*hashing*) με τυχαία ιστογράμματα [25], παρότι έχουν αποδειχθεί ιδιαίτερα αποδοτικές σε προβλήματα αναγνώρισης, έχουν αποτύχει όταν εφαρμόζονται για αναζήτηση [86].

Για οπτικά λεξικά που χρησιμοποιούνται στην αναζήτηση μεγάλης κλίμακας, η πλέον διαδεδομένη τεχνική είναι η κατασκευή επίπεδου λεξικού μέσω του προσεγγιστικού αλγορίθμου k-means (AKM) [85], όπου ο παραδοσιακός αλγόριθμος επιταχύνεται με την χρήση μιας συστάδας *τυχαίων kd-δέντρων* (*randomized k-d trees*) [94] για την εύρεση των κοντινότερων γειτόνων σε κάθε επανάληψη. Έχει αποδειχθεί ότι τα λεξικά από τον αλγόριθμο αυτό είναι πιο αποδοτικά κατά την αναζήτηση από ότι αυτά που κατασκευάστηκαν με τον ιεραρχικό αλγόριθμο k-means (HKM) [76], αλλά επίσης και ότι ο εύρωστος προσεγγιστικός αλγόριθμος k-means (RAKM) [61] είναι ακόμα γρηγορότερος. Στην προτεινόμενη προσέγγισή μας, αν και περιοριζόμαστε σε σφαιρικές κανονικές κατανομές, επιτυγχάνουμε μεγαλύτερη προσαρμοστικότητα και περιγραφική δύναμη λόγω του μοντέλου μίγματος κανονικών κατανομών, έχοντας παράλληλα ταχύτητα σύγκλισης παρόμοια με την [61].

Η ανάθεση ενός διανύσματος χαρακτηριστικών σε οπτική λέξη επιτυγχάνεται αποδοτικά μέσω προσεγγιστικών τεχνικών κοντινότερου γείτονα, ή πιο διαδεδομένη από τις οποίες κάνει χρήση μιας συστάδας *τυχαίων kd-δέντρων*(*randomized k-d trees*) [94]. Συγκεκριμένα χρησιμοποιούμε την υλοποίηση που δίνεται στη βιβλιοθήκη FLANN [75], η οποία έχει αποδειχθεί αποδοτικότερη από ανταγωνιστικές τεχνικές όπως το *locality-sensitive hashing* [23]. Απο την άλλη, οι Moosmann *et al*. [74], ισχυριζόμενοι ότι μια ιεραρχική δομή, για παράδειγμα ένα δέντρο k-means, δεν μπορεί από μόνη της να αναπαραστήσει σωστά τη ποικιλία των πολυδιάστατων χαρακτηριστικών και εισήγαγαν το *extremely randomized clustering* (ERC) *forest*. Η δομή αυτή, που αποτελείται από μια συστάδα δέντρων με τυχαιότητα στη κατασκευή τους, όταν δεν υπάρχουν ετικέτες κλάσεων στα δεδομένα εκπαίδευσης γινεται ταυτόσημη με τη συστάδα τυχαίων kd-δέντρων. Σαν τεχνική

ομαδοποίησης, όμως, η δομή ERC έχει απόδοση χαμηλότερη από τον αλγόριθμο k-means.

Η *Παραγοντική κβαντοποίηση* (*Product quantization*) [42] έχει πρόσφατα αποδειχθεί ανώτερη της FLANN για αναζήτηση κοντινότερου γείτονα, αλλά μιας και απαιτεί μεγαλύτερο χρόνο για κατασκευή της κύριας δομής, τίθεται ακατάλληλη για να χρησιμοποιείται σε κάθε επανάληψη του αλγορίθμου EM. Στην παρούσα μέθοδο, χρησιμοποιούμε την βιβλιοθήκη FLANN, αλλά ταυτόχρονα εκμεταλλευόμαστε την επαναληπτική φύση του αλγορίθμου EM, έτσι ώστε να γίνει η διαδικασία αναζήτησης *αυξητική*, προσδίδοντας ακρίβεια και μειώνοντας την ταχύτητα.

Στην προσπάθεια μετριασμού του σφάλματος κβαντοποίησης, έχει προταθεί η *πολλαπλή ανάθεση* σε οπτικές λέξεις (*multiple* ή *soft assignment*) [86]. Ερμηνεύοντας την πολλαπλή ανάθεση ως εκτίμηση πυκνότητας (*kernel density estimation*) οι Van Gemert *et al*. [108] την εφαρμόζουν σε μικρού μεγέθους λεξικά για αναγνώριση, ενώ στις δημοσιεύσεις [66], [13] διερευνώνται διάφορες τεχνικές συνάθροισης (*pooling* strategies). Η πολλαπλή ανάθεση μπορεί να εφαρμοστεί στα χαρακτηριστικά μόνο της εικόνας αναζήτησης [41], αλλά αυτή η προσέγγιση αυξάνει γενικά την πολυπλοκότητα, ειδικά σε εφαρμογές αναγνώρισης. Έτσι, οι Lehmann *et al*. [57] προσπαθούν να μεταφέρουν την λειτουργικότητα αυτή στο οπτικό λεξικό μέσω μιας λειτουργίας θολώματος, ενώ οι Mikulik *et al*. [73] μαθαίνουν οπτικά συνώνυμα για τις οπτικές λέξεις ενός ιδιαίτερα μεγάλου λεξικού, χρησιμοποιώντας ακολουθίες χαρακτηριστικών τα οποία έχουν ταιριάξει γεωμετρικά (*feature tracks*).

Η τεχνική *Hamming embedding* [39] κινείται προς την αντίθετη κατεύθυνση, καθώς προσπαθεί να κωδικοποιήσει προσεγγιστικά της θέσεις των δεδομένων μέσα στα Βορονόι κελιά ενός μικρότερου λεξικού, απαιτώντας βέβαια μεγαλύτερο χώρο στη μνήμη για τη δομή δεικτοδότησης. Με το να μάθουμε τις παραμέτρους σχήματος των κανονικών κατανομών του μοντέλου μίγματος, καταφέρνουμε στην παρούσα εργασία να επιτύχουμε μια πιο ακριβή πολλαπλή ανάθεση, ενώ παράλληλα η γενική μέθοδος ομαδοποίησης που προτείνουμε μπορεί κάλλιστα να εφαρμοστεί σε συνδυασμό με τεχνικές εκμάθησης συνωνύμων ή ενσωμάτωσης περαιτέρω πληροφορίας, αυξάνοντας την τελική απόδοση του συστήματος.

2.3 Επεκτατικά μίγματα κανονικών κατανομών

Ξεκινάμε παρουσιάζοντας συνοπτικά τη θεωρία του μοντέλου μίγματος κανονικών κατανομών και της εκμάθησης των παραμέτρων του μέσω του αλγορίθμου EM, ακολουθώντας κυρίως την ανάλυση του [11], στην ενότητα 2.3.1. Έπειτα αναπτύσσουμε τις δύο μετατροπές που προτείνουμε για διαγραφή και επέκταση στις ενότητες 2.3.2 και 2.3.3 αντίστοιχα. Τέλος, αναφερόμαστε στην αρχικοποίηση και τερματισμό του αλγορίθμου, για την συγκεκριμένη εφαρμογή των οπτικών λεξικών. Με αυτή την ενότητα ολοκληρώνεται η παρουσίαση του γενικότερου προτει-

νόμενου αλγορίθμου EGM και η εργασία συνεχίζεται με την ανάπτυξη της προσεγγιστικής εκδοχής του αλγορίθμου, AGM, στην ενότητα 2.4. Στην παρακάτω ανάλυση, χρησιμοποιούμε τους όρους *συνιστώσα* και *κέντρο* εννοώντας το ίδιο πράγμα, παρομοίως για τους όρους *μέσος όρος* και *θέση* για τις συνιστώσες.

2.3.1 Εκμάθηση παραμέτρων

Η πυκνότητα $p(\mathbf{x})$ ενός μίγματος κανονικών κατανομών είναι ένας κυρτός συνδυασμός από K D-διάστατες πυκνότητες ή *συνιστώσες*,

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{2.1}$$

για κάθε $\mathbf{x} \in \mathbb{R}^D$, όπου τα $\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ είναι οι *συντελεστές του μίγματος*, οι *μέσοι όροι* και οι*μήτρες συνδιακύμανσης* αντίστοιχα, για την k-στη συνιστώσα και τα $\boldsymbol{\pi} = \{\pi_k\}, \boldsymbol{\mu} = \{\boldsymbol{\mu}_k\}, \boldsymbol{\Sigma} = \{\boldsymbol{\Sigma}_k\}$ αναπαριστούν το σύνολο των παραμέτρων. Οι συντελεστές του μίγματος είναι κανονικοποιημένες πιθανότητες, έτσι ώστε

$$\pi_k \geq 0, \quad \sum_{k=1}^{K} \pi_k = 1 \tag{2.2}$$

για $k = 1 \ldots, K$. Ερμηνεύοντας το π_k ως την εκ των προτέρων πιθανότητα $p(k)$ της συνιστώσας k, και δεδομένης της παρατήρησης $\mathbf{x}$, η ποσότητα

$$\gamma_k(\mathbf{x}) = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \tag{2.3}$$

για $\mathbf{x} \in \mathbb{R}^D, k = 1, \ldots, K$, εκφράζει την εκ των υστέρων πιθανότητα $p(k|\mathbf{x})$. Λέμε ότι η ποσότητα $\gamma_k(\mathbf{x})$ είναι η *υπευθυνότητα*(*responsibility*) που παίρνει η συνιστώσα k για να «ερμηνεύσει» την παρατήρηση $\mathbf{x}$, ή πιο απλά η *υπευθυνότητα του* k *για το* $\mathbf{x}$. Δεδομένου ενός συνόλου ανεξάρτητων και ανεξαρτήτως κατανεμημένων παρατηρήσεων ή διανυσμάτων δεδομένων $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, η συνάρτηση λογαριθμικής πιθανότητας ενός συγκεκριμένου μοντέλου μίγματος κανονικών κατανομών με παραμέτρους $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$ δίνεται από τη σχέση

$$\ln p(X|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}. \tag{2.4}$$

Η λύση *μέγιστη πιθανοφάνειας* ή *maximum likelihood* (ML) για τα $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$ δίνεται υπολογίζοντας τις μερικές παραγώγους της συνάρτησης λογαριθμικής πιθανότητας ως προς τα $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$ αντιστοίχως και εξισώνοντάς την με το μηδέν. Κατά τη διαδικασία παραγώγισης θα πρέπει να χρησιμοποιηθούν και οι πολλαπλασιαστές

Lagrange για την παράγωγο ως προς π, καθώς πρέπει να ικανοποιείται ταυτόχρονα και ο περιορισμός της εξίσωσης (2.2) για τους συντελεστές. Για κάθε συνιστώσα $k = 1, \ldots, K$, η λύση δίνεται από τις παρακάτω εξισώσεις [11]

$$\pi_k = \frac{N_k}{N} \tag{2.5}$$

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \mathbf{x}_n \tag{2.6}$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathsf{T}}, \tag{2.7}$$

όπου $\gamma_{nk} = \gamma_k(\mathbf{x}_n)$ για $n = 1, \ldots, N$, και η ποσότητα $N_k = \sum_{n=1}^{N} \gamma_{nk}$, ερμηνεύεται ως ο *ενεργός αριθμός* (*effective number*) των δεδομένων που αντιστοιχούνται στη συνιστώσα k. Ο αλγόριθμος *expectation-maximization* (EM) περιλαμβάνει μια επαναληπτική διαδικασία μάθησης: με δεδομένο ένα αρχικό σύνολο παραμέτρων $\pi, \boldsymbol{\mu}, \Sigma$, υπολογίζει τις υπευθυνότητες γ_{nk} σύμφωνα με την (2.3) (E-step), Έπειτα, επανεκτιμούνται οι παράμετροι χρησιμοποιώντας τις σχέσεις (2.5)-(2.7), κρατώντας τις υπευθυνότητες σταθερές (M-step). Η αρχικοποίηση και ο τερματισμός του παραπάνω αλγορίθμου αναλύονται στην ενότητα 2.3.4.

Εκτός από υπερπροσαρμογή (overfitting), η λύση μέγιστης πιθανοφάνειας μπορεί επιπλέον να δημιουργήσει απροσδιοριστίες (singularities), οι οποίες εμφανίζονται όταν η μέση τιμή μιας συνιστώσας $\boldsymbol{\mu}_k$ «καταρρέει» σε ένα από τα διανύσματα δεδομένων $\mathbf{x}_n$ με $\boldsymbol{\Sigma}_k \to \mathbf{0}$, και σε αυτή την περίπτωση η συνάρτηση πιθανοφάνειας τείνει στο άπειρο. Μια εναλλακτική προσέγγιση είναι η Μπεϋσιανή (Bayesian), κατά την οποία οι παράμετροι $\pi, \boldsymbol{\mu}, \Sigma$ εκφράζονται ως τυχαίες μεταβλητές κατανομές συζυγής εκ των προτέρων. Έχοντας και ένα σύνολο παρατηρήσεων X μπορεί να υπολογιστεί η λύση *εξ των υστέρων μέγιστης πιθανότητας* (*maximum posterior* ή MAP). Κάνοντας περαιτέρω υποθέσεις ανεξαρτησίας, η από κοινού κατανομή όλων των τυχαίων μεταβλητών ανάγεται σε πρόβλημα *μεταβολικού συμπερασμού* (*variational inference*) [11]. Παρόλα αυτά, παρατηρήσαμε ότι στην πράξη το επεκτατικό μας μοντέλο που περιγράφεται στην ενότητα 2.3.3 δεν παρουσιάζει απροσδιοριστίες, έτσι δεν χρειάστηκε να εισάγουμε και εκ των προτέρων πιθανότητες για τις παραμέτρους.

Για το υπόλοιπο του κεφαλαίου, θα εστιάσουμε στην ειδική περίπτωση των *σφαιρικών* ή *ισοτροπικών* κανονικών κατανομών, με μήτρα συνδιακύμανσης $\boldsymbol{\Sigma}_k = \sigma_k^2 I$. Σε αυτή την ειδική περίπτωση, η εξίσωση μερικής παραγώγου (2.7) απλουστεύεται στη μορφή

$$\sigma_k^2 = \frac{1}{DN_k} \sum_{n=1}^{N} \gamma_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2. \tag{2.8}$$

Υπάρχουν τρείς κύριου λόγοι που μας κάνουν να επιλέξουμε ένα πιο περιορισμένο μοντέλο όπως το σφαιρικό. Πρώτων, σε σύγκριση με τον απλό αλγόριθμο

k-means, το μοντέλο μας είναι αρκετά πιο ευέλικτο, καθώς μπορεί να εκφράσει τον πληθυσμό και την έκταση κάθε κέντρο, μέσω των συντελεστών του μίγματος π_k και της τυπικής απόκλισης σ_k αντίστοιχα. Δεύτερον, το μοντέλο είναι ιδιαίτερα αποδοτικό καθώς η τυπική απόκλιση σ_k είναι απλά ένας αριθμός, ενώ η αναπαράσταση του σχήματος των κέντρων με μια πλήρη μήτρα συνδιακύμανσης $\mathbf{\Sigma}_k$ θα καθιστούσε το μοντέλο τετραγωνικό ως προς τις διαστάσεις D, πράγμα απαγορευτικό για τους χώρους πολλών διαστάσεων της εφαρμογής μας. Τρίτον, η ένωση πολλών σφαιρικών κανονικών κατανομών μπορεί να περιγράψει συνολικά ένα σύνολο δεδομένων αυθαίρετου σχήματος.

2.3.2 Διαγραφή επικαλυπτώμενων κέντρων

Ένα γνωστό πρόβλημα στις περισσότερες τεχνικές ομαδοποίησης καθώς και στα μοντέλα μιγμάτων το οποίο δεν έχει αναφερθεί έως τώρα είναι η εκτίμηση του πραγματικού αριθμού των ομάδων, συνιστωσών ή κέντρων. Στον αλγόριθμο k-means, η υπερεκτίμηση του αριθμού των κέντρων οδηγεί σε υπερκατάτμηση του χώρου. Στο μοντέλο μείγματος κανονικών κατανομών το αποτέλεσμα είναι ακόμα χειρότερο: εφόσον το μοντέλο επιτρέπει τις συνιστώσες να επικαλύπτονται στον χώρο, ο αλγόριθμος ΕΜ μπορεί να συγκλίνει σε μια κατάσταση κατά την οποία δύο συνιστώσες είναι καθολικά ευθυγραμμισμένες, μοιράζοντας τα διανύσματα δεδομένων τους και έχοντας μη μηδενικούς συντελεστές μείγματος.

Το παραπάνω πρόβλημα μας παρότρυνε να προτείνουμε μια νέα τεχνική για *διαγραφή* συνιστωσών, καθοδηγούμενοι από ένα κριτήριο *επικάλυψης*. Επίσης, αντί να εφαρμόσουμε τη προαναφερθείσα τεχνική μετά την σύγκλιση, επιλέξαμε να εκτελούμε τη διαγραφή *δυναμικά*, αρχικοποιώντας το μοντέλο με όσο το δυνατό μεγαλύτερο αριθμό συνιστωσών και στη συνέχεια διαγράφοντας όποτε χρειάζεται κατά την φάση επανεκτίμησης των παραμέτρων του μοντέλου. Πρακτικά, η ιδέα αυτή εισάγει ένα τρίτο βήμα ***P-step*** στην επαναληπτική διαδικασία του ΕΜ, ένα βήμα που εκτελείται αμέσως μετά τα δύο αρχικά E- and M-steps σε κάθε επανάληψη.

Έστω ότι p_k είναι η συνάρτηση που εκφράζει την συνεισφορά της συνιστώσας k στην κατανομή του μείγματος κανονικών κατανομών της σχέσης (2.1), όπου

$$p_k(\mathbf{x}) = \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{\Sigma}_k) \tag{2.9}$$

για $\mathbf{x} \in \mathbb{R}^D$. Τότε το p_k μπορεί να ερμηνευθεί ως μια αναπαράσταση της ίδιας της συνιστώσας k. Πρέπει εδώ να σημειωθεί ότι η ποσότητα p_k δεν εκφράζει κανονικοποιημένη κατανομή, εκτός αν ισχύει ότι $\pi_k = 1$. Ας θεωρήσουμε τώρα ότι το $\mathbf{x}$ δεν εκφράζει ένα διάνυσμα δεδομένων αλλά μια άλλη συνιστώσα του μείγματος. Αν το $\mathbf{x}$ ήταν μια τυχαία μεταβλητή με κατανομή που δίνεται από το $q(\mathbf{x})$, μπορούμε επίσης να υπολογίσουμε την ποσότητα $\mathbb{E}_{\mathbf{x}}\{p_k(\mathbf{x})\}$. Γενικότερα, ακόμα και αν το

$q(\mathbf{x})$ δεν είναι κανονικοποιημένο, μπορούμε να χρησιμοποιήσουμε την ποσότητα $\langle p_k, q\rangle$, όπου

$$\langle p, q\rangle = \int p(\mathbf{x})q(\mathbf{x})\mathrm{d}\mathbf{x} \tag{2.10}$$

είναι το L^2 *εσωτερικό γινόμενο* δύο πραγματικών και τετραγωνικά ολοκληρώσιμων συναρτήσεων p, q—χωρίς πάλι να είναι απαραίτητα κατανομές—όπου η ολοκλήρωση ορίζεται στον $\mathbb{R}^D$. Η αντίστοιχη L^2 *νόρμα* της συνάρτησης p δίνεται από τη σχέση $\|p\| = \sqrt{\langle p, p\rangle}$. Αν τα p, q οριστούν ως κανονικές κατανομές, τότε το ολοκλήρωμα της σχέσης (2.10) μπορεί να υπολογιστεί αριθμητικά σύμφωνα με το παρακάτω θεώρημα, το οποίο διατυπώνουμε στην γενική πολυμεταβλητή του περίπτωση.

Θεώρημα 1. *Έστω* $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{a}, A)$ *και* $q(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{b}, B)$ *όπου* $\mathbf{x} \in \mathbb{R}^D$*. Τότε*

$$\langle p, q\rangle = \mathcal{N}(\mathbf{a}|\mathbf{b}, A + B). \tag{2.11}$$

Απόδειξη. Από την ορισμό της πολυμεταβλητής κανονικής κατανομής,

$$\langle p, q\rangle = \frac{1}{(2\pi)^D}\frac{1}{|AB|^{1/2}}\int \exp\left\{-\frac{E(\mathbf{x})}{2}\right\}\mathrm{d}\mathbf{x}, \tag{2.12}$$

όπου

$$E(\mathbf{x}) = (\mathbf{x} - \mathbf{a})^{\mathsf{T}}A^{-1}(\mathbf{x} - \mathbf{a}) + (\mathbf{x} - \mathbf{b})^{\mathsf{T}}B^{-1}(\mathbf{x} - \mathbf{b}). \tag{2.13}$$

Με συμπλήρωση του τετραγώνου [1] παίρνουμε,

$$E(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}C^{-1}(\mathbf{x} - \boldsymbol{\mu}) + (\mathbf{a} - \mathbf{b})^{\mathsf{T}}D^{-1}(\mathbf{a} - \mathbf{b}), \tag{2.14}$$

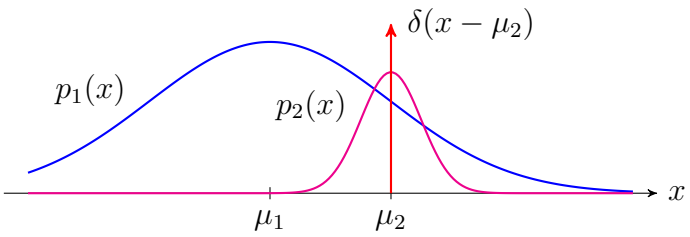
όπου $C^{-1} = A^{-1} + B^{-1}$, $D = A + B$ and $\boldsymbol{\mu} = C(A^{-1}\mathbf{a} + B^{-1}\mathbf{b})$. Το εκθετικό του δεύτερου όρου μπορεί να βγει από το ολοκλήρωμα ως σταθερά. Ο πρώτος όρος είναι το εκθετικό κομμάτι της $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, C)$, συνεπώς ολοκλήρωση ως προς $\mathbf{x}$ δίνει το συντελεστή κανονικοποίησης $(2\pi)^{D/2}|C|^{1/2}$. Συνεπώς

$$\langle p, q\rangle = \frac{1}{(2\pi)^{D/2}}\left(\frac{|C|}{|AB|}\right)^{1/2}\exp\left\{-\frac{1}{2}(\mathbf{a} - \mathbf{b})^{\mathsf{T}}D^{-1}(\mathbf{a} - \mathbf{b})\right\}. \tag{2.15}$$

Από εδώ καταλήγουμε στη σχέση (2.11) καθώς

$$\begin{aligned} |D| &= |A + B| = |A||I + A^{-1}B| = |A||I + BA^{-1}| && (2.16)\\ &= |AB||B^{-1} + A^{-1}| = |AB||C^{-1}|, && (2.17)\end{aligned}$$

όπου η τελευταία ισότητα της σχέσης (2.16) είναι γνωστός τύπος για τις ορίζουσες πινάκων [1]. □



Σχήμα 2.2: *«Δειγματοληψία» μιας απλωμένης συνιστώσας* $p_1(x) = \pi_1 \mathcal{N}(x|\mu_1, \sigma_1^2)$ *μέσω μιας μικρότερης,* $p_2(x) = \pi_2 \mathcal{N}(x|\mu_2, \sigma_2^2)$*, σε μία διάσταση. Όταν η μικρή συνιστώσα συρρικνώνεται σε ένα σημείο, η* $p_2(x)$ *καταρρέει στη συνάρτηση Dirac* $\delta(x-\mu_2)$*, και το εσωτερικό γινόμενο* $\langle p_1, p_2 \rangle$ *σε* $p_1(\mu_2)$*.*

Συνεπώς, με δεδομένες τις συνιστώσες που αναπαρίστανται από τα p_i, p_k, η *επικάλυψη* τους στο χώρο, όπως υπολογίζεται από το εσωτερικό γινόμενο $\langle p_i, p_k \rangle$, δίνεται από τη σχέση

$$\langle p_i, p_k \rangle = \pi_i \pi_k \mathcal{N}(\boldsymbol{\mu}_i | \boldsymbol{\mu}_k, (\sigma_i^2 + \sigma_k^2) I) \tag{2.18}$$

υποθέτοντας σφαιρικές κανονικές κατανομές. Η επικάλυψη μπορεί να υπολογιστεί σε χρόνο $O(D)$, απαιτώντας μονάχα μία D-διάστατη διανυσματική πράξη $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_k\|^2$, ενώ το μέτρο

$$\|p_i\|^2 = \langle p_i, p_i \rangle = \pi_i^2 (4\pi\sigma_i^2)^{-D/2} \tag{2.19}$$

υπολογίζεται σε σταθερό χρόνο $O(1)$. Αν τώρα η συνάρτηση q αναπαριστά μια οποιαδήποτε συνιστώσα ή κέντρο, η σχέση (2.18) μας οδηγεί στη γενίκευση της σχέσης (2.3) ώστε να οριστεί η ποσότητα

$$\hat{\gamma}_k(q) = \frac{\langle q, p_k \rangle}{\sum_{j=1}^{K} \langle q, p_j \rangle}, \tag{2.20}$$

έτσι ώστε $\hat{\gamma}_{ik} = \hat{\gamma}_k(p_i) \in [0, 1]$ είναι η *γενικευμένη υπευθυνότητα* της συνιστώσας k για την *συνιστώσα* i. Στην προκειμένη περίπτωση, η συνάρτηση p_i εκφράζει ένα *γενικευμένο διάνυσμα δεδομένων*, με κέντρο το μ_i, συντελεστή βάρους π_i και με χωρική έκταση ίση με σ_i . Εύκολα παρατηρεί κανείς ότι η σχέση (2.20) ανάγεται στην (2.3) όταν το q καταρρέει σε συνάρτηση Dirac, *δειγματοληπτώντας* τις συναρτήσεις των συνιστωσών p_k, όπως φαίνεται και στο παράδειγμα του Σχήματος 2.2 σε μία διάσταση.

Σύμφωνα με τους παραπάνω ορισμούς, η ποσότητα $\hat{\gamma}_{ii}$ είναι η υπευθυνότητα της συνιστώσας i για *τον εαυτό της*. Γενικότερα, έχοντας δεδομένο ένα σύνολο $\mathcal{K}$ από συνιστώσες και μία άλλη συνιστώσα $i \notin \mathcal{K}$, έστω

$$\rho_{i,\mathcal{K}} = \frac{\hat{\gamma}_{ii}}{\hat{\gamma}_{ii} + \sum_{j \in \mathcal{K}} \hat{\gamma}_{ij}} = \frac{\|p_i\|^2}{\|p_i\|^2 + \sum_{j \in \mathcal{K}} \langle p_i, p_j \rangle}. \tag{2.21}$$

Η ποσότητα $\rho_{i,\mathcal{K}} \in [0, 1]$ είναι η υπευθυνότητα της συνιστώσας i για τον εαυτό της *σε σχέση με το σύνολο* $\mathcal{K}$. Αν η ποσότητα $\rho_{i,\mathcal{K}}$ είναι μεγάλη, τότε η συνιστώσα

Algorithm 1: Διαγραφή επικαλυπτώμενων συνιστωσών (P-Step)

input : set of components $\mathcal{C} \subseteq \{1, \ldots, K\}$ at current iteration
output: updated set of components $\mathcal{C}' \subseteq \mathcal{C}$, after purging

$\mathcal{K} \leftarrow \emptyset$ // set of components to keep
Sort $\mathcal{C}$ such that $i \leq k \rightarrow \pi_i \geq \pi_k$ for $i, k \in \mathcal{C}$ // re-order components i…
foreach $i \in \mathcal{C}$ **do** // …in descending order of π_i
  if $\rho_{i,\mathcal{K}} \geq \tau$ **then** // compute $\rho_{i,\mathcal{K}}$ by (2.21)
    $\mathcal{K} \leftarrow \mathcal{K} \cup i$ // keep i if it does not overlap with $\mathcal{K}$
$\mathcal{C}' \leftarrow \mathcal{K}$ // updated components

i μπορεί να «ερμηνεύσει» τον εαυτό της καλύτερα από ότι *το σύνολο των συνιστωσών* $\mathcal{K}$. Στην αντίθετη περίπτωση, η συνιστώσα i φαίνεται να είναι περιττή. Έτσι, αν το σύνολο $\mathcal{K}$ αναπαριστά τις συνιστώσες που έχουμε ήδη αποφασίσει να *διατηρηθούν* μέχρι στιγμής, είναι λογικό να *διαγράψουμε* τη συνιστώσα i αν η ποσότητα $\rho_{i,\mathcal{K}}$ πέσει κάτω από ένα *κατώφλι επικάλυψης* $\tau \in [0, 1]$ και σε αυτή την περίπτωση λέμε ότι η συνιστώσα i *επικαλύπτεται* από το σύνολο $\mathcal{K}$.

Διατρέχουμε τις συνιστώσες σειριακά, ταξινομώντας τες πρώτα σε φθίνουσα σειρά ως προς τους συντελεστές μίγματος. Αρχίζουμε έτσι από την πολυπληθέστερη συνιστώσα, η οποία διατηρείται πάντα. Στη συνέχεια για κάθε συνιστώσα αποφασίζουμε είτε να την *διατηρήσουμε* είτε να την *διαγράψουμε*, αν έχει επικάλυψη με τις συνιστώσες που ήδη έχουμε κρατήσει. Αν θέσουμε το κατώφλι $\tau \geq \frac{1}{2}$ μπορούμε να είμαστε βέβαιοι ότι δεν θα διατηρήσουμε δύο πανομοιότυπες συνιστώσες. Η διαδικασία του βήματος *P-step* περιγράφεται στον Αλγόριθμο 1, στον οποίο υποθέτουμε ότι το σύνολο $\mathcal{C} \subseteq \{1, \ldots, K\}$ περιέχει τις συνιστώσες της τρέχουσας επανάληψης. Στην προκειμένη περίπτωση ο αριθμός K αναφέρεται στο αρχικό πλήθος των συνιστωσών. Το *πρέχων πλήθος*, $C = |\mathcal{C}| \leq K$, μειώνεται σε κάθε επανάληψη. Ο Αλγόριθμος 1 είναι τετραγωνικός ως προς το πλήθος C, στην Ενότητα 2.4 όμως παραθέτουμε μια προσεγγιστική και αποδοτική λύση.

Για να περιγράψουμε συνολικά τον αλγόριθμο EM μαζί με το P-step, πρέπει να επανεκτιμήσουμε τα π_k, $\boldsymbol{\mu}_k$, σ_k σύμφωνα με τις σχέσεις (2.5), (2.6), (2.8) μόνο για τα $k \in \mathcal{C}$ στο M-step. Παρομοίως, στο E-step, υπολογίζουμε τα $\gamma_{nk} = \gamma_k(\mathbf{x}_n)$ για όλα τα $n = 1, \ldots, N$ αλλά μοναχά για τις συνιστώσες $k \in \mathcal{C}$, σύμφωνα με τη σχέση

$$\gamma_k(\mathbf{x}) = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j \in \mathcal{C}} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}, \tag{2.22}$$

όπου το άθροισμα του παρονομαστή της σχέσης (2.3) έχει τώρα περιοριστεί στο σύνολο συνιστωσών $\mathcal{C}$.

2.3.3 Επέκταση συνιστωσών

Όταν μια συνιστώσα, έστω η i, σβηστεί, τα διανύσματα δεδομένων που «περιγράφονταν» καλύτερα από τη συνιστώσα i πριν τη διαγραφή θα πρέπει τώρα να διανεμηθούν στις γειτονικές συνιστώσες που διατηρήθηκαν. Αναμένεται οι συνιστώσες αυτές, έστω k, να είναι εκείνες με την μεγαλύτερη τιμή στην ποσότητα ρ_{ik}. Εκείνες οι συνιστώσες, θα πρέπει τώρα να *επεκταθούν* και να καλύψουν τον χώρο που περιλαμβάνει τα αντίστοιχα διανύσματα δεδομένων. Γι αυτό τον σκοπό, επαναδιατυπώνουμε τη σχέση (2.8) έτσι ώστε να *υπερεκτιμάται* η έκταση της κάθε συνιστώσας, στο βαθμό που εκείνη δεν επικαλύπτεται με τις γειτονικές της. Με αυτό τον τρόπο, οι συνιστώσες θα τείνουν να καλύψουν όσον το δυνατόν τον κενό χώρο, γεγονός που ως έναν βαθμό καθορίζει και τον ρυθμό σύγκλισης του όλου αλγορίθμου.

Συγκεκριμένα, για μία συνιστώσα k από το σύνολο $\mathcal{C}$, χρησιμοποιούμε τον αρχικό ορισμό της υπευθυνότητας $\gamma_k(\mathbf{x})$ της σχέσης (2.22) για να διαχωρίσουμε το σύνολο των διανυσμάτων δεδομένων $\mathcal{P} = \{1, \ldots, N\}$ στο σύνολο των *εσωτερικών σημείων* της

$$\underline{\mathcal{P}}_k = \{n \in \mathcal{P} : \gamma_{nk} = \max_{j \in \mathcal{C}} \gamma_{nj}\}, \tag{2.23}$$

τα οποία περιλαμβάνονται μέσα στο *Βορονόι κελί* της συνιστώσας k, και το σύνολο των *εξωτερικών σημείων*, $\overline{\mathcal{P}}_k = \mathcal{P} \backslash \underline{\mathcal{P}}_k$. Ο κανόνας επανεκτίμησης της σχέσης (2.8) μπορεί τώρα να αναλυθεί ως

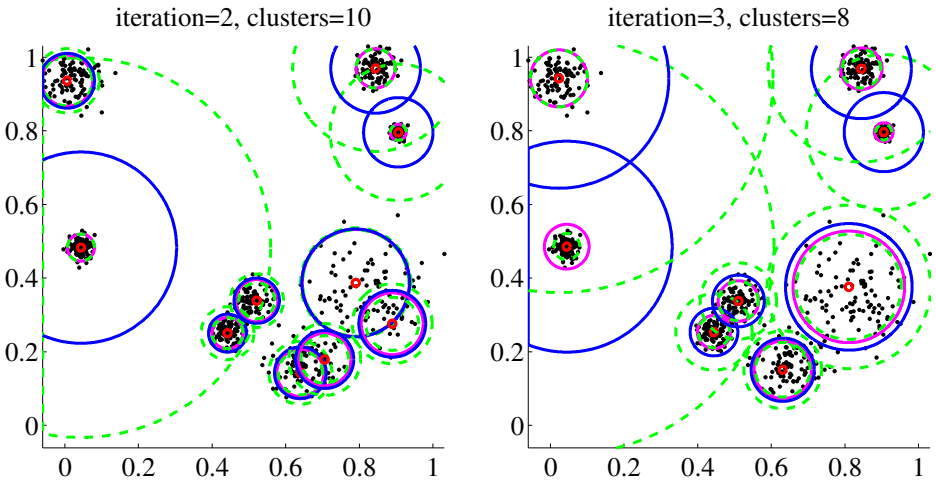
$$D\sigma_k^2 = \frac{\underline{N}_k}{N_k}\underline{\Sigma}_k + \frac{\overline{N}_k}{N_k}\overline{\Sigma}_k, \tag{2.24}$$

όπου $\underline{N}_k = \sum_{n \in \underline{\mathcal{P}}_k} \gamma_{nk}$, $\overline{N}_k = \sum_{n \in \overline{\mathcal{P}}_k} \gamma_{nk}$, και

$$\underline{\Sigma}_k = \frac{1}{\underline{N}_k} \sum_{n \in \underline{\mathcal{P}}_k} \gamma_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2, \quad \overline{\Sigma}_k = \frac{1}{\overline{N}_k} \sum_{n \in \overline{\mathcal{P}}_k} \gamma_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2. \tag{2.25}$$

Καθώς ισχύει ότι $\underline{N}_k + \overline{N}_k = N_k$, τα βάρη στον γραμμικό συνδυασμό της σχέσης (2.24) συνεχίζουν να αθροίζουν στην μονάδα. Το *εσωτερικό άθροισμα* $\underline{\Sigma}_k$ εκφράζει μια μέση σταθμισμένη απόσταση από το μέσο $\boldsymbol{\mu}_k$ για τα δεδομένα που «περιγράφονταν» καλύτερα από τη συνιστώσα k, συνεπώς ταιριάζει ακριβώς με τα διανύσματα δεδομένων του αντίστοιχου κέντρου. Αντίθετα, το *εξωτερικό άθροισμα* $\overline{\Sigma}_k$ παίζει έναν παρόμοιο ρόλο για τα δεδομένα τα οποία «περιγράφονται» καλύτερα από τα υπόλοιπα κέντρα. Είναι λογικό να ισχύει τυπικά ότι $\underline{\Sigma}_k < \overline{\Sigma}_k$ αν και η προηγούμενη ανισότητα δεν ισχύει πάντα, ιδιαίτερα σε περιπτώσεις υπερβολικής επικάλυψης των συνιστωσών. Στη μετατροπή που προτείνουμε, κάνουμε το σταθμισμένο άθροισμα να τείνει προς το εξωτερικό άθροισμα $\overline{\Sigma}_k$ στη σχέση (2.24) ως εξής,

$$D\sigma_k^2 = w_k\underline{\Sigma}_k + (1 - w_k)\overline{\Sigma}_k, \tag{2.26}$$



Σχήμα 2.3: *Επέκταση συνιστωσών για τις επαναλήψεις* 2 *και* 3 *για το παράδειγμα του Σχήματος 2.1. Μπλε κύκλοι: δύο τυπικές αποκλίσεις με επέκταση που δίνεται από τη σχέση (2.26) και* $\lambda = 0.25$*, όπως και στο Σχήμα 2.1. Ματζέντα: χωρίς επέκταση (2.8); διακεκομμένοι πράσινοι: συνεισφορές εσωτερικών και εξωτερικών αθροισμάτων.*

όπου $w_k = \frac{N_k}{N_k}(1-\lambda)$ και το $\lambda \in [0,1]$ είναι μια *παράμετρος επέκτασης*. Είναι εμφανές ότι όταν $\lambda = 0$ η σχέση (2.26) γυρίζει πίσω στην μορφή της σχέσης (2.8), ενώ όταν $\lambda = 1$ χρησιμοποιείται μοναχά το εξωτερικό άθροισμα $\overline{\Sigma}_k$.

Λόγω της εκθετικής φύσης της κανονικής κατανομής, το εξωτερικό άθροισμα $\overline{\Sigma}_k$ κυριαρχείται από τα εξωτερικά διανύσματα δεδομένων του συνόλου $\overline{\mathcal{P}}_k$ τα οποία είναι κοντινότερα στη συνιστώσα k, δίνοντας έτσι μια εκτίμηση για την μέγιστη δυνατή επέκταση της συνιστώσας k πριν ξεκινήσει η επικάλυψή της με τις γειτονικές της συνιστώσες. Το Σχήμα 2.3 παρουσιάζει την επέκταση για το παράδειγμα του Σχήματος 2.1. Παρατηρείται ότι οι εξωτερικοί πράσινοι κύκλοι συνήθως φτάνουν μέχρι τις γειτονικές συνιστώσες, ενώ οι εσωτερικοί ταιριάζουν ακριβώς στην κατανομή των δεδομένων του κάθε κέντρου.

2.3.4 Αρχικοποίηση και τερματισμός

Η στρατηγική μας που περιλαμβάνει διαγραφή συνιστωσών και επέκταση των υπολοίπων, προϋποθέτει ότι αρχικά ξεκινάμε με έναν αρχικά μεγάλο αριθμό από συνιστώσες—αν τέτοιου είδους γνώση υπάρχει—και τον μειώνουμε σταδιακά μέχρι τη σύγκλιση του αλγορίθμου στο σωστό αριθμό. Ο αλγόριθμος μας έχει παρουσιαστεί ως τώρα σε μια γενικότερη μορφή, ωστόσο η αρχικοποίηση και ο τερματισμός του είναι συνήθως θέματα που εξαρτώνται από το εκάστοτε πρόβλημα. Γι αυτό, παρακάτω θα ασχοληθούμε μόνο με το πιο συγκεκριμένο πρόβλημα της δημιουργίας οπτικών λεξικών μεγάλης κλίμακας.

Έχει πρόσφατα υποστηριχθεί [48][12] ότι τα πιο συχνά εμφανιζόμενα δεδομένα στις πυκνές περιοχές του χώρου είναι σε πολλές περιπτώσεις και εκείνα

με την λιγότερη πληροφορία, και αντίθετα ότι απομονωμένα σημεία μπορούν να είναι ιδιαίτερα πληροφοριακά. Έτσι, ακόμα και αν υπάρχει μια υπόνοια για τον κατάλληλο αριθμό των κέντρων που χρειάζονται για τις δεδομένες εικόνες και τη δεδομένη ποικιλία περιεχομένου, μια τυχαία επιλογή ενός ποσοστού των διανυσμάτων δεδομένων ως αρχικές θέσεις κέντρων, μπορεί να οδηγήσει σε μεγάλη απώλεια πληροφορίας.

Επιλέγουμε συνεπώς να αρχικοποιήσουμε τον αλγόριθμό μας με *όλα* τα διαθέσιμα διανύσματα δεδομένων σαν κέντρα, και άρα ισχύει ότι αρχικά $K = N$. Χρησιμοποιώντας προσεγγιστικούς αλγορίθμους για την αναζήτηση κοντινότερων γειτόνων, η παραπάνω επιλογή δεν είναι τόσο αναποτελεσματική όσο αρχικά δείχνει. Για την ακρίβεια, από την αρχικοποίηση αυτή επηρεάζεται μόνο η πρώτη επανάληψη του αλγορίθμου, καθώς από τη δεύτερη και μετά το ποσοστό των κέντρων που διατηρούνται είναι συνήθως της τάξεως του 10%. Οι *συντελεστές μείγματος* αρχικοποιούνται ομοιόμορφα. Οι *τυπικές αποκλίσεις* θέτονται αρχικά ίσες με την απόσταση του κοντινότερου γείτονα κάθε σημείου όπως αυτή εκτιμάται από τον προσεγγιστικό αλγόριθμο.

Η σύγκλιση στον αλγόριθμο EM συνήθως ανιχνεύεται από παρακολούθηση της τιμής τις συνάρτησης πιθανοφάνειας της σχέσης (2.4). Αυτή η πρακτική έχει νόημα στην προκειμένη περίπτωση μόνο αφότου ο αριθμός των κέντρων έχει σταθεροποιηθεί και δεν συνεχίζονται να διαγράφονται κέντρα. Παρόλα αυτά, έχουμε παρατηρήσει ότι, για κατασκευή λεξικών μεγάλης κλίμακας, η σύγκλιση δεν επιτυγχάνεται στην πράξη, καθώς τέτοια πειράματα συνήθως απαιτούν ώρες ή ακόμα και μέρες για να τρέξουν. Πιο σημαντικό σε αυτή την περιοχή είναι να μετρήσουμε την απόδοση του εκάστοτε λεξικού στο συγκεκριμένο πρόβλημα—στην περίπτωσή μας την αναζήτηση εικόνων—ως προς τον όγκο υπολογισμών που απαιτούνται.

2.4 Προσεγγιστικά μείγματα κανονικών κατανομών

Μετρώντας τις D-διάστατες διανυσματικές πράξεις και αγνοώντας τις επαναλήψεις, η πολυπλοκότητα του αλγορίθμου όπως παρουσιάστηκε μέχρι στιγμής είναι $O(NK)$. Συγκεκριμένα, η πολυπλοκότητα των E-step (2.22) και M-step (2.5), (2.6), (2.25)-(2.26) για κάθε επανάληψη είναι $O(NC)$, όπου $C = |\mathcal{C}| \leq K \leq N$ ειναι ο τρέχων αριθμός συνιστωσών, και η πολυπλοκότητα του P-step (Αλγόριθμος 1) είναι $O(C^2)$. Εμφανώς, η παραπάνω προσέγγιση δεν είναι εφαρμόσιμη στη πράξη για μεγάλες τιμές του C, ιδιαίτερα όταν το K είναι της τάξεως του αριθμού των δεδομένων εκπαίδευσης N.

Όπως συμβαίνει και στη δημοσίευση [85], η προσεγγιστική εκδοχή του προτεινόμενου αλγορίθμου ομαδοποίησης με μοντέλο μείγματος κανονικών κατανομών προϋποθέτει τη δεικτοδότηση του συνόλου $\mathcal{C}$ των κέντρων σύμφωνα με το διάνυσμα μέσου όρου τους $\boldsymbol{\mu}_k$ και έπειτα εφαρμογή προσεγγιστικής αναζήτηση κοντι-

Algorithm 2: Αυξητικοί m-κοντινότεροι γείτονες (N-step)

input : m best neighbors $\mathcal{B}(\mathbf{x}_n)$ found so far for $n = 1, \ldots, N$
output: updated m best neighbors $\mathcal{B}'(\mathbf{x}_n)$ for $n = 1, \ldots, N$

for $n = 1, \ldots, N$ **do** // for all data points
  $\mathcal{B}(\mathbf{x}_n) \leftarrow \mathcal{C} \cap \mathcal{B}(\mathbf{x}_n)$ // ignore purged neighbors
  $(\mathcal{N}, d) \leftarrow \mathsf{NN}_m(\mathbf{x}_n)$ // $\mathcal{N}$: m-NN of $\mathbf{x}_n$; d: distances to $\mathbf{x}_n$
  // (such that $d_k^2 = \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$ for $k \in \mathcal{N}$)
  for $k \in \mathcal{B}(\mathbf{x}_n) \setminus \mathcal{N}$ **do** // for all previous neighbors…
    $d_k^2 \leftarrow \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$ // …find distance after $\boldsymbol{\mu}_k$ update (M-step)
  $\mathcal{A} \leftarrow \mathcal{B}(\mathbf{x}_n) \cup \mathcal{N}$ // for all previous and new neighbors…
  for $k \in \mathcal{A}$ **do** // …compute unnormalized…
    $p_k \leftarrow (\pi_k/\sigma_k^D) \exp\{-d_k^2/(2\sigma_k^2)\}$ // …responsibility of k for $\mathbf{x}_n$
  Sort $\mathcal{A}$ such that $i \leq k \rightarrow p_i \geq p_k$ for $i, k \in \mathcal{A}$ // keep the top-ranking…
  $\mathcal{B}'(\mathbf{x}_n) \leftarrow \mathcal{A}[1, \ldots, m]$ // …m neighbors

νότερου γείτονα για κάθε ένα από τα διανύσματα δεδομένων $\mathbf{x}_n$, πριν το E-step της κάθε επανάληψης. Η διαδικασία αυτή απαιτεί πολυπλοκότητα $O(C\alpha(C))$ για τη δεικτοδότηση και $O(N\alpha(C))$ για την αναζήτηση, όπου το α εκφράζει την πολυπλοκότητα ενός ερωτήματος ως συνάρτηση του μεγέθους του δεικτοδοτημένου συνόλου. Για παράδειγμα, το $\alpha(C) = \log C$ για τις συνήθεις μεθόδους βασισμένες σε δενδρικές δομές. Οι υπευθυνότητες γ_{nk} υπολογίζονται έπειτα σύμφωνα με τη σχέση (2.22) για κάθε $n = 1, \ldots, N$, $k \in \mathcal{C}$, αλλά με τις αποστάσεις από τα κέντρα να δίνονται τώρα από τη σχέση

$$d_m^2(\mathbf{x}, \boldsymbol{\mu}_k) = \begin{cases} \|\mathbf{x} - \boldsymbol{\mu}_k\|^2, & \text{if } k \in \mathsf{NN}_m(\mathbf{x}) \\ 0, & \text{otherwise,} \end{cases} \tag{2.27}$$

όπου το $\mathsf{NN}_m(\mathbf{x}) \subseteq \mathcal{C}$ δηλώνει την προσεγγιστική γειτονιά των m κοντινότερων γειτόνων του διανύσματος δεδομένων $\mathbf{x} \in \mathbb{R}^D$. Κάθε συνιστώσα k που επιστρέφεται σαν κοντινότερος γείτονας ενός διανύσματος δεδομένων $\mathbf{x}_n$ στη συνέχεια επανεκτιμάται σταδιακά υπολογίζοντας τις συνεισφορές γ_{nk}, $\gamma_{nk}\mathbf{x}_n$, $\gamma_{nk}\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$ στα N_k (άρα και τα π_k στη σχέση (2.5)), $\boldsymbol{\mu}_k$ στη σχέση (2.6), σ_k^2 στις σχέσεις (2.25)-(2.26), αντίστοιχα, συνεισφορές οι οποίες οφείλονται στο διάνυσμα δεδομένων $\mathbf{x}_n$. Συνεπώς δεν προστίθεται περαιτέρω πολυπλοκότητα από το M-step. Παρομοίως, το P-step περιέχει ερωτήματα, ένα για κάθε μια από τις συνιστώσες $\boldsymbol{\mu}_k$, στην ίδια δομή δεικτοδότησης, διαδικασία με πολυπλοκότητα $O(C\alpha(C))$. Από όλα τα παραπάνω, καταλήγουμε στο ότι η συνολική πολυπλοκότητα ανά επανάληψη είναι $O(N\alpha(C))$.

Πηγαίνοντας ένα βήμα παραπέρα, και εμπνεόμενοι από τη δημοσίευση [61], δεν χρησιμοποιούμε τη προσεγγιστική αναζήτηση μόνο για επιτάχυνση της δια-

δικασίας ομαδοποίησης, αλλά επίσης εκμεταλλευόμαστε την επαναληπτική φύση του αλγορίθμου ομαδοποίησης και βελτιώνουμε την διαδικασία αναζήτησης. Για να το πετύχουμε αυτό, διατηρούμε μια λίστα από τους m *καλύτερους γείτονες* $\mathcal{B}(\mathbf{x}_n)$ που έχουν βρεθεί μέχρι στιγμής για κάθε διάνυσμα δεδομένων $\mathbf{x}_n$, και την επαναχρησιμοποιούμε ανάμεσα στις επαναλήψεις. Οι αποστάσεις για κάθε κοντινότερο γείτονα της τρέχουσα επανάληψης μας δίνονται από την ίδια την αναζήτηση, ενώ οι αποστάσεις για τους προηγούμενους κοντινότερους γείτονες που δεν επιστράφηκαν στη τρέχουσα αναζήτηση πρέπει να επαναϋπολογιστούν μετά την επανεκτίμηση των κέντρων (2.6) κατα το M-step. Η διαδικασία ενημέρωσης της λίστας των καλύτερων γειτόνων παρουσιάζεται στον Αλγόριθμο 2.

Ο παραπάνω *αυξητικός* αλγόριθμος για τους m-κοντινότερους γείτονες είναι μια γενίκευση της προσέγγισης της δημοσίευσης [61], το οποίο περιορίζει το $m = 1$ και το εφαρμόζει μόνο προσθετικά στον αλγόριθμο k-means. Μπορεί να θεωρηθεί και σαν ένα επιπλέον *N-step* στον συνολικά προσεγγιστικό αλγόριθμο ομαδοποίησης, βήμα που πραγματοποιείται πριν το E-step—και το οποίο υπολογίζει ως υποπροϊόν και τις αντίστοιχες υπευθυνότητες. Το επιπλέον κόστος ειναι m διανυσματικές πράξεις για υπολογισμό των αποστάσεων σε κάθε επανάληψη, αλλά είναι ένα κόστος που κερδίζουμε μειώνοντας την ακρίβεια προσέγγισης σε κάθε αναζήτηση κοντινότερων γειτόνων. Όπως αναφέρεται και στη δημοσίευση [61], το σκεπτικό είναι ότι παρακολουθώντας τους μέχρι στιγμής καλύτερους γείτονες, απαιτείται πολύ λιγότερη προσπάθεια για την αναζήτηση νέων γειτόνων.

Η παραπάνω παρατήρηση μας οδηγεί σε μια *νέα θεώρηση* της προσεγγιστικής μεθόδου μειγμάτων κανονικών κατανομών. Αντί να αντιλαμβανόμαστε τον αλγόριθμο σαν μια ακολουθία ενημέρωσης των κέντρων η οποία διακόπτεται από αναζητήσεις κοντινότερων γειτόνων, μπορούμε να τον δούμε σαν μια μοναδική αναζήτηση κοντινότερων γειτόνων ανά διάνυσμα δεδομένων, η οποία διακόπτεται από ενημερώσεις των κέντρων οι οποίες μετακινούν τους γείτονες μέχρι τη σύγκλιση. Έτσι συνολικά καταφέρνουμε να έχουμε *αυξημένη ακρίβεια σε συνδυασμό και με μεγαλύτερη ταχύτητα*. Το παραπάνω αποτέλεσμα είναι περισσότερο εμφανές στην προκειμένη περίπτωση, συγκριτικά με τη δημοσίευση [61], καθώς εδώ διατηρούνται περισσότεροι από ένας γείτονες ανάμεσα στις επαναλήψεις.

2.5 Πειράματα και συγκρίσεις

2.5.1 Πρωτόκολλο πειραμάτων

Στην παρούσα ενότητα ερευνούμε την συμπεριφορά του αλγορίθμου AGM κάτω από διαφορετικές καταστάσεις και τιμές παραμέτρων, και έπειτα συγκρίνουμε την ταχύτητα και απόδοσή του σε σχέση με τις κορυφαίες τεχνικές (state of the art) για κατασκευή οπτικών λεξικών μεγάλης κλίμακας με εφαρμογή την

αναζήτηση εικόνων, συγκεκριμένα τις τεχνικές AKM [85] και RAKM [61].

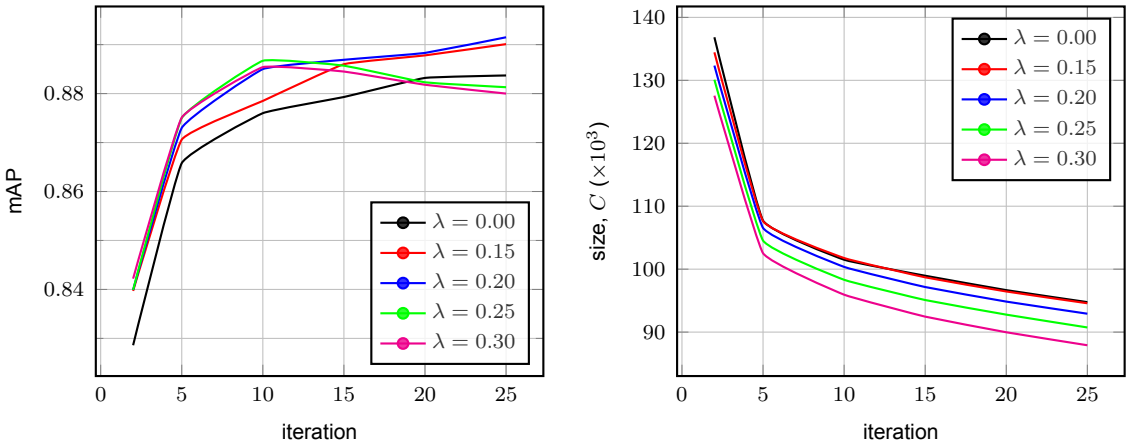
Βάσεις Εικόνων. Χρησιμοποιούμε δύο δημόσιες και ευραίως διαδεδομένες βάσης εικόνων: τη βάση *Oxford buildings*[1] [85] και τη βάση *world cities*[2] [103]. Η πρώτη αποτελείται από $5,062$ εικόνες και προσημειωμένα δεδομένα (ground truth) για 11 διαφορετικά κτίρια της πόλης της Οξφόρδης, με 5 εικόνες αναζήτησης για το κάθε ένα. Θα αναφερόμαστε στο προσημειωμένο τμήμα της δεύτερης βάσης ως *Barcelona* dataset, ένα σύνολο εικόνων που αποτελείται από 927 εικόνες ομαδοποιημένες σε 17 σημαντικές τοποθεσίες της πόλης της Βαρκελώνης, για κάθε ένα από τα οποία έχουν επίσης προσημειωθεί 5 εικόνες αναζήτησης. Το σύνολο των *εικόνων περίσπασης* (distractor images) της βάσης *world cities* αποτελείται από 2 εκατομμύρια εικόνες από διάφορες πόλεις του κόσμου. Στα παρακάτω πειράματα χρησιμοποιούμε το πρώτο εκατομμύριο ως εικόνες περίσπασης.

Οπτικά λεξικά. Αρχικά εξάγουμε τα σημεία ενδιαφέροντος και τους αντίστοιχους τοπικούς περιγραφείς SURF [10] για κάθε εικόνα. Συνολικά έχουμε 10 εκατομμύρια περιγραφείς απο τη βάση *Oxford buildings* και 550 χιλιάδες από τη βάση *Barcelona*. Οι περιγραφείς αυτοί, οι οποίοι είναι διανύσματα διάστασης $D = 64$, είναι τα δεδομένα που χρησιμοποιούμε για την κατασκευή των οπτικών λεξικών. Κατασκευάζουμε *εξειδικευμένα* (*specific*) λεξικά χρησιμοποιώντας όλους τους περιγραφείς από τη μικρότερη βάση *Barcelona*, τους οποίους έπειτα χρησιμοποιούμε για αναζήτηση στην ίδια την βάση, με σκοπό να ρυθμίσουμε τις παραμέτρους του μοντέλου και για να συγκρίνουμε την ταχύτητα του προτεινόμενου αλγορίθμου με τις άλλες τεχνικές. Κατασκευάζουμε επίσης *γενικά* (*generic*) λεξικά χρησιμοποιώντας 6.5 εκατομμύρια περιγραφείς από ένα ανεξάρτητο σύνολο 15 χιλιάδων εικόνων πόλης, και συγκρίνουμε την απόδοση το προτεινόμενου αλγορίθμου έναντι των λοιπών τεχνικών στη βάση *Oxford buildings* μαζί με εικόνες περίσπασης.

Πρωτόκολλο πειραμάτων. Μετά την κατασκευή των οπτικών λεξικών, χρησιμοποιούμε τη βιβλιοθήκη FLANN για να αναθέσουμε οπτικές λέξεις σε κάθε έναν από τους περιγραφείς των εικόνες της εκάστοτε βάσης, κρατώντας μόνο τον έναν κοντινότερο με την Ευκλείδεια απόσταση γείτονα, και έπειτα κατασκευάζουμε μια δομή δεικτοδότησης *ανεστραμμένου αρχείου* (*inverted file*) όπως περιγράφηκε στο Κεφάλαιο 1. Εφαρμόζουμε *πολλαπλή ανάθεση* (*soft assignment*) για τους περιγραφείς των εικόνων αναζήτησης μόνο, όπως και στη δημοσίευση [41], κρατώντας τους κοντινότερους 1, 3, ή 5 γείτονες. Για τις τεχνικές AKM και RAKM, οι αποστάσεις μετατρέπονται σε ομοιότητες κεντράροντας μια Κανονική κατανομή σε κάθε κοντινότερο γείτονα που επιστρέφει η FLANN, με σταθερή τυπική απόκλιση σ όπως και στη δημοσίευση [86], και έπειτα κατατάσσοντας τις εικόνες με φθίνουσα σειρά ομοιότητας. Πειραματικά βρήκαμε ότι η τιμή $\sigma = 0.01$ δίνει τα καλύτερα αποτελέσματα και στις δύο προαναφερθείσες τεχνικές. Για την προτεινόμενη

[1] `http://www.robots.ox.ac.uk/~vgg/data/oxbuildings/`

[2] `http://image.ntua.gr/iva/datasets/wc/`



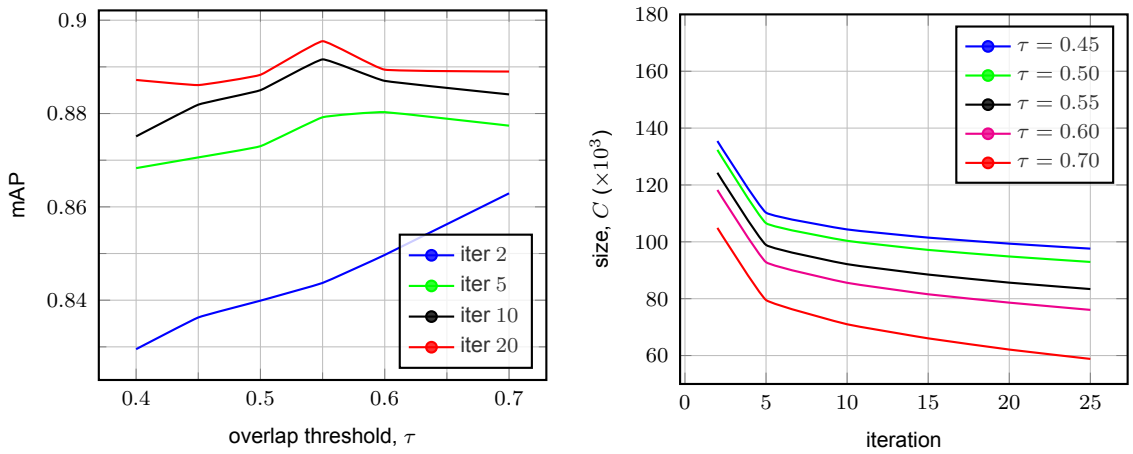
Σχήμα 2.4: *Απόδοση (mAP) για εξειδικευμένα λεξικά στη βάση* Barcelona *(αριστερά) και μεγέθη λεξικών* C *(δεξιά) ως προς τον αριθμό των επαναλήψεων κατά την εκμάθηση, για διαφορετικές τιμές του* παράγοντα επέκτασης λ *και σταθερό* $\tau = 0.5$.

τεχνική μας, AGM, χρησιμοποιούμε για κάθε ένα από τα κέντρα k τις παραμέτρους του μοντέλου μείγματος $\pi_k, \boldsymbol{\mu}_k, \sigma_k$ που μάθαμε κατά την κατασκευή, ακριβώς όπως στις γραμμές 7-10 του Αλγορίθμου 2. Έπειτα ανακτούμε τις λίστες των οπτικών λέξεων της εικόνας αναζήτησης από το ανεστραμμένο αρχείο και κατατάσσουμε τις εικόνες τις βάσης με βάση το εσωτερικό γινόμενο των κανονικοποιημένων κατά ℓ^2 ιστογραμμάτων του Bag of Words μοντέλου, χρησιμοποιώντας επίσης και βάρη τύπου *tf-idf*. Μετράμε τον χρόνο εκπαίδευσης σε D-διάστατες *διανυσματικές πράξεις* (*vector operations* per data point ή *vop*) και την απόδοση στην αναζήτηση με το μέτρο της μέσης ακρίβειας ή *mean average precision* (mAP) για όλες τις εικόνες αναζήτησης. Χρησιμοποιούμε τις δικές μας υλοποιήσεις σε C++ για τις ανταγωνιστικές τεχνικές AKM και RAKM.

2.5.2 Βελτιστοποίηση παραμέτρων

Επιλέγουμε την μικρότερη από τις δύο βάσεις εικόνων, τη βάση *Barcelona* για βελτιστοποίηση των παραμέτρων, καθώς, αν προσθέσουμε και την πολυπλοκότητα ανάθεσης οπτικών λέξεων, θα ήταν απαγορευτικό να χρησιμοποιήσουμε μία μεγαλύτερη βάση. Στην πράξη αποδείχτηκε μάλιστα ότι η τεχνική AGM ξεπερνά το state of the art σε πολύ μεγαλύτερες κλίμακες, με ακριβώς τις ίδιες τιμές παραμέτρων όπως στο μικρό πείραμα ρύθμισης. Η συμπεριφορά του επεκτατικού αλγορίθμου μειγμάτων Κανονικών κατανομών EGM ελέγχεται από τον *παράγοντα επέκτασης* λ καθώς και από το *κατώφλι επικάλυψης* τ. Η προσεγγιστική εκδοχή, AGM, εμπεριέχει μια ακόμα παράμετρο, την *μνήμη* m η οποία καθορίζει την ακρίβεια και ταχύτητα της προσεγγιστικής αναζήτησης κοντινότερου γείτονα.

Το Σχήμα 2.4 δείχνει την απόδοση (mAP) και τα μεγέθη λεξικών ως προς τον αριθμό των επαναλήψεων κατά την εκμάθηση, για διαφορετικές τιμές του *παρά-*

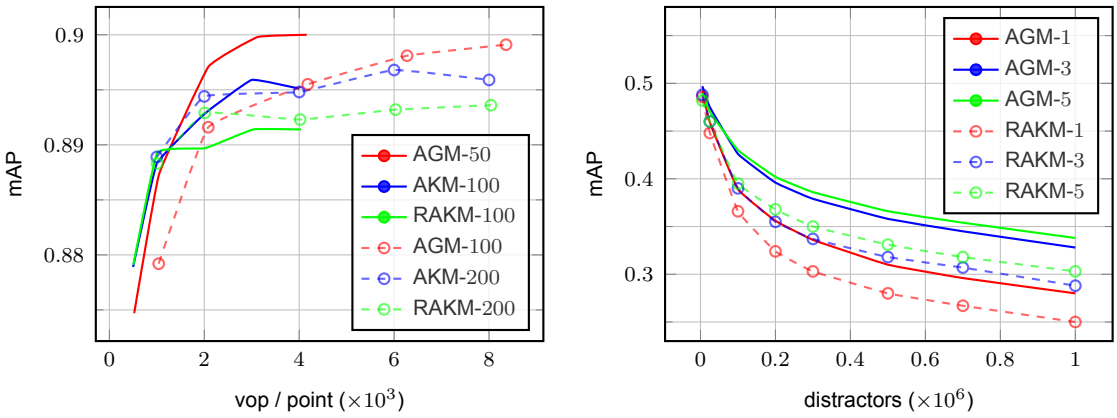


Σχήμα 2.5: *Απόδοση (mAP) για εξειδικευμένα λεξικά στη βάση* Barcelona *ως προς το* κατώφλι επικάλυψης τ *για διάφορετικό αριθμό επαναλήψεων (αριστερά) και τα διαφορετικά μεγέθη λεξικών* C *ως προς τις επαναλήψεις, για διαφορετικές τιμές του* τ *(αριστερά), διατηρώντας σταθερό το* $\lambda = 0.2$.

γοντα επέκτασης λ. Σε σύγκριση με την περίπτωση όπου $\lambda = 0$, είναι εμφανές ότι η επέκταση *επιταχύνει τον ρυθμό σύγκλισης*. Παρόλα αυτά, η επίπτωση είναι προσωρινή και για τιμές από $\lambda = 0.2$ και πάνω η απόδοση σταδιακά πέφτει, προφανών λόγω υπερβολικής επέκτασης και στη συνέχεια σβησίματος. Το *κατώφλι επικάλυψης* τ ελέγχει το σβήσιμο των κέντρων, και εκτιμούμε ότι πρέπει να παίρνει τιμές $\tau \geq 0.5$ όπως αναλύθηκε στην ενότητα 2.3.2. Επειδή οι τιμές $\rho_{i,\mathcal{K}}$ είναι κανονικοποιημένες, η επιλογή $\tau = 0.5$ δείχνει λογική, καθώς έτσι η κάθε συνιστώσα k που διατηρείται «ερμηνεύει» τον εαυτό της τουλάχιστον τόσο όσο το σύνολο $\mathcal{K}$ από τη σχέση (2.21). Το Σχήμα 2.5 δείχνει ότι στις πρώτες επαναλήψεις όσο μεγαλύτερη είναι η τιμή του τ τόσο καλύτερη και η απόδοση, αλλά συνολικά η τιμή $\tau = 0.55$ είναι φανερά βέλτιστη, παρότι είναι περισσότερο αυστηρή από την αρχική προφανή εκτίμηση. Το μέγεθος του λεξικού επηρεάζεται επίσης ιδιαίτερα από την τιμή της παραμέτρου τ. Ενδιαφέρουσα παρατήρηση είναι επίσης ότι παρότι αρχικοποιούμε τον αλγόριθμο AGM με το σύνολο των δεδομένων εκπαίδευσης, δηλαδή όλα τα 550 χιλιάδες διανύσματα, μετά την πρώτη επανάληψη διατηρούνται μονάχα 187 χιλιάδες από αυτά με τις βέλτιστες παραμέτρους. Επιλέγουμε τις τιμές $\lambda = 0.2$ και $\tau = 0.55$ για όλα τα παρακάτω πειράματα.

2.5.3 Συγκρίσεις

Χρησιμοποιούμε τη βιβλιοθήκη FLANN [75] για όλες τις μεθόδους, με 15 δέντρα και την ακρίβεια να ελέγχεται από τον αριθμό των *φύλλων* ή *checks*, δηλαδή τον αριθμό των φύλλων των δέντρων που ελέγχονται σε κάθε αναζήτηση κοντινότερου γείτονα. Χρησιμοποιούμε $1,000$ checks κατά την ανάθεση, και λιγότερα



Σχήμα 2.6: *(Αριστερά) Απόδοση (mAP) για εξειδικευμένα λεξικά στη βάση* Barcelona *ως προς τον χρόνο εκπαίδευσης, μετρημένο ως αριθμό* διανυσματικών πράξεων *(vop) ανα διάνυσμα εκπαίδευσης, για τις τεχνικές AGM, AKM και RAKM για διαφορετικά επίπεδα ακρίβειας, μετρημένα σε αριθμό* checks *της βιβλιοθήκης FLANN. Κάθε καμπύλη αποτελείται από* 5 *μετρήσεις, οι οποίες αντιστοιχούν, από τα αριστερά προς τα δεξιά, στις εμαναλήψεις* 5, 10, 20, 30 *και* 40. *(Δεξιά) Απόδοση (mAP) για τις τεχνικές AGM και RAKM στη βάση* Oxford buildings, *χρησιμοποιώντας* γενικά *λεξικά και με τη βάση να περιέχει μέχρι και* 1 *εκατομμύριο εικόνες* περίσπασης, *χρησιμοποιώντας επίσης* πολλαπλή ανάθεση *με* 1, 3 *και* 5 *κοντινότερους γείτονες για τους περιγραφείς των εικόνων αναζήτησης.*

κατά την εκπαίδευση/κατασκευή του λεξικού. Η προσέγγιση στον αλγόριθμο AGM ελέγχεται από την *μνήμη* m και απαιτούνται m checks του αλγορίθμου FLANN, συν το πολύ m ακόμα υπολογισμούς αποστάσεων στον Αλγόριθμο 2, έχοντας συνολικά $2m$ διανυσματικές πράξεις ανά σημείο και ανά επανάληψη. Γι αυτό το λόγο χρησιμοποιούμε $2m$ checks για τις τεχνικές AKM και RAKM στις συγκρίσεις. Το Σχήμα 2.6 (αριστερά) συγκρίνει τις τρεις μεθόδους ως προς την σύγκλιση για τιμές $m = 50$ και $m = 100$, όπου οι τεχνικές AKM/RAKM εκπαιδεύονται για λεξικά 80 χιλιάδων κέντρων, τιμή η οποία αποδείχθηκε η βέλτιστη πειραματικά.

Ο προτεινόμενος αλγόριθμος AGM όχι μόνο συγκλίνει τόσο γρήγορα όσο οι AKM και RAKM, αλλά έχει και καλύτερη απόδοση για τον ίδιο χρόνο εκπαίδευσης. Οι τεχνικές AKM και RAKM είναι καλύτερες μόνο στις λίγες πρώτες επαναλήψεις, πράγμα λογικό καθώς το AGM αρχικοποιείται από *όλα* τα δεδομένα εκπαίδευσης και χρειάζεται μερικές επαναλήψεις για να φτάσει σε ένα λογικό μέγεθος λεξικού. Η σχετική απόδοση του RAKM σε σχέση με το AKM δεν είναι ακριβώς εκείνη που αναμενόταν [61], αλλά πολύ πιθανό κάτι τέτοιο να οφείλεται στην *τυχαία αρχικοποίηση* των δύο αλγορίθμων, ένα ανοιχτό πρόβλημα για τον αλγόριθμο k-means από το οποίο το AGM έχει απελευθερωθεί. Είναι επίσης ενδιαφέρον ότι, σε αντίθεση με τις άλλες τεχνικές, το AGM δείχνει να βελτιώνεται σε απόδοση για μικρότερη τιμή του m, άρα και μικρότερη ακρίβεια στην εκπαίδευση.

Στην περίπτωση της *αναζήτησης εικόνων μεγάλης κλίμακας* (*large scale image search*), εκπαιδεύουμε *γενικά* λεξικά, δηλαδή λεξικά που το σύνολο περιγραφέων

Method	RAKM						AKM	AGM
Vocabulary	100K	200K	350K	500K	550K	700K	550K	857K
No distractors	0.43	0.464	0.471	0.479	**0.486**	0.476	0.485	**0.492**
20K distractors	0.412	0.427	0.439	0.44	**0.448**	0.437	0.447	**0.459**

Πίνακας 2.1: *Συγκρίσεις στην απόδοση αναζήτησης (mAP) για γενικά (global) λεξικά διαφόρων μεγεθών στη βάση* Oxford Buildings *χωρίς ή με* 20 *χιλιάδες εικόνες περίσπασης, χρησιμοποιώντας* $100/200/200$ *checks στη βιβλιοθήκη FLANN για τις τεχνικές AGM/AKM/RAKM αντίστοιχα,* 40 *επαναλήψεις για τα AKM/RAKM, και μόλις* 15 *για το AGM.*

εκπαίδευσης προέρχεται από μια ανεξάρτητη βάση 15 εικόνων και περιλαμβάνει 6.5 εκατομμύρια περιγραφείς, και μετράμε την απόδοση αναζήτησης στη βάση *Oxford buildings*, με την παρουσία (μέχρι και) ενός εκατομμυρίου εικόνων περίσπασης προερχόμενες από τη βάση *world cities*. Καθώς τα πειράματα σε αυτή την κλίμακα είναι ιδιαίτερα χρονοβόρα, επιλέγουμε πρώτα την καλύτερη από τις ανταγωνιστικές τεχνικές χρησιμοποιώντας μέχρι 20K εικόνες περίσπασης όπως φαίνεται στον Πίνακα 2.1. Χρησιμοποιούμε την τιμή $m = 100$ για αυτό το πείραμα, εκτελώντας 40 επαναλήψεις για τις τεχνικές AKM/RAKM, και μόνο 15 για το AGM. Επιλέγουμε αυτές τις τιμές, οι οποίες ευνοούν τις ανταγωνιστικές μεθόδους AKM/RAKM, καθώς στο προκείμενο πείραμα μας ενδιαφέρει η *ακρίβεια στην απόδοση αναζήτησης* (mAP) και όχι η ταχύτητα εκπαίδευσης. Από τον Πίνακα φαίνεται ότι το μέγεθος 550K είναι το καλύτερο για τη μέθοδο RAKM, με το AKM να έχει παρόμοια απόδοση όπως αναμενόταν, καθώς η κύρια διαφορά τους είναι στην ταχύτητα εκπαίδευσης. Το AGM δείχνει ήδη λίγο καλύτερο, με το μέγεθος λεξικού $C = 857$K να έχει *εξαχθεί αυτόματα*, κρατώντας πάντα τις ίδιες τιμές $\lambda = 0.2$, $\tau = 0.55$ για τις παραμέτρους όπως βρέθηκαν από την ενότητα 2.5.2.

Διατηρώντας ακριβώς τις ίδιες ρυθμίσεις και τιμές για τις παραμέτρους, επεκτείνουμε το πείραμα σε βάσεις μέ μέχρι και 1 εκατομμύριο *εικόνες περίσπασης* για τα δύο καλύτερα λεξικά, δηλαδή το AGM 857K και το RAKM 550K, χρησιμοποιώντας επίσης *πολλαπλή ανάθεση* με 1, 3 και 5 κοντινότερους γείτονες για τους περιγραφείς των εικόνων αναζήτησης. Τα αποτελέσματα φαίνονται στο Σχήμα 2.6 (δεξιά). Χωρίς περαιτέρω βελτιστοποίηση των παραμέτρων σε σχέση με τα πειράματα στη πολύ μικρότερη βάση *Barcelona*, η προτεινόμενη τεχνική AGM αποδεικνύει σταθερά καλύτερη απόδοση, με βελτίωση έως και 3.5% στο μέτρο της μέσης ακρίβειας σε βάση μεγαλύτερη από ένα εκατομμύριο εικόνες.

Κεφάλαιο 3

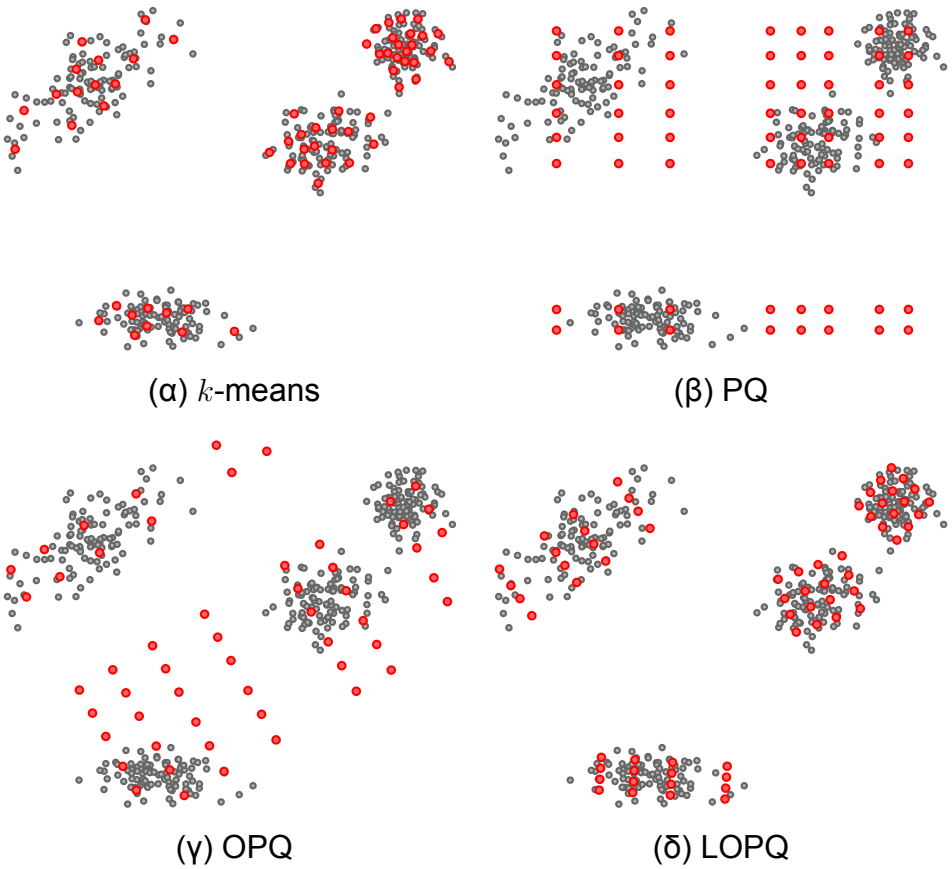
Τοπικά Βελτιστοποιημένος Παραγοντικός Κβαντισμός

3.1 Εισαγωγή

Η προσεγγιστική αναζήτηση κοντινότερων γειτόνων (Approximate nearest neighbor (ANN) search) σε χώρους πολλών διαστάσεων αποτελεί ένα πρόβλημα που συναντάται συχνά στην περιοχή της όρασης υπολογιστών και τα τελευταία χρόνια έχει αποτελέσει ιδιαίτερα ενεργό πεδίο έρευνας. Η πλειοψηφία των μεθόδων που έχουν προταθεί απαιτεί το σύνολο των δεδομένων διαθέσιμα στη μνήμη και χρησιμοποιεί αποδοτικές δομές δεδομένων, έτσι ώστε να χρειαστεί να υπολογιστούν μόνο ένα υποσύνολο από τις αποστάσεις. Ιδανικά, το μέγεθος του υποσυνόλου αυτού είναι σταθερό [75]. Στο άλλο άκρο, άλλες μέθοδοι χρησιμοποιούν μικρές σε μέγεθος δυαδικές κωδικοποιήσεις, απαιτώντας έτσι πολύ λιγότερο χώρο στη μνήμη και έχοντας επίσης και το πλεονέκτημα του πολύ γρήγορου υπολογισμού αποστάσεων στον Hamming χώρο [33, 79]. Οι μέθοδοι αυτές μπορούν να υπολογίσουν εξαντλητικά το σύνολο των αποστάσεων σε σχετικά μικρό χρόνο.

Ο παραγοντικός κβαντισμός, *Product quantization* (PQ) [42], είναι ένας εναλλακτικός τρόπος συμπίεσης μέσω κωδικοποίησης, ο οποίος είναι διακριτός όχι δυαδικός και μπορεί να χρησιμοποιηθεί είτε με εξαντλητική ή μερική αναζήτηση, χρησιμοποιώντας δομές ανεστραμμένης ή πολλαπλής δεικτοδότησης (multi-indexing [9]). Όπως ισχύει και στις περισσότερες μεθόδους κατακερματισμού (hashing) [37], και εδώ, μια όσο το δυνατόν καλύτερη προσέγγιση της υποκείμενης κατανομής των σημείων του χώρου του είναι ιδιαίτερα σημαντική για την απόδοση του αλγορίθμου αναζήτησης. Μία μέθοδος που πηγαίνει τον παραγοντικό κβαντισμό προς αυτή την κατεύθυνση είναι ο *βελτιστοποιημένος παραγοντικός κβαντισμός*, *optimized product quantization* (OPQ) [32] καθώς και ο παρόμοιος αλγόριθμος *Cartesian* k*-means* [78].

Πως όμως οι προαναφερθείσες μέθοδοι βελτιώνουν την απόδοση αναζήτη-



Σχήμα 3.1: *Τέσσερις κβαντιστές με 64 κέντρα (•) ο καθένας, εκπαιδευμένοι σε ένα τυχαίο σύνολο δισδιάστατων σημείων (∘), τα οποία ακολουθούν μια κατανομή μείγματος με 100 σημεία ανα συνιστώσα. Οι κβαντιστές των σχημάτων (γ) και (δ) εκτός της περιστροφής του χώρου, πραγματοποιούν και αναδιάρθρωση των διαστάσεων του χώρου, κάτι που δεν μπορεί να παρουσιαστεί σε αυτό το απλό παράδειγμα με τα δισδιάστατα σημεία.*

σης; Παρόλο που η κάθε μία θέτει τα δικά της κριτήρια, η βασική ιδέα είναι ότι *όλα τα bits που δίνονται στα κβαντισμένα σημεία του χώρου πρέπει να έχουν όσο το δυνατόν περισσότερη πληροφορία*. Μιας και η αναζήτηση αυτή καθ' αυτή μπορεί να εκτελεστεί γρήγορα, η προσοχή στρέφεται προς την αναπαράσταση των δεδομένων. Γι αυτό το λόγο αυτές οι μέθοδοι μπορούν να θεωρηθούν και ως *συμπίεση δεδομένων με σφάλμα* (lossy compression), που έχει ως κριτήριο την όσο το δυνατόν ελάχιστη παραμόρφωση (distortion). Ακραία παραδείγματα υιοθέτησης αυτής της θεώρησης αποτελούν οι δημοσιεύσεις [4, 15]. Η μέθοδος που παρουσιάζουμε στο τρέχον κεφάλαιο έχει περιγραφεί επίσης και στη δημοσίευση [49]

Ο αλγόριθμος k-means, ο οποίος παρουσιάζεται στο Σχήμα 3.1(α), είναι μια μέθοδος κβαντισμού διανυσμάτων όπου θέτοντας ως k τον αριθμό των κέντρων, ένα οποιοδήποτε σημείο σε χώρο $\mathbb{R}^d$ οποιασδήποτε διάστασης d, μπορεί να αναπαρασταθεί με $\log_2 k$ bits. Δυστυχώς, μια αφελής μορφή αναζήτησης κατά αυτό το κβαντισμό απαιτεί πολυπλοκότητα $O(dk)$ και συνήθως χαμηλές τιμές παρα-

μόρφωσης απαιτούν επίσης μεγάλες τιμές για τον αριθμό των κέντρων k. Ο παραγοντικός κβαντισμός (product quantization – PQ) περιορίζει τα κέντρα σε ένα m-διάστατο πλέγμα παράλληλο με τους άξονες και έτσι καταφέρνει κρατήσει την πολυπλοκότητα της αναζήτησης $O(dk)$ για εκθετικά μεγαλύτερο αριθμό κέντρων k^m. Παρόλα αυτά, όπως φαίνεται και στο Σχήμα 3.1(β), πολλά από τα κέντρα αυτά δεν υποστηρίζονται από σημεία του συνόλου, κάτι που συμβαίναι για παράδειγμα εάν οι κατανομές στους m υποχώρους του πλέγματος δεν είναι ανεξάρτητες.

Ο αλγόριθμος OPQ επιτρέπει στο πλέγμα να υποστεί μια ελεύθερη περιστροφή και ανακατάταξη των διαστάσεων, έτσι ώστε να προσεγγίζει καλύτερα την υποκείμενη κατανομή και επίσης να εξισορροπήσει όσο το δυνατόν τη διακύμανση μεταξύ των m υποχώρων, κάτι που οδηγεί και σε μια πιο εξισορροπημένη κατανομή των bit στους υπόχωρους. Όμως, όπως φαίνεται και στο σχήμα 3.1(γ), μια κατανομή μείγματος στον χώρο των σημείων είναι πιθανό να μην ωφεληθεί από μια τέτοιου είδους προσέγγιση. Επίσης, μια τέτοια ολική προσέγγιση του χώρου είναι ιδιαίτερα ευαίσθητη και μπορεί να επηρεαστεί εύκολα αν για παράδειγμα υπάρχουν κάποια λίγο σημεία που δεν ακολουθούν τη γενική κατανομή.

Η λύση που προτείνουμε στο κεφάλαιο αυτό είναι ο *τοπικά βελτιστοποιημένος παραγοντικός κβαντισμός*, *locally optimized product quantization* (LOPQ). Ακολουθώντας μια συνήθη δομή αναζήτησης που προτάθηκε στη δημοσίευση [42], αρχικά χρησιμοποιείται ένας πρώτος κβαντιστής για τη δεικτοδότηση των δεδομένων μέσω ανεστραμμένων λιστών. Έπειτα, κβαντίζονται με παραγοντικό κβαντισμό (PQ) τα *διανύσματα των διαφορών* (residuals) των σημείων. Παρατηρώντας, όμως, ότι οι κατανομές των σημείων γύρω από το κάθε κέντρο του αρχικού κβαντιστή είναι συνήθως μονοτροπικές (unimodal), προτείνουμε τη τοπική βελτιστοποίηση των παραγοντικών κβαντιστών ανά κέντρο. Η μέθοδος μας φαίνεται σχηματικά στο Σχήμα 3.1(δ). Παρότι η μέθοδος μας δεν κάνει υποθέσεις για την γενική υποκείμενη κατανομή, παρατηρείται ότι όλα τα κέντρα υποστηρίζονται από σημεία και έτσι όλα συνεισφέρουν μείωση της μέσης παραμόρφωσης.

Ενδιαφέρον είναι το γεγονός ότι από της δύο λύσεις βελτιστοποίησης που προτείνονται στη δημοσίευση [32], εστιάζουμε στη παραμετρική λύση παρότι αυτή κάνει την υπόθεση ότι η κατανομή τοπικά είναι κανονική. Παρατηρήσαμε ότι, αν χρησιμοποιηθεί η παραμετρική λύση τοπικά, αποδίδει πειραματικά το ίδιο καλά με την μη παραμετρική λύση, ενώ είναι πολλές φορές ταχύτερη.

Η μέθοδος LOPQ απαιτεί λίγο παραπάνω μνήμη σε σχέση με τον παραγοντικό κβαντισμό καθώς και λίγο παραπάνω χρόνο κατά την εκμάθηση, τη δεικτοδότηση και την αναζήτηση. Όμως όλα αυτά είναι σταθερά ως προς το μέγεθος της συλλογής σημείων. Είναι μια μέθοδος που μπορεί πολύ εύκολα να εφαρμοστεί σε πλήθος εφαρμογών και καταφέρνει να αυξήσει κατά πολύ την απόδοση αναζήτησης σε πολλά από τα συνήθη σύνολα σημείων. Για τη δεικτοδότηση συνόλων μεγέθους δισεκατομμυρίων, απαιτείτε πολυ-δεικτοδότηση (multi-index) και παρα-

κάτω θα παρουσιαστεί και ο τρόπος με τον οποίο το LOPQ μπορεί να συνδυαστεί με το multi-index.

3.2 Σχετική βιβλιογραφία και συνεισφορά

Καθώς εστιάζουμε σε μεγάλου μεγέθους συλλογές, το να βρίσκονται όλα τα διανύσματα της συλλογής χωρίς συμπίεση στη μνήμη είναι απαγορευτικό. Κατά συνέπεια, δεν θα ασχοληθούμε με μεθόδους που βασίζονται σε δενδρικές δομές όπως στη δημοσίευση [75]. Οι δυαδικές κωδικοποιήσεις είναι οι πλέον μικρές σε μέγεθος αναπαραστάσεις και προσεγγίζουν την αναζήτηση κοντινότερου γείτονα στον Hamming χώρο. Μέθοδοι όπως αυτές του spectral hashing [111], MLH [77], ITQ [33] και k-means hashing [37] εστιάζουν στην εκμάθηση βελτιστοποιημένων δυαδικών κωδικών λαμβάνοντας υπόψη την υποκείμενη κατανομή των σημείων. Η αναζήτηση στον Hamming χώρο είναι ιδιαίτερα γρήγορη, αλλά παρά την εκμάθηση η απόδοση των παραπάνω τεχνικών είναι χειρότερη από την αναμενόμενη.

Για την επίτευξη καλύτερης απόδοσης μπορούν να χρησιμοποιηθούν πολλαπλοί κβαντιστές ή πίνακες κατακερματισμού, όπως στο LSH [23], έχοντας όμως ως συνέπεια την πολλαπλή ανάθεση του κάθε σημείου στη δομή δεικτοδότησης. Οι δημοσιεύσεις [82, 114] για παράδειγμα αυξάνουν την απόδοση με πολλαπλούς κβαντιστές k-means που δημιουργούνται από διαφορετικές αρχικοποιήσεις του αλγορίθμου ή μέσω κατάτμησης πολλαπλών κέντρων που μαθαίνονται ταυτόχρονα. Παρομοίως, η πρόσφατη μέθοδος multi-index hashing [79] κερδίζει σε ταχύτητα αναζήτησης χρησιμοποιώντας πολλαπλά hash tables για τα διάφορα υποσύνολα δυαδικών κωδικών. Η μέθοδος που προτείνουμε επιτυγχάνει σημαντικά καλύτερη απόδοση *χρησιμοποιώντας μόνο ένα μικρό ποσοστό της μνήμης σε σχέση με αυτές τις τεχνικές*.

Ο παραγοντικός κβαντισμός PQ [42] παρέχει ένα αποδοτικό τρόπο κβαντισμού διανυσμάτων ο οποίος προσφέρει χαμηλότερη παραμόρφωση από ότι οι δυαδικές αναπαραστάσεις. Η μέθοδος Transform coding [14] είναι μια ιδιαίτερη περίπτωση κβαντισμού ανά διάσταση, η οποί επιπλέον αναθέτει τα bits ανάλογα με την διακύμανση ανά διάσταση. Ο βελτιστοποιημένος παραγοντικός κβαντισμός OPQ [32] και η παρόμοια μέθοδος Ck-means [78] γενικεύουν τη μέθοδο PQ βελτιστοποιώντας ταυτόχρονα την περιστροφή του χώρου, την κατάτμηση των υπόχωρων και τους υπο-κβαντιστές. Ενδιαφέρον αποτελεί το ότι η παραμετρική μέθοδος του OPQ στοχεύει ακριβώς στο αντίθετο από ότι εκείνη της δημοσίευσης [14]: προσπαθεί να εξισορροπήσει τη διακύμανση κρατώντας σταθερή την κατανομή των bit στους υποχώρους.

Παρότι στη δημοσίευση [42] προτείνεται και η μή εξαντλητική μέθοδος IVFADC η οποία βασίζεται σε έναν γενικό κβαντιστή και διανύσματα διαφοράς κβαντισμένα με PQ,οι μέθοδοι που προτείνονται στις δημοσιεύσεις [32, 78] είναι εξαντλητικές. Η

μέθοδος δεικτοδότησης του ανεστραμμένου *multi-index* [9] πετυχαίνει μια πολύ λεπτομερέστερη κατάτμηση του χώρου χρησιμοποιώντας έναν κβαντιστή ανά υπόχωρο και είναι μέθοδος συμπληρωματική ως προς την κωδικοποίηση PQ, ενώ έχει βελτιωμένη απόδοση αναζήτησης σε χρόνους συγκρίσιμους με αναζήτηση στον Hamming χώρο. Παράλληλα, η ιδέα της κατάτμησης των υποχώρων μπορεί να εκτελεστεί αναδρομικά ως προς τις διαστάσεις και να παρέχει έναν πολύ γρήγορο τρόπο για την δημιουργία των λεξικών και τον κβαντισμό [8].

Η πολύ πρόσφατη επέκταση του αλγορίθμου OPQ [31] συνδυάζει τη βελτιστοποίηση με χρήση δεικτοδότησης μέσω multi-index και είναι η μέθοδος που παρέχει την καλύτερη επίδοση σε ένα σύνολο με ένα δισεκατομμύριο σημεία, εκτελώντας όμως πάντα τις βελτιστοποιήσεις σε γενικό και όχι τοπικό επίπεδο. Η μέθοδος OPQ όμως αποδίδει σημαντικά καλύτερα σε περιπτώσεις όπου η κατανομή των σημείων είναι κανονική, κάτι το οποίο ισχύει σε μεγαλύτερο βαθμό για τα διανύσματα διαφορών παρά για τα αρχικά διανύσματα. Προτείνουμε συνεπώς να βελτιστοποιήσουμε ανεξάρτητα ανά κελί τα κέντρα ως προς την υποκείμενη κατανομή, παρά τους περιορισμούς του παραγοντικού κβαντισμού. Συγκεκριμένα η μέθοδος μας έχει τις ακόλουθες συνεισφορές:

1. Μοιράζοντας τα δεδομένα σε κελιά, βελτιστοποιούμε *τοπικά* έναν παραγοντικό κβαντιστή ανά κελί, στις κατανομές των διανυσμάτων διαφορών.

2. Δείχνουμε ότι η εκμάθηση μπορεί να εκτελεστεί αποδοτικά χρησιμοποιώντας μια απλοποιημένη εκδοχή του μοντέλου OPQ.

3. Προτείνουμε λύσεις και για απλά ανεστραμμένα αρχεία αλλά και δεικτοδότηση μέσω *multi-index*, δείχνοντας ότι η τοπική βελτιστοποίηση που προτείνουμε μπορεί να εφαρμοστεί άμεσα σε διάφορες μεθόδους δεικτοδότησης και να επιτύχει την καλύτερη υπάρχουσα απόδοση με πολύ μικρό έξτρα κόστος σε μνήμη και χρόνο.

Μια πρόσφατη σχετική δημοσίευση είναι η [24], όπου προτείνεται η τοπική περιστροφή μέσω PCA ανά κέντρο πριν την εξαγωγή περιγραφέων VLAD [44] και συνάθροισης. Όμως όπως δείχνουν και τα δικά μας πειράματα αλλά και εκείνα των δημοσιεύσεων [32, 31], η εφαρμογή PCA χωρίς βέλτιστη ανάθεση των υποχώρων πλήττει ιδιαίτερα την απόδοση κατά την αναζήτηση κοντινότερων γειτόνων σε μεγάλες κλίμακες.

3.3 Υπόβαθρο

3.3.1 Διανυσματικός Κβαντισμός

Ως *κβαντιστής* [34] ορίζεται μια συνάρτηση q η οποία απεικονίζει ένα d-διάστατο διάνυσμα $\mathbf{x} \in \mathbb{R}^d$ σε ένα διάνυσμα $q(\mathbf{x}) \in \mathcal{C}$, όπου το σύνολο $\mathcal{C}$ είναι ένα πεπερα-

σμένο υποσύνολο του $\mathbb{R}^d$, μεγέθους k. Κάθε διάνυσμα $\mathbf{c} \in \mathcal{C}$ αποκαλείται *κέντρο*, και το σύνολο $\mathcal{C}$ *λεξικό*. Άν το $\mathcal{X}$ αποτελεί ένα πεπερασμένο σύνολο σημείων στο χώρο $\mathbb{R}^d$, ο κβαντιστής q προκαλεί *παραμόρφωση* (*distortion*)

$$E = \sum_{\mathbf{x} \in \mathcal{X}} \|\mathbf{x} - q(\mathbf{x})\|^2. \tag{3.1}$$

Σύμφωνα με τον πρώτο κανόνα του Lloyd (Lloyd's first condition) [34], ανεξάρτητα από την επιλογή λεξικού, ένας κβαντιστής ο οποίος ελαχιστοποιεί την παραμόρφωση θα πρέπει να αντιστοιχεί το κάθε σημείο $\mathbf{x}$ στο κοντινότερό του κέντρο, δηλαδή

$$\mathbf{x} \mapsto q(\mathbf{x}) = \arg\min_{\mathbf{c} \in \mathcal{C}} \|\mathbf{x} - \mathbf{c}\|, \tag{3.2}$$

για κάθε $\mathbf{x} \in \mathbb{R}^d$. Συνεπώς, ένας βέλτιστος κβαντιστής θα πρέπει να ελαχιστοποιεί την παραμόρφωση E ως συνάρτηση του λεξικού $\mathcal{C}$ και μόνο. Αυτό συνήθως πραγματοποιείται με διάφορες παραλλαγές του αλγορίθμου k-means.

3.3.2 Παραγοντικός Κβαντισμός

Θεωρώντας ότι το πλήθος των διαστάσεων d είναι πολλαπλάσιο του m, το κάθε διάνυσμα $\mathbf{x} \in \mathbb{R}^d$ μπορεί να γραφτεί σαν ένωση $(\mathbf{x}^1, \ldots, \mathbf{x}^m)$ από m υποδιανύσματα, όπου το καθένα τώρα έχει διάσταση d/m. Εάν $\mathcal{C}^1, \ldots, \mathcal{C}^m$ είναι m υπολεξικά, το κάθε ένα από τα οποία βρίσκεται σε υπόχωρο $\mathbb{R}^{d/m}$ και αποτελείται από k υπο-κέντρα, ένας *παραγοντικός κβαντιστής* (*product quantizer*) [42] περιορίζει το λεξικό $\mathcal{C}$ στο Καρτεσιανό γινόμενο

$$\mathcal{C} = \mathcal{C}^1 \times \cdots \times \mathcal{C}^m, \tag{3.3}$$

δηλαδή σε ένα λεξικό k^m κέντρων της μορφής $\mathbf{c} = (\mathbf{c}^1, \ldots, \mathbf{c}^m)$ όπου κάθε υποκέντρο $\mathbf{c}^j \in \mathcal{C}^j$ για $j \in \mathcal{M} = \{1, \ldots, m\}$. Ένας βέλτιστος παραγοντικός κβαντιστής q θα πρέπει να ελαχιστοποιεί την παραμόρφωση E (3.1) σαν συνάρτηση του λεξικού $\mathcal{C}$, όπου το λεξικό $\mathcal{C}$ υπόκειται στη μορφή της εξίσωσης (3.3) [32]. Στην περίπτωση αυτή, για κάθε διάνυσμα $\mathbf{x} \in \mathbb{R}^d$, το κοντινότερο κέντρο στο λεξικό $\mathcal{C}$ δίνεται από τη σχέση

$$q(\mathbf{x}) = (q^1(\mathbf{x}^1), \ldots, q^m(\mathbf{x}^m)), \tag{3.4}$$

όπου $q^j(\mathbf{x}^j)$ είναι το κοντινότερο υπο-κέντρο του υπο-διανύσματος $\mathbf{x}^j$ in $\mathcal{C}^j$, για κάθε $j \in \mathcal{M}$ [32]. Συνεπώς ο βέλτιστος παραγοντικός κβαντιστής q σε d διαστάσεις μπορεί να αναχθεί σε m υπο-προβλήματα, κάθε ένα από τα οποία παράγει έναν από τους m βέλτιστους υπο-κβαντιστές $q^j, j \in \mathcal{M}$, σε d/m διαστάσεις. στην περίπτωση αυτή μπορούμε να ορίσουμε ότι $q = (q^1, \ldots, q^m)$.

3.3.3 Βελτιστοποιημένος Παραγοντικός Κβαντισμός

Ο Βελτιστοποιημένος Παραγοντικός Κβαντισμός [32],[78] προτείνει να εκτελεστεί βελτιστοποίηση και ως προς την κατάτμηση των υπόχωρων εκτός από την βελτιστοποίηση των λεξικών. Συγκεκριμένα, ο περιορισμός της εξίσωσης (3.3) για τα λεξικά μετατρέπεται στη μορφή

$$\mathcal{C} = \{R\hat{\mathbf{c}} : \hat{\mathbf{c}} \in \mathcal{C}^1 \times \cdots \times \mathcal{C}^m, R^\mathsf{T} R = I\}, \tag{3.5}$$

όπου ο ορθογώνιος πίνακας R διάστασης $d \times d$ επιτρέπει την μια ελεύθερη περιστροφή και αναδιάταξη των διανυσματικών στοιχείων. Συνεπώς η παραμόρφωση E τώρα βελτιστοποιείται σαν συνάρτηση του λεξικού $\mathcal{C}$, με τον περιορισμό ότι το $\mathcal{C}$ έχει τη μορφή της εξίσωσης (3.5). Η βελτιστοποίηση ως προς R και $\mathcal{C}^1, \ldots, \mathcal{C}^m$ μπορεί να εκτελεστεί είτε συγχρόνως όπως στη δημοσίευση του Ck-means [78] και το μη παραμετρικό μοντέλο OPQ$_{\text{np}}$ της δημοσίευσης [32], ή σειριακά, όπως στο παραμετρικό μοντέλο OPQ$_{\text{p}}$ της δημοσίευσης [32].

3.3.4 Εξαντλητική αναζήτηση

Έχοντας κατασκευάσει έναν παραγοντικό κβαντιστή $q = (q^1, \ldots, q^m)$, υποθέτουμε ότι το κάθε διάνυσμα-σημείο $\mathbf{x} \in \mathcal{X}$ αναπαρίσταται ως $q(\mathbf{x})$ και κωδικοποιείται ως ένα σύνολο $(i^1, \ldots, i^m)$ από m δείκτες υπο-κέντρων (3.4), κάθε ένας από τους οποίους προέρχεται από το σύνολο δεικτών $\mathcal{K} = \{1, \ldots, k\}$. Η PQ-*κωδικοποίηση* αυτή απαιτεί $m \log_2 k$ bits ανά σημείο.

Αν το $\mathbf{y}$ είναι το διάνυσμα του σημείου *αναζήτησης* (query vector), η (τετραγωνισμένη) Ευκλείδεια απόσταση του από κάθε σημείο $\mathbf{x} \in \mathcal{X}$ μπορεί να προσεγγιστεί από τη σχέση

$$\delta_q(\mathbf{y}, \mathbf{x}) = \|\mathbf{y} - q(\mathbf{x})\|^2 = \sum_{j=1}^{m} \|\mathbf{y}^j - q^j(\mathbf{x}^j)\|^2, \tag{3.6}$$

όπου $q^j(\mathbf{x}^j) \in \mathcal{C}^j = \{\mathbf{c}_1^j, \ldots, \mathbf{c}_k^j\}$ για κάθε $j \in \mathcal{M}$. Οι αποστάσεις $\|\mathbf{y}^j - \mathbf{c}_i^j\|^2$ υπολογίζονται και αποθηκεύονται για κάθε $i \in \mathcal{K}$ και $j \in \mathcal{M}$ πριν από την αναζήτηση, με συνέπεια η αποτίμηση της σχέσης (3.6) να αντιστοιχεί σε μονάχα $O(m)$ πράξεις μεταφοράς και άθροισης, επιτρέποντας την εκτέλεση εξαντλητικής αναζήτησης σε εκατομμύρια σημεία. Ο υπολογισμός με τη σχέση αυτή αντιστοιχεί στον ασύμμετρο υπολογισμό απόστασης (asymmetric distance computation ή ADC) της σχέσης [42].

3.3.5 Δεικτοδότηση

Μετά τον κβαντισμό του σημείου $\mathbf{x} \in \mathbb{R}^d$ από τον κβαντιστή q, το *διάνυσμα διαφοράς* ή *residual* vector ορίζεται ως

$$r_q(\mathbf{x}) = \mathbf{x} - q(\mathbf{x}), \tag{3.7}$$

όπου $\|r_q(\mathbf{x})\|^2$ είναι η συνεισφορά του σημείου $\mathbf{x}$ στην συνολική παραμόρφωση της εξίσωσης (3.1). Η μη εξαντλητική αναζήτηση προϋποθέτει επιπλέον την χρήση ενός πρώτου *γενικού κβαντιστή* (*coarse quantizer*) Q αποτελούμενου από K κέντρα, ή *κελιά*. Κάθε σημείο $\mathbf{x} \in \mathcal{X}$ κβαντίζεται (εξαντλητικά) με χρήση της συνάρτησης $Q(\mathbf{x})$, ενώ έπειτα το διάνυσμα διαφοράς $r_Q(\mathbf{x})$ κβαντίζεται με έναν παραγοντικό κβαντιστή q. Για κάθε κελί συντηρείται μια ανεστραμμένη λίστα από τα σημεία τα οποία αντιστοιχούνται σε αυτό μαζί με τα PQ-κβαντισμένα διανύσματα διαφοράς.

Κατά την αναζήτηση, το διάνυσμα αναζήτησης $\mathbf{y}$ κβαντίζεται πρώτα μέσω του γενικού κβαντιστή στα w κοντινότερά του κελιά. Σε κάθε ένα από αυτά υπολογίζονται οι προσεγγιστικές ασύμμετρες αποστάσεις μεταξύ του διανύσματος διαφοράς του σημείου αναζήτησης και των σημείων της ανεστραμμένης λίστας του κελιού σύμφωνα με την εξίσωση (3.6). Η διαδικασία αυτή συμβολίζεται ως αναζήτηση IVFADC στη δημοσίευση [42].

3.3.6 Ανακατάταξη δευτέρου επιπέδου

Σε συνδυασμό με την ασύμμετρη απόσταση ADC ή την αναζήτηση IVFADC μπορούν να χρησιμοποιηθούν και δευτέρου επιπέδου διανύσματα διαφορών, τα οποία μπορούν και πάλι να κωδικοποιηθούν μέσω PQ σε m' υπο-κβαντιστές. Στην περίπτωση αυτή όμως πρέπει αναγκαστικά να γίνει ολική ανακατασκευή των σημείων της βάσης μέσω των υπο-κβαντιστών για τον υπολογισμό της απόστασης και συνεπώς τα δεύτερου επιπέδου διανύσματα διαφοράς χρησιμοποιούνται μονάχα για αναδιάταξη, όπως προτείνεται στη δημοσίευση [46].

3.3.7 Πολυ-δεικτοδότηση – multi-index

Στη δεικτοδότηση τύπου multi-index η ιδέα του παραγοντικού κβαντισμού εφαρμόζεται και στη φάση του γενικού κβαντιστή. Κατασκευάζεται ένα δευτέρου επιπέδου ανεστραμμένο multi-index [9], το οποί αποτελείται από δύο *κβαντιστές υποχώρων* Q^1, Q^2, ο καθένας από τους οποίους βρίσκεται στον χώρο $\mathbb{R}^{d/2}$ και έχει K υπο-κέντρα. Ένα *κελί* στην περίπτωση αυτή είναι ένα ζεύγος από υπο-κέντρα. Συνολικά έχουμε K^2 κελιά, τα οποία μπορούν να δομηθούν πάνω σε ένα δισδιάστατο *πλέγμα*, παράγοντας μια πολύ λεπτομερή κατάτμηση του χώρου $\mathbb{R}^d$. Για

κάθε σημείο $vx = (\mathbf{x}^1, \mathbf{x}^2) \in \mathcal{X}$, τα υπο-διανύσματα $\mathbf{x}^1, \mathbf{x}^2 \in \mathbb{R}^{d/2}$ κβαντίζονται ξεχωριστά (και με εξαντλητικό τρόπο) μέσω την κβαντιστών στους δείκτες $Q^1(\mathbf{x}^1)$ και $Q^2(\mathbf{x}^2)$ αντίστοιχα. Και πάλι, για κάθε κελί διατηρείται μια ανεστραμμένη λίστα με τα αντίστοιχα σημεία.

Αν $\mathbf{y} = (\mathbf{y}^1, \mathbf{y}^2)$ είναι ένα διάνυσμα αναζήτησης, πρώτα υπολογίζονται οι (τετραγωνισμένες) Ευκλείδειες αποστάσεις για κάθε ένα από τα υπο-διανύσματα $\mathbf{y}^1, \mathbf{y}^2$ σε όλα τα υπο-κέντρα των Q^1, Q^2 αντίστοιχα. Η απόσταση του σημείου αναζήτησης $\mathbf{y}$ από ένα κελί μπορεί να υπολογιστεί με μία μονάχα πράξη άθροισης, καθ' αντιστοιχία με την εξίσωση (3.6) για $m = 2$. Τα κελιά διατρέχονται σε αύξουσα σειρά απόστασης ως προς το διάνυσμα $\mathbf{y}$ με χρήση του αλγορίθμου *multi-sequence* που προτείνεται στη δημοσίευση [9] μέχρι να συγκεντρωθεί ένας προκαθορισμένος αριθμός από T σημεία.

3.4 Τοπικά βελτιστοποιημένος παραγοντικός κβαντισμός

Θα αναλύσουμε δύο λύσεις όσων αφορά τη δεικτοδότηση: τα ιδιαίτερα διαδεδομένα ανεστραμμένα αρχεία πάνω σε έναν γενικό (coarse) κβαντιστή, καθώς και το δευτέρου επιπέδου ανεστραμμένο multi-index. Η ενότητα 3.4.1 μελετά τον προτεινόμενο LOPQ κβαντισμό στην πρώτη περίπτωση, όπου πολύ απλά βελτιστοποιούμε τοπικά τα δεδομένα κάθε κελιού μετά την ανάθεσή τους, κατασκευάζοντας έναν παραγοντικό κβαντιστή ανά κελί πάνω στα διανύσματα διαφορών. Η βελτιστοποίηση ανά κελί αναλύεται στην ενότητα 3.4.2 και ακολουθεί κατά βάση τις μεθόδους που περιγράφονται στις δημοσιεύσεις [32, 78]. Η ίδια διαδικασία χρησιμοποιείται και την ενότητα 3.4.4, όπου αναλύεται ο κβαντισμός LOPQ για την περίπτωση του multi-index.

3.4.1 Αναζήτηση σε ανεστραμμένο αρχείο

Αν $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ είναι το σύνολο από n σημεία δεδομένων στον χώρο $\mathbb{R}^d$, βελτιστοποιούμε έναν γενικο (coarse) κβαντιστή Q, ο οποίος αποτελείται από τα K κέντρα ή κελιά $\mathcal{E} = \{\mathbf{e}_1, \ldots, \mathbf{e}_K\}$. Για κάθε κελί $i \in \mathcal{K} = \{1, \ldots, K\}$, κατασκευάζουμε μια ανεστραμμένη λίστα $\mathcal{L}_i$ η οποία περιέχει τα κβαντισμένα σημεία που αντιστοιχούνται στο κελί $\mathbf{e}_i$,

$$\mathcal{L}_i = \{j \in \mathcal{N} : Q(\mathbf{x}_j) = \mathbf{e}_i\} \tag{3.8}$$

όπου $\mathcal{N} = \{1, \ldots, n\}$ και συλλέγουμε τα διανύσματα διαφορών τους στο σύνολο

$$\mathcal{Z}_i = \{\mathbf{x} - \mathbf{e}_i : \mathbf{x} \in \mathcal{X},\ Q(\mathbf{x}) = \mathbf{e}_i\}. \tag{3.9}$$

Για κάθε κελί $i \in \mathcal{K}$, βελτιστοποιούμε τοπικά έναν PQ κβαντιστή πάνω στο σύνολο των διανυσμάτων διαφορών $\mathcal{Z}_i$, με τη διαδικασία που περιγράφεται στην ενότητα 3.4.2, μαθαίνοντας έναν ορθογώνιο πίνακα R_i και έναν παραγοντικό κβαντιστή q_i. Τα διανύσματα διαφορών αρχικά περιστρέφονται σύμφωνα με τη σχέση $\hat{\mathbf{z}} \leftarrow R_i^\mathsf{T}\mathbf{z}$ για κάθε $\mathbf{z} \in \mathcal{Z}_i$ και στη στη συνέχεια κβαντίζονται μέσω του κβαντιστή PQ δίνοντας $q_i(\hat{\mathbf{z}}) = q_i(R_i^\mathsf{T}\mathbf{z})$.

Κατά τη στιγμή της αναζήτησης, το διάνυσμα αναζήτησης $\mathbf{y}$ ανατίθεται στα w κοντινότερα κελιά του $\mathcal{A}$ από το σύνολο $\mathcal{E}$. για κάθε κελί $\mathbf{e}_i \in \mathcal{A}$, υπολογίζεται το διάνυσμα διαφορών $\mathbf{y}_i = \mathbf{y} - \mathbf{e}_i$ και περιστρέφεται κατά $\hat{\mathbf{y}}_i \leftarrow R_i^\mathsf{T}\mathbf{y}_i$. Έπειτα, υπολογίζονται οι ασύμμετρες αποστάσεις $\delta_{q_i}(\hat{\mathbf{y}}_i, \hat{\mathbf{z}}_p)$ στα διανύσματα διαφορών $\hat{\mathbf{z}}_p$ για κάθε $p \in \mathcal{L}_i$, σύμφωνα με τη σχέση (3.6), χρησιμοποιώντας τον τοπικό παραγοντικό κβαντιστή q_i. Οι υπολογισμοί είναι εξαντλητικοί ως προς τα σημεία της λίστας $\mathcal{L}_i$, αλλά εκτελούνται στον συμπιεσμένο χώρο χρησιμοποιώντας πίνακες ανάθεσης (lookup tables) και χωρίς να χρειάζεται η ανακατασκευή των σημείων.

Ανάλυση πολυπλοκότητας.. Για να μελετήσουμε ξεχωριστά το κέρδος από τις δύο ποσότητες που βελτιστοποιούμε, παρουσιάζουμε αποτελέσματα βελτιστοποιώντας μονάχα την περιστροφή με σταθερούς (όχι τοπικούς) υπο-κβαντιστές, καθώς και βελτιστοποιώντας και την περιστροφή αλλά και τους κβαντιστές τοπικά. Οι δύο περιπτώσεις αυτές θα αναφέρονται ως LOR+PQ και LOPQ, αντίστοιχα. Στην δεύτερη περίπτωση, έχουμε ένα επιπλέον κόστος σε μνήμη της τάξεως του $O(K(d^2 + dk))$ σε σχέση με τη μέθοδο IVFADC [42]. Παρομοίως, οι τοπικές περιστροφές των διανυσμάτων διαφορών του σημείου αναζήτησης προσθέτουν ένα κόστος χρόνου $O(wd^2)$. Παρατηρείται εύκολα ότι και τα δύο έξτρα κόστη είναι σταθερά, ανεξάρτητα από το μέγεθος της συλλογής n.

3.4.2 Τοπική Βελτιστοποίηση

Έστω $\mathcal{Z} \in \{\mathcal{Z}_1, \ldots, \mathcal{Z}_K\}$ το σύνολο από διανύσματα διαφορών των σημείων που αντιστοιχούν σε ένα κελί του συνόλου $\mathcal{E}$. Σε αντίθεση με τη δημοσίευση [42], κβαντίζουμε μέσω παραγοντικού κβαντισμού τα διανύσματα αυτά, βελτιστοποιώντας τοπικά και την κατάτμηση των υπόχωρων και τους υπο-κβαντιστές ανά κελί. Με παραμέτρους τα m και k, το πρόβλημα αυτό εκφράζεται ως πρόβλημα ελαχιστοποίησης της παραμόρφωσης σαν συνάρτηση του ορθοκανονικού πίνακα περιστροφής $R \in \mathbb{R}^{d\times d}$ και των υπο-λεξικών $\mathcal{C}^1, \ldots, \mathcal{C}^m \subset \mathbb{R}^{d/m}$ ανά κελί,

$$\begin{aligned} \text{minimize} \quad & \sum_{\mathbf{z}\in\mathcal{Z}} \min_{\hat{\mathbf{c}}\in\hat{\mathcal{C}}} \|\mathbf{z} - R\hat{\mathbf{c}}\|^2 \\ \text{subject to} \quad & \hat{\mathcal{C}} = \mathcal{C}^1 \times \cdots \times \mathcal{C}^m \\ & R^\mathsf{T}R = I, \end{aligned} \tag{3.10}$$

όπου $|\mathcal{C}^j| = k$ για κάθε $j \in \mathcal{M} = \{1, \ldots, m\}$. Με δεδομένη τη λύση για τον πίνακα $R, \mathcal{C}^1, \ldots, \mathcal{C}^m$, το λεξικό $\mathcal{C}$ δίνεται από τη σχέση (3.5). Για κάθε $j \in \mathcal{M}$, κάθε ύπο-

λεξικό $\mathcal{C}^j$ ορίζουν έναν υπο-κβαντιστή q^j μέσω της σχέσης

$$\mathbf{x} \mapsto q^j(\mathbf{x}) = \arg\min_{\hat{\mathbf{c}}^j \in \mathcal{C}^j} \|\mathbf{x} - \hat{\mathbf{c}}^j\| \tag{3.11}$$

για κάθε $\mathbf{x} \in \mathbb{R}^{d/m}$, παρομοίως με τη σχέση (3.2). Όλοι μαζί οι υπο-κβαντιστές ορίζουν έναν παραγοντικό κβαντιστή $q = (q^1, \ldots, q^m)$ σύμφωνα με τη σχέση (3.4). Η τοπική βελτιστοποίηση τότε μπορεί να εκφραστεί σαν μια αντιστοίχιση $\mathcal{Z} \mapsto (R, q)$. Σύμφωνα με τις δημοσιεύσεις [32, 78], υπάρχουν δύο μέθοδοι λύσεις τις οποίες θα περιγράψουμε εν συντομία, εστιάζοντας περισσότερο στην παραμετρική μέθοδο OPQ_p.

Παραμετρική μέθοδος βελτιστοποίησης. (OPQ_p [32]) Θεωρώντας ότι τα διανύσματα διαφορών $\mathcal{Z}$ γεννούνται από μια d-διάστατη κανονική κατανομή $\mathcal{N}(\mathbf{0}, \Sigma)$ με μέση τιμή μηδέν, μπορούμε να ελαχιστοποιήσουμε το θεωρητικό κάτω όριο παραμόρφωσης σα συνάρτηση μονάχα του πίνακα περιστροφής R [32]. Ο πίνακας R βελτιστοποιείται δηλαδή ανεξάρτητα και πριν την βελτιστοποίηση των υποκβαντιστών, η οποία έπεται και εκτελείται ακριβώς όπως στην περίπτωση του PQ, μαθαίνοντας δηλαδή έναν ανεξάρτητο κβαντιστή μέσω k-means ανά υπόχωρο.

Με δεδομένο τον $d \times d$ θετικά ορισμένο πίνακα συνδιακύμανσης Σ υπολογισμένο στο $\mathcal{Z}$, η λύση για τον πίνακα R μπορεί να βρεθεί σε κλειστή μορφή με δύο βήματα. Αρχικά περιστρέφουμε τα δεδομένα κατά $\hat{\mathbf{z}} \leftarrow R^\mathsf{T}\mathbf{z}$ και για κάθε $\mathbf{z} \in \mathcal{Z}$ παίρνουμε ένα μπλόκ-διαγώνιο (block-diagonal) πίνακα συνδιακύμανσης $\hat{\Sigma}$, με το j-ο διαγώνιο μπλοκ να είναι ο υπο-πίνακας $\hat{\Sigma}_{jj}$ του j-ου υπόχωρου, για κάθε $j \in \mathcal{M}$. Θέλουμε δηλαδή τις κατανομές των υπόχωρων να είναι ανά δύο *ανεξάρτητες*. Για να το επιτύχουμε αυτό, μπορούμε να διαγωνοποιήσουμε τον πίνακα Σ ως $U\Lambda U^\mathsf{T}$.

Έπειτα, θέλουμε τις ορίζουσες $|\hat{\Sigma}_{jj}|$ να είναι ίσες για κάθε $j \in \mathcal{M}$, θέλουμε δηλαδή την διακύμανση (variance) να είναι όσο το δυνατόν *μοιρασμένη* ανάμεσα στους υπόχωρους. Για να το επιτύχουμε αυτό, χρησιμοποιούμε τον αλγόριθμο *ανάθεσης ιδιοτιμών* (*eigenvalue allocation*) [32]. Συγκεκριμένα, το σύνολο $\mathcal{B}$ από m θυρίδες B^j αρχικοποιείται ως $B^j = \emptyset$, $j \in \mathcal{M}$, όπου η κάθε θυρίδα έχει χωρητικότητα ίση με $d^* = d/m$. Οι ιδιοτιμές του πίνακα Λ διατρέχονται με φθίνουσα σειρά, $\lambda_1 \geq \cdots \geq \lambda_d$. και κάθε ιδιοτιμή λ_s, $s = 1, \ldots, d$, ανατίθεται διαδοχικά στην θυρίδα B^* με την ελάχιστη μέχρι στιγμής διακύμανση που δεν έχει γεμίσει, δηλαδή $B^* \leftarrow B^* \cup s$ μέ

$$B^* = \arg\min_{\substack{B \in \mathcal{B} \\ |B| < d^*}} \prod_{s \in B} \lambda_s, \tag{3.12}$$

μέχρις ότου όλες οι θυρίδες να γεμίσουν. Έπειτα, οι θυρίδες καθορίζουν μια αναδιάταξη των διαστάσεων: εάν το διάνυσμα $\mathbf{b}^j \in \mathbb{R}^{d^*}$ περιέχει στοιχεία της θυρίδας B^j (με οποιαδήποτε σειρά) για κάθε $j \in \mathcal{M}$ και $\mathbf{b} = (\mathbf{b}^1, \ldots, \mathbf{b}^m)$, τότε το διάνυσμα $\mathbf{b}$ μπορεί να θεωρηθεί ως μια αναδιάταξη (permutation) π του συνόλου $\{1, \ldots, d\}$. Εάν P_π είναι ο πίνακας αναδιάταξης του π, τότε ο πίνακας UP_π^T αναπαριστά μια

ανακατάταξη των ιδιοδιανυσμάτων του πίνακα συνδιακύμανσης Σ και αυόν θεωρούμε ως τελική λύση R.

Μη παραμετρική μέθοδος. (OPQ_{np} [32] ή *Ck-means* [78]) ονομάζεται μια παραλλαγή του αλγορίθμου k-means, η οποία εκτελείται σε όλου τους m υπόχωρους ταυτόχρονα, παρεμβάλλοντας σε κάθε επανάληψη βήματα για περιστροφή των δεδομένων και τη βελτιστοποίηση του πίνακα R μεταξύ των δύο παραδοσιακών βημάτων *ανάθεσης* και *μετακίνησης* των κέντρων. Συγκεκριμένα *ευθυγραμμίζονται* τα κέντρα στα δεδομένα,

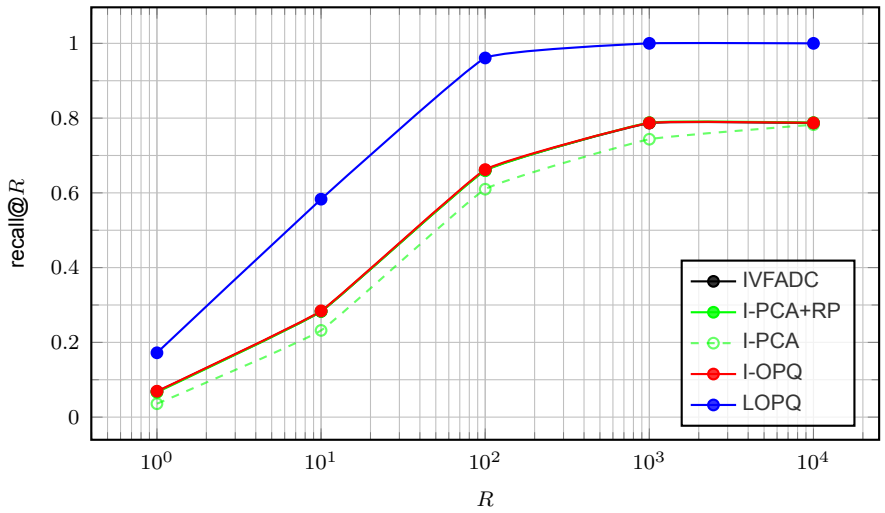
$$\begin{aligned}
&\text{rotate:} && \hat{\mathbf{z}} \leftarrow R^{\mathsf{T}}\mathbf{z}, \;\; \forall \mathbf{z} \in \mathcal{Z} \\
&\text{assign:} && \mathcal{A}_i^j \leftarrow \{\hat{\mathbf{z}}^j \in \hat{\mathcal{Z}}^j : q^j(\hat{\mathbf{z}}^j) = \hat{\mathbf{c}}_i^j\}, \;\; \forall i,j \\
&\text{update:} && \hat{\mathbf{c}}_i^j \leftarrow \frac{1}{|\mathcal{A}_i^j|} \sum_{\hat{\mathbf{z}}^j \in \mathcal{A}_i^j} \hat{\mathbf{z}}^j, \;\; \forall i,j \\
&\text{align:} && \min_R \|Z - RY\|_F^2 \;\; \text{s.t.} \;\; R^{\mathsf{T}}R = I,
\end{aligned} \tag{3.13}$$

όπου $\hat{\mathcal{Z}}^j$ είναι το σύνολο από τα j-α υπο-διανύσματα $\hat{\mathbf{z}}^j$ των περιστραμμένων δεδομένων $\hat{\mathbf{z}}$ για κάθε $\mathbf{z} \in \mathcal{Z}$, και όπου το i-ο κέντρο $\hat{\mathbf{c}}_i^j$ του υπο-λεξικού $\mathcal{C}^j$ του υπο-κβαντιστή q^j δίνεται από τη σχέση (3.11).

Τέλος, οι $d \times |\mathcal{Z}|$ πίνακες Z, Y περιέχουν σαν στήλες αντιστοίχως όλα τα διανύσματα $\mathbf{z} \in \mathcal{Z}$ και τα περιστραμμένα και κβαντισμένα ομόλογά τους $q(\hat{\mathbf{z}})$, όπως αυτά δίνονται από τη σχέση (3.4). Με $\|\cdot\|_F$ συμβολίζουμε την Φρομπέρνια νόρμα. Η βελτιστοποίηση ως προς R είναι το *orthogonal procrustes problem* [92] και έχει ως λύση τη $R \leftarrow UV^{\mathsf{T}}$, όπου οι πίνακες U, V βρίσκονται μέσω SVD των ZY^{T} ως $U\Lambda^{1/2}V^{\mathsf{T}}$ [33, 32]. Στην πράξη, η μέθοδος OPQ_{p} είναι πολύ πιο γρήγορη από την OPQ_{np}. Καθώς στην περίπτωσή μας πρέπει να βελτιστοποιήσουμε τοπικά χιλιάδες κβαντιστές, το να εκτελέσουμε την εκπαίδευση μέσω OPQ_{np} είναι μη πρακτικό, έτσι τη μη παραμετρική μέθοδο τη χρησιμοποιούμε μόνο σε ένα μικρό πείραμα στην ενότητα 3.5.2 και κατά τα άλλα εστιάζουμε στην παραμετρική μέθοδο OPQ_{p}, στην οποία θα αναφερόμαστε και ως I-OPQ στη συνέχεια.

3.4.3 Παράδειγμα

Για να παρουσιάσουμε το κέρδος από την τοπική βελτιστοποίηση, εκτελέσαμε πειράματα σε ένα συνθετικό σύνολο σημείων το οποίο ονομάζουμε SYNTH1M και το οποίο περιέχει 1 εκατομμύριο 128-διάστατα σημεία και 10K σημεία αναζήτησης, το οποί κατασκευάσαμε παίρνοντας 1000 δείγματα από 1000 συνιστώσες ενός ανισοτροπικού μείγματος Κανονικής κατανομής. Για όλες τις μεθόδους πραγματοποιείται μη εξαντλητική αναζήτηση μέσω ανεστραμμένου αρχείου και ενός γενικού κβαντιστή πάνω στα διανύσματα διαφορών κωδικοποιημένα μέσω ενός παραγοντικού κβαντιστή, όπως περιγράφεται στην ενότητα 3.4.1. Παρόλα αυτά, εκτός από την περίπτωση του LOPQ, όλες οι βελτιστοποιήσεις είναι ολικές (global). Για να είναι η σύγκριση δίκαιη, τόσο εδώ όσο και στην ενότητα 3.5, ως I-OPQ συμβολίζουμε



Σχήμα 3.2: *Ανάκληση στα πρώτα R δείγματα (recall@R) για τη συλλογή SYNTH1M—το μέτρο recall@R ορίζεται στην υποενότητα 3.5.1. Για όλες τις μεθόδους χρησιμοποιούμε τις τιμές $K = 1024$ και $w = 8$. Επίσης για όλους τους παραγοντικούς κβαντιστές τις τιμές $m = 8$ και $k = 256$. Οι καμπύλες για τις μεθόδους IVFADC, I-OPQ και I-PCA+RP συμπίπτουν σχεδόν παντού.*

την δικιά μας μη-εξαντλητική προσαρμογή του αλγορίθμου της δημοσίευσης [32]. Η μέθοδος IVFADC (PQ) [42] χρησιμοποιεί την φυσική σειρά των διαστάσεων και κανενός άλλου είδους βελτιστοποίηση ως προς την ανάθεση διαστάσεων σε υποχώρους.

Το σχήμα 3.2 παρουσιάζει αποτελέσματα σε αναζήτηση κοντινότερου γείτονα (ANN search). Σε αυτή τη συνθετική και πολυ-τροπική (multi-modal) κατανομή, η μέθοδος I-OPQ δεν καταφέρνει να παράγει καλύτερα αποτελέσματα από την βασική μη βελτιστοποιημένη μέθοδο IVFADC. Η περίπτωση όπου ευθυγραμμίζουμε τα δεδομένα μέσω PCA και στη συνέχεια αναθέτουμε τις διαστάσεις κρατώντας την φθίνουσα σειρά των ιδιοτιμών, χωρίς αναδιάταξη δηλαδή, αναφέρεται ως I-PCA. Η στρατηγική αυτή αποδεικνύεται χειρότερη από τη αρχική κατάταξη χωρίς ευθυγράμμιση, καθώς χωρίς αναδιάταξη οι μεγαλύτερες d/m ιδιοτιμές ανατίθενται όλες σε έναν υπόχωρο, κάτι που έρχεται σε αντίθεση με τη λογική εξισορρόπησης της διακύμανσης στους υπόχωρους της μεθόδου I-OPQ. Απο την άλλη, αν αναδιατάξουμε τις διαστάσεις μετά την ευθυγράμμιση τυχαία, μέθοδος που συμβολίζεται ως I-PCA+RP, το πρόβλημα αυτό εξαλείφεται. Παρατηρείται όμως εύκολα, ότι η προτεινόμενη μέθοδος με την τοπική βελτιστοποίηση LOPQ ξεπερνά όλες τις άλλες μεθόδους, ως και 30%.

3.4.4 Αναζήτηση με multi-index

Αν και η τοπική βελτιστοποίηση στην περίπτωση δεικτοδότησης με ανεστραμ-

μένα αρχεία μοιάζει απλή, όταν η δεικτοδότηση εκτελείται με multi-index η αναγωγή δεν είναι τόσο προφανής καθώς η μνήμη και η χρόνος που θα χρειαζόταν για την τοπική βελτιστοποίηση σε κάθε κελί είναι απαγορευτικός, σε αντίθεση με την περίπτωση της ενότητας 3.4.1. Προτείνουμε συνεπώς την βελτιστοποίηση ανά κελί ανεξάρτητα στους δύο αρχικούς κβαντιστές υποχώρων, καθώς και τον ανεξάρτητο κβαντισμό των δύο υπο-διανυσμάτων διαφοράς. Ονομάζουμε την διαδικασία αυτή *παραγοντική βελτιστοποίηση* (*product optimization*) και τη συμβολίζουμε ως Multi-LOPQ.

Παραγοντική Βελτιστοποίηση

Οι δύο κβαντιστές υποχώρων Q^1, Q^2 του mutli-index αποτελούνται από K κέντρα έκαστος, κατασκευάζονται όπως περιγράφεται στη δημοσίευση [9] και έχουν ως λεξικά τα $\mathcal{E}^j = \{\mathbf{e}^j_1, \dots, \mathbf{e}^j_K\}$ για το $j = 1, 2$. Το κάθε σημείο $\mathbf{x} = (\mathbf{x}^1, \mathbf{x}^2) \in \mathcal{X}$ κβαντίζεται στο κελί $Q(\mathbf{x}) = (Q^1(\mathbf{x}^1), Q^2(\mathbf{x}^2))$. Για κάθε κελί $(\mathbf{e}^1_{i^1}, \mathbf{e}^2_{i^2})$ του πλέγματος $\mathcal{E} = \mathcal{E}^1 \times \mathcal{E}^2$, όπου $i^1, i^2 \in \mathcal{K}$ κατασκευάζεται μια ανεστραμμένη λίστα $\mathcal{L}_{i^1 i^2}$.

Οι κβαντιστές Q^1, Q^2 χρησιμοποιούνται επίσης για τα διανύσματα διαφορών, όπως και στη μέθοδο Multi-D-ADC [9]. Για κάθε σημείο $\mathbf{x} = (\mathbf{x}^1, \mathbf{x}^2) \in \mathcal{X}$, κβαντίζονται μέσω παραγοντικού κβαντιστή τα διανύσματα διαφορών $\mathbf{x}^j - Q^j(\mathbf{x}^j)$ όπου το $j = 1, 2$. Όμως, δεδομένου ότι τα λεξικά που ανήκουν στο χώρο $\mathbb{R}^d$ μέσω των κβαντιστών Q^1, Q^2 είνα ιδιαίτερα μεγάλα και παράγουν μια πολύ ψιλη κατάτμηση του χώρου (το δισδιάστατο πλέγμα αποτελείται από K^2 κελιά), δεν μπορούμε να βελτιστοποιήσουμε ανά κελί – το συνολικό κόστος χώρου, για παράδειγμα, θα ήταν $O((d^2 + dk)K^2)$. Αυτό που κάνουμε είναι το να βελτιστοποιούμε ανά υπόχωρο: αντίστοιχα με τη σχέση (3.9), έστω ότι το σύνολο

$$\mathcal{Z}^j_i = \{\mathbf{x}^j - \mathbf{e}^j_i : \mathbf{x} \in \mathcal{X},\ Q^j(\mathbf{x}^j) = \mathbf{e}^j_i\}. \tag{3.14}$$

περιέχει τα διανύσματα διαφορών των σημείων $\mathbf{x} \in \mathcal{X}$ των οποίων το j-στο υποδιάνυσμα κβαντίζεται στο κελί $\mathbf{e}^j_i$ for $i \in \mathcal{K}$ and $j = 1, 2$. Στη συνέχεια, βελτιστοποιούμε τοπικά κάθε σύνολο $\mathcal{Z}^j_i$ με τη διαδικασία που περιγράφεται στην ενότητα 3.4.2, παίρνοντας έναν πίνακα περιστροφής R^j_i και έναν παραγοντικό κβαντιστή q^j_i.

Αν $\mathbf{x} = (\mathbf{x}^1, \mathbf{x}^2) \in \mathcal{X}$ είναι ένα σημείο κβαντισμένο στο κελί $(\mathbf{e}^1_{i^1}, \mathbf{e}^2_{i^2}) \in \mathcal{E}$, τα υποδιανύσματα διαφοράς $\mathbf{z}^j = \mathbf{x}^j - \mathbf{e}^j_{i^j}$ έχουν περιστραφεί και κβαντιστεί παραγοντικά ανεξάρτητα στους δύο υπόχωρους σύμφωνα με τη σχέση $q^j_{i^j}(\hat{\mathbf{z}}^j) = q^j_{i^j}((R^j_{i^j})^\mathsf{T}\mathbf{z}^j)$ όπου το $j = 1, 2$.

Αν $\mathbf{y}$ είναι ένα σημείο αναζήτησης, οι περιστροφές $\hat{\mathbf{y}}^j_{i^j} = (R^j_{i^j})^\mathsf{T}(\mathbf{y}^j - \mathbf{e}^j_{i^j})$ υπολογίζονται μια φορά, όταν ο αλγόριθμος multi-sequence φέρει το αντίστοιχο κελί $i^j = 1, ..., K$ και $j = 1, 2$, ενώ έπειτα αποθηκεύονται για να ξαναχρησιμοποιηθούν. Για κάθε σημείο με δείκτη p που περιέχεται στο κελί $(\mathbf{e}^1_{i^1}, \mathbf{e}^2_{i^2}) \in \mathcal{E}$ με σχετικά διανύσματα διαφοράς $\hat{\mathbf{z}}^j_p$ όπου $j = 1, 2$, υπολογίζεται η ασύμμετρη απόσταση

(asymmetric distance)

$$\|\hat{\mathbf{y}}^1_{i^1} - q^1_{i^1}(\hat{\mathbf{z}}^1_p)\|^2 + \|\hat{\mathbf{y}}^2_{i^2} - q^2_{i^2}(\hat{\mathbf{z}}^2_p)\|^2 \tag{3.15}$$

και τα σημεία κατατάσσονται με αύξουσα σειρά αυτής.

Θεωρώντας συνολικά το χώρο $\mathbb{R}^d$, αυτού του είδους η βελτιστοποίηση είναι τοπική ανά κελί, αλλά πιό περιορισμένη από την τοπική βελτιστοποίηση που περιγράφεται στην ενότητα 3.4.2. Η περιστροφή στο κελί $(\mathbf{e}^1_{i^1}, \mathbf{e}^2_{i^2}) \in \mathcal{E}$ για παράδειγμα, περιορίζεται σε μορφή block-diagonal με μπλοκ $R^1_{i^1}, R^2_{i^2}$, κρατώντας τις περιστροφές μέσα στους υπόχωρους. Αντίθετα, η μέθοδος OMulti-D-OADC [31] δεν βάζει τέτοιους περιορισμούς στον πίνακα περιστροφής, χρησιμοποιώντας, βέβαια, μόνο ένα για όλα τα κελιά.

Ανάλυση Πολυπλοκότητας

Σε σύγκριση με τη μέθοδο Multi-D-ADC [9], η παραπάνω μνήμη που απαιτείται παραμένει (ασυμπτωτικά) ίδια όπως στην ενότητα 3.4.1, είναι δηλαδή ίση με $O(K(d^2 + dk))$. Ο παραπάνω χρόνος κατά την αναζήτηση είναι $O(Kd^2)$ στη χειρότερη περίπτωση, αλλά πολύ μικρότερος στην πράξη. Συνολικά βελτιστοποιούμε δύο φορές περισσότερα κελιά, αλλά η διάσταση των υπο-χώρων είναι η μισή.

3.4.5 Αναδιάταξη δευτέρου επιπέδου

Η ιδέα της κωδικοποίησης διανυσμάτων διαφορών για την προσέγγιση του αρχικού διανύσματος μπορεί να αναχθεί και σε διανύσματα διαφορών μεγαλύτερης τάξης. Η επέκταση αυτή αναλύεται στη δημοσίευση [46] για δευτέρας τάξης διανύσματα διαφοράς και αποτελεί μια μέθοδο συμπληρωματική στην τοπική βελτιστοποίηση. Παρόλα αυτά επιλέγουμε να μην βελτιστοποιήσουμε τοπικά και τα διανύσματα διαφοράς δευτέρου επιπέδου.

Στην περίπτωση χρήσης ενός γενικού (coarse) κβαντιστή Q με λεξικό $\mathcal{E}$, για κάθε διάνυσμα διαφοράς $\mathbf{z} \in \mathcal{Z}_i$ του κελιού $\mathbf{e}_i \in \mathcal{E}$, το τοπικά περιστραμμένο διάνυσμα διαφοράς $\hat{\mathbf{z}} = R_i^\mathsf{T}\mathbf{z}$ κωδικοποιείται ως $q_i(\hat{\mathbf{z}})$ από τον τοπικό παραγοντικό κβαντιστή q_i με τη διαδικασία που περιγράφεται στην ενότητα 3.4.1. Τώρα όμως, και το *δευτέρου επιπέδου διάνυσμα διαφοράς* $\hat{\mathbf{z}}' = \hat{\mathbf{z}} - q_i(\hat{\mathbf{z}})$ κωδικοποιείται επίσης ως $q'_i(\hat{\mathbf{z}}')$, όπου ο κβαντιστής q'_i είναι επίσης ένας (τοπικός) παραγοντικός κβαντιστής με m' υπο-κβαντιστές που έχει εκπαιδευτεί στο σύνολο των δευτέρου επιπέδου διανυσμάτων διαφορών $\hat{\mathbf{z}}'$ έτσι ώστε $\mathbf{z} \in \mathcal{Z}_i$.

Αν $\mathbf{y}$ είναι το σημείο αναζήτησης, για κάθε κελί $\mathbf{e}_i \in \mathcal{E}$ που επισκεπτόμαστε, μπορούμε τώρα να υπολογίσουμε την *δευτέρας τάξης* ασύμμετρη απόσταση $\delta_{q_i,q'_i}(\hat{\mathbf{y}}_i, \hat{\mathbf{z}}_p)$ ανάμεσα στα περιστραμμένα διανύσματα διαφορών $\hat{\mathbf{y}}_i = R_i^\mathsf{T}(\mathbf{y} - \mathbf{e}_i)$ και $\hat{\mathbf{z}}_p$ για κάποιο $p \in \mathcal{L}_i$, όπου

$$\delta_{q,q'}(\mathbf{y}, \mathbf{x}) = \|\mathbf{y} - q(\mathbf{x}) - q'(r_q(\mathbf{x}))\|^2. \tag{3.16}$$

Σε αντίθεση με την εξίσωση (3.6), όμως, ο παραπάνω υπολογισμός απαιτεί ανακατασκευή των διανυσμάτων και δεν μπορεί να εκτελεστεί μέσω lookup-tables. Γι αυτό, η προσέγγιση αυτή έχει νόημα μόνο αν χρησιμοποιηθεί σε δεύτερο στάδιο, για αναδιάταξη ενός μόνο μικρού ποσοστού των πιο κοντινών σημείων σύμφωνα με τα διανύσματα διαφοράς πρώτης τάξης. Η αναδιάταξη αυτού του είδους μπορεί εύκολα να προσαρμοστεί και για την περίπτωση του multi-index όπως αυτό περιγράφεται στην ενότητα 3.4.4.

3.5 Πειράματα

3.5.1 Πρωτόκολλο πειραμάτων

Σύνολα σημείων

Εκτελούμε πειράματα σε τέσσερα δημόσια σύνολα σημείων. Τρία από αυτά είναι δημοφιλή στην βιβλιογραφία της προσεγγιστικής αναζήτησης κοντινότερων γειτόνων: Τα σύνολα SIFT1M, GIST1M [42] and SIFT1B [46][1]. Το σύνολο SIFT1M περιέχει 1 εκατομμύριο 128-διάστατα διανύσματα περιγραφέων SIFT και 10K διανύσματα αναζήτησης, το σύνολο GIST1M περιέχει 1 εκατομμύριο 960-διάστατα διανύσματα GIST και 1000 διανύσματα αναζήτησης, ενώ το σύνολο SIFT1B περιέχει 1 δισεκατομμύριο διανύσματα SIFT και άλλα 10K διανύσματα για αναζήτηση.

Καθώς η προτεινόμενη μέθοδος LOPQ είναι ιδιαίτερα αποδοτική σε πολυτροπικές (multi-modal) κατανομές σημείων, εκτελούμε επιπλέον πειράματα και στο γνωστό σύνολο MNIST[2] όπως και στο συνθετικό σύνολο SYNTH1M που κατασκευάσαμε και περιγράψαμε στην ενότητα 3.4.3. Το σύνολο MNIST περιέχει 70 χιλιάδες εικόνες από χειρόγραφα ψηφία, το κάθε ένα από τα οποία αναπαρίσταται ως ένα 784-διάστατο διάνυσμα από τις φωτεινότητες των pixel. Όπως και στις δημοσιεύσεις [32, 31], και εδώ επιλέγουμε τυχαία 1000 από τα διανύσματα των ψηφίων ως διανύσματα αναζήτησης και χρησιμοποιούμε τα υπόλοιπα ως δεδομένα.

Αξιολόγηση

Όπως συνηθίζεται στη σχετική βιβλιογραφία [42, 32, 46, 9, 79, 78], μετράμε την απόδοση αναζήτηση μέσω του μέτρου ανάκλησης στη θέση R ή *recall@*R. Το μέτρο αυτό δίνει το ποσοστό των διανυσμάτων αναζήτησης των οποίον ο πραγματικός κοντινότερος γείτονας έχει αναχθεί στις R πρώτες θέσεις. Ή αλλιώς, το

[1] `http://corpus-texmex.irisa.fr/`

[2] `http://yann.lecun.com/exdb/mnist/`

μέτρο recall@R είναι το υποσύνολο των διανυσμάτων αναζήτησης για τα οποία ο κοντινότερος γείτονας θα έχει βρεθεί, αν υπολογίσουμε εξαντλητικά τις Ευκλείδειες αποστάσεις στις R πιο πιθανές θέσεις. Ιδιαίτερα σημαντικό είναι το μέτρο Recall@1 το οποίο είναι ταυτόσημο με το μέτρο ακρίβειας (*precision*) [75].

Αναδιάταξη

Ακολουθώντας τη μεθοδολογία της δημοσίευσης [46], τα διανύσματα διαφορών δευτέρας τάξης (second-order residuals) μπορούν να χρησιμοποιηθούν για αναδιάταξη, συμπληρωματικά σε όλες τις παραλλαγές της μεθόδου LOPQ, για να είναι η σύγκριση δίκαια όμως, εφαρμόζουμε τη μέθοδο αυτή μόνο στην περίπτωση του απλού ανεστραμμένου αρχείου. Έχουμε λοιπόν μια ακόμα παραλλαγή, τη LOPQ+R, κατά την οποία βελτιστοποιούμε τοπικά δευτέρου επιπέδου υπο-κβαντιστές. Παρόλα αυτά, δε βελτιστοποιούμε και τους πίνακες περιστροφής τοπικά για να αποφύγουμε το επιπλέον κόστος κατά την αναζήτηση σε σχέση με τη μέθοδο [46].

Ρυθμίσεις και υλοποίηση

Εκτελούμε πάντα την αναζήτηση μη εξαντλητικά, είτε χρησιμοποιώντας απλά ανεστραμμένα αρχεία είτε multi-index. Σε κάθε περίπτωση, θέτουμε το $k = 256$, απαιτούμε δηλαδή 8 bits για κάθε υπο-κβαντιστή. Επίσης, χρησιμοποιούμε 64-bit κωδικούς θέτοντας $m = 8$ σε όλες τις περιπτώσεις που δεν υπάρχει περαιτέρω σημείωση. Για τη μεγάλη συλλογή SIFT1B εκτελούμε πειράματα επίσης και με 128-bit κωδικούς, με $m = 16$, εκτός από όταν χρησιμοποιούμε αναδιάταξη, όπου θέτουμε $m = m' = 8$ όπως στη δημοσίευση [46]. Για όλες τις μεθόδους με multi-index, η παράμετρος T αναφέρεται στον συνολικό αριθμό διανυσμάτων που φέρνουμε μέσω του αλγορίθμου multi-sequence προτού σταματήσουμε την αναζήτηση.

Όλα τα αποτελέσματα που ακολουθούνται από παραπομπή έχουν αντιγραφεί απευθείας από τις σχετικές δημοσιεύσεις. Για τα υπόλοιπα χρησιμοποιούμε δικές μας υλοποιήσεις σε Matlab και C++. Για τον αλγόριθμο k-means καθώς και για εξαντλητική ανάθεση μέσω κοντινότερων γειτόνων χρησιμοποιούμε τη βιβλιοθήκη `yael`[3]. Εκτελούμε όλα μας τα πειράματα σε έναν 8-πύρινο υπολογιστή με 64GB RAM.

Μέθοδοι σύγκρισης στα σύνολα MNIST, SIFT1M, GIST1M

Εκτελούμε συγκρίσεις με τρεις από τις μεθόδους που παρουσιάστηκαν στην ενότητα 3.4.3. Σε όλες τις περιπτώσεις χρησιμοποιούμε απλά ανεστραμμένα αρχεία και έναν γενικό (coarse) κβαντιστή και παραγοντικό κβαντισμό στα διανύ-

[3] `https://gforge.inria.fr/projects/yael`

σματα διαφορών, ενώ όλες οι βελτιστοποιήσεις είναι γενικές (global). Πιο συγκεκριμένα, οι μέθοδοι που συγκρίνουμε έιναι οι IVFADC [42], η δικιά μας παραλλαγή I-PCA+RP, καθώς και η δικιά μας μη-εξαντλητική προσαρμογή της μεθόδου OPQ [32], χρησιμοποιώντας είτε OPQ_{p} είτε OPQ_{np} για την βελτιστοποίηση. Οι προαναφερθείσες μη-εξαντλητικές παραλλαγές δεν είναι μόνο πιο γρήγορες από τις αντίστοιχες εξαντλητικές, αλλά και καλύτερες σε απόδοση. Καθώς η μέθοδος OPQ_{np} απαιτεί πολύ χρόνο κατά την εκμάθηση, παρουσιάζουμε αποτελέσματα με αυτή μόνο στη μικρή συλλογή MNIST. Σε όλες τις άλλες περιπτώσεις, ως I-OPQ συμβολίζουμε τη μέθοδο OPQ_{p}. Δε συγκρίνουμε με τις μεθόδους transform coding [14] ή ITQ [33], καθώς η μέθοδος I-OPQ έχει καλύτερη απόδοση από αυτές [32].

Μέθοδοι σύγκρισης στο σύνολο SIFT1B

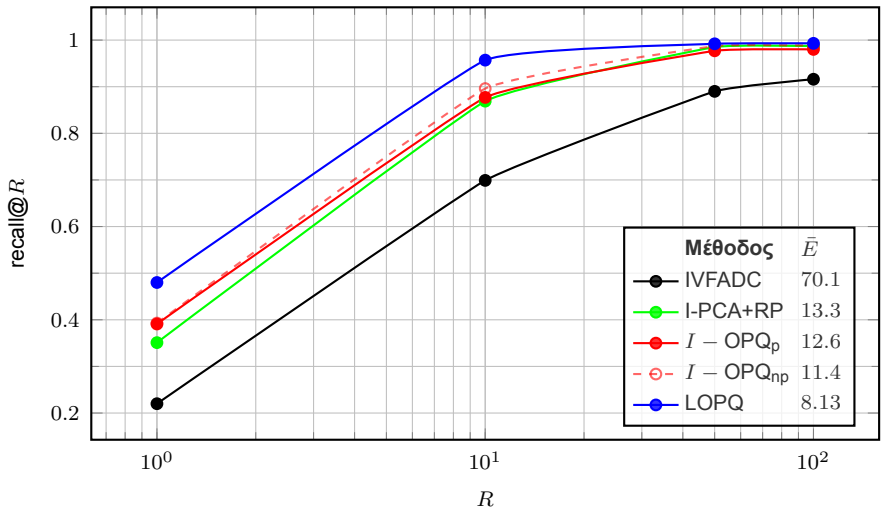
Μετά από τα πειράματα με απλό ανεστραμμένο αρχείο στα οποία συγκρίνουμε την προτεινόμενη μέθοδο κυρίως με τις μεθόδους IVFADC και I-OPQ, εστιάζουμε στη χρήση του multi-index για τα μεγάλης κλίμακας πειράματα, όπου και συγκρίνουμε την μέθοδό μας με τις μεθόδους Multi-D-ADC [9] και OMulti-D-OADC [31], μέθοδοι που δίνουν τα καλύτερα αποτελέσματα (state-of-the-art) στην αναζήτηση κοντινότερου γείτονα σε μεγάλη κλίμακα. Και οι δύο μέθοδοι εκτελούν παραγοντικό κβαντισμό στα διανύσματα διαφορών σε κάθε έναν από τους δύο κβαντιστές υποχώρων του multi-index. Η μέθοδος OMulti-D-OADC χρησιμοποιεί επίσης τον αλγόριθμο OPQ_{np} για να βελτιστοποιήσει συνολικά την περιστροφή των δεδομένων και πριν από τη δεικτοδότηση για την κατασκευή των υπόχωρων του multi-index, αλλά και για τα διανύσματα διαφορών έπειτα. Παρουσιάζουμε επίσης αποτελέσματα και για αρκετές άλλες πρόσφατες μεθόδους, όπως τη μέθοδο IVFADC με αναδιάταξη (IVFADC+R) [46] και τις μέθοδους Ck-means [78], KLSH-ADC [82], multi-index hashing (Multi-I-Hashing) [79] και joint inverted indexing (Joint-ADC) [114].

3.5.2 Αποτελέσματα στα σύνολα MNIST, SIFT1M, GIST1M

Το σύνολο MNIST είναι το πρώτο στο οποίο παρουσιάζουμε αποτελέσματα. Είναι το μόνο σύνολο σημείων για το οποίο παρουσιάζουμε αποτελέσματα και για τη μέθοδο OPQ_{np}, καθώς υπερτερεί κατά πολύ της μεθόδου OPQ_{p} στο σύνολο αυτό [32]. Όπως προτείνεται και στη δημοσίευση [32], εκτελούμε 100 επαναλήψεις για τη βελτιστοποίηση του OPQ_{np} χρησιμοποιώντας την υλοποίηση που δίνουν οι ίδιοι [4].

Στο σχήμα 3.3 παρουσιάζουμε αποτελέσματα που συγκρίνουν την ανάκληση

[4] `http://research.microsoft.com/en-us/um/people/kahe/cvpr13/`

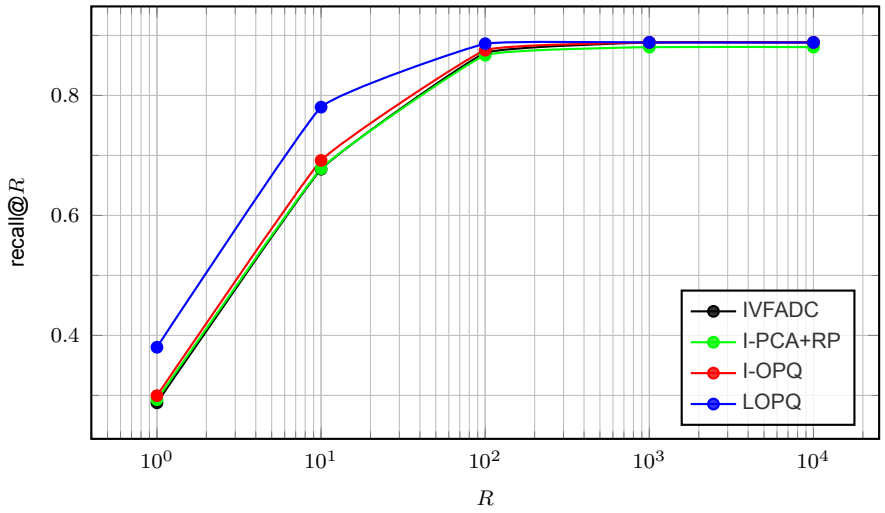


Σχήμα 3.3: *Ανάκληση στα πρώτα R δείγματα (recall@R) για τη συλλογή MNIST με $K = 64$, το οποίο επιλέχθηκε ως βέλτιστο και $w = 8$. $\bar{E} = E/n$: μέση παραμόρφωση ανά σημείο.*

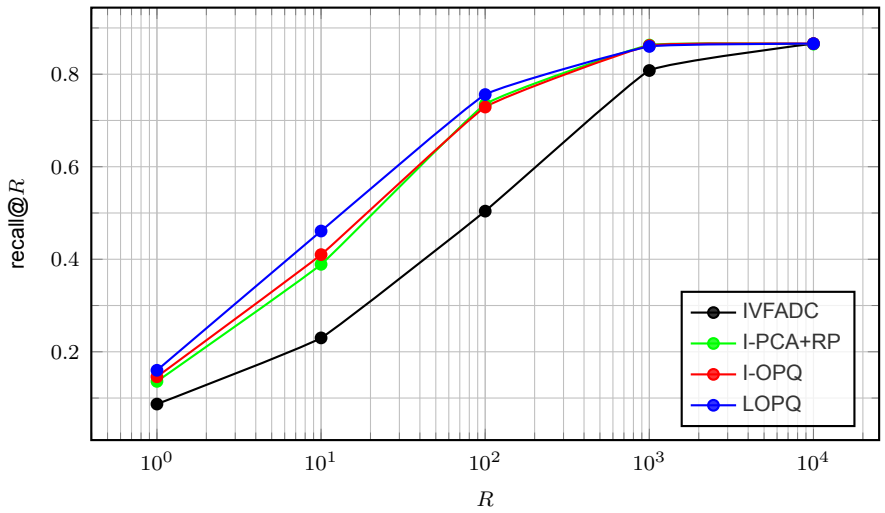
και την παραμόρφωση. Μια σημαντική παρατήρηση είναι ότι το κέρδος της μεθόδου OPQ$_{np}$ σε σχέση με την OPQ$_p$ είναι πλύ μικρό τώρα που η ολική βελτιστοποίηση γίνεται στα διανύσματα διαφορών. Μια πιθανή εξήγηση για αυτό είναι ότι τα διανύσματα διαφορών αναμένεται να ακολουθούν μια κατανομή πιο κοντά σε μονοτροπική, πιο κοντά δηλαδή στην υπόθεση που απαιτεί η μέθοδος OPQ$_p$. Επίσης, η απόδοση της πολύ απλής μεθόδου I-PCA+RP με τυχαία ανάθεση των διαστάσεων μετά το PCA αποδίδει σχεδόν το ίδιο με τις μεθόδους I-OPQ. Παρόλα αυτά, είναι ξεκάθαρο ότι η προτεινόμενη μέθοδος μας LOPQ που βελτιστοποιεί τοπικά τα διανύσματα διαφορών υπερτερεί κατά πολύ όλων των άλλων μεθόδων.

Αποτελέσματα για τα σύνολα σημείων SIFT1M και GIST1M απεικονίζονται στα Σχήματα 3.4 και 3.5 αντίστοιχα, όπου τώρα η μέθοδος OPQ πλέον αναφέρεται στη παραμετρική OPQ$_p$. Όπως και στη δημοσίευση [32], χρησιμοποιούμε τη βέλτιστη για κάθε σύνολο διάταξη των διαστάσεων για τη βασική μέθοδο IVFADC [42], τη φυσική και δομική δηλαδή σειρά αντίστοιχα για το σύνολα SIFT1M και GIST1M. Και στα δύο αυτά σύνολα, η προτεινόμενη μέθοδος LOPQ επιτυγχάνει ξεκάθαρα υψηλότερη απόδοση σε σχέση με τις μεθόδους που εκτελούν τις βελτιστοποιήσεις ολικά.

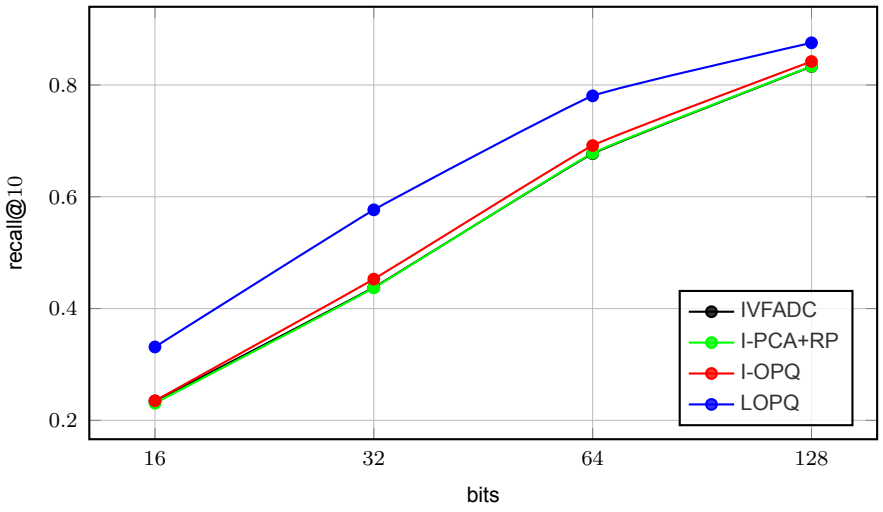
Τα Σχήματα 3.6 και 3.7 παρουσιάζουν το μέτρο της ανάκλησης recall@10 ως προς την ανάθεση bit ανά σημείο (μεταβάλλοντας την παράμετρο m) και ως προς το μέγεθος της γειτονιάς w του soft assignment, αντιστοίχως. Η μέθοδος LOPQ έχει καλύτερη απόδοση σε όλες τις περιπτώσεις, με τη διαφορά να αυξάνεται όσο μικραίνουμε τα bit ανάθεσης και κάνοντας περισσότερο soft assignment, κάτι που σημαίνει καλύτερες μετρήσεις στην απόσταση, άρα και συνολικά χαμηλότερη παραμόρφωση.



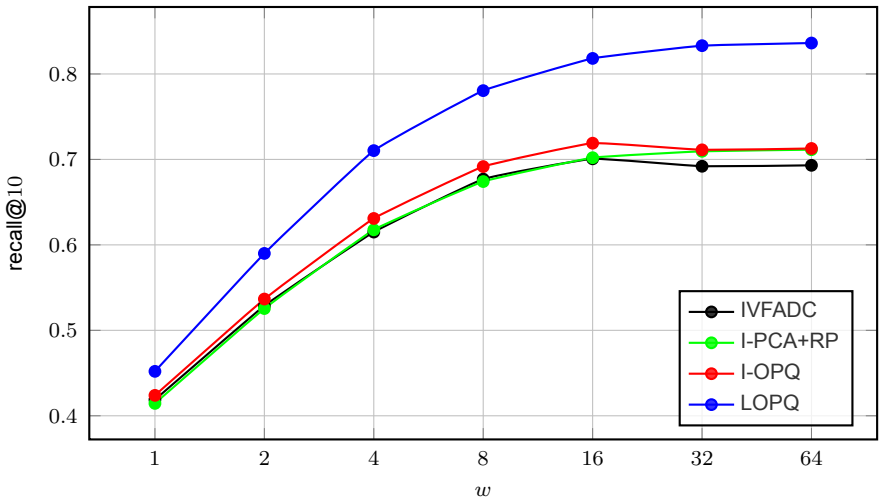
Σχήμα 3.4: *Ανάκληση στα πρώτα R δείγματα (recall@R) για τη συλλογή SIFT1M με παραμέτρους $K = 1024$, $w = 8$.*



Σχήμα 3.5: *Ανάκληση στα πρώτα R δείγματα (recall@R) για τη συλλογή GIST1M με παραμέτρους $K = 1024$, $w = 16$.*



Σχήμα 3.6: *Ανάκληση στα πρώτα R δείγματα (recall@R) για τη συλλογή SIFT1M ως πρός το μέγεθος των κωδικών ανα σημείο (bit allocation per point), με παραμέτρους $K = 1024$ και $w = 8$. Για $16, 32, 64$ και 128 bits, η παράμετρος m παίρνει τιμές $2, 4, 8$ και 16 αντιστοίχως.*



Σχήμα 3.7: *Ανάκληση στα πρώτα R δείγματα (recall@R) για τη συλλογή SIFT1M ως προς την παράμετρο w, με $K = 1024$ και $m = 8$.*

Μέθοδος	$R = 1$	$R = 10$	$R = 100$
Ck-means [78]	–	–	0.649
IVFADC	0.106	0.379	0.748
IVFADC [46]	0.088	0.372	0.733
I-OPQ	0.114	0.399	0.777
Multi-D-ADC [9]	0.165	0.517	0.860
LOR+PQ	0.183	0.565	0.889
LOPQ	**0.199**	**0.586**	**0.909**

Πίνακας 3.1: *Ανάκληση στα πρώτα $\{1, 10, 100\}$ δείγματα, για το σύνολο SIFT1B με κωδικούς μεγέθους 64-bit, $K = 2^{13} = 8192$ και $w = 64$. Για τη μέθοδο Multi-D-ADC, το $K = 2^{14}$ και $T = 100$K.*

3.5.3 Αποτελέσματα στο σύνολο SIFT1B

Για κωδικοποιήσεις μεγέθους 64-bit code ($m = 8$) αποτελέσματα παρουσιάζονται στον Πίνακα 3.1 και περιλαμβάνουν τις μεθόδους I-OPQ, Ck-means [78], Multi-D-ADC [9] και IVFADC χωρίς ανακατάταξη (re-ranking), καθώς όπως αναφέρεται στην δημοσίευση [46] η ανακατάταξη δέν δίνει βελτίωση για αυτό το μέγεθος κώδικα. Όλες οι μέθοδοι χρησιμοποιούν απλό ανεστραμμένο αρχείο εκτός από τις μεθόδους Multi-D-ADC η οποία χρησιμοποιεί multi-index και Ck-means η οποία είναι εξαντλητική. Για τη μέθοδο IVFADC αναπαράγουμε τα νούμερα από τη δημοσίευση [46] και αναφέρουμε επίσης τα αποτελέσματα της δικιάς μας υλοποίησης της μεθόδου. Για να δείξουμε και το κέρδος από την τοπική βελτιστοποίηση των περιστροφών και των υπο-κβαντιστών, παρουσιάζουμε στα αποτελέσματα και την υπο-βέλτιστη παραλλαγή LOR+PQ η οποία συζητήθηκε στην ενότητα 3.4.1. Και οι δύο μέθοδοι LOR+PQ και LOPQ είναι ξεκάθαρα ανώτερες από όλες τις άλλες μεθόδους, με κέρδος που φτάνει στο 18% σε σχέση με τη μέθοδο I-OPQ και 7% σε σχέση με τη Multi-D-ADC για το μέτρο recall@10, παρότι η δεύτερη χρησιμοποιεί πιο προηγμένη/ανώτερη μορφή δεικτοδότησης (multi-index).

Για κωδικοποιήσεις μεγέθους 128-bit, παρουσιάζουμε τα αποτελέσματα μας σε σχέση με το state-of-the-art στον Πίνακα 3.2 και το Σχήμα 3.8. Οι μέθοδοι που προτείνουμε εδώ περιλαμβάνουν λύσεις με απλό ανεστραμμένο αρχείο και ανακατάταξη με διανύσματα διαφορών δευτέρας τάξης (LOPQ+R) αλλά και με multi-index (Multi-LOPQ). Σε σύγκριση μέ τις άλλες μεθόδους με ανακατάταξη, η LOPQ+R έχει σαφές πλεονέκτημα σε απόδοση σε σύγκριση με τη IVFADC+R, όπου ακολουθούμε το μοντέλο $m = m' = 8$ το οποίο έχει αποδειχθεί ότι δίνει την καλύτερη απόδοση [46]. Για όλες τις άλλες μεθόδους και παραλλαγές χρησιμοποιούμε $m = 16$. Οι μέθοδοι Multi-I-Hashing [79], KLSH-ADC [79] και Joint-ADC [114] είναι κατά πολύ κατώτερες για $R = 100$, αν και η τελευταία απαιτεί 4 φορές περισσότερη μνήμη.

T	Μέθοδος	$R=1$	10	100
	Multi-I-Hashing [79]	–	–	0.420
	KLSH-ADC [82]	–	–	0.894
	Joint-ADC [114]	–	–	0.938
20K	IVFADC+R [46]	0.262	0.701	0.962
	LOPQ+R	**0.350**	**0.820**	**0.978**
10K	Multi-D-ADC [9]	0.304	0.665	0.740
	OMulti-D-OADC [31]	0.345	0.725	**0.794**
	Multi-LOPQ	**0.430**	**0.761**	0.782
30K	Multi-D-ADC [9]	0.328	0.757	0.885
	OMulti-D-OADC [31]	0.366	0.807	**0.913**
	Multi-LOPQ	**0.463**	**0.865**	0.905
100K	Multi-D-ADC [9]	0.334	0.793	0.959
	OMulti-D-OADC [31]	0.373	0.841	**0.973**
	Multi-LOPQ	**0.476**	**0.919**	**0.973**

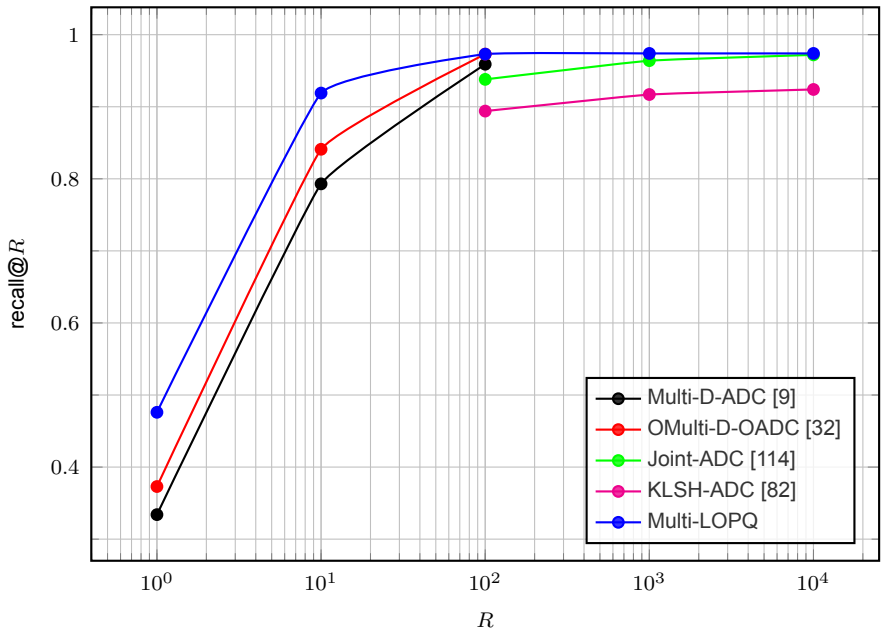
Πίνακας 3.2: *Ανάκληση στα πρώτα $\{1, 10, 100\}$ δείγματα, για το σύνολο SIFT1B με κωδικούς μεγέθους 128-bit, $K = 2^{13} = 8192$ ($K = 2^{14}$) για απλό ανεστραμμένο αρχείο (multi-index). Για τις μεθόδους IVFADC+R και LOPQ+R, $m' = 8$, $w = 64$. Τα αποτελέσματα των μεθόδων Joint-ADC και KLSH-ADC τα πήραμε από τη δημοσίευση [114]. Οι γραμμές με αναφορά αναφέρονται στα ακριβή δημοσιευμένα νούμερα των αντίστοιχων δημοσιεύσεων.*

Για να πάρουμε τα αποτελέσματα που είναι πλέον το state-of-the-art, χρησιμοποιούμε δεικτοδότηση multi-index, και με αυτό τον τρόπο καταφέρνουμε να έχουμε και ιδιαίτερα μικρούς χρόνους αναζήτησης. Συγκρίνουμε με την μέθοδο που είχε τα καλύτερα δημοσιευμένα νούμερα OMulti-D-OADC [31] παίρνοντας τη μέθοδο Multi-D-ADC [9]. Η απόδοση που καταφέρνει η μέθοδος μας Multi-LOPQ είναι αξεπέραστη και θέτει το νέο state-of-the-art, τα καλύτερα δημοσιευμένα νούμερα απόδοσης για τη βάση SIFT1B, έχοντας κέρδος 10% σε σχέση με τη μέθοδο OMulti-D-OADC στο πλέον σημαντικό μέτρο για την αναζήτηση κοντινότερου γείτονα, την ακρίβεια (precision ή recall@1).

3.5.4 Ανάλυση της ταχύτητας αναζήτησης και της μνήμης

Η μέθοδος LOPQ απαιτεί λίγο παραπάνω μνήμη σε σχέση με τον παραγοντικό κβαντισμό καθώς και λίγο παραπάνω χρόνο κατά την εκμάθηση, τη δεικτοδότηση και την αναζήτηση. Όμως όλα αυτά είναι σταθερά ως προς το μέγεθος της συλλογής σημείων.

Μνήμη.. Ο επιπλέον χώρος που απαιτείται σε σχέση με τη μέθοδο IVFADC (αντίστοιχα Multi-D-ADC) αναφέρεται στους τοπικούς πίνακες περιστροφών και



Σχήμα 3.8: *Ανάκληση στο σύνολο SIFT1B με κωδικούς μεγέθους* 128*-bit και* $T = 100$*K, όπως παρουσιάζεται και στον πίνακα 3.2.*

τους υπο-κβαντιστές ανά κελί. Για τους πίνακες απαιτείται επιπλέον μνήμη Kd^2 (αντίστοιχα $2K(d/2)^2$) για απλό ανεστραμμένο αρχείο (αντίστοιχα multi-index). Στην πράξη, αυτός ο επιπλέον χώρος είναι περίπου 500MB για το SIFT1B. Για τα κέντρα των υπο-κβαντιστών, ο επιπλέον χώρος είναι Kdk σε όλες τις περιπτώσεις. Πρακτικά είναι 2GB για το SIFT1B και τη μέθοδο Multi-LOPQ με $K = 2^{14}$.

Χρόνος αναζήτησης.. Ο επιπλέον χρόνος που απαιτείται κατά την αναζήτηση σε σχέση με τη μέθοδο IVFADC (αντίστοιχα Multi-D-ADC) είναι μοναχά ο χρόνος που απαιτείται για την περιστροφή του σημείου αναζήτησης σε κάθε κελί, είναι δηλαδή w ($2K$ worst-case) γινόμενα ενός πίνακα $d \times d$ ($\frac{d}{2} \times \frac{d}{2}$) και ενός d ($\frac{d}{2}$)-διάστατου διανύσματος για το απλό ανεστραμμένο αρχείο (αντίστοιχα multi-index). Στην πράξη, τα γινόμενα που απαιτούνται στην περίπτωση του multi-index είναι πολύ λιγότερα και ο μέσος επιπλέον χρόνος που απαιτείται στη συλλογή SIFT1B για τη μέθοδο Multi-LOPQ είναι $0.776, 1.92, 4.04$ms αντίστοιχα για $T = 10$K, 30K, 100K.

Η μέθοδος Multi-D-ADC απαιτεί 49ms για $T = 100$K [9], συνεπώς συνολικά για τη μεθοδό μας απαιτούνται 53ms. Ή, σε μονάχα 7ms για $T = 10$K [9], η προτεινόμενη μέθοδος Multi-LOPQ ξεπερνά κατά by 5% το καλύτερο προηγούμενο δημοσιευμένο αποτέλεσμα στο σύνολο SIFT1B, με κωδικούς μεγέθους $128bit$ απαιτώντας λιγότερο από 8ms.

Κεφάλαιο 4

Γεωγραφική και οπτική ομαδοποίηση

4.1 Εισαγωγή

Στην τρέχων κεφάλαιο θα παρουσιάσουμε τους *χάρτες σκηνών* και θα δείξουμε ότι μια εκ των προτέρων ομαδοποίηση των εικόνων της συλλογής μπορεί να βελτιώσει την απόδοση της οπτικής αναζήτησης, ενώ παράλληλα ένα κριτήριο παραμόρφωσης (distortion) μπορεί να εγγυηθεί την ανάκτηση ακόμα και απομονωμένων εικόνων από μη δημοφιλής τοποθεσίες όπως σε ένα γενικό σύστημα αναζήτησης εικόνων. Προτείνουμε μια λύση που παρότι μπορεί να δουλέψει σε συλλογές εκατομμυρίων εικόνων, μπορεί και να ανακτήσει τις μη δημοφιλής εικόνες.

Αρχίζοντας από μια μεγάλη συλλογή εικόνων με γνωστή τη γεωγραφική τους θέση ή geo-tag, τις ομαδοποιούμε πρώτα γεωγραφικά, κατασκευάζοντας *γεωγραφικές ομάδες*. Ο σκοπός αυτής της διαδικασίας είναι να αναγνωριστούν εικόνες που πιθανώς απεικονίζουν όψης της ίδιας σκηνής. Δύο εικόνες που έχουν τραβηχτεί με απόσταση 2 χιλιομέτρων, για παράδειγμα, είναι απίθανο να απεικονίζουν το ίδιο κτίριο. Στη συνέχεια, χρησιμοποιώντας γρήγορες δομές δεικτοδότησης για αποτελεσματικότητα, υπολογίζουμε τις οπτικές αποστάσεις μεταξύ των εικόνων μίας γεωγραφικής ομάδας και τις ομαδοποιούμε κατάλληλα ώστε εικόνες που απεικονίζουν την ίδια σκηνή ή τοποθεσία να δημιουργήσουν *οπτικές ομάδες*. Με δεδομένη μία οπτική ομάδα, ευθυγραμμίζουμε τις εικόνες που περιέχει ως προς μία εικόνα αναφοράς εκτιμώντας το μεταξύ τους ομογραφικό ταίριασμα και κατασκευάζουμε έναν δισδιάστατο *χάρτη σκηνής* ενώνοντας τα τοπικά χαρακτηριστικά των εικόνων της οπτικής ομαδας που μοιάζουν, συμπιέζοντας την περιττή πληροφορία. Επεκτείνουμε τις διαδικασίες της δεικτοδότησης, της αναζήτησης και χωρικού ταιριάσματος ώστε να λειτουργούν με χάρτες σκηνής αντί για εικόνες. Έτσι, όχι μόνο μειώνουμε την απαιτούμενη μνήμη, αλλά επίσης αυξάνουμε και τα επίπεδα ανάκλησης σχετικών εικόνων.

Εκτελούμε πειράματα σε μία δύσκολη συλλογή ενός εκατομμυρίου αστικών εικόνων ο οποία περιέχει εικόνες από 22 Ευρωπαϊκές πόλεις. Η όλη διαδικασία ομαδοποίησης και εξόρυξης είναι ιδιαίτερα αποτελεσματική και καθόλα αυτοματοποιημένη. Χρειάστηκαν για αυτήν περίπου δύο μέρες επεξεργασία, χρησιμοποιώντας ένα μόνο οκταπύρηνο μηχάνημα και με δεδομένη την απλή βασική δομή δεικτοδότησης. Κατά τη διαδικασία της αναζήτησης, απαιτούνται milliseconds για να φιλτραριστούν οι σχετικοί χάρτες σκηνών, ενώ το γεωμετρικό ταίριασμα απαιτεί λίγα δευτερόλεπτα. Αρκεί να βρεθεί μονάχα *ένα* επαληθευμένο ταίριασμα από τη συλλογή για να έχουμε μια εκτίμηση της τοποθεσίας της εικόνας αναζήτησης.

4.2 Σχετική βιβλιογραφία

4.2.1 Αναγνώριση Τοποθεσίας

Σε μια από τις πρώτες δημοσιεύσεις για ταίριασμα πολλαπλών όψεων σε αστικά τοπία, οι Johansson και Cipolla [47] προσπαθούν να εκτιμήσουν τους ομογραφικούς μετασχηματισμούς μεταξύ ζευγών εικόνων και να παράγουν αυτόματη *εκτίμηση της άποψης της κάμερας* (*pose estimation*). Χρησιμοποιώντας ακμές και γωνίες σαν περιγραφείς των εικόνων, η τεχνική που προτείνουν, όπως και η μετέπειτα τεχνική των Robertson και Cipolla [89] περιορίζονται σε απλές γεωμετρικές δομές όπως προσόψεις κτιρίων. Περιγραφείς SIFT χρησιμοποιούνται στις προσέγγισεις των [68], Zhang και Kosecka [115] και η αναζήτηση και το ταίριασμα εκτελείται απευθείας στον χώρο των περιγραφέων με σκοπό να ανακληθεί η κοντινότερη όψη από μια μικρή βάση εικόνων, παρέχοντας μια χοντροκομμένη *εκτίμηση θέσης* σε αστικά περιβάλλοντα. Μια λεπτομερέστερη εκτίμηση της όψης της κάμερας παράγεται ακολούθως με χρήση του αλγορίθμου RANSAC με μοντέλο 7 ή 8 βαθμών ελευθερίας, ενώ ακριβής εκτίμηση της θέσης της κάμερας στον τρισδιάστατο χώρο (*localization* in 3D) απαιτεί τριγωνοποίηση με χρήση της εικόνας αναζήτησης και δύο ακόμα όψεων αναφοράς.

Στη δημοσίευση [70] χρησιμοποιούνται οι περιοχές MSER μαζί με γρήγορη αναζήτηση κοντινότερων εικόνων, ένα μοντέλο το οποίο επεκτείνουν οι Steinhoff *et al*. [100] για να παράγουν εκτίμηση της άποψης σε ελάχιστο χρόνο, σχεδόν αρκετά γρήγορα για χρήση σε συστήματα πραγματικού χρόνου, όπως *συνεχής εκτίμηση θέσης* σε κινητές συσκευές, με ακρίβεια αντίστοιχη με εκείνη του GPS. Σε αυτή την περίπτωση όμως η βάση εικόνων περιορίζεται σε 600 εικόνες αναφοράς από αστικό περιβάλλον που καλύπτουν μερικά οικοδομικά τετράγωνα. Οι Schindler *et al*. [91] είναι από τους πρώτους που χρησιμοποιούν δεικτοδότηση ανεστραμμένων αρχείων μέσω ιεραρχικών λεξικών [76] για αναγνώριση τοποθεσίας στην *κλίμακα μιας πόλης*, με την τεχνική τους να εφαρμόζεται σε συλλογές μέχρι $30,000$ εικόνων που καλύπτουν περίπου $20km$ από όψεις κατα μήκος δρό-

μων (streetside views).

Οι Hayes και Efros [36] είναι από τους πρώτους που επιχειρούν εκτίμηση τοποθεσίας *παγκόσμιας κλίμακας* ψέχνοντας σε μια συλλογή έξι εκατομμυρίων εικόνων με γνωστή τοποθεσία που προέρχονται από τον ιστότοπο Flickr. Το τίμημα είναι ότι οι εικόνες αναπαρίστανται με περιγραφείς από ολόκληρη την εικόνα (global features), όπως για παράδειγμα ιστογράμματα χρώματος και υφής ή περιγραφείς GIST [80]. Η ακρίβεια ταιριάσματος απέχει πολύ από εκείνη των τοπικών χαρακτηριστικών με συνέπεια η έξοδος του αλγορίθμου να είναι ένας πιθανοτικός χάρτης γεωγραφικής θέσης (*geolocation probability map*). Σε μετέπειτα δημοσίευση οι Kalogerakis *et al*. [53] αναπτύσσουν περισσότερο την προηγούμενη ιδέα και εκμεταλλεύονται επίσης τον χρόνο που τραβήχτηκε η κάθε εικόνα, κάτι που συμβαίνει και στη δημοσίευση [22]. Η έξοδος παραμένει πιθανοτικός χάρτης και η συγκεκριμένη τεχνική δουλεύει με *ακολουθίες εικόνων* και όχι με μια μοναδική εικόνα αναζήτησης. Πρόσφατα έχουν αναπτυχθεί ακριβέστεροι αλγόριθμοι για αναγνώριση θέσης οι οποίες ναι μεν μπορούν να εφαρμοστούν σε παγκόσμια κλίμακα, περιορίζονται όμως μόνο σε εικόνες ορόσημων (landmarks). Μερικές από αυτές εξετάζονται στις παρακάτω υποενότητες.

4.2.2 Αναγνώριση Ορόσημων

Οι Kennedy *et al*. [54] είναι σίγουρα από τους πρώτους που ασχολήθηκαν με την εξόρυξη δημοφιλών τοποθεσιών και ορόσημων σε μεγάλη κλίμακα, ξεκινώντας από 10^7 εικόνες του Flickr που περιείχαν επίσης μεταδεδομένα θέσης, χρηστών και τοποθεσίας. Παρότι η ομαδοποίηση που πραγματοποιούν με βάση τα δημοφιλή tags και τις τοποθεσίες βοηθά στην κατασκευή *γεωγραφικών χαρτών για τα tags* (*tag maps*) για αυθαίρετες περιοχές στον κόσμο, η μετέπειτα οπτική ομαδοποίηση δεν είναι αποδοτική, κυρίως λόγω των περιγραφέων που χρησιμοποιούνται, οι οποίοι προέρχονται από ολόκληρη την εικόνα. Παρομοίως, οι Crandall *et al*. [22], ανιχνεύουν δημοφιλείς γεωγραφικές περιοχές, περιοχές δηλαδή με υψηλή πυκνότητα εικόνων και εξάγουν με αυτόματο τρόπο τα ονόματα των ορόσημων από τα τοπικά tags. Οι σχετικές εικόνες χρησιμοποιούνται έπειτα σαν δεδομένα εκπαίδευσης σε ένα πρόβλημα εκμάθησης. Δυστυχώς η αυτόματη αυτή επιλογή των εικόνων χωρίς επίβλεψη καταλήγει να είναι θορυβώδης, με αποτέλεσμα τα οπτικά χαρακτηριστικά από μόνα του να έχουν κακή απόδοση, συγκρίσιμη σε μερικές περιπτώσεις με την τυχαία επιλογή. Οι Li *et al*. [63] βελτιώνουν λίγο την απόδοση χρησιμοποιώντας ταξινομητές SVM πολλαπλών κλάσεων (multi-class SVM). Το πρόβλημα προσεγγίζεται σαν *αναγνώριση αντικειμένων*, ένα ιδιαίτερα δύσκολο πρόβλημα για 30 εκατομμύρια εικόνες, από τις οποίες τα 2 δύο εκατομμύρια είναι επισημειωμένα σε μια από τις 500 πιθανές κατηγορίες. Οι τεχνικές με δομές δεικτοδότησης αποδίδουν καλύτερα από τη συγκεκριμένη προσέγγιση με

εκμάθηση.

Οι Simon *et al*. [95] εστιάζουν στην οπτική ομαδοποίηση χωρίς δεδομένα τοποθεσίας, ακολουθώντας μια πιό δομημένη προσέγγιση βελτιστοποίησης μέσω της οποίας διαλέγουν μια σειρά από *χαρακτηριστικές όψεις* (*canonical views*) και κατασκευάζουν μια *περίληψη της σκηνής* για εξερεύνηση. Δυστυχώς η τεχνική αυτή δεν μπορεί να εφαρμοστεί σε πάνω από 10^4 εικόνες. Μια άλλη παρόμοια ιδέα είναι οι *Ιστοί εικόνων* (*Image webs*) [38], όπου οι συγγραφείς αντιμετωπίζουν το μεγάλο υπολογιστικό κόστος με παράλληλα συστήματα. Οι Chum και Matas [20] επεκτείνονται σε ομαδοποίηση εικόνων *παγκόσμιας κλίμακας* επίσης χωρίς να χρησιμοποιούν δεδομένα τοποθεσίας, βασιζόμενοι σε μεθόδους κατακερματισμού για να αναγνωρίσουν εικόνες σχεδόν διπλότυπες (near-duplicates). Η προσέγγιση αυτή οδηγεί σε μεγάλη αύξηση της απόδοσης, με την προϋπόθεση βέβαια ότι ανακαλύπτονται δημοφιλείς τοποθεσίες με μεγάλο αριθμό σχετικών εικόνων.

Οι Quack *et al*. [87] διαιρούν τον γεωγραφικό χάρτη σε επικαλυπτώμενες τετραγωνικές περιοχές ενδιαφέροντος. Όπως και στη δημοσίευση [54] και σε αντίθεση με τις δημοσιεύσεις [95, 20], εκτελούν την οπτική ομαδοποίηση ανεξάρτητα σε κάθε περιοχή για αποδοτικότητα. Εκτελούν, όμως, εξαντλητικά σε όλα τα πιθανα ζεύγη εικόνων της περιοχής εκτιμήσεις ομογραφίας, χάνοντας ουσιαστικά το υπολογιστικό τους προβάδισμα. Αν και η εξόρυξη οροσήμων, αντικειμένων και γεγονότων γίνεται μια φορά και όχι κατά την διάρκεια της αναζήτησης (offline), η αναγνώριση τοποθεσίας μιας νέας εικόνας αναζήτησης είναι περιορισμένη και αργή, καθώς η αναζήτηση εκτελείται με εξαντλητικό/γραμμικό τρόπο. Οι Gammeter *et al*. [28] βελτιώνουν την προαναφερθείσα προσέγγιση χρησιμοποιώντας δεικτοδότηση ανεστραμμένων αρχείων, αλλά η διαδικασία της εξόρυξης παραμένει τετραγωνική ως προς τον αριθμό των εικόνων σε κάθε γεωγραφική περιοχή. Παρόλα αυτά, προσθέτουν μια ανάστροφη αναζήτηση μέσω άρθρων της Wikipedia, όπου αντικείμενα ενδιαφέροντος ανιχνεύονται αυτόματα και επισημειώνονται στις εικόνες. Τέλος, οι Zheng *et al*. πραγματοποιούν στη δημοσίευση τους [117] έναν παρόμοιο συνδυασμό γεωγραφικής και οπτικής ομαδοποίησης, καθώς και ανάστροφη αναζήτηση μέσω ταξιδιωτικών άρθρων που περιέχουν και ονόματα ορόσημων. Και σε αυτή την περίπτωση δεν χρησιμοποιείται καμία δομή δεικτοδότησης κατά την εξόρυξη και το ιδιαίτερα υψηλό υπολογιστικό κόστος αντιμετωπίζεται πάλι με χρήση παράλληλων υπολογιστικών συστημάτων.

4.2.3 Ανακατασκευή τρισδιάστατων σκηνών

Μια παρεμφερής και ιδιαίτερα ενδιαφέρουσα εφαρμογή είναι η ανακατασκευή με μεθόδους όρασης και η πλοήγηση σε τρισδιάστατες σκηνές δεδομένης μιας συλλογής από όψεις της ίδιας σκηνής. Στοχεύοντας στην ανακατασκευή από μικρές προσωπικές συλλογές, οι Schaffalitzky and Zisserman [90] προτείνουν μια από

τις πρώτες μεθόδους. Χρησιμοποιούνται τοπικά χαρακτηριστικά τα οποία έπειτα διαμορφώνουν, και έτσι τα ταιριάσματα ζευγών εικόνων συνδυάζονται σε μια συνολική όψη της συλλογής. Αυτού το είδους οι τεχνικές, που εμπίπτουν στην κατηγορία *structure from motion*, επεκτάθηκαν παραπέρα από τους Snavely *et al*. [98] ώστε να μπορούν να εφαρμοστούν σε συλλογές της τάξεως των 10^3 εικόνων, που προέρχονται από αναζητήσεις με κείμενο στο Flickr.

Δουλεύοντας σε συλλογές παρομοίου μεγέθους, οι Li *et al*. [62] προσπαθούν να επιταχύνουν τη διαδικασία ανακατασκευής επιστρατεύοντας μια ιεραρχική μέθοδο, κατασκευάζοντας τελικά έναν γράφο από *εικονικές όψεις* (*iconic scene graph*). Δυστυχώς με την χρήση ολικών περιγραφέων, η αύξηση στην ταχύτητα συνοδεύεται από πτώση στην απόδοση. Οι Snavely *et al*. [99] από την άλλη εφαρμόζουν την ιδέα ενός *σκελετικού γράφου* (*skeletal graph*) για την σύντμηση της οπτικής πληροφορίας με σκοπό την επιτάχυνση την διαδικασίας ανακατασκευής. Σε μία ακραία πρόσφατη εφαρμογή των παραπάνω μεθόδων, οι Agarwal *et al*. [3] επιτυγχάνουν την ανακατασκευή ολόκληρων πόλεων, από συλλογές της τάξεως των 10^5 εικόνων από το Flickr. Για να το επιτύχουν αυτό, χρησιμοποιούν μια μαζικά παράλληλη αρχιτεκτονική και εκμεταλλεύονται τεχνικές cloud computing.

Μια ενδιαφέρουσα παρατήρηση είναι ότι παρότι οι παραπάνω εφαρμογές είναι ίσως οι πλέον υπολογιστικά απαιτητικές, καμία από αυτές δεν αξιοποιεί πληροφορία τοποθεσίας για την καθοδήγηση και υποβοήθηση της διαδικασίας ομαδοποίησης. Η παράβλεψη αυτή είναι σημαντική, όχι μόνο γιατί με τη χρήση της τοποθεσίας κάθε υποπρόβλημα θα ήταν μικρότερο, αλλά και γιατί οι εικόνες με geo-tag αναπαριστούν εξωτερικές τοποθεσίες πιο συχνά από τις εικόνες που επιστρέφονται από μία γενική αναζήτηση κειμένου στο Flickr με τη λέξη "`rome`". Τέλος, παρά την τεράστια προσπάθεια που απαιτείται για την ανακατασκευή του μοντέλου, η έξοδος σε καμία από τις παραπάνω περιπτώσεις δεν χρησιμοποιείται για την βελτίωση της αναζήτησης ή αναγνώριση τοποθεσίας μιας νέας εικόνας.

4.2.4 Δομές Δεικτοδότησης για βάσης μεγάλης κλίμακας

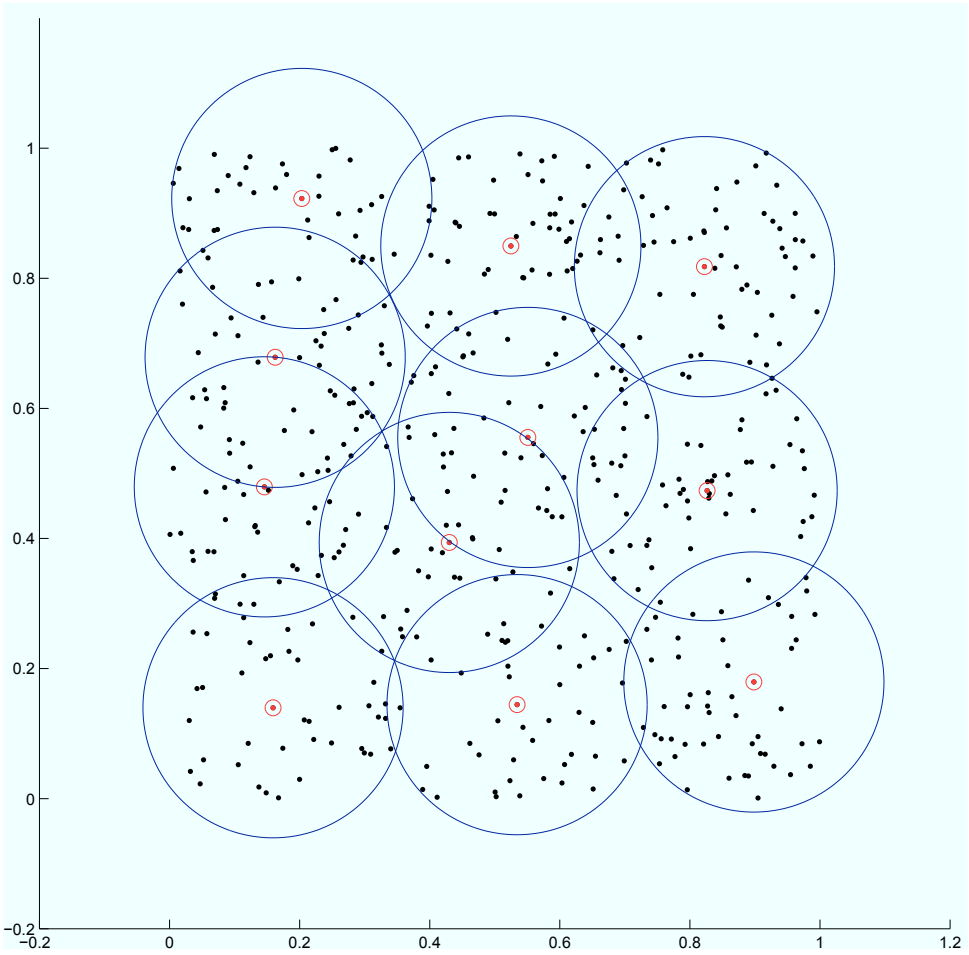
Από τις παραπάνω σχετικές δουλείες είναι προφανές ότι το ταίριασμα με τοπικά χαρακτηριστικά μπορεί να παρέχει ακριβή αναγνώριση τοποθεσίας, αλλά η επιτυχής εφαρμογή του σε μεγάλη κλίμακα εξαρτάται κατά πολύ από την αποτελεσματικότητα της δομής δεικτοδότησης που χρησιμοποιείται καθώς και από την μέθοδο αναζήτησης συνολικότερα. Με χρήση της αναπαράστασης *bag of words*, οι Sivic και Zisserman [96] έδειξαν πως τεχνικές που προέρχονται από την αναζήτηση κειμένου όπως τα λεξικά, η δεικτοδότηση με ανεστραμμένα αρχεία και τα βάρη τύπου TF-IDF μπορούν να εφαρμοστούν για την αναζήτηση εικόνων. Οι Nister and Stewenius [76] επέκτειναν το παραπάνω μοντέλο για ιεραρχικά λεξικά, δημιουργώντας ένα *δέντρο λεξικών* (*vocabulary tree*) το οποίο χρησιμοποιείται επίσης για

την ανάθεση των περιγραφέων σε οπτικές λέξεις. Στη συνέχει οι Philbin *et al*. [85] έδειξαν ότι το «επίπεδο» λεξικό μέσω k-means, όντας πιο ευέλικτο, στην πράξη αποδίδει καλύτερα από το δέντρο λεξικών. Έτσι κατασκεύασαν ενα μεγάλης κλίμακας λεξικό ενός εκατομμυρίου λέξεων χρησιμοποιώντας μια προσεγγιστική εκδοχή του αλγορίθμου k-means, η οποία περιλαμβάνει τα τυχαία kd-δέντρα των Silpa-Anan και Hartley [94] για την ανάθεση των διανυσμάτων στα κέντρα σε κάθε επανάληψη. Επίσης, εκμεταλλεύτηκαν το *σχήμα* των τοπικών περιγραφέων ώστε να επιταχύνουν και τη διαδικασία του γεωμετρικού ταιριάσματος.

Οι Chum *et al*. [19] πάνε ένα βήμα παραπέρα και εκμεταλλεύονται τις ομοιότητες μεταξύ των εικόνων της βάσης για να αυξήσουν την ανάκληση (recall) προτείνοντας τεχνικές *επέκτασης αναζήτησης* (*query expansion*). Παρόμοια και με τη δημοσίευση [3], η επέκταση αυτή είναι ένα είδος ομαδοποίησης την ώρα της αναζήτησης. Προϋποθέτει την ύπαρξη πολλαπλών και διαφορετικών όψεων της ίδιας σκηνής στη βάση, κάτι το οποίο ισχύει για συλλογές εικόνων από το Flickr με δεδομένα τοποθεσίας. Πιο πρόσφατες δημοσιεύσεις για δεικτοδότηση εικόνων περιλαμβάνουν τις δουλειές των Jegou *et al*. [41], Perdoch *et al*. [83] και Jegou *et al*. [44], οι οποίες εστιάζουν σε διάφορες όψεις της γεωμετρικής συνάφειας, των οπτικών λεξικών και της χρήσης μνήμης αντίστοιχα. Επιπλέον, οι Chum *et al*. [18] ασχολούνται με αναζήτηση μικρών αντικειμένων, ενώ στη δημοσίευση μας [7] καταφέρνουμε να εισάγουμε την *συνολική γεωμετρία* των εικόνων στη δομή δεικτοδότησης. Γενικά, αν και όλες οι πρόσφατες τεχνικές είναι ιδιαίτερα γρήγορες, υπάρχει ιδιαίτερη σχέση (trade-off) μεταξύ της ακρίβειας δεικτοδότησης και των απαιτήσεων σε μνήμη. Οι επιλογές μας σε αυτόν τον τομέα παρουσιάζονται και αναλύονται στην ενότητα 5.2.

4.3 Ομαδοποίηση όψεων

Όπως και σε άλλες πρόσφατες σχτικές δημοσιεύσεις, ακολουθούμε μια μέθοδο ομαδοποίησης σε δύο επίπεδα, πρώτα σύμφωνα με την τοποθεσία (latitude, longitude) και έπειτα σύμφωνα με την οπτική ομοιότητα των εικόνων (τον αριθμό των επαληθευμένων χαρακτηριστικών μετά από γεωμετρικό ταίριασμα). Παρακάτω θα αναφερόμαστε στα δύο αυτά επίπεδα με τους όρους *γεωγραφική ομαδοποίηση* (*geo-clustering*) και *οπτική ομαδοποίηση* (*visual clustering*) αντίστοιχα. Ο σκοπός της οπτικής ομαδοποίησης είναι η ανίχνευση εικόνων που αναπαριστούν διάφορες *όψεις* (*views*) της ίδιας *σκηνής* (*scene*). Το τελικό αποτέλεσμα θα είναι λοιπόν ένα σύνολο από *ομάδες όψεων* (*view clusters*) και στην όλη διαδικασία θα αναφερόμαστε με τον όρο *ομαδοποίηση όψεων* (*view clustering*). Η βασική ιδέα πίσω από την προσέγγιση δύο επιπέδων είναι ότι όψεις της ίδιας σκηνής δεν αναμένονται σε εικόνες που έχουν τραβηχθεί πολύ μακριά η μία από την άλλη και έτσι η γεωγραφική ομαδοποίηση βοηθά στον περιορισμό της υπολογιστικής πολυπλοκότητας



Σχήμα 4.1: *Παράδειγμα του αλγορίθμου ομαδοποίησης KVQ σε ένα σύνολό από* $n = 500$ *τυχαία δισδιάστατα δεδομένα, παρμένα από ομοιόμορφη κατανομή στο* $[0, 1]^2$*, με ακτίνα* $r = 0.2$*. Ο αλγόριθμος καταλήγει σε* 11 *από τα αρχικά δεδομένα για κέντρα, σημεία τα οποία απεικονίζονται με κόκκινους κύκλους.*

της επακόλουθης οπτικής ομαδοποίησης. Για την ομαδοποίηση χρησιμοποιούμε τον αλγόριθμο *kernel vector quantization* (KVQ) των Tipping και Schölkopf [101]. Σε αυτή την ενότητα, αρχικά θα συνοψίζουμε τις βασικές ιδιότητες του αλγορίθμου KVQ. Στη συνέχεια θα αναλύσουμε τη συγκεκριμένη διαδικασία που προτείνουμε για ομαδοποίηση δύο επιπέδων και θα παρουσιάσουμε παραδείγματα από γεωγραφικές και οπτικές ομάδες που προέρχονται από τη συλλογή μας με αστικές εικόνες. Τέλος, θα αναλύσουμε τις επιλογές μας σε σχέση με άλλες πιθανές λύσεις.

4.3.1 Kernel Vector Quantization

Έστω (X, d) ένας μετρικός χώρος και $D \subseteq X$ ένα πεπερασμένο σύνολο σημείων με αριθμό στοιχείων $|D| = n$, του οποίου τα στοιχεία συμβολίζονται ως $D = \{x_1, \ldots, x_n\}$. Σκοπός του αλγορίθμου είναι να επιλεγεί ένα υποσύνολο $Q(D) \subseteq D$ όσο το δυνατόν μικρότερο, το οποίο υπακούει στον περιορισμό ότι όλα τα σημεία

του D δεν είναι αρκετά μακριά από κάποιο σημείο του Q. Εάν

$$B_r(x) = \{y \in X : d(x, y) < r\} \tag{4.1}$$

είναι η ανοικτή μπάλα (open ball) στον χώρο X με ακτίνα r κεντραρισμένη στο σημείο x, και $\mathbf{1}_A : X \to \{0, 1\}$ δηλώνει τη συνάρτηση δείκτη (indicator function) του συνόλου $A \subseteq X$, μπορούμε να ορίσουμε τη *συνάρτηση πυρήνα* (*kernel function*) $k : X \times X \to \mathbb{R}$ ως[1]

$$k(x, y) = \mathbf{1}_{B_r(x)}(y) \tag{4.2}$$

η οποία δείχνει το κατά πόσο τα σημεία $x, y \in X$ βρίσκονται σε απόσταση μικρότερη από r, όπου το $r > 0$ είναι πρακτικά μια παράμετρος *κλίμακας* της εισόδου. Για ένα σημείο $x \in X$, μπορούμε να ορίσουμε τον εμπειρικό χάρτη πυρήνα ή *empirical kernel map*

$$\boldsymbol{\phi}(x) = (k(x_1, x), \ldots, k(x_n, x))^{\mathsf{T}}. \tag{4.3}$$

Η βασική παρατήρηση είναι ότι, αν υφίσταται ένα *διάνυσμα από βάρη* $\mathbf{w} \in \mathbb{R}^n$ με στοιχεία w_j τέτοια ώστε για κάθε $x \in D$,

$$\mathbf{w}^{\mathsf{T}}\boldsymbol{\phi}(x) > 0 \tag{4.4}$$

τότε όλα τα σημεία $x \in D$ βρίσκονται σε απόσταση μικρότερη από r από κάποιο σημείο $x_j \in D$ με θετικά αντίστοιχο βάρος $w_j > 0$. Για να επιτευχθεί μία λύση όσο το δυνατόν *αραιότερη* για τα βάρη $\mathbf{w}$ που ικανοποιούν τη σχέση (4.4), συνήθως χρησιμοποιούμε την ℓ_1 νόρμα στο χώρο $\mathbb{R}^n$ και προκύπτει το ακόλουθο πρόβλημα βελτιστοποίησης:

$$\min_{\mathbf{w} \in \mathbb{R}^n} \quad \|\mathbf{w}\|_1 \tag{4.5}$$

$$\text{subject to} \quad \mathbf{w}^{\mathsf{T}}\boldsymbol{\phi}(x) \geq 1 \quad \forall x \in D. \tag{4.6}$$

Για ένα σημείο $x \in D$, μπορούμε να ορίσουμε την *ομάδα* $C(x) = D \cap B_r(x) = \{y \in D : d(x, y) < r\}$ ως το σύνολο των σημείων $y \in D$ τα οποία βρίσκονται σε απόσταση μικρότερη από r ως προς το σημείο x. Αν ορίσουμε και το διάνυσμα $\boldsymbol{\gamma} \in \mathbb{R}^n$ με στοιχεία $\gamma_j = |C(x_j)|^{-1} = \|\boldsymbol{\phi}(x_j)\|_1^{-1}$, καταλήγουμε στο παρακάτω πρόβλημα γραμμικής βελτιστοποίησης:

$$\min_{\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^n} \quad \boldsymbol{\gamma}^{\mathsf{T}}(\boldsymbol{\alpha} + \boldsymbol{\beta}) \tag{4.7}$$

$$\text{subject to} \quad \mathbf{K}(\boldsymbol{\alpha} - \boldsymbol{\beta}) \geq 1 \tag{4.8}$$

$$\boldsymbol{\alpha}, \boldsymbol{\beta} \geq 0, \tag{4.9}$$

όπου το διάνυσμα βαρών $\mathbf{w}$ έχει αναλυθεί σύμφωνα με τη σχέση $\mathbf{w} = \boldsymbol{\alpha} - \boldsymbol{\beta}$ και $\mathbf{K}$ είναι η μήτρα πυρήνα ή *Gram* matrix, με στοιχεία $K_{ij} = k(x_i, x_j)$. Έχοντας την

[1] Χρησιμοποιούμε απλοποιημένη σημειογραφία για τις ανοικτές μπάλες. Στη δημοσίευση [101], επιτρέπεται η απόσταση $d(x, y) = r$.

βέλτιστη λύση $\mathbf{w}^\star = \boldsymbol{\alpha}^\star - \boldsymbol{\beta}^\star$ με στοιχεία $w_j^\star$, το *λεξικό* $Q(D)$ ενός συνόλου στοιχείων D ορίζεται ως[2]

$$Q(D) = \{x_j \in D : w_j^\star > 0\}. \tag{4.10}$$

Είναι εμφανές ότι τα κέντρα του λεξικού $Q(D) \subseteq D$ είναι σημεία του αρχικού συνόλου δεδομένων. Θα αναφερόμαστε σε αυτά και ως *κέντρα ομάδων*. Είναι επίσης εμφανές ότι η *μέγιστη παραμόρφωση* (*maximal distortion*) είναι φραγμένη εκ των άνω από την ακτίνα r καθώς ισχύει $\max_{y \in C(x)} d(x, y) < r$ για κάθε $x \in Q(D)$. Επίσης, η *συλλογή ομάδων* (*cluster collection*)

$$\mathcal{C}(D) = \{C(x) : x \in Q(D)\} \tag{4.11}$$

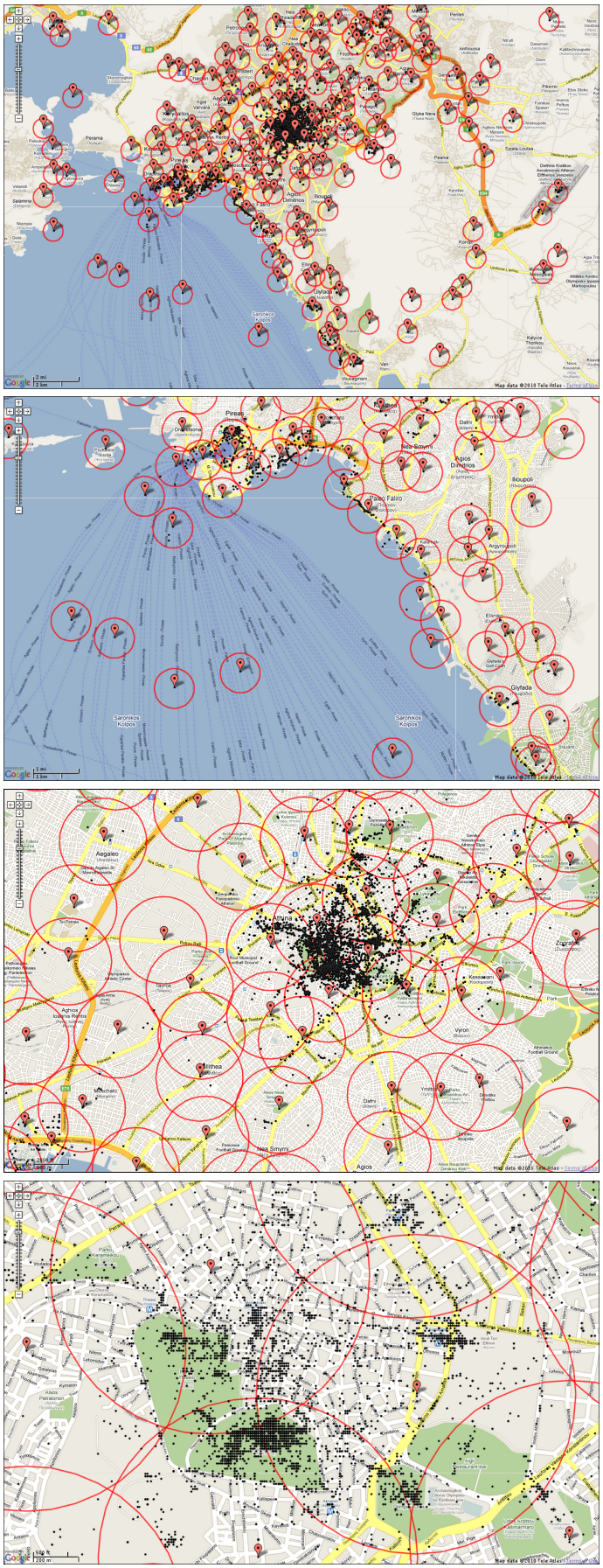
αποτελεί ένα *κάλυμμα* του συνόλου D καθώς $D = \bigcup_{x \in Q(D)} C(x)$. *Δεν* αποτελεί όμως μια διαμέριση, καθώς δεν ισχύει στη γενική περίπτωση ότι $C(x) \cap C(y) \neq \emptyset$ για τα $x, y \in D$. Αυτό σημαίνει πρακτικά ότι οι ομάδες έχουν *επικάλυψη*. Κάτι τέτοιο είναι ιδιαίτερα χρήσιμο στην περίπτωση της γεωγραφικής ομαδοποίησης, όπου θέλουμε να μην χωριστούν όψεις της ίδιας σκηνής. Για την οπτική ομαδοποίηση, η ιδιότητα αυτή είναι επίσης χρήσιμη, στις περιπτώσεις σταδιακής μετάβασης ανάμεσα σε όψεις οι οποίες σε αντίθετη περίπτωση θα διαχωρίζονταν. Σε αντίθεση με άλλους αλγόριθμους ομαδοποίησης, όπως για παράδειγμα τον k-means, ο αριθμός των ομάδων καθορίζεται αυτόματα και ελέγχεται από την τιμή της ακτίνας ή της μέγιστης παραμόρφωσης r.

Η παραπάνω λύση δεν είναι η βέλτιστη ως προς το μέγεθος του λεξικού $|Q(D)|$. Μια τέτοια απαίτηση θα οδηγούσε σε πρόβλημα συνδυαστικής βελτιστοποίησης (combinatorial optimization), άρα στην παραπάνω λύση υπάρχει πλεονασμός κέντρων ως κάποιο βαθμό. Γι αυτό το λόγο, μετά την εύρεση της βέλτιστης λύσης ακολουθεί μια διαδικασία «*κλαδέματος*» (*pruning*), κατά το οποίο αφαιρούνται με τυχαία σειρά τα πλεονάζοντα κέντρα x του συνόλου $Q(D)$, έτσι ώστε τα κέντρα που παραμένουν να συνεχίζουν να αποτελούν ένα κάλυμμα του συνόλου D[3]. Αφαιρείται δηλαδή κάθε κέντρο $x \in Q(D)$ έτσι ώστε $C(x) \subseteq \bigcup_{y \in Q(D) \setminus \{x\}} C(y)$. Θεωρούμε ότι το στάδιο του «κλαδέματος» πραγματοποιείται πάντα, συνεπώς θα συμβολίζουμε ως $Q(D)$, $\mathcal{C}(D)$ το τελικό λεξικό και την τελική συλλογή ομάδων, αντίστοιχα, μετά το κλάδεμα.

Για να παρουσιαστούν οι παραπάνω ιδιότητες, στο Σχήμα 4.1 απεικονίζεται το λεξικό που παράγεται από ένα συνθετικό παράδειγμα για τυχαία δισδιάστατα σημεία, όπως και στη δημοσίευση [101]. Ο αριθμός των τελικών ομάδων καθορίζεται αυτόματα μέσω της μέγιστης παραμόρφωσης r, η οποία στην προκειμένη περίπτωση έχει επιλεχθεί έτσι ώστε να παραχθεί ένα σχετικά μικρό τελικό λεξικό.

[2] Στην πράξη θέλουμε τα βάρη w_j να είναι μεγαλύτερα από ένα μικρό θετικό κατώφλι.

[3] Τα κέντα που κλαδεύονται είναι συνήθως λιγότερο από το 5% του λεξικού.



Σχήμα 4.2: *Χάρτης της Αθήνας με τις γεωγραφικές ομάδες για διάφορα επίπεδα μεγέθυνσης.*

4.3.2 Γεωγραφική ομαδοποίηση

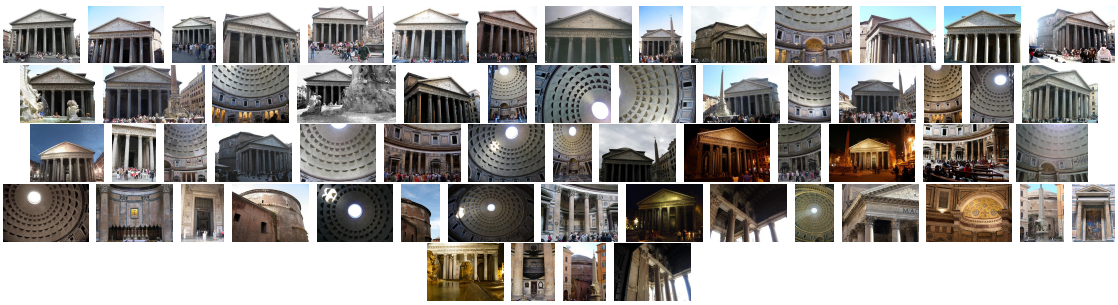
Με δεδομένο ένα σύνολο φωτογραφιών[4], θα αναπαριστούμε κάθε φωτογραφία $p \in P$ από το ζεύγος (ℓ_p, F_p), όπου ℓ_p είναι η γεωγραφικές συντεταγμένες όπου τραβήχτηκε η φωτογραφία (latitude και longitude) και F_p είναι το σύνολο των οπτικών χαρακτηριστικών της εικόνας και περιλαμβάνει τις θέσεις και το σχήμα των χαρακτηριστικών, μαζί με τις αντίστοιχες οπτικές λέξεις τους, που εξάγονται σύμφωνα με τη διαδικασία που περιγράφεται στην ενότητα 5.2.1. Κατά την γεωγραφική ομαδοποίηση εφαρμόζουμε τον αλγόριθμο KVQ στο σύνολο P στον μετρικό χώρο $(\mathcal{P}, d_g)$ με παράμετρο κλίμακας r_g, όπου $\mathcal{P}$ είναι το σύνολο όλων των φωτογραφιών και ως μετρική d_g χρησιμοποιούμε την $d_g : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ η οποία ορίζεται ως η *γεωδεσική* απόσταση ή *great circle distance*[5] μεταξύ δύο οποιονδήποτε σημείων στην επιφάνεια της γης. Ορίζουμε ως $d_g(p, q)$ την γεωδεσική απόσταση μεταξύ των γεωγραφικών θέσεων των φωτογραφιών $p, q \in \mathcal{P}$. Για μια φωτογραφία $p \in P$, μπορούμε να ορίσουμε μια *γεωγραφική ομάδα* (*geo-cluster*) ως $C_g(p) = \{q \in P : d_g(p, q) < r_g\}$, το σύνολο δηλαδή των φωτογραφιών $q \in P$ που βρίσκονται σε γεωγραφική απόσταση μικρότερη από r_g από το p. Έχοντας το τεικό λεξικό $Q_g(P)$, μπορούμε να ορίσουμε με παρόμοιο τρόπο τη *συλλογή γεωγραφικών ομάδων* (*geo-cluster collection*) ως $\mathcal{C}_g(P) = \{C_g(p) : p \in Q_g(P)\}$.

Στην πράξη χρησιμοποιούμε έναν χωρικό *κβαντισμό* των γεωγραφικών συντεταγμένων σε ομοιόμορφο πλέγμα, και κρατάμε μόνο ένα δείγμα από κάθε κβαντισμένη τιμή για την εφαρμογή του αλγορίθμου KVQ. Το μέγεθος του πλέγματος είναι πολύ μικρό σε σχέση με την ακτίνα r_g και έτσι επηρεάζονται ελάχιστα οι γεωγραφικές ομάδες. Με αυτό τον τρόπο όμως, η υπολογιστική πολυπλοκότητα μειώνεται κατά πολύ και καταλήγει να εξαρτάται από το μέγεθος του πλέγματος και όχι από το μέγεθος της συλλογής $|P|$. Γενικότερα το κόστος της γεωγραφικής ομαδοποίησης είναι αμελητέο σε σχέση με τα επόμενα βήματα. Χρειάζονται για παράδειγμα λίγα δευτερόλεπτα για να εκτελεστεί η γεωγραφική ομαδοποίηση σε ένα σύνολο P, με μέγεθος της τάξεως $|P| = 10^5$ φωτογραφιών. Για περαιτέρω επιτάχυνση του αλγορίθμου, μπορούν να χρησιμοποιηθούν δομές δεικτοδότησης στις δισδιάστατες χωρικές συντεταγμένες, για παράδειγμα με τη χρήση kd-δέντρων, έτσι ώστε να ανακτώνται οι χωρικά κοντινότεροι γείτονες σε λογαριθμικό χρόνο.

Στο Σχήμα 4.2, απεικονίζεται ο χάρτης της πόλης στη Αθήνας μαζί με όλες τις γεωγραφικές ομάδες που εξήχθησαν, σε τέσσερα επίπεδα μεγέθυνσης, για ακτίνα $r_g = 700m$. Οι μαύρες κουκκίδες, οι κόκκινοι markers και οι κόκκινοι κύκλοι αντιστοιχούν σε φωτογραφίες, κέντρα και όρια ομάδων αντίστοιχα. Παρατηρείται εύκολα η μεγάλη πυκνότητα φωτογραφιών στο ιστορικό κέντρο και συγκεκριμένα στην περιοχή της Ακρόπολης. Η επικάλυψη βοηθάει να διατηρηθούν τέτοιες πυ-

[4] Θα χρησιμοποιούμε τους όρους *φωτογραφία*, *εικόνα* και *όψη* εναλλακτικά στην παρακάτω ανάλυση.

[5] `http://en.wikipedia.org/wiki/Great-circle\_distance`



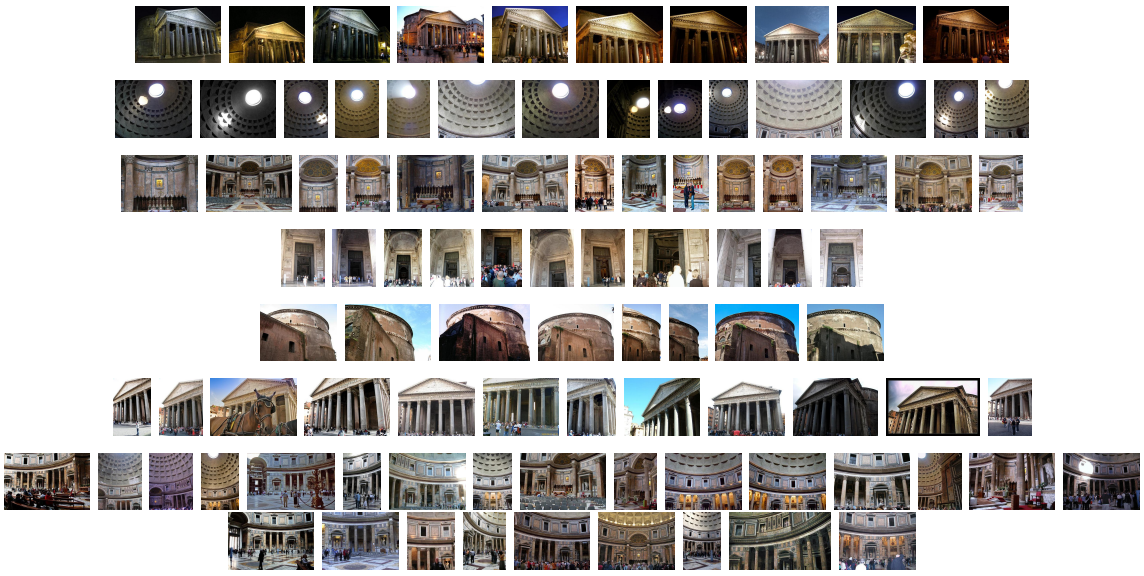
Σχήμα 4.3: *Φωτογραφίες που ανήκουν στα κέντρα των πλέον πολυπληθών ομάδων από το Πάνθεον της Ρώμης.*

κνές περιοχές στην ίδια ομάδα για το μετέπειτα οπτικό ταίριασμα. Φωτογραφίες που έχουν τραβηχτεί ακόμα και $1km$ μακριά μπορούν να συμπεριληφθούν στην ίδια ομάδα. Ο συνολικός αριθμός και η θέση των κέντρων εξάγεται αυτόματα από τα δεδομένα.

4.3.3 Οπτική ομαδοποίηση

Όπως και στη δημοσίευση [95], θα λέμε ότι δύο φωτογραφίες $p, q \in P$ είναι *συνδεδεμένες* αν απεικονίζουν τουλάχιστον ένα κοινό αντικείμενο, πιθανώς και υπο διαφορετικές οπτικές γωνίες. Μπορούμε να ορίσουμε, λοιπόν, μια *σκηνή* ως ένα υποσύνολο $S \subseteq P$ από συνδεδεμένες φωτογραφίες. Ισχύει λοιπόν ότι για κάθε ζεύγος φωτογραφιών $p, q \in S$, μπορούμε να ταιριάξουμε οπτικά τα κοινά τους αντικείμενα μέσω κάποιο μοντέλο τρισδιάστατης γεωμετρίας, ανεξάρτητα από την οπτική γωνία. Για το ταίριασμα αυτό χρησιμοποιούμε τα τοπικά οπτικά χαρακτηριστικά και περιγραφείς, όπως περιγράφεται στην ενότητα 5.2.1. Η έξοδος του οπτικού ταιριάσματος είναι συνεπώς ο αριθμός των *inliers* $I(p, q)$, δηλαδή ο αριθμός των αντιστοιχιών που επαληθεύτηκαν γεωμετρικά, ανάμεσα στα σύνολα χαρακτηριστικών F_p, F_q των φωτογραφιών p, q αντίστοιχα.

Μπορούμε τώρα να εφαρμόσουμε τον αλγόριθμο ομαδοποίησης KVQ ανεξάρτητα σε κάθε γεωγραφική ομάδα $G \in \mathcal{C}_g(P)$ στον χώρο $(\mathcal{P}, d_v)$ με παράμετρο κλίμακας r_v. Καθώς το μέγεθος $I(F_p, F_q)$ είναι ένα μέτρο ομοιότητας, κάθε φθίνουσα συνάρτηση θα μπορεί να χρησιμοποιηθεί ως μετρική, για παράδειγμα η $d_v(p, q) = \exp\{-I(F_p, F_q)\}$. Η ακριβής συνάρτηση της μετρικής $d_v(p, q)$ δεν είναι σημαντική. Στην πράξη, η παράμετρος κλίμακας καθορίζει ένα κατώφλι $\tau = -\log r_v$ στον αριθμό των inliers. Αν $Q_v(G)$ είναι το λεξικό που θα προκύψει, μπορούμε να ορίσουμε την *οπτική ομάδα* (*visual cluster*) ως $C_v(p) = \{q \in G : d_v(p, q) < r_v\}$ για κάθε $p \in G$ καθώς και την *συλλογή οπτικών ομάδων* (*visual cluster collection*) ως $\mathcal{C}_v(G) = \{C_v(p) : p \in Q_v(G)\}$, παρομοίως με την γεωγραφική ομαδοποίησης. Η διαδικασία της οπτικής ομαδοποίησης πραγματοποιείται σε όλες τις γεωγραφικές ομάδες και τελικά το ολοκληρωμένο λεξικό $Q(P)$ για ολόκληρη τη συλλογή εικόνων είναι η ένωση $Q(P) = \bigcup_{G \in \mathcal{C}_g(P)} Q_v(G)$. Τέλος, το σύνολο όλων των *ομάδων*



Σχήμα 4.4: *Φωτογραφίες σε μερικές από τις οπτικές ομάδες για το Πάνθεον. Η πρώτη φωτογραφία (από τα αριστερά) σε κάθε γραμμής/ομάδες αντιστοιχεί στο κέντρο της ομάδας.*

όψεων (*view clusters*) $\mathcal{C}(P)$ ορίζεται παρομοίως ως $\mathcal{C}(P) = \{C_v(p) : p \in Q(P)\}$.

Στην όλη διαδικασία της ομαδοποίησης, το πλέον υπολογιστικά «βαρύ» κομμάτι είναι ο υπολογισμός των αποστάσεων/ομοιοτήτων σε όλα τα ζεύγη των εικόνων, διαδικασία που είναι τετραγωνική στο μέγεθος του συνόλου. Για την γεωγραφική ομαδοποίηση δεν είναι ιδιαίτερα μεγάλο πρόβλημα, αλλά είναι ιδιαίτερα σημαντικό για την οπτική ομαδοποίηση. Για να το ξεπεράσουμε, προτείνουμε την *δεικτοδότηση των εικόνων ανά γεωγραφική ομάδα* (geo-cluster specific indexing). Συγκεκριμένα, χρησιμοποιούμε μια ανεστραμμένη δομή για τις οπτικές λέξεις και για τις γεωγραφικές ομάδες. Με δεδομένη μια εικόνα αναζήτησης $q \in G$, επιστρέφουμε όλες τις εικόνες $p \in G$ που ταιριάζουν γεωμετρικά, με $I(F_p, F_q) > \tau$ σε σταθερό χρόνο ο οποίος είναι πρακτικά μικρότερος από ένα δευτερόλεπτο. Η όλη διαδικασία γίνεται έτσι γραμμική ως προς το $|G|$.

Για να φανεί το αποτέλεσμα της οπτικής ομαδοποίησης σε ένα σύνολο εικόνων, παρουσιάζουμε ένα παράδειγμα από εικόνες του Πάνθεον της Ρώμης, ακολουθώντας τη μορφή των παραδειγμάτων στις δημοσιεύσεις [95] και [87]. Συγκεκριμένα, αρχικά επιλέγουμε όλες τις φωτογραφίες που τραβήχτηκαν στο κέντρο της Ρώμης από το Flickr. Έπειτα απομονώνουμε ενα αρχικό υποσύνολο των φωτογραφιών αυτών οι οποίες έχουν στα μεταδεδομένα τους τον όρο `pantheon` και επεκτείνουμε το σύνολο με όσες άλλες εικόνες της βάσης έχουν οπτική ομοιότητα με το αρχικό υποσύνολο. Καταλήγουμε, έτσι, σε ένα σύνολο $1,146$ εικόνων τις οποίες θεωρούμε για το παράδειγμα ότι αποτελούν μια γεωγραφική ομάδα. Εκτελούμε οπτική ομαδοποίηση και παίρνουμε 258 οπτικές ομάδες. Το μέσο μέγεθος οπτικής ομάδας είναι 30 εικόνες και η κάθε εικόνα εμπεριέχεται σε 4 οπτικές ομάδες κατά μέσω όρο, λόγω της επικάλυψης.

Στο Σχήμα 4.3 φαίνονται οι φωτογραφίες που αντιστοιχούν στα κέντρα των πλέον πολυπληθέστερων οπτικών ομάδων. Σε αντίθεση με τα αντίστοιχα αποτελέσματα της δημοσίευσης [95], εδώ δεν έχουμε ως σκοπό την περίληψη ή την επιλογή κανονικών όψεων και δεν υπάρχει έτσι απαίτηση για *ορθογωνιότητα* μεταξύ των κέντρων των ομάδων. Η μέγιστη απόσταση όμως των εικόνων ανά ομάδα είναι τόση ώστε επιτρέπεται να τις ευθυγραμμίσουμε όλες σε έναν χάρτη σκηνής. Το Σχήμα 4.4 απεικονίζει φωτογραφίες σε μια επιλογή οπτικών ομάδων. Λόγω της αυστηρής διαδικασίας ταιριάσματος, οι εικόνες σε κάθε ομάδα είναι αρκετά παρόμοιες. Ακόμα και η τελευταία ομάδα στο κάτω μέρος του σχήματος, αν και αρχικά μπορεί να δείχνει απλωμένη, με προσεκτικότερη παρατήρηση φαίνεται ότι όλες οι εικόνες της *συνδέονται*—μοιράζονται δηλαδή τουλάχιστον ένα κοινό αντικείμενο—με την πρώτη εικόνα, δηλαδή το κέντρο της ομάδας ή *εικόνα αναφοράς*.

4.3.4 Συζήτηση

Στη σχετική βιβλιογραφία ακολουθούνται διάφορες στρατηγικές ομαδοποίησης. Για παράδειγμα οι Crandall *et al*. [22] και οι Li *et al*. [63] χρησιμοποιούν τον αλγόριθμο mean-shift για την γεωγραφική ομαδοποίηση και εξάγουν περιοχές με μεγάλη πυκνότητα που αντιστοιχούν σε δημοφιλής τοποθεσίες. Σε άλλες τεχνικές ακολουθεί επίσης ένα δεύτερο επίπεδο οπτικής ομαδοποίησης, με χρήση διαφόρων μεθόδων, όπως k-means ([54]) και συσσωρευτική (agglomerative) ομαδοποίηση ([87],[28],[117]). Όσων αφορά τη γεωγραφική ομαδοποίηση, στις δημοσιεύσεις [54] και [117] οι συγγραφείς χρησιμοποιούν τον ίδιο αλγόριθμο όπως και στην οπτική ομαδοποίηση, ενώ στις [87] και [28] οι τοποθεσίες απλώς κβαντίζονται σε ορθογώνια πλακίδια. Σε μερικές άλλες δημοσιεύσεις [62, 95, 20] εκτελείται μόνο οπτική ομαδοποίηση, μέθοδος η οποία προφανώς δε μπορεί να επεκταθεί σε μεγάλη κλίμακα.

Το μεγάλο μειονέκτημα των συγχωνευτικών αλγορίθμων καθώς και του k-means είναι ότι δεν προσφέρουν κάποιου είδους έλεγχο για την μέγιστη απόσταση στο εσωτερικό των ομάδων. Κάτι τέτοιο είναι ιδιαίτερα σημαντικό καθώς μπορεί να οδηγήσει σε γεωγραφικές ομάδες με τις φωτογραφίες να έχουν τραβηχτεί από πολύ μακριά, ή οπτικές ομάδες με εικόνες που ταιριάζουν με μικρό αριθμό από inliers. Καθώς ο αλγόριθμος k-means απαιτεί διανυσματικό χώρο για να τρέξει, δεν μπορεί να χρησιμοποιεί τον αριθμό των inliers σαν μέτρο ομοιότητας. Αντίθετα, η αλγόριθμος KVQ μπορεί να ελέγξει τη μέγιστη παραμόρφωση και μπορεί να εφαρμοστεί σε μετρικούς χώρους.

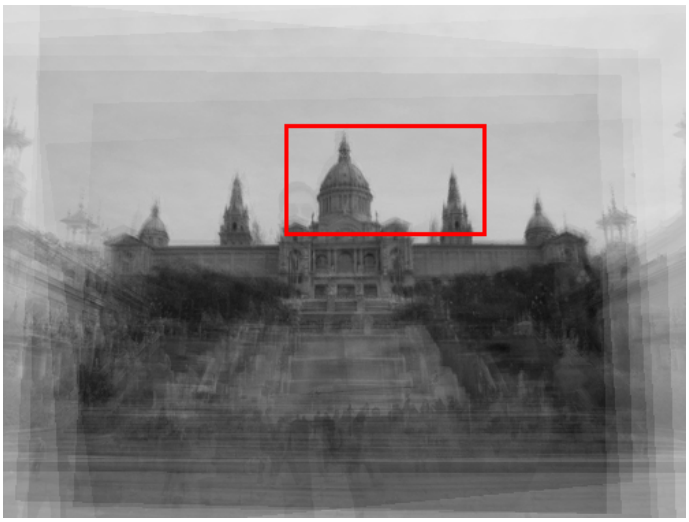
Ο αλγόριθμος Mean-shift [16], ο οποίος χρησιμοποιείται στις δημοσιεύσεις [22] και [63], έχει μια παρόμοια ιδιότητα για τον έλεγχο της παραμόρφωσης: το άνω όριο είναι η παράμετρος εύρους (bandwidth) της συνάρτησης πυρήνα ή η *κλίμακα παρατήρησης*. Παρόλα αυτά, το mean-shift απαιτεί *αρχικοποίηση* και για παρά-

δειγμα στη δημοσίευση [22] χρησιμοποιείται χωρικός διαμερισμός (*bucketing*). Ο αλγόριθμος KVQ δεν απαιτεί αρχικοποίηση, κάτι ιδιαίτερα θεμιτό καθώς ο διαμερισμός του προϋποθέτει την ύπαρξη διανυσματικού χώρου και συνεπώς δεν μπορεί να εφαρμοστεί για την αρχικοποίηση της οπτικής ομαδοποίησης Τα σταθερά πλακίδια της δημοσίευσης [87] ελέγχουν επίσης την κλίμακα/παραμόρφωση στη γεωγραφική ομαδοποίηση, αλλά ο αλγόριθμος KVQ έχει επίσης το πλεονέκτημα ότι προσαρμόζεται στα δεδομένα.

Τέλος, το πρόβλημα της τετραγωνικής πολυπλοκότητας για τις αποστάσεις μεταξύ εικόνων εμφανίζεται σε πολλές σχετικές δουλειές. Εμφανίζεται για παράδειγμα στη δημοσίευση των Quack *et al*. [87] οι οποίοι χρησιμοποιούν σχετικά μικρά γεωγραφικά πλακίδια με πλευρά $200m$ καθώς απαιτούν εξαντλητική εκτίμηση ομογραφίας μεταξύ όλων των ζευγαριών εικόνων στο πλακίδιο. Φυσικά, η προσέγγιση αυτή δεν μπορεί να περιγράψει σκηνές με έκταση μεγαλύτερη από $200m$, κάτι ιδιαίτερα σύνηθες. Το ίδιο τετραγωνικό κόστος εμφανίζεται και στις δημοσιεύσεις [28, 117, 95] και ειδικά στη δημοσίευση [54] είναι ο λόγος που δεν χρησιμοποιούνται τοπικά οπτικά χαρακτηριστικά. Εμείς χρησιμοποιούμε μεγαλύτερες γεωγραφικές ομάδες με μέγιστη ακτίνα $r_g = 700m$, και παρόλα αυτά καταφέρνουμε να έχουμε μια ιδιαίτερα γρήγορη υλοποίηση, η οποία μπορεί να μην είναι τόσο γρήγορη όπως των Chum και Matas [20], αλλά προσφέρει επιπλέον το πλεονέκτημα της γεωγραφικής ομαδοποίησης. Έτσι το κόστος μειώνεται δραματικά και απαιτείται απλά μία αναζήτηση ανά εικόνα σε κάθε γεωγραφική ομάδα. Αντιθέτως, η δημοσίευση [20] βασίζεται αρχικά σε κατακερματισμό (hashing), μέθοδος η οποία πάσχει ως προς την ανάκληση και καταλήγει να συλλέγει μόνο δημοφιλείς τοποθεσίες. Απομονωμένες τοποθεσίες έχουν πολύ μικρή πιθανότητα να ανακαλυφθούν.

4.4 Χάρτες Σκηνής

Μέχρι στιγμής, αυτό που ξέρουμε είναι ότι η εικόνα που αντιστοιχεί στο κέντρο κάθε ομάδας μοιράζεται τουλάχιστον ένα αντικείμενο με όλες τις εικόνες της ομάδας. Τη χρησιμοποιούμε, λοιπόν, ως *εικόνα αναφοράς* της ομάδας και ευθυγραμμίζουμε πάνω της όλες τις άλλες εικόνες της ομάδας, υπολογίζοντας τον μεταξύ τους ομογραφικό μετασχηματισμό, με τη διαδικασία του περιγράφεται λεπτομερώς την ενότητα 5.2. Συλλέγουμε όλα τα ευθυγραμμισμένα οπτικά χαρακτηριστικά και κατασκευάζουμε μια πιο συμπαγή αναπαράσταση την οποία ονομάζουμε *χάρτη σκηνής* (*scene map*), καθώς είναι πρακτικά ένας δισδιάστατος χωρικός χάρτης των χαρακτηριστικών που συνδέονται με τις διάφορες όψης της ίδιας σκηνής. Δίνεται μέσω αυτού η δυνατότητα να ταιριάξουμε γεωμετρικά μια εικόνα αναζήτησης με έναν ολόκληρο χάρτη σκηνής. Συνεπώς μπορούμε να χρησιμοποιήσουμε τους χάρτες σκηνών κατευθείαν για αναζήτηση *αντί για εικόνες*. Με αυτό τον τρόπο



Σχήμα 4.5: *Κατασκευή χάρτη σκηνής από 10 εικόνες του Palau Nacional, Montjuic, Barcelona.*

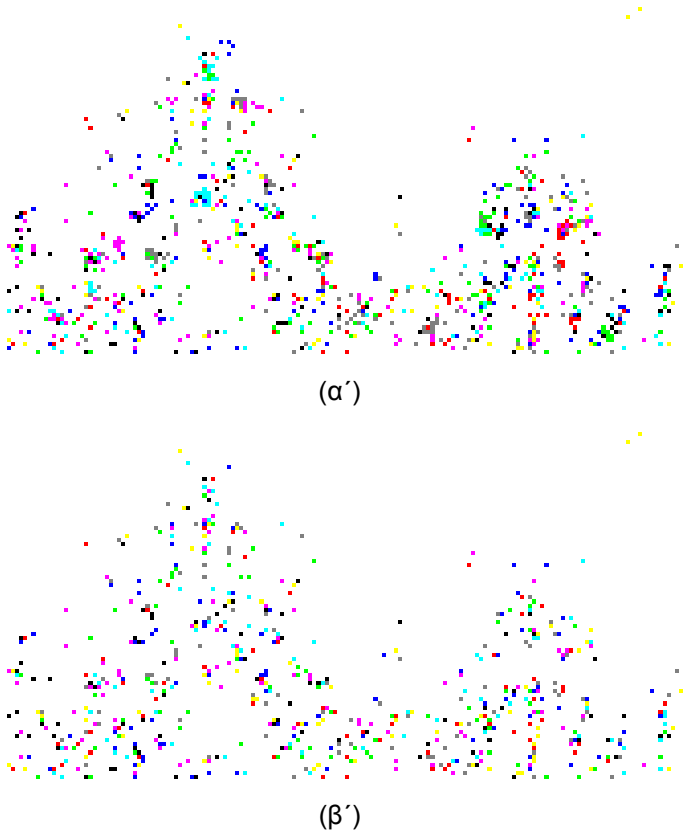
εξοικονομούνται μνήμη και υπολογισμοί κατά την αναζήτηση, το ταίριασμα γίνεται πιο εύρωστο καθώς οι inliers αυξάνονται και τέλος αυξάνεται η ανάκτηση, μιας και για κάθε ταιριασμένο χάρτη σκηνής επιστρέφουμε όλες τις όψεις που αυτός περιέχει. Θα παρουσιάσουμε τώρα την κατασκευή των χαρτών σκηνής και παρακάτω θα συζητηθεί το προτεινόμενο μοντέλο σε σχέση με την υπάρχουσα σχετική βιβλιογραφία.

4.4.1 Κατασκευή χαρτών σκηνής

Για κάθε εικόνα αναφοράς $p \in Q(P)$ και την αντίστοιχη ομάδα όψεων $C_v(p)$ κατασκευάζουμε μια συλλογή χαρακτηριστικών $F(p)$, την ένωση δηλαδή των χαρακτηριστικών από όλες τις εικόνες $q \in C_v(p)$ αφότου τις ευθυγραμμίσουμε με την εικόνα αναφοράς. Πιο συγκεκριμένα, έστω ότι H_{qp} είναι η εκτιμώμενη ομογραφία από την εικόνα q στην εικόνα p και ότι υποθέτουμε ότι κάθε οπτικό χαρακτηριστικό αναπαρίσταται από το ζεύγος (x, w) με το x να είναι το διάνυσμα θέσης και w η οπτική λέξη. Η συλλογή κατασκευάζεται ως

$$F(p) = \bigcup_{q \in C_v(p)} \{(H_{qp}x, w) : (x, w) \in F_q\}. \tag{4.12}$$

Εδώ το διάνυσμα θέσης x υποτίθεται ως τρισδιάστατο, περιέχοντας τις ομογενείς συντεταγμένες της δισδιάστατης θέσης. Ο *χάρτης σκηνής* $S(p)$ είναι μια αραιή αναπαράσταση της συλλογής $F(p)$ έτσι ώστε μια εικόνα αναζήτησης θα ταιριάζει (ιδανικά) με τον χάρτη, όποτε ταιριάζει με οποιαδήποτε από τις εικόνες που περιέχει. Κάτι τέτοιο ευνοείται πάλι από κβαντισμό διανυσμάτων με χρήση KVQ και μάλιστα μπορούμε να διαιρέσουμε τη παραπάνω συλλογή σε έναν αριθμό υποπροβλημάτων. Συγκεκριμένα, μπορούμε να διαμερίσουμε τη συλλογή $F(p)$ στα



(α΄)

(β΄)

Σχήμα 4.6: *Λεπτομέρεια του σύννεφου σημείων για τον χάρτη σκηνής του Montjuic, που αντιστοιχεί στην επισημειωμένη περιοχή του Σχήματος 4.5, (α)* πριν *και (β)* μετά *τον κβαντισμό διανυσμάτων. Τα διάφορα χρώματα εκφράζουν διαφορετικέ΄ς οπτικές λέξεις, modulo 9.*

ανεξάρτητα σύνολα $F_w(p) = \{(x, u) \in F(p) : u = w\}$, ένα για κάθε οπτική λέξη w, και να εφαρμόσουμε τον αλγόριθμο KVQ ανεξάρτητα σε κάθε $F_w(p)$ στο χώρο $(\mathbb{R}^2, d_2)$ με παράμετρο κλίμακας r_x. Στη μορφή αυτή, τα διανύσματα θέσης είναι οι δισδιάστατες συντεταγμένες των σημείων και ως μετρική χρησιμοποιείται η Ευκλείδεια απόσταση d_2. Τέλος, ενώνουμε όλα τα εξαχθέντα λεξικά $Q_x(F_w(p))$ σε ένα χάρτη σκηνής, $S(p) = \bigcup_{w \in \mathcal{W}} Q_x(F_w(p))$. Θέτουμε την παράμετρο κλίμακας $r_x = \theta$, όπου θ ειναι το κατώφλι σφάλματος που χρησιμοποιείται στο γεωμετρικό ταίριασμα. Συνεπώς, ένα χαρακτηριστικό f θα ανήκει στη χωρική ομάδα $C_x(f')$ ενός άλλου χαρακτηριστικού f' όποτε τα f, f' είναι inliers κατά το γεωμετρικό ταίριασμα.

Για να παρουσιαστεί ένα παράδειγμα κατασκευής χάρτη σκηνής, χρησιμοποιούμε μία οπτική ομάδα που περιέχει 30 φωτογραφίες του Palau Nacional, Montjuic, Barcelona, 10 από της οποίες φαίνονται ευθυγραμμισμένες στο Σχημα 4.5. Από τα συνολικά $11,623$ χαρακτηριστικά των 30 εικόνων, μόνο τα $9,924$ από αυτά παραμένουν στο χάρτη σκηνής μετά τον κβαντισμό, δίνοντας ένα ποσοστό συμπίεσης γύρω στο 15%. Όσων αφορά τον αριθμό καταχωρίσεων στα ανεστραμμένα

αρχεία, τον αριθμό δηλαδή των διαφορετικών (unique) οπτικών λέξεων, τα αντίστοιχα νούμερα ειναι $11,165$, $8,616$, και 23% αντίστοιχα. Μια λεπτομέρεια από το σύννεφο χαρακτηριστικών του συγκεκριμένου χάρτη σκηνής παρουσιάζεται στο Σχήμα 4.6, όπου φαίνεται καθαρά ότι τα χαρακτηριστικά είναι λιγότερα μετά την εφαρμογή του αλγορίθμου KVQ ανα οπτική λέξη.

4.4.2 Συζήτηση

Η παραπάνω αναπαράσταση έχει ομοιότητες με διάφορα άλλα μοντέλα της βιβλιογραφίας σε διαφορετικές εφαρμογές. Για παράδειγμα ο Lowe [67] εκτελεί *ομαδοποίηση όψεων* με τα τοπικά χαρακτηριστικά, συνδέοντας παρόμοια χαρακτηριστικά που ταίριαξαν σε κοντινές όψεις του ίδιου αντικειμένου, και εφαρμόζει την αναπαράσταση αυτή σε τρισδιάστατη ανακατασκευή αντικειμένων. Οι Simon *et al*. [95] οργανώνουν τα ταιριασμένα χαρακτηριστικά από πολλές εικόνες σε *ακολουθίες* (tracks), όπου τελικά η κάθε ακολουθία αντιστοιχεί σε ένα τρισδιάστατο σημείο της σκηνής. Χρησιμοποιούν αυτή την αναπαράσταση για να παράγουν μια περίληψη της όλης σκηνής μέσω εξαγωγής *κανονικών όψεων*. Οι Gammeter *et al*. [28] εκτελούν μια παρόμοια ευθυγράμμιση σε οπτικά κέντρα, με σκοπό να απομονώσουν τις περιοχές των εικόνων που απεικονίζουν ορόσημα. Στην αναζήτηση εικόνων, οι Chum *et al*. [19] συλλέγουν τις ταιριασμένες εικόνες μετά από μια αναζήτηση και παράγουν ένα *λανθάνων μοντέλο* της σκηνής παίρνοντας τον μεσο όρο των διανυσμάτων συχνότητας οπτικών λέξεων. Το μοντέλο αυτό χρησιμοποιείται από την πλευρά της εικόνας αναζήτησης για *επέκταση ερωτήματος*. Οι Leibe *et al*. [58] κατασκευάζουν ένα σύνολο από *κατανομές χωρικής εμφάνισης* για χρήση στην αναγνώριση αντικειμένων.

Συγκρίνοντας το μοντέλο μας με εκείνο της δημοσίευσης [19], παρατηρείται εύκολα ότι το δεύτερο δεν κωδικοποιεί τις θέσεις των χαρακτηριστικών και ότι κατασκευάζεται δυναμικά κατά την αναζήτηση από την πλευρά της εικόνας αναζήτησης, ενώ οι χάρτες σκηνής εξάγονται από την όλη συλλογή εικόνων και είναι στατικοί. Επίσης, σε αντίθεση με την αντικειμενοστραφή προσέγγιση της δημοσίευσης [28] εμείς θέλουμε να κρατηθεί πληροφορία από ολόκληρες τις εικόνες. Τέλος, όπως και στις δημοσιεύσεις [67, 95] όπου τα ταιριασμένα χαρακτηριστικά συνδυάζονται σε συνδεδεμένες συνιστώσες, έτσι και εμείς χρειαζόμαστε μια πιο πυκνή αναπαράσταση, πιο συμπιεσμένη από το να αποθηκεύονται ανεξάρτητα όλα τα σημεία των όψεων. Παράλληλα όμως θέλουμε να μπορούμε να ελέγχουμε το μέγεθος και τη έκταση των συνιστωσών αυτών, έτσι ώστε να συμπεριφέρονται σαν απλά χαρακτηριστικά σε μια εικόνα. Ένας τρόπος για να επιτευχθεί αυτό είναι να διατηρηθεί ένα ελάχιστο υποσύνολο $S(p) \subseteq F(p)$ τέτοιο ώστε κανένα χαρακτηριστικό της συλλογής $F(p)$ δεν είναι μακριά από τον κοντινότερο γείτονα του στον χάρτη σκηνής $S(p)$, κάτι που ενισχύει την επιλογή του KVQ και σε αυτή την

περίπτωση.

4.5 Πειράματα

4.5.1 Η συλλογή εικόνων European Cities 1M

Εκτελούμε τα πειράματα των προτεινόμενων τεχνικών σε μία απαιτητική συλλογή ενός εκατομμυρίου αστικών εικόνων, την οποία ονομάζουμε *European Cities 1M*[6]. Αποτελείται από συνολικά $1,037,574$ εικόνες του Flickr με πληροφορία τοποθεσίας (geo-tag) από 22 Ευρωπαϊκές πόλεις και είναι ένα υποσύνολο της συλλογής των εικόνων της εφαρμογής VIRaL. Από αυτές, $1,081$ εικόνες από την Βαρκελώνη έχουν επισημειωθεί σε 35 ομάδες εικόνων της ίδιας σκηνής, κτιρίου ή ορόσημου. Γνωστά ορόσημα της πόλης της Βαρκελώνης απεικονίζονται σε 17 από τις ομάδες, ενώ οι υπόλοιπες 18 απεικονίζουν σκηνές ή κτίρια στην περιοχή του κέντρου της πόλης. Δείγματα από το επισημειωμένο υποσύνολο φαίνονται στα Σχήματα 4.7 και 4.8, τα οποία απεικονίζουν εικόνες από τις ομάδες ορόσημων (landmarks) και μη-ορόσημων (non-landmarks) αντίστοιχα. Παρακάτω, θα αναφερόμαστε στις τοποθεσίες που δεν περιλαμβάνουν ορόσημα ως *σκηνές*. Καθώς μόνο ένα υποσύνολο των επισημειωμένων εικόνων είναι ορόσημα, η επισημείωση δεν μπορούσε να βασιστεί σε tags μόνο και τελικά παράχθηκε με ένα μείγμα οπτικής επέκτασης ερωτήματος και χειροκίνητου καθαρίσματος. Έχουμε επίσης αναθέσει γεωγραφικές συντεταγμένες και τα σχετικά άρθρα τις Wikipedia, όποτε υπάρχουν, σε κάθε μια από τις ομάδες. Έτσι, η συλλογή μπορεί να χρησιμοποιηθεί περαιτέρω για αξιολόγηση του γεωγραφικού εντοπισμού και και την αναγνώρισης οροσήμων. Από κάθε ομάδα έχουν επιλεχθεί πέντε εικόνες ως εικόνες αναζήτησης. Αν η ομάδα περιέχει λιγότερες από 5 εικόνες, κάτι σύνηθες για τις ομάδες των σκηνών, τότε όλες οι εικόνες τις ομάδας χρησιμοποιούνται ως εικόνες αναζήτησης. Έχουμε λοιπόν συνολικά 157 εικόνες αναζήτησης. Ο Πίνακας 4.1 παρουσιάζει τα ονόματα των ορόσημων που έχουν επιλεχτεί για την αξιολόγηση και το μέγεθος της αντίστοιχης ομάδας για κάθε ένα από τα ορόσημα, καθώς και για τις σκηνές που περιέχονται στο επισημειωμένο υποσύνολο.

Η συλλογή του ενός εκατομμυρίου εικόνων περιέχει συνολικά $128,715$ εικόνες από την πόλη της Βαρκελώνης. Καθώς οι $1,081$ επισημειωμένες εικόνες είναι υποσύνολο αυτών, εξαιρούμε τις υπόλοιπες εικόνες της πόλης από τη συλλογή για τα πειράματα μας, για να είμαστε σίγουροι ότι καμία άλλη εικόνα της συλλογής δεν απεικονίζει την ίδια σκηνή με τις επισημειωμένες. Οι υπόλοιπες $908,859$ εικόνες χρησιμοποιούνται ως εικόνες *περίσπασης* (distractors). Οι περισσότερες από αυτές απεικονίζουν αστικά τοπία όπως και οι επισημειωμένες σκηνές, κάτι που κάνει

[6] Η συλλογή αυτή είναι δημοσίως διαθέσιμη στο διαδίκτυο στη διεύθυνση `image.ntua.gr/iva/datasets/ec1m/`.



Σχήμα 4.7: *Εικόνες αναζήτησης από τις* 17 *ομάδες του επισημειωμένου υποσυνόλου που αντιστοιχούν σε ορόσημα.*



Σχήμα 4.8: *Εικόνες αναζήτησης από τις* 18 *ομάδες του επισημειωμένου υποσυνόλου που αντιστοιχούν σε σκηνές.*

Landmark	Group size	Non-landmark	Group size
La Pedrera(a)	129	Scene1	5
Park Guell(a)	50	Scene2	3
Museu Nat. d' Art	17	Scene3	22
Columbus Monument	18	Scene4	2
Carrer B.I.-El Gotic	36	Scene5	30
Port Vell	18	Scene6	5
Sagrada Familia	29	Scene7	4
Casa Batllo	16	Scene8	3
Arc de Triomf	20	Scene9	17
La Pedrera(b)	71	Scene10	14
Hotel Arts	106	Scene11	22
Hosp. de San Pau(a)	116	Scene12	7
Hosp. de San Pau(b)	73	Scene13	4
Park Guell(b)	17	Scene14	2
Torre Agbar	93	Scene15	2
Placa de Catalunya	48	Scene16	5
Cathedral (side)	70	Scene17	4
		Scene18	3

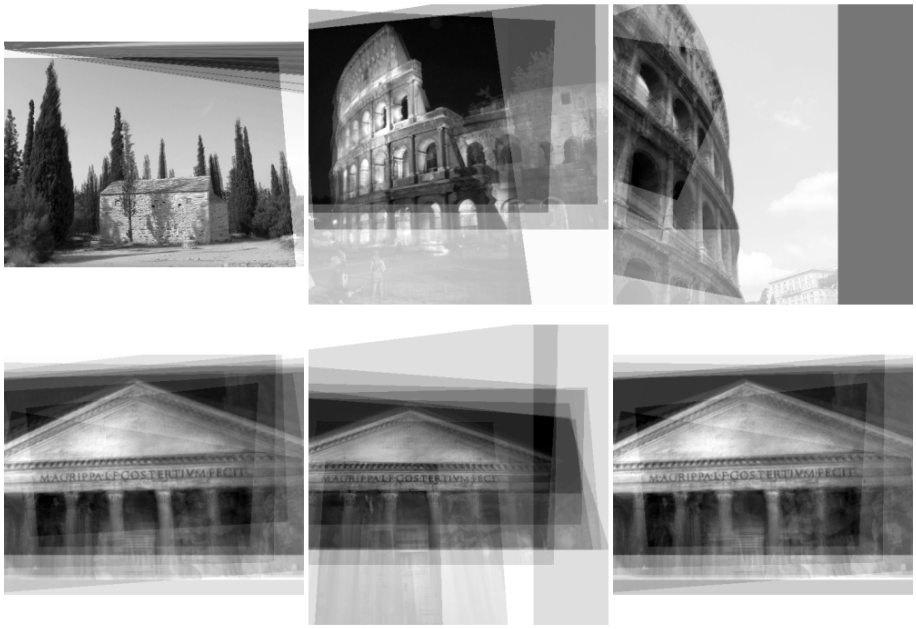
Πίνακας 4.1: *Ονόματα και μεγέθη των ομάδων που αντιστοιχούν σε ορόσημα (*17 *ομάδες) και σκηνές (*18 *ομάδες) του επισημειωμένου υποσυνόλου από τη Βαρκελώνη.*

αυτή τη συλλογή εικόνων περίσπασης ιδιαίτερα δύσκολη.

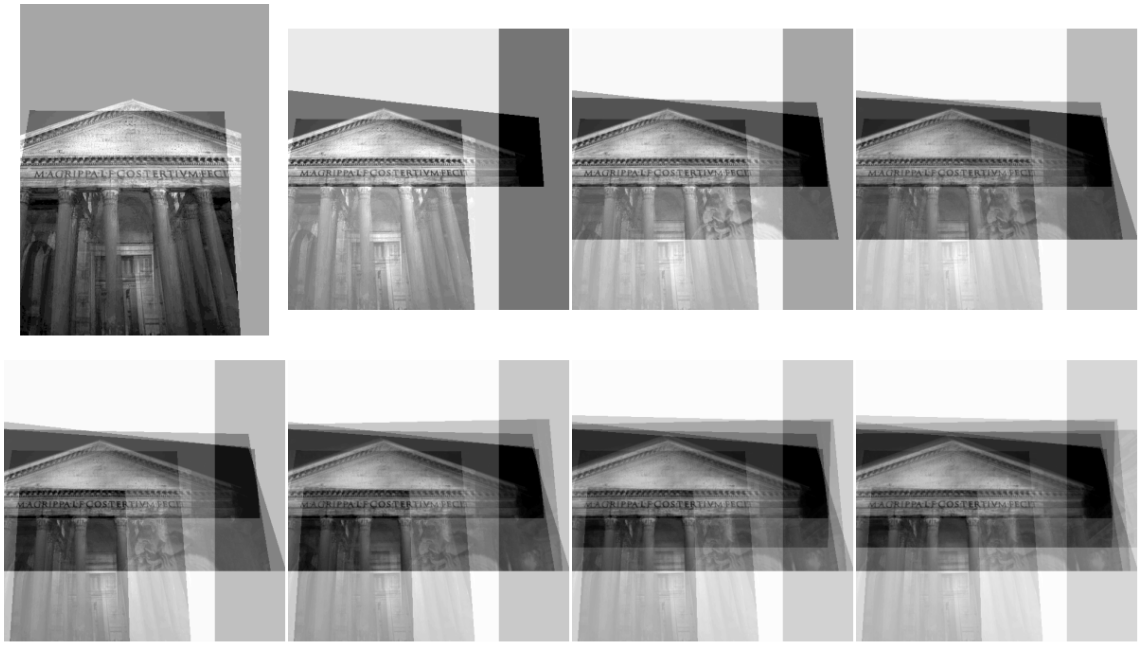
Το Σχήμα 4.9 παρουσιάζει 6 παραδείγματα από χάρτες σκηνής, που περιέχουν συνολικά 42 εικόνες. Τα παραδείγματα αυτά είναι από το σύνολο των εικόνων περίσπασης. Οι 42 αυτές εικόνες αντικαθίστανται στη δομή δεικτοδότησης από 6 συμπιεσμένους χάρτες σκηνής. Το Σχήμα 4.10 παρουσιάζει τη σταδιακή κατασκευή ενός τέτοιου χάρτη, όπου οι εικόνες τις αντίστοιχης οπτικής ομάδας ευθυγραμμίζονται η μία μετά την άλλη πάνω στην εικόνα αναφοράς

4.5.2 Πρωτόκολλο πειραμάτων

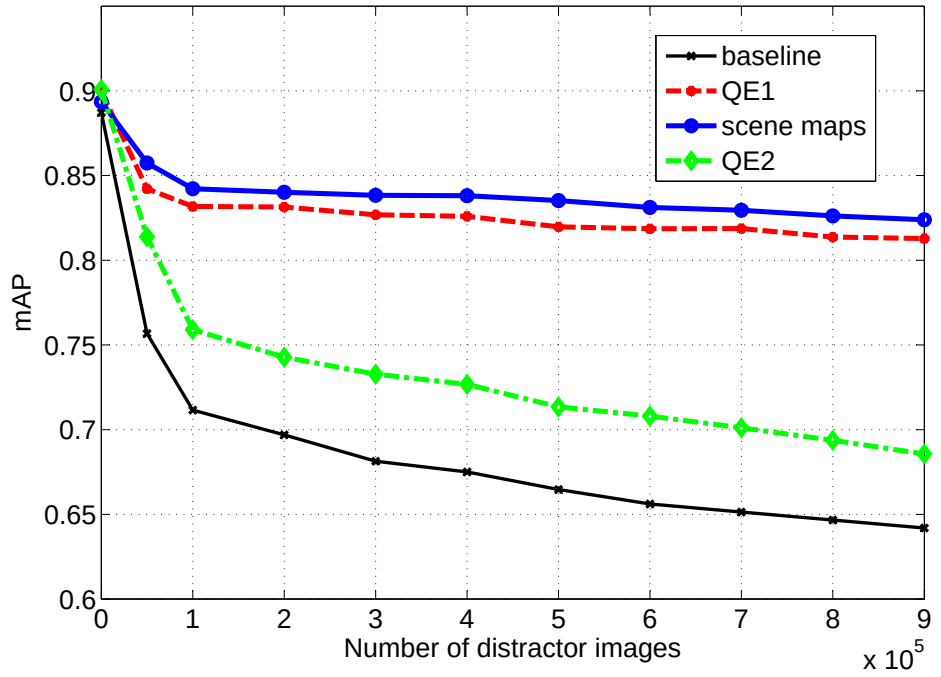
Για όλα τα πειράματα χρησιμοποιούμε το μεσαίο μέγεθος εικόνων του Flickr, δηλαδή 500×500 pixels στην μέγιστη περίπτωση. Εξάγουμε SURF χαρακτηριστικά και περιγραφείς [10] και κρατάμε το πολύ $1,000$ χαρακτηριστικά ανά εικόνα. Κατασκευάζουμε ένα γενικό (generic) οπτικό λεξικό μεγέθους 75K από περιγραφείς αστικών εικόνων που δεν είναι υποσύνολο της βάσης αξιολόγησης των πειραμάτων. Μεγαλύτερα λεξικά δεν είχαν καλύ απόδοση κατά την κατασκευή χαρτών σκηνής. Για την κατασκευή του οπτικού λεξικού και την ανάθεση οπτικών λέξεων χρησιμοποιήθηκε ο προσεγγιστικός αλγόριθμος k-means [85] και η βιβλιοθήκη



Σχήμα 4.9: *Παραδείγματα χαρτών σκηνών.* 42 *εικόνες χρησιμοποιήθηκαν για να κατασκευαστούν οι* 6 *χάρτες σκηνής του παραδείγματος.*



Σχήμα 4.10: *Η διαδικασία κατασκευής ενός χάρτη σκηνής. Οι εικόνες τις αντίστοιχης οπτικής ομάδας ευθυγραμμίζονται πάνω στην εικόνα αναφοράς ακολουθιακά.*

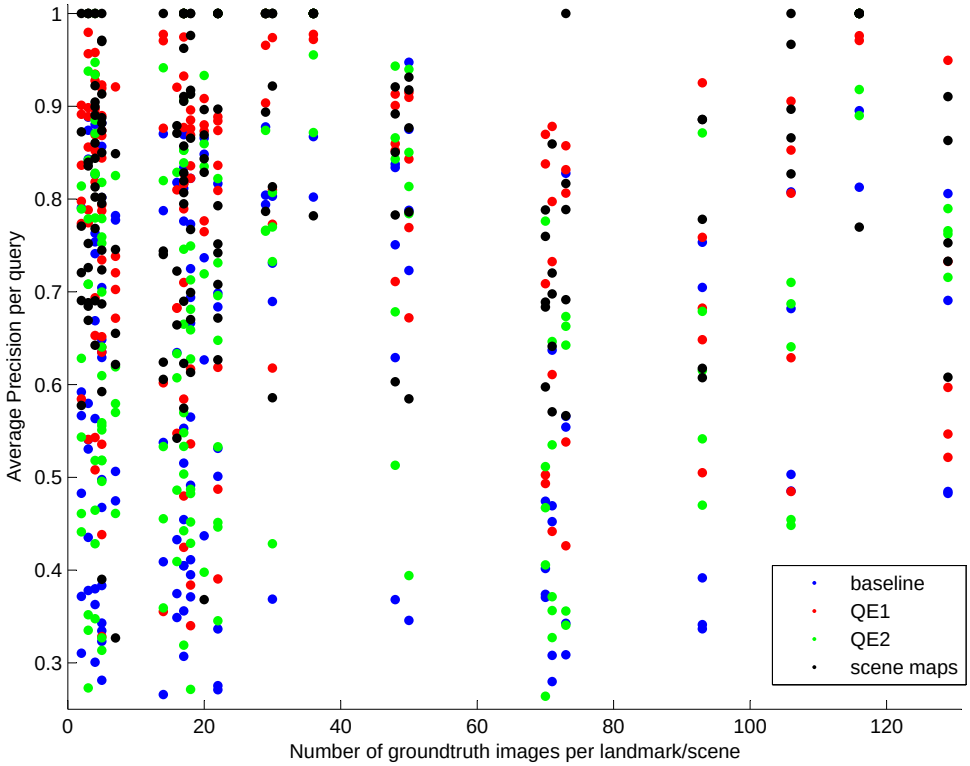


Σχήμα 4.11: *Σύγκριση του μέτρου Mean Average Precision για τις τέσσερις μεθόδους στη συλλογή* European Cities 1M *για διαφορετικό αριθμό εικόνων περίσπασης.*

FLANN των Muja και Lowe [75] αντίστοιχα. Η υλοποίηση μας του μοντέλου bag of words χρησιμοποιεί την ομοιότητα μέσω τομης ιστογραμμάτων, σε διανύσματα κανονικοποιημένα κατά L_1-νόρμα και βάρη TF-IDF. Λεπτομέρειες για τη δεικτοδότηση και το γεωμετρικό ταίριασμα κατά τη διάρκεια της οπτικής ομαδοποίησης και της κατασκευής των χαρτών σκηνής έχουν παρουσιαστεί στις ενότητες 4.3 και 4.4 αυτού του κεφαλαίου, αντίστοιχα. Μετράμε την απόδοση της αναζήτησης με το μέτρο της μέσης ακρίβειας ή mean Average Precision (mAP).

4.5.3 Αξιολόγηση της οπτικής αναζήτησης

Η διαδικασίες εξόρυξης και ομαδοποίησης που οδηγούν στην αναζήτηση μέσω χαρτών σκηνής είναι πλήρως αυτοματοποιημένες. Η γεωγραφική ομαδοποίηση στις εικόνες της βάσης European Cities 1M παίρνει κάτω από 5 λεπτά και παράγει συνολικά $1,677$ γεωγραφικές ομάδες. Η οπτική ομαδοποίηση παράγει συνολικά $493,693$ οπτικές ομάδες. Ο χρόνος της ομαδοποίηση είναι περίπου 22 λεπτά, όμως η αναζήτηση με όλες τις εικόνες που απαιτείται για την εξαγωγή των πυρήνων οπτικής ομοιότητας απαιτεί 52 ώρες και είναι με διαφορά η πλέον χρονοβόρα διαδικασία. Η κατασκευή των χαρτών σκηνής χρειάζεται επιπλέον 5 ώρες. Σημαντική παρατήρηση είναι ότι $351,391$ από τις οπτικές ομάδες είναι *μονές* εικόνες, περιπτώσεις στις οποίες δεν κατασκευάζεται ουσιαστικά χάρτης σκηνής. Για μεγαλύτερες βάσεις από περισσότερες πόλεις οι παραπάνω χρόνοι θα αυξάνο-



Σχήμα 4.12: *Μέση ακρίβεια για κάθε αναζήτηση ως προς το μέγεθος της αντίστοιχης ομάδας.*

νται γραμμικά, ενώ μπορούμε φυσικά να μειώσουμε τον χρόνο χρησιμοποιώντας παράλληλα συστήματα. Το ανεστραμμένο αρχείο για την αναζήτηση με χάρτες σκηνής απαιτεί 1.20GB μνήμης, αντί για 1.61GB που απαιτούνται για τη βασική αναζήτηση, παρέχοντας έτσι μια συμπίεση της τάξεως του 25%. Εκτελούμε όλα τα πειράματα με τις δικές μας υλοποιήσεις σε C++, σε ένα τετραπύρηνο μηχάνημα 2GHz με 8GB μνήμης. Ο συνολικός αριθμός tags της βάσης είναι $7,764,264$ και το λεξικό περιέχει $188,989$ και $181,752$ όρους, πριν και μετά από την εφαρμογή stoplist, αντίστοιχα. $2,396,926$ tags αντιστοιχούν στους όρους που «κόβονται» με τη stoplist.

Μέθοδος	Μέσος χρόνος αναζήτησης	mAP
Baseline BoW	1.03s	0.642
QE1	20.30s	0.813
QE2	2.51s	0.686
Scene maps	1.29s	0.824

Πίνακας 4.2: *Μέσος χρόνος αναζήτησης και το μέτρο απόδοσης μέσης ακρίβειας (mAP) για τις τέσσερις μεθόδους, στη βάση* European Cities 1M *μαζί με όλες τις εικόνες περίσπασης.*

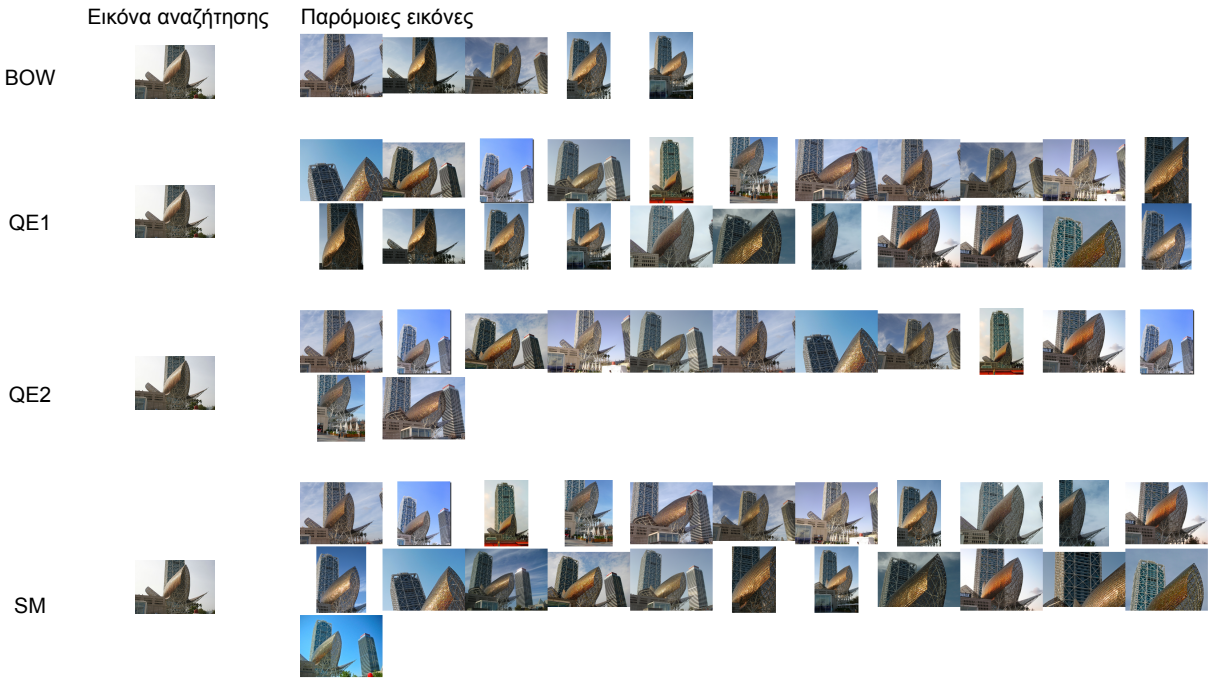
Για να αξιολογήσουμε την απόδοση της προτεινόμενης μεθόδου για οπτική αναζήτηση, υπολογίζουμε το μέτρο μέσης ακρίβειας (mean average precision ή mAP) στη προαναφερθείσα βάση εικόνων *European Cities 1M*. Συγκρίνουμε την αναζήτηση μέσω χαρτών σκηνής απέναντι στη βασική αναζήτηση με το μοντέλο bag of words καθώς και δύο μεθόδους *επέκτασης ερωτήματος* (*query expansion*). Η πρώτη (QE1) είναι η «αφελής» επαναληπτική μέθοδος, όπου το κάθε επιβεβαιωμένο αποτέλεσμα της αρχικής αναζήτησης χρησιμοποιείται για μια νέα αναζήτηση και ενώνονται στη συνέχεια όλες οι λίστες αποτελεσμάτων. Στα πειράματά μας, η επέκταση αυτής της μορφής εκτελέστηκε 3 φορές επαναληπτικά για κάθε αρχική αναζήτηση. Στη δεύτερη μέθοδο (QE2) κατασκευάζουμε ένα χάρτη σκηνής από τα επιβεβαιωμένα αποτελέσματα της αρχικής αναζήτησης και επαναλαμβάνουμε τη διαδικασία για μία ακόμα φορά. Σε όλες τις μεθόδους χρησιμοποιείται το ίδιο πρωτόκολλο για την ανακατάταξη μέσω γεωμετρικού ταιριάσματος, όπως αυτό περιγράφεται στην ενότητα 4.4. Οι μετρήσεις για τη μέση ακρίβεια mAP για όλες τις 157 επισημειωμένες εικόνες αναζήτησης και για τις τέσσερις μεθόδους, αυξάνοντας σταδιακά τον αριθμό των εικόνων περίσπασης, απεικονίζονται στο Σχήμα 4.11. Παρατηρείται ότι η προτεινόμενη μέθοδος που χρησιμοποιεί τους χάρτες σκηνής (SM) αποδίδει καλύτερα από όλες τις άλλες μεθόδους.

Όπως φαίνεται στον Πίνακα 4.2, η μέθοδός μας δε διαφέρει ιδιαίτερα ως προς τη ταχύτητα αναζήτησης από τη βασική μέθοδο, η οποία είναι με διαφορά η πιό γρήγορη. Ο χρόνος φιλτραρίσματος της προτεινόμενης μεθόδου είναι είναι μικρότερος και από τη βασική, καθώς οι χάρτες χαρακτηριστικών είναι λιγότερες από τις αρχικές εικόνες και το ανεστραμμένο αρχείο περιέχει λιγότερες καταχωρήσεις. Ο χρόνος γεωμετρικού ταιριάσματος όμως είναι λίγο μεγαλύτερος, καθώς οι χάρτες χαρακτηριστικών περιέχουν περισσότερα τοπικά χαρακτηριστικά από ότι οι εικόνες. Γενικά, ο χρόνος φιλτραρίσματος εξαρτάται μόνο από τον αριθμό των σχετικών χαρτών σκηνής, ενώ ο χρόνος για το γεωμετρικό ταίριασμα είναι σταθερός. Έτσι, ο χρόνος αναζήτησης δεν αναμένεται να αυξηθεί ιδιαίτερα με την αύξηση της βάσης σε πολλά εκατομμύρια εικόνες. Σημαντική παρατήρηση είναι επίσης ότι και οι δύο μέθοδοι επέκτασης ερωτήματος απαιτούν αρκετά μεγαλύτερο χρόνο, ενώ παράλληλα έχουν χαμηλότερη απόδοση. Η μέθοδος QE2 αντιστοιχεί περίπου σε χρόνο δύο βασικών αναζητήσεων και κατασκευής ενός χάρτη χαρακτηριστικών, ενώ η μέθοδος QE1 απαιτεί πολλές βασικές αναζητήσεις, κάτι που κάνει το χρόνο αναζήτησης απαγορευτικό.

Το επισημειωμένο υποσύνολο της βάσης που χρησιμοποιήθηκε περιέχει ομάδες εικόνων της ίδιας σκηνής διαφορετικού μεγέθους. Οι μικρές ομάδες αντιστοιχούν συνήθως σε σκηνές (non-landmark scenes), ενώ οι μεγαλύτερες σε γνωστά ορόσημα (landmarks). Στις περιπτώσεις που οι παρόμοιες εικόνες είναι πάρα πολλές είναι δύσκολο να επιτευχθούν υψηλές τιμές ανάκλησης. Η γεωμετρική ανακατάταξη εφαρμόζεται μόνο στις εικόνες με μεγάλες τιμές ομοιότητας από το ανε-



Σχήμα 4.13: *Παράδειγμα αναζήτησης, με τις παρόμοιες εικόνες που επιστρέφονται από τις τέσσερις μεθόδους, στην περίπτωση μιας μη δημοφιλούς τοποθεσίας. Η εικόνα αναζήτησης φαίνεται στα αριστερά και οι παρόμοιες εικόνες στα δεξιά. Κάθε γραμμή αντιστοιχεί σε κάθε μια από τις μεθόδους.*



Σχήμα 4.14: *Παράδειγμα αναζήτησης, με τις παρόμοιες εικόνες που επιστρέφονται από τις τέσσερις μεθόδους, στην περίπτωση μιας δημοφιλούς τοποθεσίας. Η εικόνα αναζήτησης φαίνεται στα αριστερά και οι παρόμοιες εικόνες στα δεξιά. Κάθε γραμμή αντιστοιχεί σε κάθε μια από τις μεθόδους.*

Ορόσημο	Μέθοδος			
	Baseline	QE1	QE2	Scene maps
La Pedrera(a)	0.326	0.588	0.377	0.901
Park Guell(a)	0.795	0.794	0.812	0.847
Museu Nat. d' Art	0.590	0.702	0.602	0.637
Columbus Monument	0.505	0.658	0.558	0.698
Carrer B.I.-El Gotic	0.449	0.917	0.555	0.739
Port Vell	0.332	0.746	0.380	0.480
Sagrada Familia	0.857	0.889	0.864	0.881
Casa Batllo	0.759	0.792	0.767	0.798
Arc de Triomf	0.840	0.889	0.847	0.882
La Pedrera(b)	0.651	0.921	0.939	0.903
Hotel Arts	0.560	0.773	0.573	0.633
Hosp. de San Pau(a)	0.317	0.580	0.423	0.838
Hosp. de San Pau(b)	0.421	0.776	0.502	0.709
Park Guell(b)	0.500	0.886	0.526	0.634
Torre Agbar	0.310	0.617	0.378	0.630
Placa de Catalunya	0.794	0.853	0.798	0.812
Cathedral (side)	0.487	0.864	0.546	0.972

Πίνακας 4.3: *Μέση ακρίβεια (mAP) ανά ορόσημο για τις τέσσερις μεθόδους. Για κάθε ορόσημο χρησιμοποιούνται* 5 *εικόνες αναζήτησης.*

στραμμένο αρχείο, κάτι που οδηγεί στο να «χάνονται» αρκετές σχετικές εικόνες με τη βασική μέθοδο. Στο Σχήμα 4.12 φαίνονται οι τιμές μέσης ακρίβειας για κάθε αναζήτηση ως προς το μέγεθος τις σχετικής ομάδας, του αριθμού δηλαδή των επισημειωμένων παρόμοιων εικόνων. Παρατηρείται εύκολα ότι η αναζήτηση με χάρτες σκηνής μπορούν να επιτύχουν σχεδόν πλήρη ανάκτηση ακόμα και για σκηνές με πάνω από 100 σχετικές εικόνες. Για τις ίδιες σκηνές, ακόμα και η πανίσχυρη μέθοδος επέκτασης QE1 δεν καταφέρνει να επιτύχει πλήρη ανάκτηση, καθώς μερικές από τις σχετικές εικόνες που είχαν χαθεί με την αρχική αναζήτηση πριν την επέκταση δεν μπορούν μετά να ανακτηθούν. Επίσης, σχεδόν ολική ανάκτηση παρατηρείται και τις μικρές ομάδες, όπου συνήθως όλες οι σχετικές εικόνες περιέχοντα σε μικρό αριθμό από χάρτες σκηνής, συνήθως έναν ή δύο.

Τα Σχήματα 4.13 και 4.14 δείχνουν δύο περιπτώσεις αναζήτησης για σκηνές και ορόσημα, αντίστοιχα. Φαίνεται η εικόνα αναζήτησης μαζί με τα καλύτερα αποτελέσματα μετά και από γεωμετρική ανακατάταξη. Οι Πίνακες 4.3 και 4.4 περιέχουν τις μετρήσεις της μέσης ακρίβειας (mAP) για κάθε επισημειωμένη ομάδα ορόσημων και σκηνών, αντίστοιχα.

Σκηνές	Μέθοδος			
	Baseline	QE1	QE2	Scene maps
Σκηνή1	0.618	0.648	0.654	0.884
Σκηνή2	0.667	0.847	0.730	1.000
Σκηνή3	0.399	0.458	0.451	0.880
Σκηνή4	1.000	1.000	1.000	1.000
Σκηνή5	1.000	1.000	1.000	1.000
Σκηνή6	0.800	0.969	0.848	0.802
Σκηνή7	0.876	0.979	0.940	1.000
Σκηνή8	1.000	1.000	1.000	1.000
Σκηνή9	0.339	0.557	0.357	0.754
Σκηνή10	0.351	0.482	0.428	0.687
Σκηνή11	0.557	0.843	0.575	0.633
Σκηνή12	0.577	0.857	0.639	0.755
Σκηνή13	0.681	0.846	0.746	1.000
Σκηνή14	0.875	1.000	0.880	0.885
Σκηνή15	1.000	1.000	1.000	1.000
Σκηνή16	0.791	0.883	0.798	0.812
Σκηνή17	1.000	1.000	1.000	1.000
Σκηνή18	0.800	0.972	0.810	1.000

Πίνακας 4.4: *Μέση ακρίβεια (mAP) ανά σκηνή για τις τέσσερις μεθόδους. Για κάθε μία χρησιμοποιούνται* 5 *εικόνες αναζήτησης ή όλες αν οι εικόνες τις σκηνής είναι λιγότερες από* 5.

Κεφάλαιο 5

Αναγνώριση τοποθεσίας και σκηνής

5.1 Εισαγωγή

Στο προηγούμενο κεφάλαιο παρουσιάστηκαν οι *χάρτες σκηνών*. Στο κεφάλαιο αυτό θα παρουσιάσουμε το σύστημα αναζήτησης εικόνων, το οποίο μπορεί να χρησιμοποιηθεί για αυτόματο γεωγραφικό εντοπισμό εικόνων, καθώς και αναγνώρισης οροσήμων ή σημείων ενδιαφέροντος, όπου αυτό καθίσταται εφικτό. Επεκτείνουμε τις διαδικασίες της δεικτοδότησης, της αναζήτησης και χωρικού ταιριάσματος ώστε να λειτουργούν με χάρτες σκηνής αντί για εικόνες. Έτσι, όχι μόνο μειώνουμε την απαιτούμενη μνήμη, αλλά επίσης αυξάνουμε και τα επίπεδα ανάκλησης σχετικών εικόνων. Τέλος θα παρουσιαστεί η διαδικτυακή εφαρμογή μας, VIRaL[1]—Visual Image Retrieval and Localization—η οποία παρέχει δημόσια πρόσβαση στις προαναφερθείσες τεχνολογίες μέσω ενός ενοποιημένου γραφικού περιβάλλοντος.

Εκτελούμε πειράματα και πάλι στη βάση European Cities 1M και αρκεί να βρεθεί *μονάχα ένα* επαληθευμένο ταίριασμα από τη συλλογή για να έχουμε μια εκτίμηση της τοποθεσίας της εικόνας αναζήτησης. Στο χρήστη μπορεί να παρουσιαστεί η εκτιμώμενη τοποθεσία μαζί με τις επαληθευμένες εικόνες της συλλογής πάνω στο χάρτη. Επίσης, μπορούν να παρουσιαστούν και τα ονόματα των οροσήμων ή σημείων ενδιαφέροντος που ίσως απεικονίζονται στην εικόνα ερωτήματος, πληροφορία που εξάγεται αναλύοντας τα σχετικά επισημειωμένα κείμενα (τίτλοι, tags) από τις επαληθευμένες εικόνες της συλλογής και αντιπαραβάλλοντας τα με σχετικές πληροφορίες από τη βάση Geonames[2] και άρθρα της Wikipedia με γνωστή γεωγραφική θέση. Όποτε παρουσιάζεται στην εικόνα ερωτήματος ένα ορόσημο ή σημείο ενδιαφέροντος, εμφανίζεται επιπλέον και ένας σύνδεσμος που

[1] `viral.image.ntua.gr`

[2] `www.geonames.org`

οδηγεί αντίστοιχο άρθρο.

5.2 Οπτικό ταίριασμα και δεικτοδότηση

Στην παρούσα ενότητα περιγράφουμε όλες τις διαδικασίες που χρησιμοποιούμε για το ταίριασμα, την ευθυγράμμιση και τη δεικτοδότηση των εικόνων ή των χαρτών χαρακτηριστικών. Αυτές περιλαμβάνουν (i) τις βασικές (*baseline*) τεχνικές για οπτική αναπαράσταση, ομοιότητα, δεικτοδότηση, αναζήτηση και γεωμετρικό ταίριασμα. Με αυτές, μπορεί να κατασκευαστεί ένα ολοκληρωμένο σύστημα αναζήτησης αλλά και αναγνώρισης τοποθεσίας, σύστημα το οποίο θεωρούμε ως βάση για συγκρίσεις στην ενότητα 4.5 με τα πειράματα. Επίσης, η υλοποίηση της διαδικτυακής μας εφαρμογής, του VIRaL, αποτελείται σε πολλά σημεία από τις βασικές αυτές δομές. Οι βασικές διαδικασίες περιγράφονται περιληπτικά παρακάτω, καθώς έχουν περιγραφεί αναλυτικότερα στην ενότητα 1.3 του εισαγωγικού κεφαλαίου. (ii) Τη διαδικασία δεικτοδότησης των εικόνων ανά γεωγραφική ομάδα (*geo-cluster specific* indexing), η οποία χρησιμοποιείται για υπολογισμούς αποστάσεων κατά τη διάρκεια της οπτικής ομαδοποίησης (ενότητα 4.3.3). (iii) Τη διαδικασία οπτικής ευθυγράμμισης που απαιτείται για την κατασκευή του χάρτη σκηνής (ενότητα 4.4). (iv) Τις διαδικασίες ομοιότητας, δεικτοδότησης και γεωμετρικού ταιριάσματος των *χαρτών σκηνής* (*scene maps*), οι οποίες αποτελούν επέκταση των βασικών δομών και μπορούν να χρησιμοποιηθούν για ταίριασμα είτε εικόνων είτε χαρτών σκηνής.

5.2.1 Βασικές διαδικασίες

Κατά τις βασικές διαδικασίες όλες οι εικόνες μεταχειρίζονται ανεξάρτητα. Πρώτα αναπαρίστανται από τοπικά χαρακτηριστικά και περιγραφείς, οι οποίοι στη συνέχεια κβαντίζονται και αντιστοιχούνται σε οπτικές λέξεις μέσω ενός οπτικού λεξικού. Περισσότερες πληροφορίες για τα χαρακτηριστικά και τα λεξικά που χρησιμοποιήθηκαν στα πειράματα δίνονται παρακάτω στην ενότητα 4.5. Με δεδομένες τις οπτικές λέξεις, παράγουμε την αναπαράσταση bag of words για κάθε εικόνα και μετράμε την ομοιότητα με το μέτρο της τομής ιστογραμμάτων (histogram intersection), εισάγοντας επίσης βάρη TF-IDF. Έπειτα δεικτοδοτούμε τις εικόνες μέσω ενός ανεστραμμένου αρχείου, έτσι ώστε ο χρόνος αναζήτησης να είναι μικρός, και υπο-γραμμικός σε σχέση με τον αριθμό των εικόνων της βάσης. Η κατάταξη των εικόνων της βάσης που παράγεται από το ανεστραμμένο αρχείο (στάδιο φιλτραρίσματος) εξαρτάται μόνο από την «εμφάνιση» των τοπικών χαρακτηριστικών, δηλαδή τις τιμές των περιγραφέων και δε λαμβάνεται καθόλου υπόψη η χωρική διάταξη ή γεωμετρία των χαρακτηριστικών.

Οι εικόνες που επιστρέφονται από το στάδιο φιλτραρίσματος ως εκείνες με τη μεγαλύτερη ομοιότητα με την εικόνα αναζήτησης αποτελούν τη λίστα με τις εικό-

νες που θα εξεταστούν και ως προς τη γεωμετρία (shortlist), για να διαπιστωθεί αν όντως μοιράζονται κάποιο αντικείμενο με την εικόνα αναζήτησης ή αποτελούν δύο όψεις της ίδιας σκηνής. Για το γεωμετρικό ταίριασμα χρησιμοποιούμε μια παραλλαγή του fast spatial matching [85] και ένα μοντέλο μετασχηματισμού ομοιότητας, τεσσάρων βαθμών ελευθερίας. Το μοντέλο αυτό μπορεί να κάνει υποθέσεις από μια αντιστοιχία (*single correspondence* assumption).

Συγκεκριμένα, για κάθε ζευγάρι της εικόνας αναζήτησης με κάθε μία από της εικόνες της λίστας, πρώτα παράγονται *πιθανές αντιστοιχίες* (*tentative correspondences*) μεταξύ των χαρακτηριστικών του ζεύγους που μοιράζονται κοινές οπτικές λέξεις. Με δεδομένη μία τέτοια αντιστοιχία, χρησιμοποιούμε τη θέση, την κλίμακα και τον προσανατολισμό των δύο χαρακτηριστικών της αντιστοιχίας για να υπολογίσουμε τους μετασχηματισμούς ομοιότητας T_1, T_2 που μετασχηματίζουν τα δύο χαρακτηριστικά στο μοναδιαίο κύκλο με κέντρο το κέντρο των αξόνων. Μπορούμε λοιπόν να κατασκευάζουμε μια αρχική υπόθεση μετασχηματισμού ως $T_2^{-1}T_1$. Έπειτα μετράμε τον αριθμό των γεωμετρικά επιβεβαιωμένων αντιστοιχιών ή inliers και επαναλαμβάνουμε για την επόμενη υπόθεση μετασχηματισμού χρησιμοποιώντας την επόμενη πιθανή αντιστοιχία. Όταν βρεθεί ένας νέος μέγιστος αριθμός inliers, μπορούμε μέσω ελαχίστων τετραγώνων να υπολογίσουμε έναν αφινικό μετασχηματισμό απο αυτούς και να αποθηκεύσουμε το μέχρι τότε καλύτερο μοντέλο—μια λογική που μοιάζει με το απλό μοντέλο του Locally Optimized RANSAC (LO-RANSAC) [17]. Παρατηρήσαμε ότι εικόνες που έχουν τουλάχιστον $\tau = 10$ inliers με την εικόνα αναζήτησης συνήθως απεικονίζουν το ίδιο αντικείμενο ή σκηνή.

Η διαδικασία δεικτοδότησης των εικόνων ανά γεωγραφική ομάδα (*Geo-cluster specific indexing*) είναι μία απλή παραλλαγή της βασικής διαδικασίας δεικτοδότησης, κατά την οποία το ανεστραμμένο αρχείο περιέχει δεικτες και για τις οπτικές λέξεις, αλλά και για τις γεωγραφικές ομάδες. Χρησημοποιούμες αυτή τη δομή κατά τη διάρκεια της οπτικής ομαδοποίησης, όπου εκτελούμε μια αναζήτηση με κάθε εικόνα της γεωγραφικής ομάδας και συλλέγουμε τις ταιριασμένες εικόνες ποτ δίνουν $I(p, q) > \tau$ inliers. Μιας και οι γεωγραφικές ομάδες είναι αρκετά μικρότερες σε μέγεθος σε σχέση με ολόκληρη τη συλλογή, η αναζήτηση στα προσαρμοσμένα ανεστραμμένα αρχεία εκτελείτε πολύ πιο γρήγορα. Οι άσχετες εικόνες είναι επίσης λιγότερες, συνεπώς μπορούμε να μικρύνουμε το μέγεθος της λίστας των εικόνων που θα εξεταστούν ως προς τη γεωμετρία, κάνοντας και τη διαδικασία αυτή ταχύτερη. Στην πράξη είδαμε ότι ο χρόνος αναζήτησης γίνεται σταθερός και κατά μέσο όρο μία τάξη μεγέθους μικρότερος από την βασική μέθοδο. Είναι επίσης πλέον ανεξάρτητος του μεγέθους της βάσης των εικόνων.

5.2.2 Ευθυγράμμιση όψεων

Για να κατασκευάσουμε τον χάρτη σκηνών από τις εικόνες μίας ομάδας όψεων, πρέπει πρώτα να ευθυγραμμίσουμε τα αντίστοιχα χαρακτηριστικά. Για να το επιτύχουμε αυτό, πρώτα εκτιμούμε έναν μετασχηματισμό ομογραφίας μεταξύ των ταιριασμένων εικόνων. Η ευθυγράμμιση εκτελείται σε κάθε ομάδα όψεων μεταξύ της εικόνας αναφοράς της ομάδας και κάθε μίας απο τις εικόνες της ομάδας. Είναι δηλαδή *γραμμική* ως προς τον αριθμό των εικόνων της ομάδας και μάλιστα δεν αρχίζει από το μηδέν: για κάθε ζεύγος ταιριασμένων εικόνων (p, q) μιας γεωγραφικής ομάδας, έχουμε δεδομένο τον καλυτερο αφινικό μετασχηματισμό A_{qp} που μετασχηματίζει το q στο p από το στάδιο της οπτικής ομαδοποίησης. Συνεπώς, απαιτείται μοναχά ένα τελικό στάδιο τοπικής βελτιστοποίησης για να εκτιμηθεί ένα μοντέλο ομογραφίας.

Πιό συγκεκριμένα, αν p είναι η εικόνα αναφοράς μιας ομάδας όψεων $C_v(p)$, ευθυγραμμίζουμε στο p όλες τις εικόνες $q \in C_v(p)$. Αρχίζοντας από το αποθηκευμένο αφινικό μοντέλο A_{qp}, εκτελούμε ένα μόνο βήμα της «επαναληπτικής» (“iterative”) μεθόδου LO-RANSAC. Χρησιμοποιούνται δηλαδή όλα τα χαρακτηριστικά με σφάλμα μικρότερο από ένα κατώφλι $K\theta$ για την εκτίμηση μοντέλου ομογραφίας μέσω του αλγορίθμου Direct Linear Transformation (DLT) [35]. Μειώνουμε σταδιακά το κατώφλι και επαναλαμβάνουμε μέχρι να γίνει ίσο με θ. Στην πράξη είδαμε ότι τρεις επαναλήψεις είναι αρκετές για τα πειράματά μας. Αποθηκεύουμε τον τελικό ομογραφικό μετασχηματισμό που μετασχηματίζει την εικόνα q στην εικόνα αναφοράς p και τον συμβολίζουμε ωςH_{qp}.

5.2.3 Δεικτοδότηση χαρτών σκηνής

Έχοντας υπολογίσει όλους τους χάρτες σκηνής, κατασκευάζουμε μία ξεχωριστή δομή δεικτοδότησης για αυτούς. Αν και ο μέσος χάρτης σκηνής είναι αρκετά μεγαλύτερος από μια εικόνα, μοιράζονται την ίδια αναπαράσταση, ένα σύνολο, δηλαδή, από τοπικά χαρακτηριστικά. Μπορούμε συνεπώς να μεταχειριστούμε τους χάρτες σκηνών ως εικόνες για τη δεικτοδότηση και την αναζήτηση. Κατά τη διαδικασία κατασκευής του, ο κάθε χάρτης σκηνής $S(p)$ μπορεί να μας παρέχει τα περιεχόμενα υποσύνολα $Q_x(F_w(p)$ που αντιστοιχούν σε κάθε οπτική λέξη w. Οι αριθμοί των στοιχείων των υποσυνόλων αυτών δίνουν λοιπόν κατευθείαν το διάνυσμα συχνότητας λέξεων (term frequency vector) για τον χάρτη $S(p)$. Δεικτοδοτούμε όλους τους χάρτες ανά οπτική λέξη σε ένα ανεστραμμένο αρχείο. Κατά τη διάρκεια της αναζήτησης, υπολογίζουμε ένα αντίστοιχο διάνυσμα για την εικόνα ερωτήματος και ανακτούμε τους κοντινότερους χάρτες σκηνής με μέτρο ομοιότητας την τομή ιστογραμμάτων και βάρη TDF-IF.

Όπως και στη βασική διαδικασία, δημιουργείται μια λίστα με τις εικόνες που θα εξεταστούν και ως προς τη γεωμετρία, πάλι χρησιμοποιώντας και εδώ το μοντέλο

της υπόθεσης από μια αντιστοιχία (single correspondence assumption). Όπως αναφέρθηκε και παραπάνω, η διαδικασία γεωμετρικού ταιριάσματος είναι γραμμική ως προς τις αρχικές αντιστοιχίες, συνεπώς για τους χάρτες χαρακτηριστικών θα είναι πιο αργή από ότι ανάμεσα σε δύο εικόνες. Για ταχύτητα, τερματίζουμε τη διαδικασία αν βρεθούν περισσότεροι από τ_h inliers και θεωρούμε ότι ο συγκεκριμένος χάρτης σκηνής έχει ταιριάξει. Παρομοίως, τερματίζουμε νωρίτερα τη διαδικασία και στην περίπτωση όπου έχουν βρεθεί λιγότεροι από τ_ℓ inliers μετά από τον έλεγχο ενός προκαθορισμένου ποσοστού των υποθέσεων. Τέλος, απορρίπτουμε μια υπόθεση αν οι inliers που έχουν βρεθεί για ένα προκαθορισμένο ποσοστό αντιστοιχιών είναι λιγότεροι από τ_ℓ.

Όταν ένας χάρτης σκηνής $S(p)$ ταιριάξει με την εικόνα αναζήτησης θεωρούμε ότι όλες οι εικόνες της αντίστοιχης ομάδας όψεων $q \in C_v(p)$ ταιριάζουν με αυτήν. Με αυτό τον τρόπο καταφέρνουμε να αυξήσουμε την ανάκληση. Θεωρούμε όλες τις εικόνες μιας ταιριασμένης σκηνής στην ίδια θέση κατάταξης ως προς την εικόνα αναζήτησης, κυρίως για να αποφύγουμε το εξαντλητικό ταίριασμα όλων των εικόνων με την εικόνα ερωτήματος, μια επιλογή που επηρεάζει λίγο την ακρίβεια.

5.2.4 Συζήτηση

Η βασική μας διαδικασία αναζήτησης είναι πολύ κοντά σε εκείνη της δημοσίευσης [85], μέθοδος που καταφέρνει να έχει πολύ καλή απόδοση, έχοντας παράλληλα και πολύ μικρές απαιτήσεις μνήμης. Η δεικτοδότηση των χαρτών σκηνής όμως βοηθάει και στα δύο παραπάνω. Έχει ομοιότητες με το λανθάνον μοντέλο επέκτασης ερωτήματος της δημοσίευσης [19], χωρίς όμως να εκτελείται κατά την διάρκεια της αναζήτησης και να αυξάνει το υπολογιστικό της κόστος. Υπολογίζεται στατικά, μία μόνο φορά για ολόκληρη τη βάση χωρισμένη σε γεωγραφικές ομάδες και όχι δυναμικά για την εικόνα αναζήτησης όπως στη δημοσίευση [19], όπου η επέκταση ερωτήματος μπορεί να ολισθήσει σε μη σχετικές περιοχές. Σημαντικό είναι επίσης ότι η επέκταση ερωτήματος είναι άχρηστη σε περιπτώσεις όπου οι σχετικές εικόνες στη βάση είναι λίγες (ή μόνο μία) και η αρχική αναζήτηση δεν επιτύχει. Στην ενότητα 4.5 όπου περιγράφονται τα σχετικά πειράματα, συγκρίνουμε την προτεινόμενη μέθοδο με δύο τεχνικές επέκτασης ερωτήματος.

5.3 Αναγνώριση τοποθεσίας και ορόσημων

Καθώς οι εικόνες που επιστρέφονται από μία αναζήτηση απεικονίζουν κομμάτι της ίδιας σκηνής με την εικόνα ερωτήματος, είναι πολύ πιθανό να έχουν τραβηχτεί σε κοντινές τοποθεσίες. Επίσης, όποτε κάποια από τις παρόμοιες εικόνες συσχετίζεται με κάποιο γνωστό ορόσημο ή τοποθεσία ενδιαφέροντος, μπορούμε να υποθέσουμε ότι η συσχέτιση αυτό ισχύει και για την εικόνα αναζήτησης. Παρα-

κάτω αναλύουμε τις ιδέες αυτές για την αυτόματη εξαγωγή της τοποθεσίας και των περιεχόμενων ορόσημων/σημείων ενδιαφέροντος αντίστοιχα. Στη συνέχεια σχολιάζουμε τις επιλογές μας σε σχέση με άλλες υπάρχουσες λύσεις.

5.3.1 Αναγνώριση τοποθεσίας

Μετά την ανάκτηση ενός συνόλου εικόνων της βάσης που ταιριάζουν με την εικόνα αναζήτησης, εκμεταλλευόμαστε τις γνωστές γεωγραφικές τους τοποθεσίες (geo-tags) για να αναγνωρίσουμε την τοποθεσία όπου τραβήχτηκε η εικόνα αναζήτησης. Φυσικά, τα geo-tags των εικόνων της βάσης μπορούν να έχουν διαφορετικά επίπεδα ακρίβειας, και κάποια από αυτά μπορεί να είναι τελείως λάθος. Κάνουμε όμως την υπόθεση ότι από το σύνολο των παρόμοιων εικόνων που επιστράφηκαν, υπάρχει ένα υποσύνολο αυτών με σωστή την γεωγραφική τους τοποθεσία, κάτι που σημαίνει ότι οι μεταξύ τους τοποθεσίες θα είναι κοντά η μία στην άλλη. Συνεπώς, μπορούμε να εκτελέσουμε συγχωνευτική ομαδοποίηση (agglomerative clustering) στις γεωγραφικές συντεταγμένες των παρόμοιων εικόνων, τερματίζοντας όταν η ελάχιστη απόσταση μέσα σε μία από τις ομάδες είναι μεγαλύτερη από ένα κατώφλι. Εάν υπάρχει ομάδα η οποία περιέχει περισσότερες τοποθεσίες (φωτογραφίες) από όλες τις άλλες, τότε επιστρέφουμε το κέντρο αυτής της ομάδας σαν μία εκτίμηση της τοποθεσίας της εικόνας αναζήτησης.

Εφαρμόζουμε τον αλγόριθμο των *αμοιβαίων κοντινότερων γειτόνων* (*reciprocal nearest neighbor* ή RNN) [58] για την ομαδοποίηση, χρησιμοποιώντας το κριτήριο μέσου όρου και την Ευκλείδεια απόσταση—η πιο ακριβής γεωδεσική απόσταση δεν είναι απαραίτητη εδώ, καθώς όλες οι τοποθεσίες υποτίθεται ότι είναι ιδιαίτερα κοντά μεταξύ τους. Στην πράξη θέτουμε το κατώφλι τερματισμού στα $200m$ για να μπορούμε να αναπαραστήσουμε την περιοχή γύρω από ένα κτίριο, ορόσημο ή γενικότερα σκηνή. Η επιλογή της συγχωνευτικής ομαδοποίησης είναι ιδανική στην προκειμένη περίπτωση, καθώς επιτρέπει την έκταση των ομάδων να προσαρμοστεί στο πώς κατανέμονται οι τοποθεσίες γύρω από τις απεικονιζόμενες σκηνές, αλλά παράλληλα δεν αφήνει δύο ομάδες να συγχωνευτούν αν είναι μακριά μεταξύ τους. Ο αριθμός των ομάδων εξάγεται αυτόματα από τα δεδομένα, ενώ οι υπολογισμοί είναι ιδιαίτερα γρήγοροι για να γίνονται κατά τη διάρκεια της αναζήτησης. Η επιλογή της μεγαλύτερης ομάδας κάνει την εκτίμηση μας πιο εύρωστη και με αυτό τον τρόπο οι παρόμοιες εικόνες με λάθος geo-tag δε συμμετέχουν καθόλου στην εκτίμηση και την αναγνώριση της τοποθεσίας.

5.3.2 Συνήθη tags χρηστών

Η αναγνώριση ορόσημων ή σημείων ενδιαφέροντος βασίζεται στα tags των χρηστών και τους τίτλους που έχουν οι φωτογραφίες της συλλογής[3]. Οι τίτλοι αποδείχθηκαν συνήθως πιο αξιόπιστοι, αλλά και τα tags μπορούν να βοηθήσουν, παρότι πολλές φορές περιέχουν θόρυβο. Για μεγαλύτερη αποτελεσματικότητα, αναπαριστούμε τους όρους μέσω ενός λεξικού και εξάγουμε ένα σύνολο από συνήθη tags μέσω αυτής της αναπαράστασης. Αρχικά φιλτράρουμε το σύνολο των tags των εικόνων της βάσης, αφαιρώντας τους όρους που εμπίπτουν σε μία χειροκίνητα φτιαγμένη *stoplist*, μια λίστα δηλαδή που περιέχει όρους πολύ γενικούς (π.χ. `paris`, `france`, `holidays`), όρους που περιγράφουν τις συνθήκες της φωτογράφισης (π.χ. `night shot`, `black and white`), ή όρους που είναι πρακτικά άσχετοι με το περιεχόμενο της φωτογραφίας (π.χ. `nikon`, `geo-tagged`).

Έπειτα κατασκευάζουμε το λεξικό στατικά, αρχικοποιόντας το με τα δεδομένα του web service για Wikipedia Search[4] του ιστοτόπου Geonames. Για κάθε πόλη της συλλογής, έχουμε συλλέξει όλες τις καταχωρίσεις της βάσης Geonames για την γεωγραφική περιοχή του κέντρου, όπως επίσης κάναμε και για την κατασκευή της συλλογής (βλέπε ενότητα 4.5.1). Κάθε καταχώρηση αντιστοιχεί σε ένα ορόσημο ή σημείο ενδιαφέροντος της πόλης και έτσι δημιουργούμε ουσιαστικά μια ομάδα για κάθε ένα από αυτά. Για να είμαστε εύρωστοι ως προς ορθογραφικά λάθη ή ιδιαιτερότητες τις εκάστοτε γλώσσας, συγκρίνουμε τις συμβολοσειρές μέσω της απόστασης Levenshtein [60]. Ξεκινώντας με τη λίστα όλων των tags της βάσης, ακολουθιακά επιλέγουμε τυχαία ένα από αυτά, τη αφαιρούμε από τη λίστα και το συγκρίνουμε με το υπάρχον σύνολο ομάδων. Αν βρίσκεται σε απόσταση μικρότερη από T ως προς κάποιο από αυτά, το εισάγουμε στην αντίστοιχη ομάδα, αλλιώς του αφήνουμε να δημιουργήσει μία νέα, δικιά του ομάδα. Η παραπάνω διαδικασία επαναλαμβάνεται μέχρι να αδειάσει η λίστα από τα tags.

Η διαδικασία αυτή είναι παρόμοια με την ομαδοποίηση *canopy* [71], εδώ όμως χρησιμοποιούμε ένα μόνο κατώφλι και μία συγκεκριμένη αρχική κατάσταση για τις ομάδες. Αυτή η ιδιαίτερη αρχικοποίηση είναι ο λόγος για τον οποίο δεν επιστρατεύουμε την ομαδοποίηση KVQ, που κατά τα άλλα θα ταίριαζε και σε αυτό το πρόβλημα καλά. Όλα τα tags αντιστοιχούνται σε μία μόνο ομάδα από το κέντρο της οποίας απέχουν το πολύ T και εκείνα τα οποία σχετίζονται με γνωστά ορόσημα αναπαρίστανται με την επίσημη ονομασία που τους δίνει ο ιστότοπος Geonames. Στη συνέχεια συσχετίζουμε κάθε εικόνα με τις λέξεις του λεξικού που αντιστοιχούν τα tags της. Έπειτα, κατά την ώρα της αναζήτησης, συλλέγουμε όλες τις λέξεις από τις παρόμοιες εικόνες και φτιάχνουμε το σύνολο των *συνήθη tags* από τις λέξεις που έχουν τουλάχιστον δύο εμφανίσεις. Για την διαδικασία που εκτελείται

[3] Οι τίτλοι και τα tags των χρηστών είναι μεταδεδομένα τα οποία, όπως και την τοποθεσία, βάζουν αυτόματα οι χρήστες του ιστοτόπου Flickr.

[4] `www.geonames.org/export/wikipedia-webservice.html\#wikipediaSearch`

κατά την αναζήτηση δεν απαιτούνται καθόλου συγκρίσεις συμβολοσειρών.

5.3.3 Αναγνώριση ορόσημων

Θεωρούμε ως ορόσημο (landmark) ή σημείο ενδιαφέροντος (point of interest ή POI) όποιο αντικείμενο του λεξικού έχει αντιστοιχηθεί σε κάποιο άρθρο της Wikipedia, το οποίο είναι επίσης και γεωγραφικά τοποθετημένο μέσα στα γεωγραφικά όρια των πόλεων της συλλογής. Για να κατασκευάσουμε μια λίστα από τέτοια αντικείμενα, χρησιμοποιούμε και πάλι τη βάση των Geonames καθώς και την αντίστοιχη υπηρεσία του ιστοτόπου της Wikipedia[5]. Οι δύο αυτές υπηρεσίες είναι παρόμοιες και συνήθων μοιράζονται το 90% των καταχωρίσεων, είναι δηλαδή καταχωρήσεις με την ίδια σελίδα Wikipedia. Υπάρχουν όμως και περιπτώσεις διαφορών, γι αυτό το λόγο συγχωνεύουμε τις δύο λίστες σε μία. Για κάθε αντικείμενο κρατάμε το όνομα του άρθρου, τον σύνδεσμο (url) και τις γεωγραφικές συντεταγμένες.

Μετά από μια επιτυχημένη αναζήτηση, έχοντας και την εκτιμώμενη τοποθεσία, επιλέγουμε μια λίστα με τα άρθρα που έχουν γεωγραφικές συντεταγμένες μέσα σε μιά ακτίνα από αυτήν—πρακτικά $200m$, όπως και στην εκτίμηση τοποθεσίας. Κάθε τίτλος των άρθρων της λίστας συγκρίνεται με τα συνήθη tags και τους τίτλους των παρόμοιων εικόνων που επιστράφηκαν από την αναζήτηση και, χρησιμοποιώντας και πάλι την απόσταση Levenshtein, κάθε άρθρο παίρνει βάρος ίσο με την ελάχιστη απόσταση που βρέθηκε. Κατατάσσουμε τα άρθρα με αύξουσα σειρά ως προς την απόσταση, και επιλέγουμε εκείνα με την μικρότερη, αν επίσης αυτή είναι και μικρότερη από T, ώστε να αποτελέσουν το σύνολο των *προτεινόμενων tags* για την εικόνα αναζήτησης. Όταν στο σύνολο αυτό περιέχονται και ορόσημα ή σημεία ενδιαφέροντος, σύνδεσμοι που οδηγούν στα αντίστοιχα άρθρα της Wikipedia εμφανίζονται αυτόματα στη σελίδα αποτελεσμάτων της εφαρμογής VIRaL.

5.3.4 Συζήτηση

Στη βιβλιογραφία έχουν ακολουθηθεί διάφορες προσεγγίσεις για επεξεργασία των tag και για αναγνώριση ορόσημων. Στις δημοσιεύσεις [22, 63] ακολουθείται μια διαδικασία μάθησης και η γεωγραφική ομαδοποίηση χρησιμοποιείται μόνο για την κατασκευή της γνώσης εκπαίδευσης. Ταξινομητές εκπαιδεύονται με αυτή τη γνώση χρησιμοποιώντας το κείμενο, οπτικά χαρακτηριστικά ή και τα δύο. Η μέθοδος αυτή βέβαια περιορίζεται όταν η συλλογή περιέχει αρκετό θόρυβο και η αναγνώριση φτάνει μέχρι βάσεις μεγέθους 500 ορόσημων, χωρίς επίσης να υποστηρίζεται η αναγνώριση της τοποθεσίας. Αντίθετα, η προτεινόμενη λύση μπορεί να υποστηρίξει περίπου $8,500$ ορόσημα και σημεία ενδιαφέροντος, από την πα-

[5] de.wikipedia.org/wiki/Wikipedia:WikiProjekt_Georeferenzierung/Wikipedia-World/en

ρούσα συλλογή που περιλαμβάνει 23 πόλεις. Ενδιαφέρουσα παρατήρηση είναι επίσης ότι στις δημοσιεύσεις [22] και [63] αναλύονται μόνο τα tags και όχι οι τίτλοι, πεδίο πιό αξιόπιστο, που από ότι είδαμε αυξάνει πολύ και την ακρίβεια.

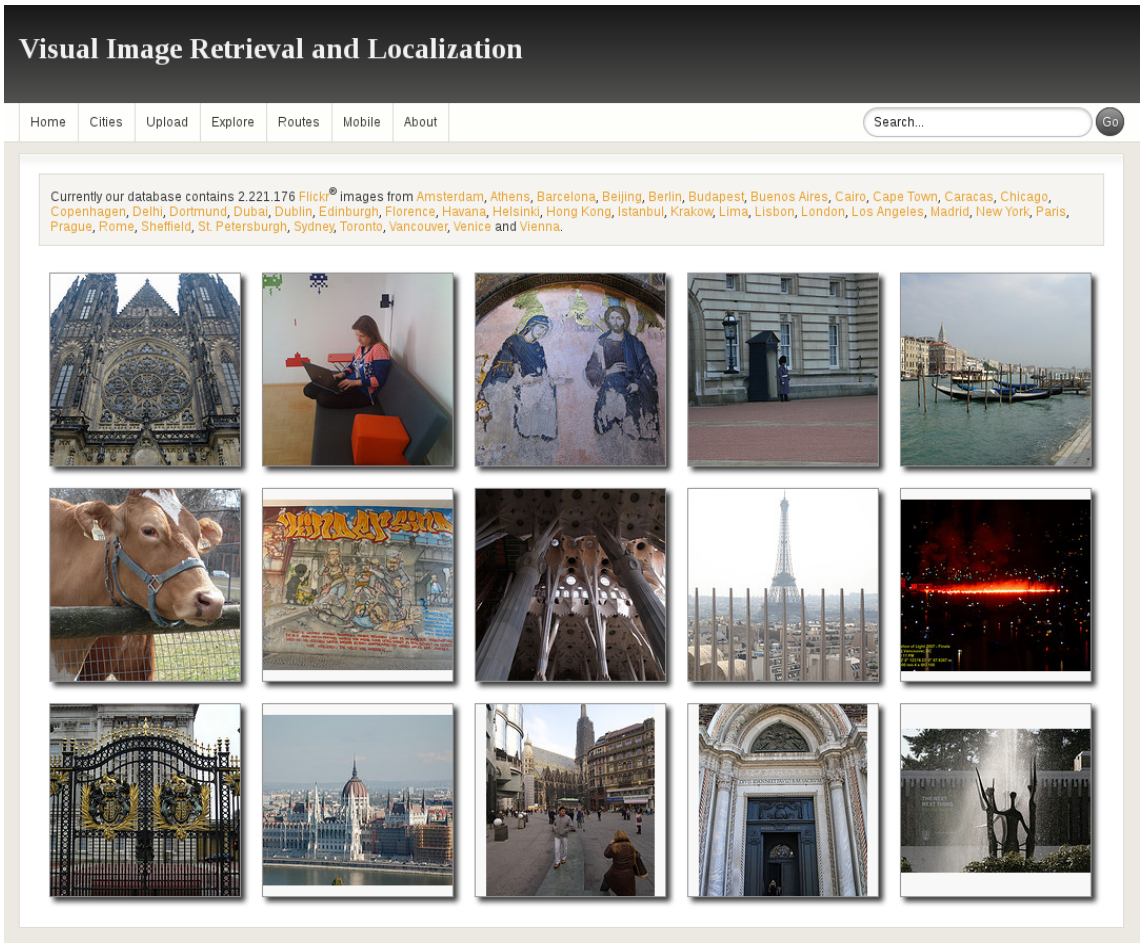
Οι Quack *et al*. [87] ακολουθούν μια εξαντλητική διαδικασία: για κάθε οπτική ομάδα εξάγουν τα tags, στέλνουν αντίστοιχα ερωτήματα στη μηχανή αναζήτησης Google για τα σχετικά άρθρα της Wikipedia, και έπειτα κατεβάζουν τις φωτογραφίες των άρθρων και τις ταιριάζουν με τις εικόνες τις συλλογής τους για να επαληθεύσουν την αντιστοιχία. Οι εικόνες των αντίστοιχων άρθρων, όμως, δεν είναι ιδιαίτερα αξιόπιστες και η όλη διαδικασία εξόρυξης που εφαρμόζουν είναι ιδιαίτερα αργή. Εμείς χρησιμοποιούμε την τοποθεσία που εξάγαμε για κάθε εικόνα αναζήτησης και έτσι περιορίζουμε πολύ την αναζήτηση κειμένου μόνο στη γύρω περιοχή κατά τη διάρκεια της αναζήτησης. Όπως και στην προτεινόμενη μέθοδο, οι θέσεις των άρθρων χρησιμοποιούνται στη δημοσίευση [28] αλλά για στατική ανάθεση των άρθρων στις οπτικές ομάδες και όχι κατά τη διάρκεια της αναζήτησης.

Οι Zheng *et al*. [117] επίσης αντιστοιχούν ορόσημα στισ οπτικές ομάδες στατικά και μάλιστα υποστηρίζουν περίπου $5,500$ ορόσημα από $1,300$ πόλεις σε 144 χώρες, ψάχνοντας όμως σε ένα μικρό υποσύνολο των αντιπροσωπευτικών εικόνων κάθε ομάδας. Αντίθετα, μέσω την χαρτών σκηνής, εμείς μπορούμε να ψάχνουμε στο σύνολο των εικόνων της συλλογής και να αναγνωρίσουμε *οποιοδήποτε* σημείο ενδιαφέροντος μέσα στις περιοχές που υποστηρίζονται καθώς και να εκτιμήσουμε τοποθεσία για *οποιαδήποτε* φωτογραφία, είτε περιέχει ορόσημα είτε όχι.

5.4 Η εφαρμογή VIRaL

Οι προτεινόμενες μέθοδοι μπορούν και εξερευνηθούν μέσω της διαδικτυακής μας εφαρμογής VIRaL. Η βάση δεδομένων της εφαρμογής περιέχει σήμερα πάνω από 2.2 εκατομμύρια εικόνες από το Flickr, τραβηγμένες σε 40 πόλεις από όλον τον κόσμο, μαζί με τα μεταδεδομένα τους (δηλαδή την τοποθεσία, tags των χρηστών, τίτλους και περιγραφές). Η συλλογή περιέχει μόνο εικόνες τραβηγμένες σε ένα γεωγραφικό παράθυρο στο κέντρο των πόλεων και ένα υποσύνολο αυτής της συλλογής αποτελεί και τη βάση που χρησιμοποιούμε για τα πειράματα μας, υποσύνολο που περιγράφεται αναλυτικότερα στην ενότητα 4.5.1.

Σε κάθε αναζήτηση, το αποτέλεσμα είναι μια λίστα από οπτικά παρόμοιες εικόνες και για το οπτικό ταίριασμα ακολουθούνται οι διαδικασίες που περιγράφονται στην ενότητα 5.2.1. Μέσω του γραφικού περιβάλλοντος ο χρήστης μπορεί επίσης να δει το γεωμετρικό ταίριασμα μεταξύ της εικόνας αναζήτησης και κάθε παρόμοιας εικονας. Επίσης η εικόνα αναζήτησης τομοθετείται στον χάρτη, στις εκτιμώμενες γεωγραφικές συντεταγμένες και παρουσιάζονται τα συνήθη και τα προτεινόμενα tags που εξάγουμε με τη διαδικασία που περιγράφεται στην ενότητα 5.3).



Σχήμα 5.1: *Η αρχική σελίδα της διαδικτυακής εφαρμογής VIRaL παρουσιάζει ένα τυχαίο υποσύνολο από εικόνες της συλλογής.*

Τα προτεινόμενα tags συνοδεύονται από συνδέσμους στα αντίστοιχα άρθρα της Wikipedia, όπου αυτά υπάρχουν.

5.4.1 Αναγνώριση τοποθεσίας μέ το VIRaL

Στην αρχική σελίδα της διαδικτυακής εφαρμογής (Σχήμα 5.1) παρουσιάζει ένα τυχαίο υποσύνολο της συλλογής. Ο χρήστης μπορεί να περιηγηθεί στη συλλογή με τρεις συνολικά τρόπους: α) μέσω της αρσικης σελίδας, β) μέσω της σελίδας πόλεων (*Cities*), σελίδα που παρουσιάζει τυχαίες εικόνες από κάθε πόλη γ) μέσω της εφαρμογής *Explore*. Κάνοντας κλικ σε οποιαδήποτε από τις εικόνες θα ξεκινήσει αναζήτηση με αυτή στο σύστημα.

Στα Σχήματα 5.2 και 5.4 παρουσιάζεται η σελίδα αποτελεσμάτων. Πάνω αριστερά στη σελίδα φαίνεται ο χάρτης, ο οποίος περιλαμβάνει έναν μπλε marker για κάθε παρόμοια εικόνα που επέστρεψε η αναζήτηση και έναν κόκκινο marker για την εκτιμώμενη τοποθεσία της εικόνας αναζήτησης. Οι γκρι markers υποδηλώνουν οπτικά όμοιες, αλλά λάθος γεωγραφικά τοποθετημένες εικόνες της συλλογής, εικόνες που τελικα δεν συμμετέχουν στην εξαγωγή της τοποθεσίας. Στο πάνω

δεξιά μέρος της σελίδας βλέπουμε την εικόνα αναζήτησης, μαζί με το σύνολο των συνήθη και προτεινόμενων tags. Στις πιό κάτω γραμμές, το VIRaL παρουσιάζει τις παρόμοιες εικόνες με φθίνουσα σειρά ομοιότητας. Η τιμή ομοιότητας που παρουσιάζεται είναι ο αριθμός των inliers, κανονικοποιημένος στο $[0, 1]$ με χρήση μιας σιγμοειδούς συνάρτησης. Στο παράδειγμα του Σχήματος 5.2, τα συνήθη tags είναι `terreiro do paço, praça do municipio, monument, stevie0020, arch`. Το τελικό σύνολο από προτεινόμενα tags ειναι `Praça do Comércio` και `Lisboa`. Και στιε δύο προτάσεις δίνονται και οι αντίστοιχοι σύνδεσμοι των άρθρων της Wikipedia. Οι προτάσεις είναι σωστές, όπως φαίνεται και από το περιεχόμενο των συνδεδεμένων άρθρων στο Σχήμα 5.3.

Μέσω της διαδικτυακής εφαρμογής ο χρήστης μπορεί επίσης να εκτελέσει αναζήτηση με μία δικιά του εικόνα από το δίσκο του ή δίνοντας ένα URL. Φυσικά, για να πάρει σωστά αποτελέσματα, η εικόνα πρέπει να έχει τραβηχτεί σε μία από τις υποστηριζόμενες πόλης της συλλογής του VIRaL. Έχουμε ρυθμίσει την εφαρμογή έτσι ώστε να παρέχει υψηλή ακρίβεια και να επιστρέφει όσο το δυνατόν λιγότερες λάθος παρόμοιες εικόνες. Για να ενισχύσουμε και την ανάκληση περισσότερο, έχουμε προσθέσει και μία μέθοδο επέκτασης ερωτήματος, τη μέθοδο που αναφέρεται ως QE1 στην ενότητα 4.5, η οποία παράγει ένα σύνολο εικόνων που είναι *όμοιες με τις παρόμοιες* (*Similar of Similar*). Το σύνολο αυτό κατασκευάζεται αμέσως κατά την ώρα της αναζήτησης καθώς οι παρόμοιες εικόνες για κάθε εικόνα της συλλογής έχουν συλλεχθεί και είναι γνωστές από πριν. Στο Σχήμα 5.5 φαίνεται αυτό το σύνολο για την αναζήτηση του Σχήματος 5.2.

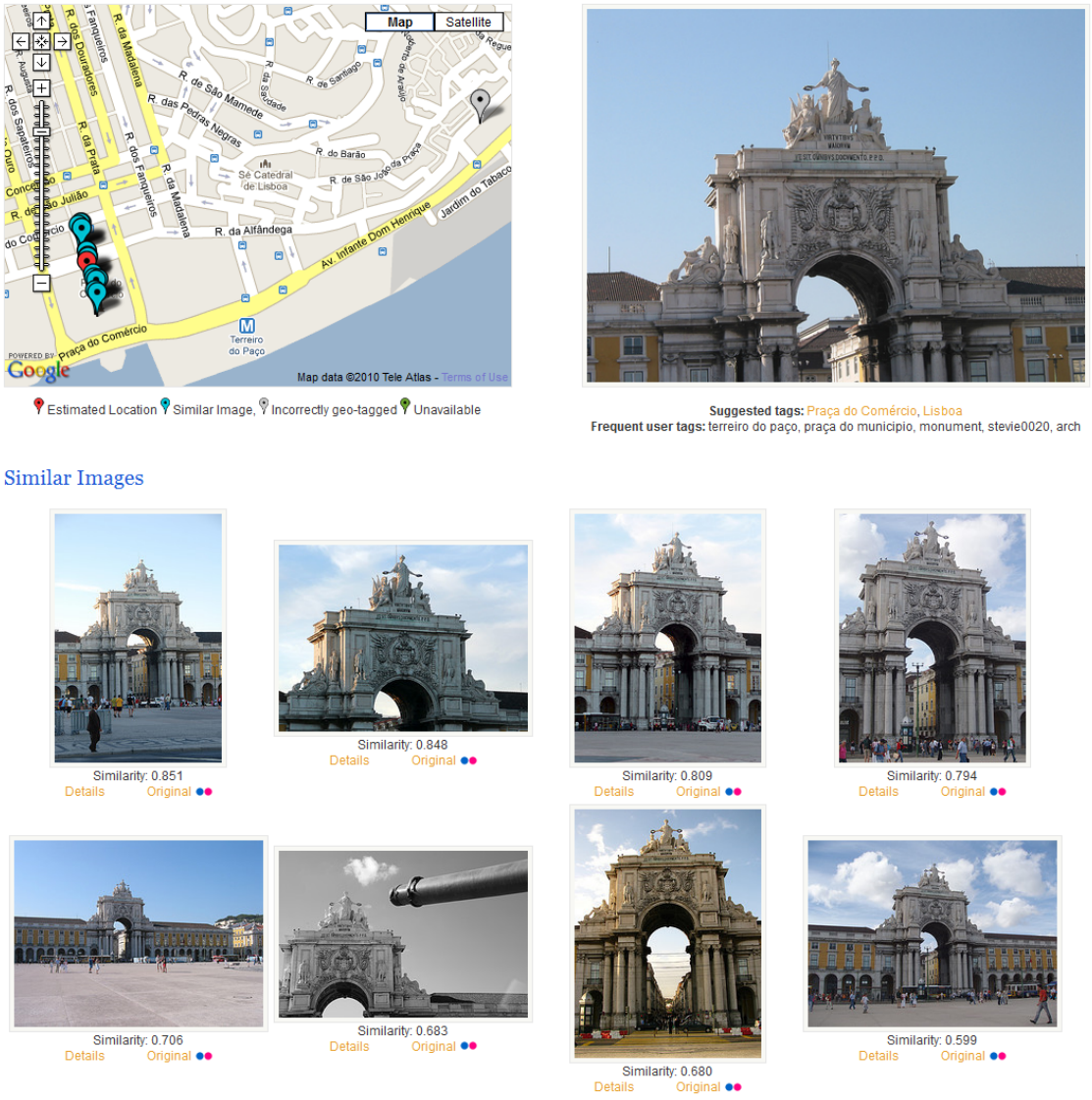
Μαζί με τις παρόμοιες εικόνες, είναι δυνατόν για τον χρήστη να δει οπτικοποιημένο και το κάθε αποτέλεσμα γεωμετρικού ταιριάσματος, κάνοντας κλικ στον σύνδεσμο *Details* που βρίσκεται κάτω από κάθε παρόμοια εικόνα. Όπως φείνεται στο Σχήμα 5.6, το τμήμα των εικόνων που περιέχει τα τοπικά χαρακτηριστικά που ταίριαξαν είναι μέσα στα μπλε ορθογώνια.

Στο Σχήμα 5.7 παρουσιάζεται μια περίπτωση όπου ένα κτίριο του Άμστερνταμ έχει σωστό τοποθετηθεί στο χάρτη από τις παρόμοιες εικόνες, αλλά δεν αντιστοιχεί σε κάποιο γνωστό ορόσημο ή σημείο ενδιαφέροντος. ΤΑ δύο προτεινόμενα tags είναι τώρα `Sint Antoniesbreestraat`, το όνομα της οδού, και `Zwanenburgwal`, το όνομα του καναλιού.

5.4.2 Οι εφαρμογές Viral Explore και Routes

Η μηχανή αναζήτησης του VIRaL πλαισιώνεται επίσης από δύο ακόμα εφαρμογές. Η εφαρμογή *VIRaL Explore*(Σχήμα 5.8) επιτρέπει την πλοήγηση σε ολόκληρη τη συλλογή φωτογραφιών του VIRaL, η οποία παρουσιάζεται στον παγκόσμιο χάρτη μέσω της εφαρμογής Google Maps[6]. Οι φωτογραφίες είναι ιεραρχικά

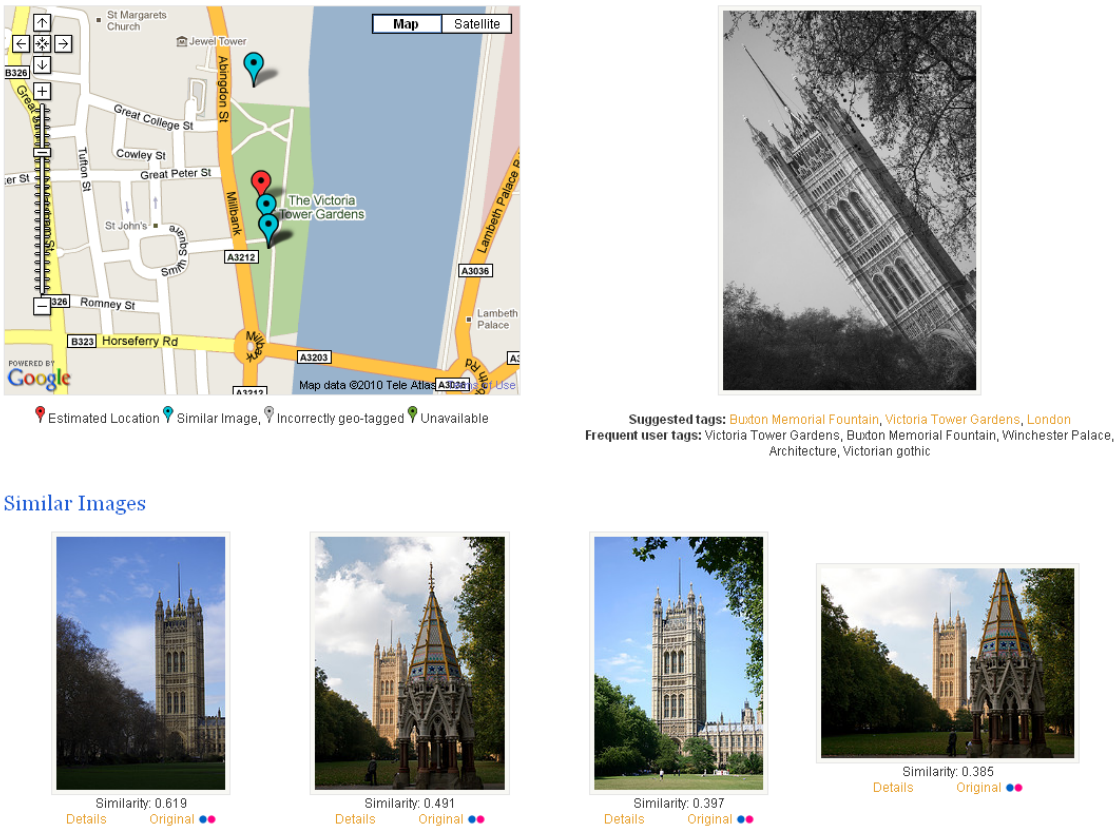
[6] `http://maps.google.com`



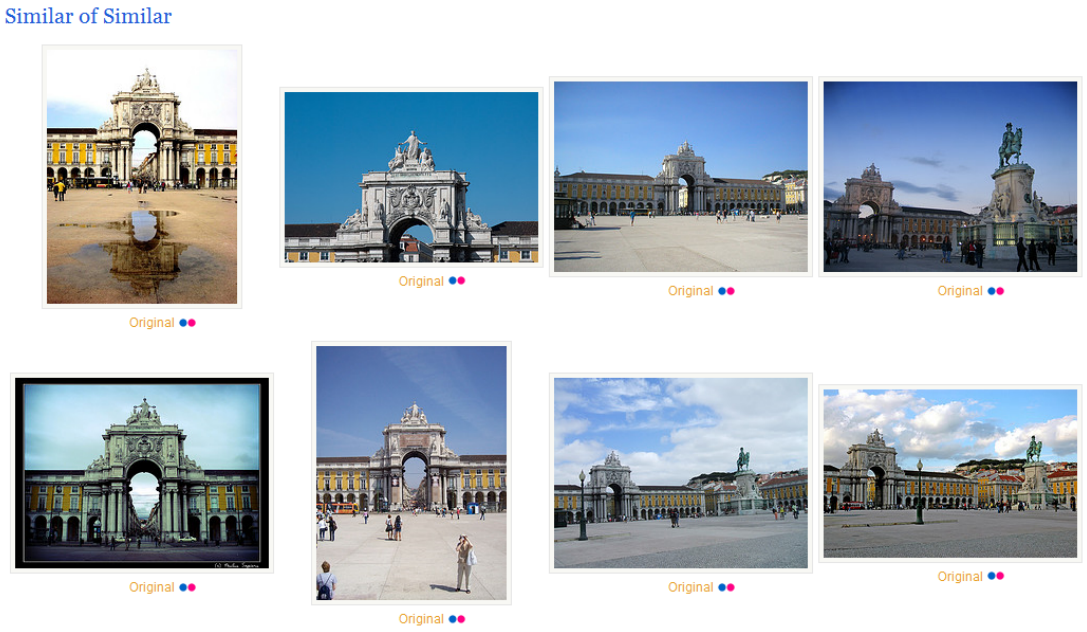
Σχήμα 5.2: *Αποτελέσματα μιας επιτυχημένης αναζήτησης. Πάνω αριστερά: ο χάρτης με μπλε markers για κάθε παρόμοια εικόνα και έναν κόκκινο marker για την εκτιμώμενη τοποθεσία. Πάνω δεξιά: η εικόνα αναζήτησης μαζί με τα συνήθη και προτεινόμενα tags. Κάτω γραμμές: οι οπτικά παρόμοιες εικόνες με φθίνουσα σειρά ομοιότητας.*



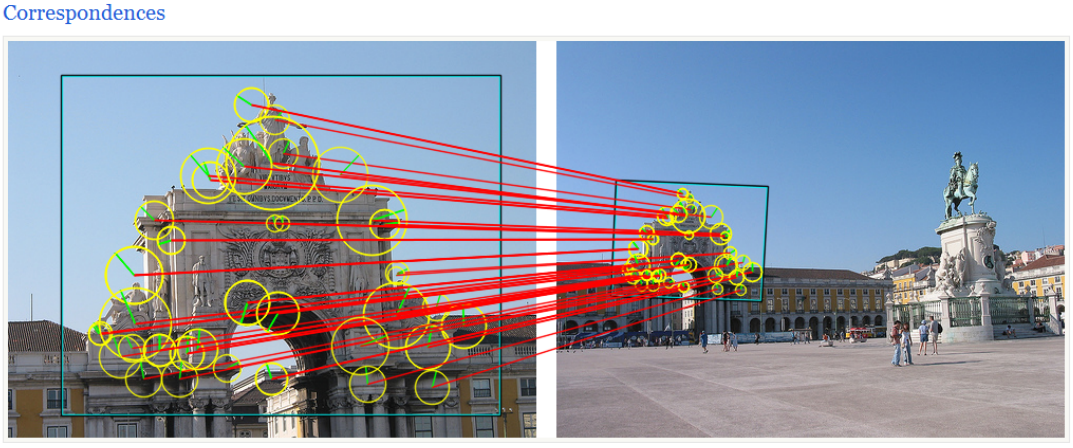
Σχήμα 5.3: *Το άρθρο της Wikipedia που προτείνεται για την εικόνα αναζήτησης του Σχήματος 5.2.*



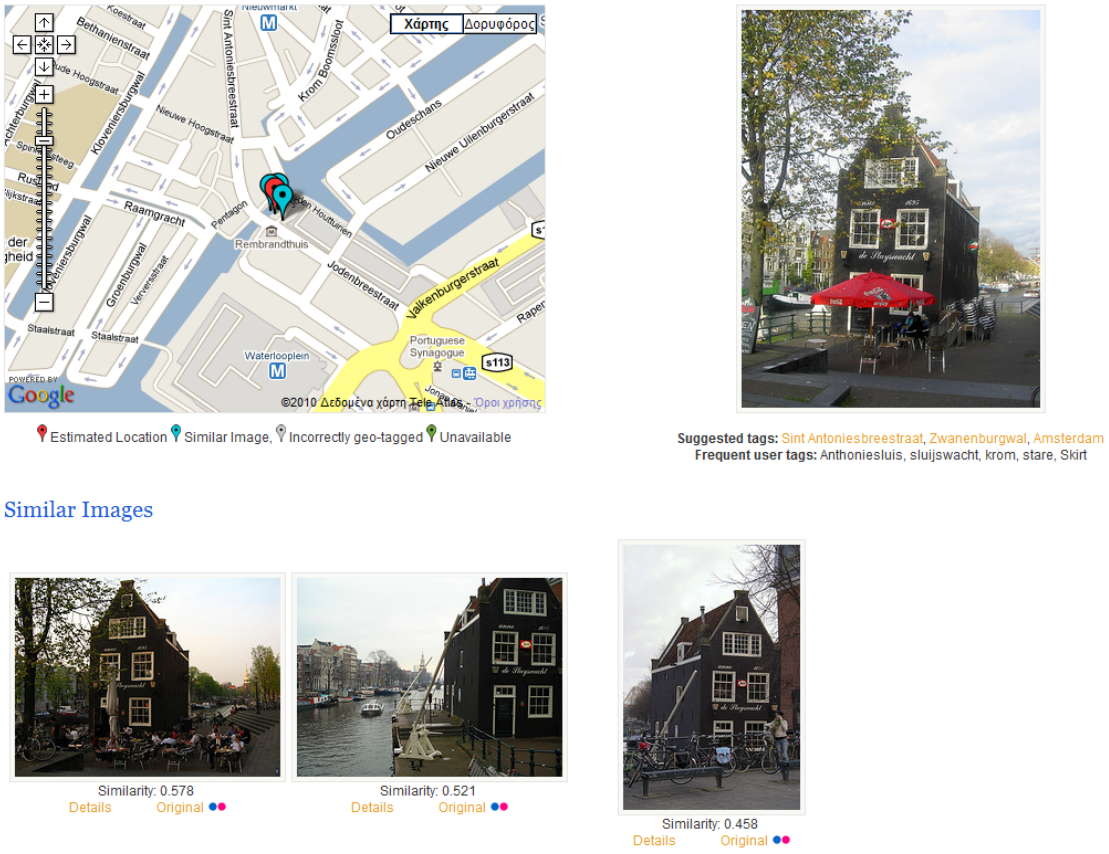
Σχήμα 5.4: *Άλλη μια σελίδα αποτελεσμάτων με επιτυχημένη αναγνώριση τοποθεσίας και αντικειμένων της εικόνας αναζήτησης που φαίνεται πάνω δεξιά.*



Σχήμα 5.5: *Το σύνολο των Similar of Similar εικόνων για την εικόνα αναζήτησης του Σχήματος 5.2.*



Σχήμα 5.6: *Αντιστοιχίες μεταξύ της εικόνας αναζήτησης (αριστερά) και μίας παρόμοιας εικόνας (δεξιά). Τα τοπικά χαρακτηριστικά που είναι inliers απεικονίζονται σαν κίτρινοι κύκλοι με κλίμακα και προσανατολισμό και οι αντιστοιχίες ως κόκκινες γραμμές. Τα μπλε ορθογώνια δείχνουν την κοινή περιοχή των δύο εικόνων.*



Σχήμα 5.7: *Εκτίμηση τοποθεσίας και αναγνώριση για μια εικόνα αναζήτησης που δεν περιέχει κάποιο ορόσημο.*

ομαδοποιημένες ανάλογα με την πόλη, την κλίμακα του χάρτη και την οπτική ομοιότητα, έτσι ώστε σε κάθε ομάδα όλες οι φωτογραφίες να απεικονίζουν το ίδιο κτίριο, σκηνή ή σημείο ενδιαφέροντος. Για την ακρίβεια, για την γεωγραφική και οπτική ομαδοποίηση των φωτογραφιών ακολουθείται η διαδικασία που περιγράφηκε αναλυτικά στην ενότητα 4.3.

Σε κάθε κλίμακα του χάρτη εμφανίζεται ένα εικονίδιο για κάθε ένα από τα πιο δημοφιλή σημεία της περιοχής. Επιλέγοντας ένα εικονίδιο, ο χρήστης μπορεί να δει τις σχετικές φωτογραφίες και να ξεκινήσει με αυτές αναζήτηση στο VIRaL. Επίσης τα εικονίδια που αντιστοιχούν σε αξιοθέατο ή σημείο ενδιαφέροντος συνοδεύονται με σύνδεσμο στη Wikipedia.

Η εφαρμογή *VIRaL Routes* (Σχήμα 5.9 παρέχει ένα νέο μοναδικό τρόπο πλοήγησης στις προσωπικές ταξιδιωτικές συλλογές φωτογραφιών. Ο χρήστης παρέχει τις φωτογραφίες του από κάποιο ταξίδι σε μια από τις πόλεις που υποστηρίζονται. Η εφαρμογή τις ομαδοποιεί σύμφωνα με την οπτική ομοιότητα και αναγνωρίζει τη γεωγραφική θέση ορισμένων από αυτές. Έπειτα, χρησιμοποιώντας τον χρόνο κατά τον οποίο τραβήχτηκαν οι φωτογραφίες, παρουσιάζει μια διαδρομή στον χάρτη, σημειώνοντας τα μέρη και τα αξιοθέατα που επισκέφθηκε ο χρήστης, σε ένα περιβάλλον παρόμοιο με του VIRaL Explore.



Σχήμα 5.8: *Η εφαρμογή VIRaL Explore.*

Στην παρούσα φάση, το VIRaL Routes παρουσιάζει μια συλλογή από ταξίδια σε 16 πόλεις, που προέρχονται από προσωπικές συλλογές των μελών της ομάδας μας.

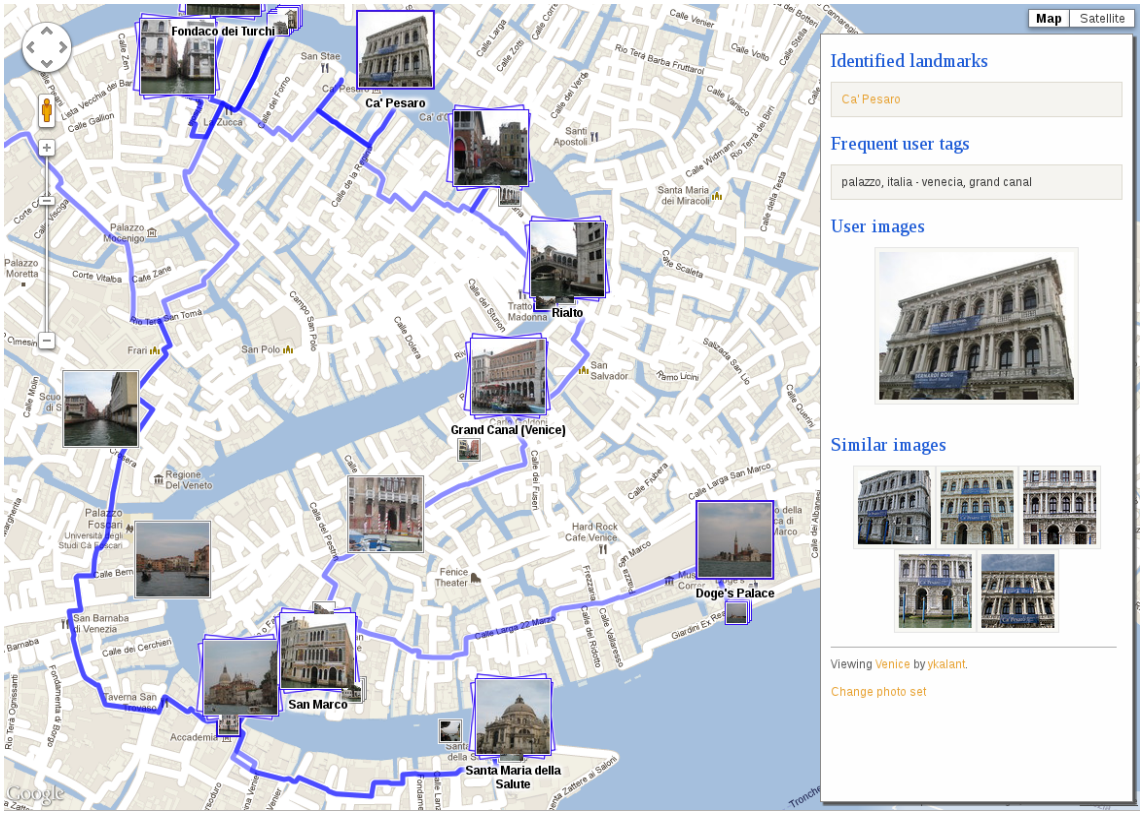
Το VIRaL είναι προσβάσιμο και από κινητές συσκευές μέσω της εφαρμογής *FindMyPhoto*[7] που διατίθεται για λειτουργικό σύστημα Android μέσω της πλατφόρμας Google Play [8]. Η εφαρμογή δεν απαιτεί GPS και μπορεί να αναγνωρίσει φωτογραφίες ακόμη και σε εσωτερικούς χώρους (π.χ. μουσεία). Η αναγνώριση πραγματοποιείται είτε στο σημείο λήψης εφόσον υπάρχει πρόσβαση σε δίκτυο, είτε εκ των υστέρων. Η γεωγραφική θέση, τα προτεινόμενα tags και τυχόν σημεία ενδιαφέροντος που αναγνωρίζονται μπορούν να αποθηκευθούν ως μεταδεδομένα στην ίδια τη φωτογραφία.

5.5 Πειράματα

5.5.1 Αξιολόγηση της εκτίμησης τοποθεσίας

Όλες οι εικόνες της βάσης *European Cities 1M* έχουν γνωστή τοποθεσία (geotag). Συνεπώς, με δεδομένο το αποτέλεσμα της αναζήτησης μπορεί να παραχθεί

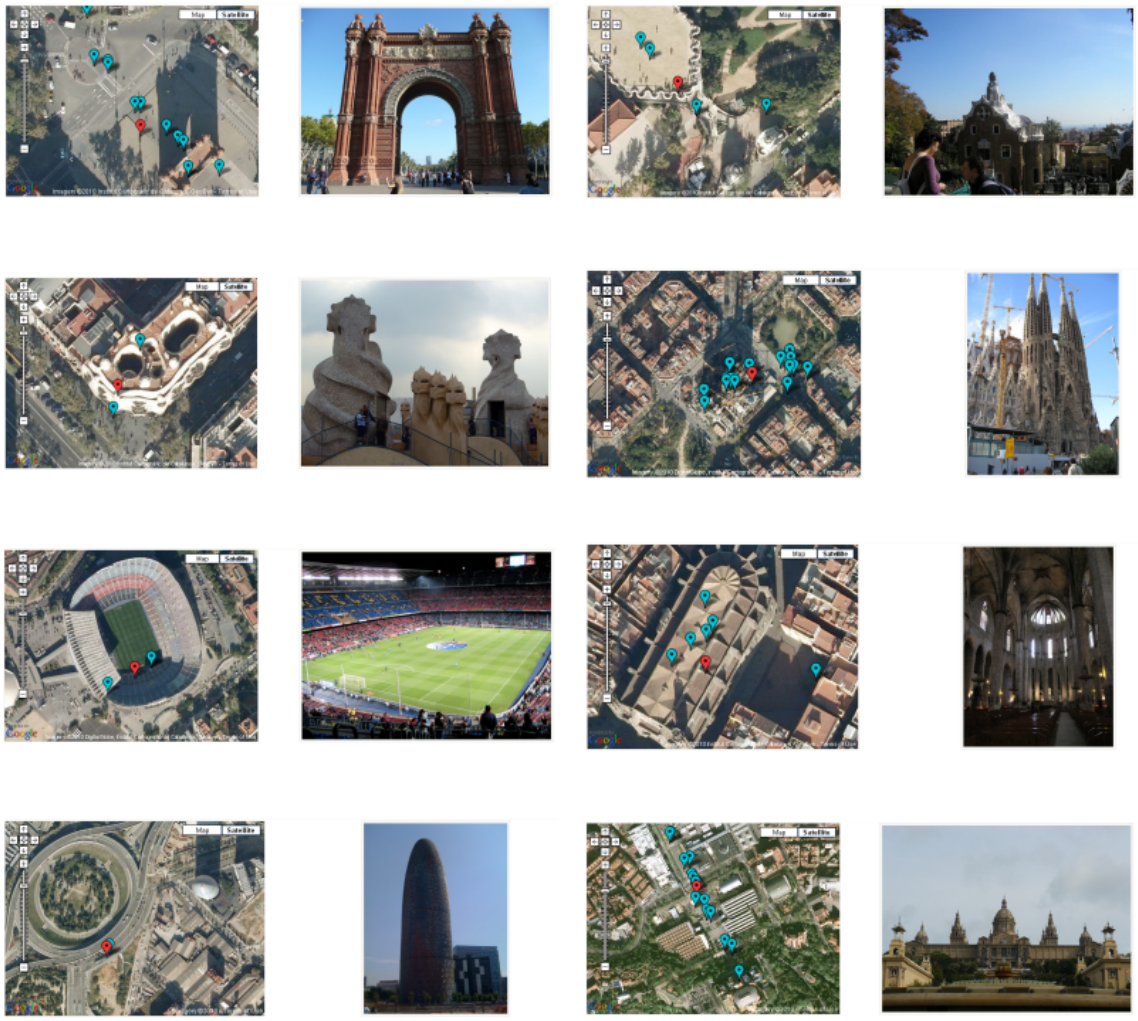
[7] `http://viral.image.ntua.gr/\~mobile`
[8] `https://play.google.com/`



Σχήμα 5.9: *Η εφαρμογή VIRaL Routes.*

Πίνακας 5.1: *Ποσοστά σωστού εντοπισμού για διάφορα κατώφλια απόστασης για τις τέσσερις μεθόδους.*

Μέθοδος	Κατώφλι απόστασης		
	$< 50m$	$< 100m$	$< 150m$
Baseline BoW	82.5%	91.6%	94.2%
QE1	86.3%	93.5%	96.2%
QE2	86.7%	93.3%	96.5%
Scene maps	87.8%	94.2%	97.1%



Σχήμα 5.10: *Δείγματα από εικόνες αναζήτησης και οι εκτιμώμενες τοποθεσίες τους στο χάρτη. Για κάθε ζεύγος φαίνεται ο χάρτης στα αριστερά και η εικόνα αναζήτησης δεξιά. Μπλε marker: Παρόμοια εικόνα. Κόκκινος marker: εκτίμηση τοποθεσίας.*

μια εκτίμηση της τοποθεσίας όπου τραβήχτηκε η εικόνα αναζήτησης σύμφωνα με τη μέθοδο που περιγράφεται στην ενότητα 5.3. Για την αξιολόγηση της εκτίμησης που παρέχουμε, τη συγκρίνουμε με τις γνωστές τοποθεσίες της κάθε επισημειωμένης ομάδας. Η ακρίβεια εντοπισμού συγκριτικά με τη βασική κα τις άλλες μεθόδους παρουσιάζεται στον Πίνακα 5.1. Όπως παρατηρείται από τα αποτελέσματα, η ακρίβεια εντοπισμού είναι υψηλή ακόμα και για τη βασική αναζήτηση. Και πάλι, όμως, η προτεινόμενη μέθοδος με χρήση χαρτών σκηνής αποδίδει τα υψηλότερα ποσοστά σωστού εντοπισμού.

Στο Σχήμα 5.10 παρουσιάζονται δείγματα από εκτιμήσεις εντοπισμού σε διάφορες εικόνες αναζήτησης. Στις πρώτες 6 περιπτώσεις επιτυγχάνεται σωστή εκτίμηση. Οι δύο τελευταίες περιπτώσεις είναι δείγματα μη σωστής εκτίμησης τοποθεσίας χρησιμοποιώντας ως «σωστή» τοποθεσία την επισημειωμένη, που είναι η ακριβής θέση του ορόσημου. Η τελική εκτίμηση είναι μακριά από την θέση του ορόσημου και η εκτίμηση μας επηρεάζεται από το γεγονός ότι τα geo-tag των εικόνων των χρηστών συνήθως εκφράζουν τη θέση από όπου τραβήχτηκε η φωτογραφία. Συνεπώς η εκτίμησης θέσης του ορόσημου είναι λάθος, αλλά η εκτίμηση της θέσης από όπου τραβήχτηκε η φωτογραφία είναι σωστή.

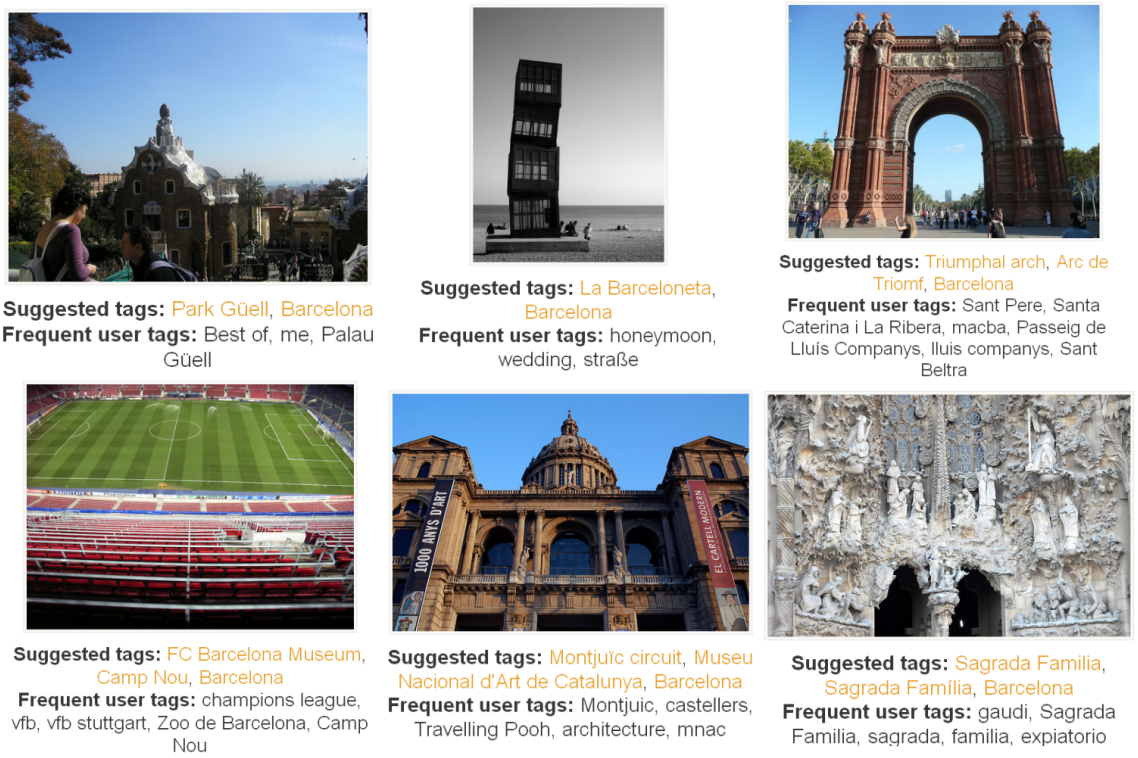
5.5.2 Αξιολόγηση της αναγνώρισης ορόσημων και σημείων ενδιαφέροντος

Καθώς πολλοί φωτογράφοι παίρνουν φωτογραφίες από γνωστά ορόσημα, μπορούμε να υποθέσουμε ότι μερικές από τις επισημειωμένες εικόνες απεικονίζουν γνωστά ορόσημα τα οποία έχουν και αντίστοιχες σελίδες στη Wikipedia. Με δεδομένα λοιπόν τα μεταδεδομένα των εικόνων της βάσης European Cities 1M και για την ακρίβεια τους τίτλους και τα tags των εικόνων, μπορούμε να εφαρμόσουμε τη μέθοδο που προτείνεται στην ενότητα 5.3 για την ανάλυση τους και την αναγνώριση των ορόσημων που απεικονίζουν.

Η απόδοση της προτεινόμενης μεθόδου παρουσιάζεται στον Πίνακα 5.2, όπου βλέπουμε τα ποσοστά των συνδέσμων στη Wikipedia που προτάθηκαν σωστά. Τα πειράματα εκτελούνται για 17 από τις επισημειωμένες ομάδες, για κάθε μία από τις οποίες ξέρουμε το ορόσημο που απεικονίζεται και τον σύνδεσμο στο αντίστοιχο άρθρο τη Wikipedia. Θεωρούμε έναν σύνδεσμο σωστό, αν συμπίπτει με τον επισημειωμένο σύνδεσμος. Όπως φαίνεται από τον πίνακα, η αναγνώριση είναι αποδοτική τόσο και με χρήση των χαρτών σκηνής όσο και με την χρήση επέκτασης ερωτήματος. Δείγματα εικόνων αναζήτησης μαζί με τα συνήθη και τα προτεινόμενα tags τους παρουσιάζονται στο Σχήμα 5.11. Σε όλες τις περιπτώσεις που παρουσιάζονται η αναγνώριση είναι σωστή.

Πίνακας 5.2: *Ποσοστό σωστών προτάσεων άρθρων της Wikipedia για κάθε ορόσημο και συνολικός μέσος όρος για τις τέσσερις μεθόδους.*

Ορόσημο	Μέθοδος			
	Baseline	QE1	QE2	Scene maps
La Pedrera(a)	100%	100%	100%	100%
Park Guell(a)	100%	100%	100%	100%
Museu Nat. d' Art	40%	100%	60%	80%
Columbus Monument	100%	100%	100%	100%
Carrer del Bisbe Irurit-El Gotic	100%	100%	100%	100%
Port Vell	80%	100%	80%	100%
Sagrada Familia(b)	100%	100%	100%	100%
Casa Batllo	100%	100%	100%	100%
Arc de Triomf	100%	100%	100%	100%
La Pedrera(b)	60%	100%	80%	80%
Hotel Arts	40%	40%	40%	60%
Hospital de Sant Pau(a)	100%	100%	100%	100%
Hospital de Sant Pau(b)	80%	80%	80%	100%
Park Guell(b)	100%	100%	100%	100%
Torre Agbar	100%	100%	100%	100%
Placa de Catalunya	100%	100%	100%	100%
Cathedral (side)	80%	80%	80%	80%
Average	**87%**	**95%**	**90%**	**95%**



Σχήμα 5.11: *Δείγματα εικόνων αναζήτησης μαζί με τα συνήθη και τα προτεινόμενα tags. Τα ορόσημα αναγνωρίζονται επιτυχώς και σε κάθε περίπτωση παρέχεται και σύνδεσμος για το αντίστοιχο άρθρο της Wikipedia.*

Κεφάλαιο 6

Συμπληρωματικές εργασίες

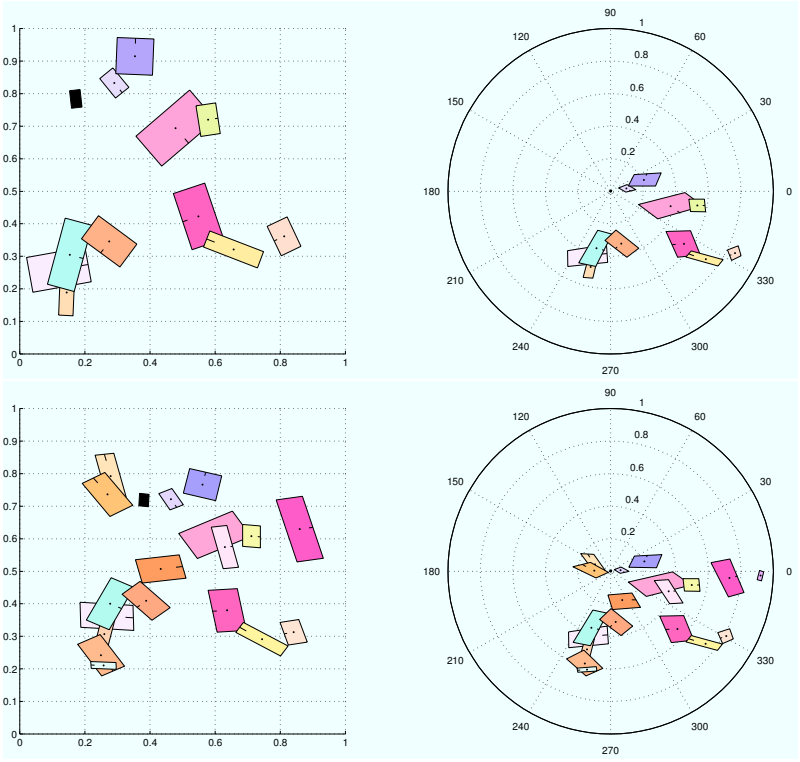
Στο κεφάλαιο αυτό αναφέρουμε συνοπτικά επιπλέον δημοσιεύσεις και εφαρμογές που αναπτύχθηκαν κατά τη διάρκεια της διατριβής αλλά δεν περιγράφηκαν αναλυτικά στα προηγούμενα κεφάλαια. Οι μέθοδοι που προτείνονται περιλαμβάνουν εφαρμογές της αναζήτησης μεγάλης κλίμακας, πέρα από την αναγνώριση σκηνών. Ειδικές τέτοιες περιπτώσεις είναι η μεγάλης κλίμακας αναζήτηση και αναγνώριση λογότυπων και η αυτόματη αναγνώριση και αναζήτηση προϊόντων ρουχισμού. Κάθε μια από αυτές έχει τις ιδιαιτερότητές της. Θα παρουσιαστεί επιπλέον συνοπτικά μια ιδιαίτερα ενδιαφέρουσα προσέγγιση για την εισαγωγή ολικής γεωμετρίας τοπικών σημείων στη δομή δεικτοδότησης μέσω χαρτών χαρακτηριστικών.

6.1 Χάρτες Χαρακτηριστικών

Η έλευση των τοπικών χαρακτηριστικών με διακριτική ικανότητα, λόγω της εξαγωγής τοπικών περιγραφέων [68, 10], έχει κάνει δυνατή την δεικτοδότηση της εμφάνισης των χαρακτηριστικών για μεγάλες συλλογές εικόνων και σε μικρούς χρόνους εκτέλεσης ενός ερωτήματος. Αντιθέτως, λιγότερα είναι τα επιτεύγματα στην δεικτοδότηση της γεωμετρικής πληροφορίας, καθώς συνήθως μία προτιμότερη λύση είναι η ανακατάταξη των εικόνων σε ένα δεύτερο στάδιο το οποίο ως πιο κοστοβόρο εφαρμόζεται μόνο στις πιο όμοιες εικόνες. Όμως η χρήση γεωμετρικής πληροφορίας φαίνεται απαραίτητη.

Η εκμετάλλευση του τοπικού σχήματος των ανιχνευθέντων περιοχών [10, 68, 6, 109], *e.g*. κλίμακα, περιστροφή, αφινικό σχήμα, μπορεί να οδηγήσει σε λύσεις οι οποίες βασίζονται σε μονές αντιστοιχίες μεταξύ χαρακτηριστικών για την εκτίμηση γεωμετρικών μετασχηματισμών ανάμεσα σε ζευγάρια εικόνων ή και μεταξύ χαρακτηριστικών της ίδιας εικόνας.

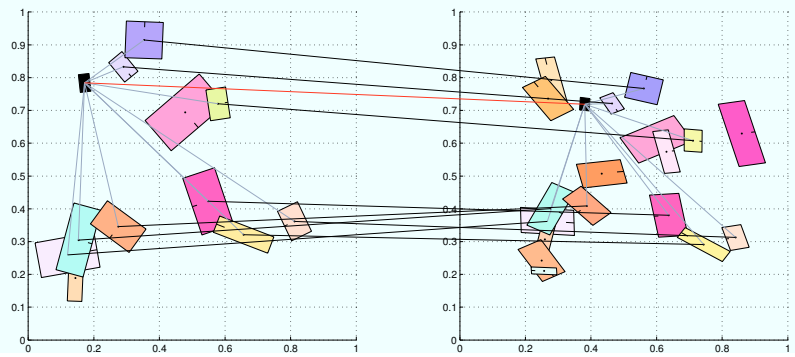
Το σημείο εκκίνησης για αυτή την δουλεία είναι το [85] όπου το γεωμετρικό ταίριασμα γίνεται με μία ειδική περίπτωση του αλγορίθμου RANSAC [26]. Το σχήμα



Σχήμα 6.1: *Πάνω αριστερά: Ένα τυχαίο σύνολο περιοχών. Κάτω αριστερά: Το ίδιο σύνολο μετασχηματισμένο αφινικά, όπου οι θέσεις των περιοχών και το σχήμα τους έχουν διαστρεβλωθεί, και έχουν εισαχθεί νέες περιοχές. Δεξιά: Τα αντίστοιχα αναμορφωμένα σύνολα. Οι πηγές είναι οι δύο μαύρες περιοχές στα αριστερά.*

των περιοχών σε αντιστοιχία χρησιμοποιείται για να δημιουργήσει υποθέσεις από μονές αντιστοιχίες σε αντίθεση με τις δυάδες, τριάδες ή τετράδες αντιστοιχιών που χρειάζεται ο παραδοσιακός αλγόριθμος, ανάλογα με τον μοντέλο μετασχηματισμού σε κάθε περίπτωση. Αυτή η ιδέα προέρχεται και από το [68] κι έχει διερευνηθεί περισσότερο , *e.g*. από τους Köser *et al*. [56]. Εμείς θα κανονικοποιήσουμε εκ των προτέρων τις θέσεις των χαρακτηριστικών ως προς το τοπικό σχήμα όλων των υπολοίπων. Χρησιμοποιούμε αυτή την πληροφορία σε μία κατάλληλη δομή δεικτοδότησης ώστε να είναι διαθέσιμη για γρήγορο ταίριασμα κατά την στιγμή του ερωτήματος. Ονομάζουμε την αναπαράσταση που δημιουργούμε *χάρτη χαρακτηριστικών* και μπορεί κανείς να την δει ως έναν τοπικό περιγραφέα ο οποίος περιγράφει ολικά το σύνολο χαρακτηριστικών σε ένα τοπικό σύστημα αναφοράς. Το ταίριασμα μεταξύ τέτοιων αναπαραστάσεων τελικά ανάγεται σε απλό εσωτερικό γινόμενο.

Οπτικά, ένας χάρτης χαρακτηριστικών μπορεί να κατανοηθεί ως η καταχώρηση των αναμορφωμένων χαρακτηριστικών σε χωρικά κυτία, όπως δεξιά στο Σχήμα 6.1. Υπάρχει ένας διαφορετικός χάρτης για κάθε πηγή: μπορούμε να φανταστούμε τον χάρτη της κάθε πηγής σαν ένα *τοπικό* περιγραφέα, ο οποίος κωδικοποιεί *ολικά* το σύνολο χαρακτηριστικών αναμορφωμένο σε ένα *τοπικό* σύστημα συντεταγμέ-



Σχήμα 6.2: *Inliers μεταξύ δύο συνόλων τοπικών χαρακτηριστικών. Ο κάθε ένας αντιστοιχεί σε ένα μη μηδενικό όρο του εσωτερικού γινομένου των αντίστοιχων χαρτών χαρακτηριστικών. Οι μαύρες γραμμές ενώνουν τους inliers. Οι κόκκινες γραμμές ενώνουν τις πηγές. Οι γκρίζες γραμμές ενώνουν τις πηγές με τους inliers.*

νων. Καλά ευθυγραμμισμένα σύνολα περιοχών είναι πιθανόν να έχουν χάρτες με μεγάλο βαθμό επικάλυψης.

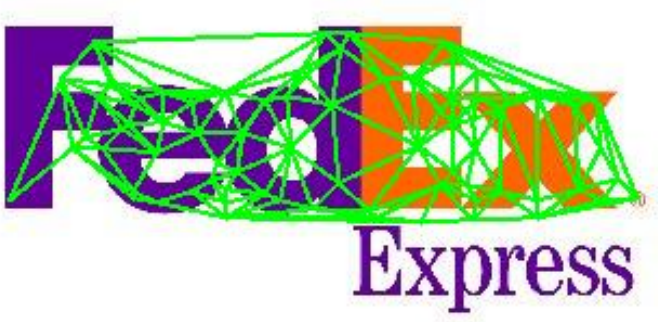
Επιστρέφοντας στο παράδειγμα του Σχήματος 6.1, οι inliers των δύο αναμορφωμένων συνόλων είναι εκείνα τα χαρακτηριστικά τα οποία πέφτουν στο ίδιο κυτίο του συνδυαστικού ιστογράμματος. Αυτοί οι inliers φαίνονται σαν αντιστοιχίες με μαύρο χρώμα στο Σχήμα 6.2, το οποίο ουσιαστικά απεικονίζει όλες τις αντιστοιχίες για τους χάρτες που έχουν σαν πηγές τα δύο μαύρα χαρακτηριστικά.

Χρησιμοποιούμε την μέθοδο των τυχαίων αντιμεταθέσεων για να καταλήξουμε σε μία πιο συμπαγή αναπαράσταση αλλά και προτείνουμε έναν τρόπο για επιλογή χαρακτηριστικών χωρίς επίβλεψη ο οποίος θα μας επιτρέψει να εφαρμόσουμε την μέθοδο των χαρτών σε πολύ μεγαλύτερη κλίμακα. Οι παραπάνω ιδέες περιγράφονται αναλυτικά στις δημοσιεύσεις [7, 105].

6.2 Ανίχνευση λογότυπων

Τα λογότυπα καταλαμβάνουν συνήθως ένα μικρό τμήμα της εικόνας και αντιπροσωπεύονται από ένα πολύ μικρο ποσοστό των τοπικών χαρακτηριστικών μιας εικόνας αναζήτησης. Η τοπική γεωμετρία των χαρακτηριστικών αυτών καθίσταται συνεπώς ιδιαίτερα σημαντική. Πειραματιστήκαμε με τριγωνοποιήσεις των θέσεων των χαρακτηριστικών στην εικόνα σε πολλαπλές κλίμακες, δεικτοδοτώντας όχι σημεία, αλλά *τρίγωνα* σημείων σε πίνακες κατακερματισμού. Καταφέρνουμε έτσι να έχουμε μικρό αποτύπωμα στη μνήμη για κάθε λογότυπο σε μια γρήγορη δομή δεικτοδότησης και έτσι μπορούμε να αποφανθούμε για την ύπαρξη των χιλιάδων διαφορετικών λογότυπων της βάσης μας στην εικόνα αναζήτησης αποτελεσματικά, σε πραγματικό χρόνο.

Μια τριγωνοποίηση Delaunay όλων των τοπικών χαρακτηριστικών από μια



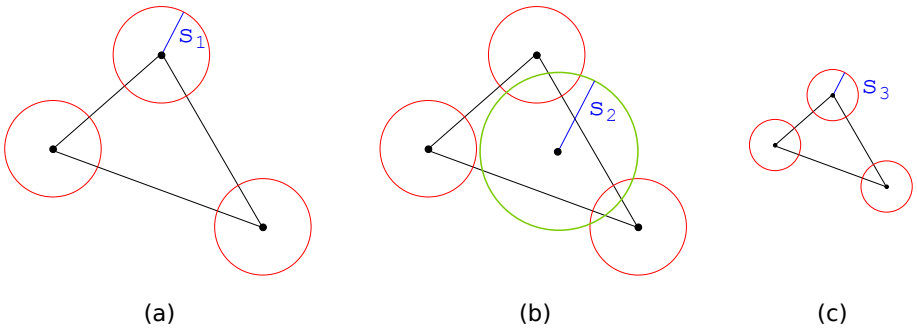
Σχήμα 6.3: *Τριγωνοποίηση Delaunay στο σύνολο όλων των τοπικών χαρακτηριστικών που εξήχθησαν από ένα λογότυπο της εταιρίας FedEx.*

εικόνα λογότυπου της εταιρίας FedEx παρουσιάζονται στο σχήμα 6.3.

Επιλέγουμε να εξάγουμε τριγωνοποιήσεις σε πολλαπλές κλίμακες. Για να γίνει πιο κατανοητό το πως οι πολλαπλές τριγωνοποιήσεις σε πολλαπλές κλίμακες αυξάνουν την ευρωστία του ταιριάσματος δύο τριγώνων, ας θεωρήσουμε τα τρίγωνα του Σχήματος 6.4(a) όπου οι κόμβοι είναι τοπικά χαρακτηριστικά εξαγμένα στην κλίμακα s_1. Στο Σχήμα 6.4(b), προσθέτουμε ένα σημείο περίσπασης σε κλίμακα s_2. Στην προσέγγισή μας, το επιπλέον σημείο θα επιρρεάσει την τριγωνοποίηση μόνο αν οι δύο κλίμακες είναι κοντά, δηλαδή μόνο αν η διαφορά $|s_1 - s_2|$ είναι μικρή. Το Σχήμα 6.5 δείχνει τις τριγωνοποιήσεις σε πολλαπλές κλίμακας των τοπικών χαρακτηριστικών για το λογότυπο του Σχήματος 6.3.

Για τη δεικτοδότηση χρησιμοποιούμε τις οπτικές λέξεις των σημείων κάθε τριγώνου για να δεικτοδοτούμε τρίγωνα αντί για τοπικά χαρακτηριστικά. Το *χωρικό ταίριασμα* και η αναζήτηση μπορεί να γίνει πλέον όπως και στο μοντέλο bag-of-words, όπου όμως τώρα έχουμε τρίγωνα αντί για σημεία. Τα τρίγωνα που ταιριάζουν μεταξύ δύο παρόμοιων λογότυπων φαίνονται στο Σχήμα 6.6.

Με την προτεινόμενη μέθοδο, καταφέρνουμε να αναγνωρίσουμε σε ελάχιστο χρόνο, χιλιάδες λογότυπα σε μια εικόνα αναζήτησης, ξεπερνώντας σε απόδοση τη μέθοδο bag-of-words.



Σχήμα 6.4: *(a) Ένα τρίγωνο με όλα τα σημεία των γωνιών του να έχουν εξαχθεί από σε παρόμοια κλίμακα, έστω s_1. (b) Έστω ότι προστίθεται ένα επιπλέον σημείο εξαγμένο σε κλίμακα s_2. Το σημείο αυτό θα επιρρεάσει την τριγωνοποίηση μόνο αν $|s_1 - s_2| < w$, όπου το w είναι μια παράμετρος κλίμακας. (c) Ένα τρίγωνο με όλα τα σημεία των γωνιών του να έχουν εξαχθεί από σε παρόμοια κλίμακα, έστω s_3. Το τρίγωνο αυτό θα ταιριάξει με το τρίγωνο του Σχήματος (a) αν όλες οι τρείς οπτικές λέξεις των σημείων τους είναι ίδιες.*



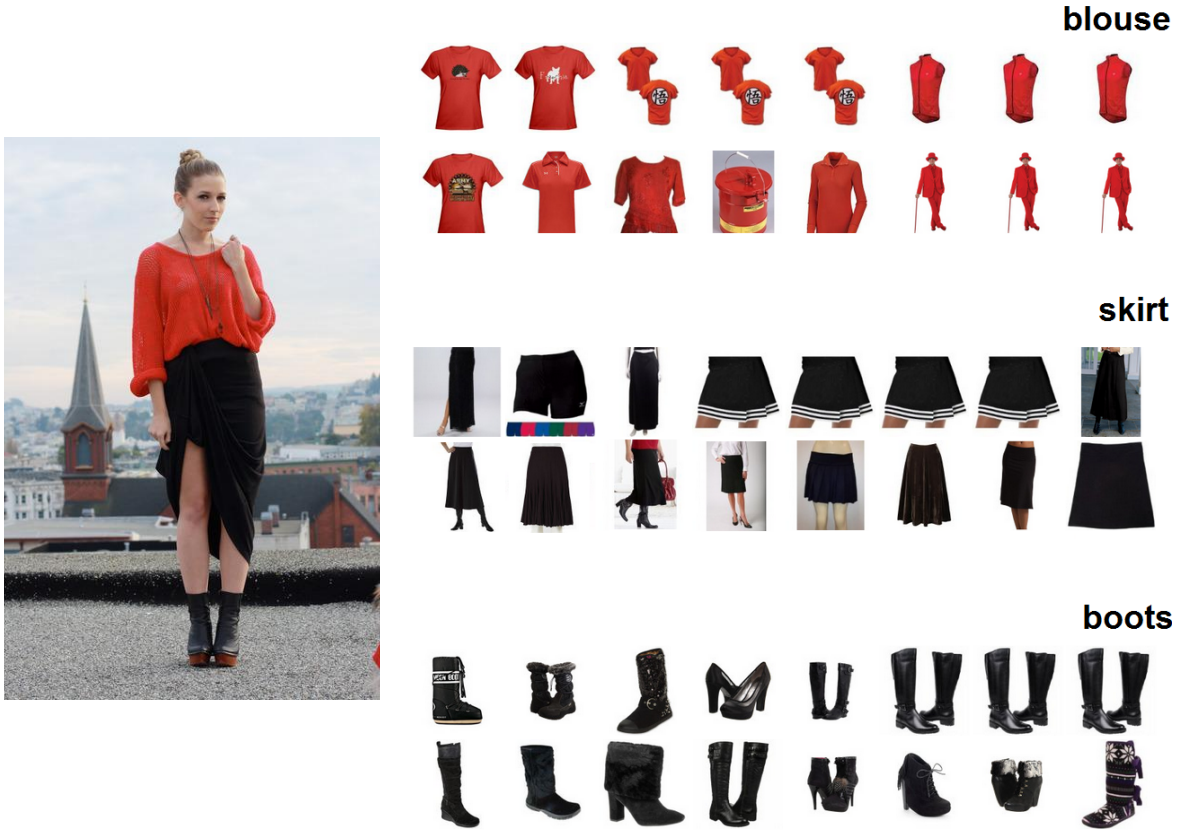
Σχήμα 6.5: *Τριγωνοποιήσεις τοπικών χαρακτηριστικών σε διάφορες κλίμακες.*



Σχήμα 6.6: *Τρίγωνα που έχουν ταιριάξει μεταξύ δύο παρόμοιων λογότυπων.*

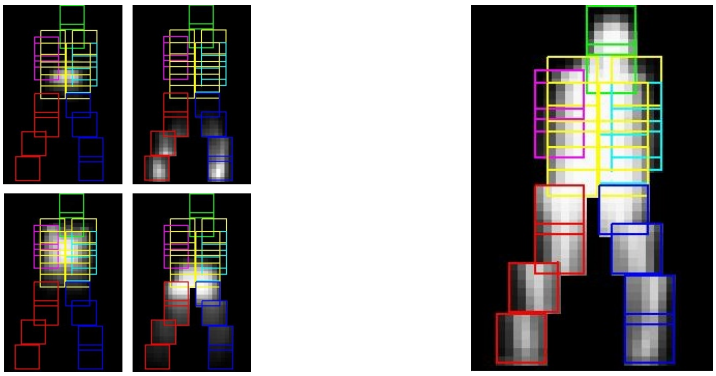
6.3 Αναγνώριση, κατάτμηση και μεγάλης κλίμακας αναζήτηση προιόντων ρουχισμού

Μια εμπορικά σημαντική εφαρμογή της οπτικής αναζήτησης είναι η αναζήτηση προιόντων. Εστιάσαμε στην αναζήτηση σχετικών προϊόντων ρουχισμού, ένα πεδίο αναζήτησης στο οποίο παρατηρούνται διάφορες ιδιαιτερότητες: Η εικόνα αναζήτησης ιδανικά είναι μια καθημερινή είκόνα, που απεικονίζει τα προιόντα σε χρήση/φορεμένα, ενώ οι εικόνες της βάσης τα αποικονίζουν απλωμένα, σε λευκό συνήθως φόντο. Το παράδειγμα του Σχήματος 6.7 παρουσιάζει ένα βασικό παράδειγμα χρήσης της ζητούμενης εφαρμογής αναζήτησης. Η εικόνα αναζήτησης φαίνεται στα αριστερά, ενώ οι εικόνες των προτεινόμενων με τον αλγόριθμό μας προϊόντων στα δεξιά.



Σχήμα 6.7: *Αριστερά: Η εικόνα αναζήτησης, μια καθημερινή εικόνα που παρουσιάζει το στύλ ρουχισμού του ατόμου που θέλουμε να "αντιγράψουμε". Δεξιά: Προτάσεις προϊόντων ρουχισμού, βασισμένες σε οπτική ανάλυση και αναγνώριση κλάσεων ρουχισμού.*

Καθώς για τα προϊόντα ξέρουμε την κλάση ρουχισμού που ανήκουν, προτείνουμε να εκτελείται η οπτική αναζήτηση *ανα κλάση ρουχισμού* για καλύτερη ακρίβεια. Στην εικόνα αναζήτησης, δηλαδή, αρχικά αναγνωρίζουμε τη σιλουέτα/στάση του ανθρώπου και εκτιμούμε τις κλάσεις ρουχισμού που εμφανίζονται. Έπειτα για κάθε κλάση εκτελούμε οπτική αναζήτηση ξεχωριστά, χρησιμοποιόντας το χρώμα



Σχήμα 6.8: *Χωρικοί χάρτες πιθανότητας των κλάσεων, κβαντισμένοι και κανονικοποιημένοι σε μια γενική πόζα. Για την εκμάθησή τους χρησιμοποιήθηκε ένα σχετικό επισημειωμένο σύνολο εικόνων. Αριστερά: Χάρτες για τις κλάσεις ζώνη, μπότες, μπλούζα και φούστα (από αριστερά προς τα δεξιά και από πάνω προς τα κάτω). Δεξιά: Ο συνολικός χάρτης* πιθανότητας εμφάνισης ρουχισμού *όπως προέκυψε μετά την ένωση όλων των επιμέρους χαρτών.*



Σχήμα 6.9: *Αριστερά: Η εικόνα αναζήτησης μετά την εξαγωγή της πόζας. Μέση: Η υπερκατάτμηση της πόζας. Δεξιά: Ομαδοποίηση με οπτικά κριτήρια. Οι περιοχές αυτές είναι οι υποψήφιες για την ανίχνευση κλάσεων ρουχισμού.*

και την υφή των περιοχών.

Δεδομένης μιας πόζας στην εικόνα αναζήτησης, προτείνουμε μια νέα μέθοδο για την αναγνώριση των κλάσεων ρουχισμού, εκπαιδεύοντας χωρικά μοντέλα ανα κλάση χρησιμοποιώντας τη γνώση από μια σχετική επισημειωμένη συλλογή. Μερικά από τα μοντέλα παρουσιάζονται στο Σχήμα 6.8. Για να αναγνωρίσουμε τα μοντέλα, αρχικά εκτελούμε υπερ-κατάτμηση της πόζας και έπειτα ομαδοποιούμε οπτικά τα κομμάτια χρησιμοποιόντας τον αλγόριθμο που παρουσιάσαμε στο Κεφάλαιο 2. Η διαδικασία φαίνεται στο Σχήμα 6.9. Τα αποτελέσματα της εφαρμογής αυτής, όπως επίσης και συγκριτικές μετρήσεις μπορούν να βρεθούν στη δημοσίευση [50].

Κεφάλαιο 7

Συμπεράσματα

Στην παρούσα εργασία παρουσιάστηκε η συνεισφορά μας σε δύο υποπεριοχές της αναζήτησης εικόνων, την κατασκευή οπτικών λεξικών και την αναζήτηση σε μεγάλες κλίμακες μέσω μιας νέας συμπαγούς αναπαράστασης σκηνών, καθώς και στη γενικότερη περιοχή της αναζήτησης κοντινότερου γείτονα σε πολύ μεγάλη κλίμακα. Προτείνουμε μεθόδους οι οποίες ξεπερνάνε το state-of-the-art της περιοχής τους, δίνοντας σε πολλές περιπτώσεις τα καλύτερα δημοσιευμένα αποτελέσματα για μεγάλα σύνολα εικόνων. Η διατριβή εστίασε σε εφαρμογές πολύ μεγάλης κλίμακας, εφαρμογές ιδιαίτερης και όλο και περισσότερο αυξανόμενης σημασίας στον ψηφιακό κόσμο. Ιδιαίτερα σημαντικά θεωρούμε τα αποτελέσματα αναζήτησης κοντινότερων γειτόνων σε μια δημόσια συλλογή *ενός δισεκατομμυρίου* πολυδιάστατων σημείων, την μεγαλύτερη διαθέσιμη συλλογή σημείων σήμερα.

Όσων αφορά τα οπτικά λεξικά, προτείνουμε μια νέα μέθοδο ομαδοποίησης που συνδυάζει την περιγραφική δύναμη των μοντέλων μείγματος κανονικών κατανομών με τις ιδιότητες που απαιτούνται κατά την κατασκευή μεγάλης κλίμακας οπτικών λεξικών για αναζήτηση εικόνων. Η ομαδοποίηση 10^7 διανυσμάτων σε 10^6 ομάδες σε έναν χώρο 10^2 διαστάσεων είναι αποδεδειγμένα ένα δύσκολο πρόβλημα. Η ανάθεση μίας ή και περισσότερων οπτικών λέξεων στους 10^3 περιγραφείς για κάθε μία από τις 10^6 εικόνες ξανά και ξανά για διάφορες παραμέτρους και ανταγωνιστικές μεθόδους σε ένα μηχάνημα αποδείχθηκε μεγαλύτερο πρόβλημα. Καταφέραμε, παρόλα αυτά να ρυθμίσουμε τις παραμέτρους του αλγορίθμου μας για οπτικά λεξικά σε μία μικρή βάση εικόνων και να έχουμε ανταγωνιστική απόδοση σε μια άλλης τάξης μεγέθους βάση, με παραμέτρους που δουλεύουν ακόμα και στο πρώτο συνθετικό δισδιάστατο παράδειγμα. Για όλες τις εναλλακτικές τεχνικές, απαιτείται η ρύθμιση τουλάχιστον του μεγέθους του λεξικού.

Ακόμα και με τη χρήση σφαιρικών συνιστωσών, η πρόσθετη περιγραφικότητα των μειγμάτων κανονικών κατανομών δείχνει να αναπαριστά καλύτερα την υποκείμενη πληροφορία και να βελτιώνει την απόδοση στην αναζήτηση εικόνων. Ταυτόχρονα, η διαδικασία εκπαίδευσης είναι εξίσου γρήγορη με τον προσεγγιστικό αλ-

γόριθμο k-means, και ως προς τον αριθμό των επαναλήψεων αλλά και ως προς την πολυπλοκότητα ανά επανάληψη.

Η λύση που προτείνουμε φαίνεται να μη δημιουργεί απροσδιοριστίες ούτε αλληλεπικαλυπτώμενες συνιστώσες, δύο ιδιαίτερα αρνητικές συνέπειες της εκτίμησης μέγιστης πιθανοφάνειας σε μοντέλα μειγμάτων κανονικών κατανομών. Θα ήταν όμως ενδιαφέρον μία προσέγγιση με *μεταβολικές* (*variational*) μεθόδους [11] και να μελετηθεί η συμπεριφορά της διαγραφής και της επέκτασης συνιστωσών με διαφορετικές εκ των προτέρων κατανομές (priors). Μια άλλη κατεύθυνση θα μπορούσε να είναι η αναζήτηση οπτικών συνωνύμων, ώστε τελικά να έχουμε ομάδες αυθαίρετου σχήματος χωρίς μεγάλο πλήθος παραμέτρων, είτε με είτε χωρίς δεδομένα εκμάθησης όπως στη δημοσίευση [73].

Όσων αφορά την προσεγγιστική αναζήτηση κοντινότερου γείτονα, στον πυρήνα της προτεινόμενης μεθόδου LOPQ υπάρχει η ιδέα ότι κανένα κέντρο κβαντισμού δεν πρέπει να μην καλύπτει δεδομένα, αλλά πρέπει να συνεισφέρει στην συνολική μείωση της παραμόρφωσης. Μοιραία καταλήγουμε στο ότι πρέπει η βελτιστοποίηση να εκτελείται τοπικά. Η ιδέα αυτή συνδυάζεται εύκολα με άλλες νέες προόδους του τομέα και καταφέρνει να δώσει αποτελέσματα πολύ υψηλότερα από τα ήδη δημοσιευμένα χωρίς μεγάλο επιπλέον κόστος σε χρόνο και μνήμη.

Η μέθοδος LOPQ μοιάζει να ακολουθεί το μοντέλο του αλγορίθμου βελτιστοποίησης παραμέτρων ενός μείγματος κατανομών: βελτιστοποιούνται πρώτα τα κέντρα των συνιστωσών και έπειτα με δεδομένα αυτά οι υπο-κβαντιστες του παραγοντικού κβαντισμού. Αποτελεί επίσης μια απλή μέθοδο για την εκμάθηση ενός μη γραμμικού μοντέλου για τα δεδομένων λύνοντας τοπικά υπο-προβλήματα με γραμμικό τρόπο. Η ταυτόχρονη βελτιστοποίηση του γενικού και των τοπικών κβαντιστών είναι μια πιθανή μελλοντική κατεύθυνση, η οποία βέβαια κρύβει αρκετές δυσκολίες όσων αφορά την υπολογιστική πολυπλοκότητα της εκμάθησης. Επίσης ενδιαφέρουσα κατεύθυνση είναι η εξερεύνηση της σχέσης της μεθόδου με δενδρικές δομές, ώστε να στοχεύσουμε στην συμπίεση *συνόλων* σημείων ολικά όπως στη δημοσίευση [4], ενώ ταυτόχρονα να μπορούμε να ψάξουμε μη-εξαντλητικά και γρήγορα.

Για την αναζήτηση σε μεγάλες κλίμακες, προτείνουμε μια μέθοδο που μπορεί να εφαρμοστεί σε συλλογές εκατομμυρίων εικόνων, η οποία καταφέρνει να ενισχύει την απόδοση της αναζήτησης με ταυτόχρονη μείωση των απαιτούμενων πόρων και κυρίως της μνήμης. Συνδυάσαμε την γεωγραφική πληροφορία που πλέον πολλές μεγάλες συλλογές παρέχουν απλόχερα με τις γρήγορες δομές δεικτοδότησης και επιτύχαμε γρήγορη και αποτελεσματική αναζήτηση και γεωεντοπισμό, χρησιμοποιώντας μόνο από τη οπτική πληροφορία των εικόνων αναζήτησης. Προτείναμε επίσης μια νέα αναπαράσταση σκηνών που απαιτεί αρκετά λιγότερη μνήμη ακόμα και από τη βασική αναζήτηση, έχοντας παράλληλα τη διακριτική ικανότητα που απαιτείται ώστε να αναζητούνται επιτυχώς ακόμα και με-

μονωμένες εικόνες της συλλογής. Η διαδικασίες εξόρυξης και ομαδοποίησης της προτεινόμενης μεθόδου είναι πιο γρήγορες ακόμα και από άλλες σχετικές υλοποιήσεις που χρησιμοποιούν παράλληλα συστήματα και δεν εκμεταλλεύονται την γεωγραφική πληροφορία.

Παρουσιάσαμε τέλος την διαδικτυακή εφαρμογή μας *VIRaL*[1], η οποία μπορεί να εκτελέσει γρήγορη αναζήτηση και γεω-εντοπισμό σε μία συλλογή με περισσότερες από δύο εκατομμύρια εικόνες. Σε αυτή μπορεί ο χρήστης να εξερευνήσει το σύνολο της συλλογής αποτελεσματικά με χρήση των ομάδων όψεων που προτείναμε και στο μέλλον θα προσπαθήσουμε να ενσωματώσουμε και τους χάρτες σκηνών στη διαδικασία αναζήτησης.

Τέλος, πρόσφατα έχουμε εφαρμόσει με επιτυχία τεχνικές παρόμοιες με τις προαναφερθείσες στην αναζήτηση από συλλογές πολιτισμικής κληρονομιάς μεγάλης κλίμακας [55], και τα αποτελέσματα παρουσιάζονται στην διαδικτυακή εφαρμογή *Vieu*[2].

Συνοψίζοντας, η παρούσα διατριβή περιέχει νέους αλγορίθμους που είτε ορίζουν είτε στέκονται δίπλα στο state-of-the-art των αντίστοιχων επιστημονικών περιοχών. Μερικοί από αυτούς αποτελούν γενικά εργαλεία για μεγάλης κλίμακας αναζήτηση κοντινότερου γείτονα ή ομαδοποίηση και άλλοι αναφέρονται σε πιο συγκεκριμένες εφαρμογές της οπτικής αναζήτησης. Συνολικά, η διατριβή καλύπτει ένα ευρύ φάσμα της περιοχής αυτής και, μιας και εστιάζει στην αναζήτηση μεγάλης κλίμακας, καθίσταται ιδιαίτερα επίκαιρη σε μια χρονική περίοδο κατά την οποία ο όγκος της διαθέσιμης οπτικής πληροφορίας πολλαπλασιάζεται.

[1] `http://viral.image.ntua.gr`
[2] `http://vieu.image.ntua.gr`

Βιβλιογραφία

[1] Karim M. Abadir and Jan R. Magnus. *Matrix Algebra*. Cambridge University Press, 2005.

[2] A. Agarwal and B. Triggs. Hyperfeatures-multilevel local coding for visual recognition. In *ECCV*, 2006.

[3] Sameer Agarwal, Noah Snavely, Ian Simon, Steven M. Seitz, and Richard Szeliski. Building Rome in a day. In *ICCV*, 2009.

[4] Relja Arandjelovic and Andrew Zisserman. Extremely low bit-rate nearest neighbor search using a set compression tree. Technical report, 2013.

[5] Y. Avrithis, Y. Kalantidis, G. Tolias, and E. Spyrou. Retrieving landmark and non-landmark images from community photo collections. In *in Proceedings of ACM Multimedia (Full paper) (MM 2010)*, Firenze, Italy, October 2010.

[6] Y. Avrithis and K. Rapantzikos. The medial feature detector: Stable regions from image boundaries. In *in Proceedings of International Conference on Computer Vision (ICCV 2011)*, Barcelona, Spain, November 2011.

[7] Y. Avrithis, G. Tolias, and Y. Kalantidis. Feature map hashing: Sub-linear indexing of appearance and global geometry. In *ACM Multimedia*, Firenze, Italy, October 2010.

[8] Yannis Avrithis. Quantize and conquer: A dimensionality-recursive solution to clustering, vector quantization, and image retrieval. In *ICCV*. 2013.

[9] Artem Babenko and Victor Lempitsky. The inverted multi-index. In *CVPR*, 2012.

[10] H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded up robust features. In *ECCV*, 2006.

[11] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[12] Oren Boiman, Eli Shechtman, and Michal Irani. In defense of nearest-neighbor based image classification. In *CVPR*, 2008.

[13] Y-Lan Boureau, Nicolas Le Roux, Francis Bach, Jean Ponce, and Yann Lecun. Ask the locals: Multi-way local pooling for image recognition. In *ICCV*, 2011.

[14] Jonathan Brandt. Transform coding for fast approximate nearest neighbor search in high dimensions. In *CVPR*, 2010.

[15] Vijay Chandrasekhar, Yuriy Reznik, Gabriel Takacs, David M Chen, Sam S Tsai, Radek Grzeszczuk, and Bernd Girod. Compressing feature sets with digital search trees. In *ICCV Workshops*. IEEE, 2011.

[16] Y. Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine intelligence*, 17(8):790–799, 1995.

[17] O. Chum, J. Matas, and J. Kittler. Locally optimized RANSAC. In *DAGM*, page 236. Springer Verlag, 2003.

[18] O. Chum, M. Perdoch, and J. Matas. Geometric min-hashing: Finding a (thick) needle in a haystack. In *CVPR*, 2009.

[19] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *ICCV*, 2007.

[20] Ondrej Chum and Jiri Matas. Large-scale discovery of spatially related images. *PAMI*, 32(2):371–377, Feb 2010.

[21] Ondrej Chum, Andrej Mikulik, Michal Perdoch, and Jirı Matas. Total recall ii: Query expansion revisited. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 889–896. IEEE, 2011.

[22] David Crandall, Lars Backstrom, Daniel Huttenlocher, and Jon Kleinberg. Mapping the world's photos. In *WWW*, 2009.

[23] M. Datar, N. Immorlica, P. Indyk, and V.S Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *Symposium on Computational Geometry*, pages 253–262. ACM New York, NY, USA, 2004.

[24] J. Delhumeau, PH. Gosselin, H. Jegou, and P. Perez. Revisiting the VLAD image representation. In *ACM Multimedia*, Oct 2013.

[25] Wei Dong, Zhe Wang, Moses Charikar, and Kai Li. Efficiently matching sets of features with random histograms. In *ACM Multimedia*, pages 179–188, 2008.

[26] M.A. Fischler and R.C Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[27] Brian Fulkerson, Andrea Vedaldi, and Stefano Soatto. Localizing objects with smart dictionaries. In *ECCV*, 2008.

[28] Stephan Gammeter, Lukas Bossard, Till Quack, and Luc V. Gool. I know what you did last summer: Object-level auto-annotation of holiday snaps. In *ICCV*, 2009.

[29] Efstratios Gavves and Cees GM Snoek. Landmark image retrieval using visual synonyms. In *Proceedings of the international conference on Multimedia*, pages 1123–1126. ACM, 2010.

[30] Efstratios Gavves, Cees GM Snoek, and Arnold WM Smeulders. Visual synonyms for landmark image retrieval. *Computer Vision and Image Understanding*, 116(2):238–249, 2012.

[31] Tiezheng Ge, Kaiming He, Qifa Ke, and Jian Sun. Optimized product quantization. Technical report, 2013.

[32] Tiezheng Ge, Kaiming He, Qifa Ke, and Jian Sun. Optimized product quantization for approximate nearest neighbor search. In *CVPR*, 2013.

[33] Yunchao Gong and Svetlana Lazebnik. Iterative quantization: A procrustean approach to learning binary codes. In *CVPR*, 2011.

[34] Robert Gray. Vector quantization. *ASSP Magazine, IEEE*, 1(2):4–29, 1984.

[35] R. Hartley and A. Zisserman. *Multiple View Geometry*. Cambridge university press Cambridge, UK, 2000.

[36] James Hays and Alexei A. Efros. IM2GPS: Estimating geographic information from a single image. In *CVPR*, 2008.

[37] Kaiming He, Fang Wen, and Jian Sun. K-means hashing: an affinity-preserving quantization method for learning binary compact codes. In *CVPR*, 2013.

[38] K. Heath, N. Gelfand, M. Ovsjanikov, M. Aanjaneya, and L. J. Guibas. Image webs: Computing and exploiting connectivity in image collections. In *CVPR*, 2010.

[39] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*, 2008.

[40] H. Jegou, M. Douze, and C. Schmid. Improving bag-of-features for large scale image search. *International Journal of Computer Vision*, pages 1–21, 2010.

[41] H. Jegou, M. Douze, and C. Schmid. Improving bag-of-features for large scale image search. *IJCV*, 87(3):316–336, 2010.

[42] H. Jegou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *PAMI*, 33(1):117–128, 2011.

[43] H. Jegou, M. Douze, C. Schmid, and P. Perez. Aggregating local descriptors into a compact image representation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3304–3311, 2010.

[44] H. Jegou, M. Douze, C. Schmid, and P. Perez. Aggregating local descriptors into a compact image representation. In *CVPR*, 2010.

[45] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Packing bag-of-features. In *International Conference on Computer Vision*, 2009.

[46] Herve Jégou, Romain Tavenard, Matthijs Douze, and Laurent Amsaleg. Searching in one billion vectors: Re-rank with source coding. In *ICASSP*, 2011.

[47] B. Johansson and R. Cipolla. A system for automatic pose-estimation from a single image in a city scene. In *Proc. IASTED Int. Conf. Signal Processing, Pattern Recognition and Applications*, 2002.

[48] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *ICCV*, 2005.

[49] Y. Kalantidis and Y. Avrithis. Locally optimized product quantization for approximate nearest neighbor search. In *in Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR 2014)*, Columbus, Ohio, June 2014. IEEE.

[50] Y. Kalantidis, L. Kennedy, and L.-J. Li. Getting the look: Clothing recognition and segmentation for automatic product suggestions in everyday photos. In *in Proceedings of International Conference on Multimedia Retrieval (ICMR) (ICMR 2013)*, Dallas, TX, April 2013. ACM.

[51] Y. Kalantidis, LG. Pueyo, M. Trevisiol, R. van Zwol, and Y. Avrithis. Scalable triangulation-based logo recognition. In *in Proceedings of ACM International Conference on Multimedia Retrieval (ICMR 2011)*, Trento, Italy, April 2011.

[52] Y. Kalantidis, G. Tolias, Y. Avrithis, M. Phinikettos, E. Spyrou, P. Mylonas, and S. Kollias. Viral: Visual image retrieval and localization. *Multimedia Tools and Applications*, 2011.

[53] Evangelos Kalogerakis, Olga Vesselova, James Hays, Alexei A. Efros, and Aaron Hertzmann. Image sequence geolocation with human travel priors. In *ICCV*, 2009.

[54] L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury. How flickr helps us make sense of the world: Context and content in community-contributed media collections. In *ACM Multimedia*, volume 3, pages 631–640, 2007.

[55] I. Kollia, Y. Kalantidis, K. Rapantzikos, and A. Stafylopatis. Improving semantic search in digital libraries using multimedia analysis. 2012.

[56] Kevin K□"oser, Christian Beder, and Reinhard Koch. Conjugate rotation: Parameterization and estimation from an affine feature correspondence. In *CVPR*, 2008.

[57] Alain Lehmann, Bastian Leibe, and Luc Van Gool. PRISM: Principled implicit shape model. In *BMVC*, 2009.

[58] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *IJCV*, 77(1):259–289, 2008.

[59] B. Leibe, K. Mikolajczyk, and B. Schiele. Efficient clustering and matching for object class recognition. In *BMVC*, 2006.

[60] V.I. Levenshtein. Binary codes capable of correcting spurious insertions and deletions of ones. *Problems of Information Transmission*, 1(1):8–17, 1965.

[61] Darui Li, Linjun Yang, Xian-Sheng Hua, and Hong-Jiang Zhang. Large-scale robust visual codebook construction. In *ACM Multimedia*, 2010.

[62] Xiaowei Li, Changchang Wu, Christopher Zach, Svetlana Lazebnik, and Jan-Michael Frahm. Modeling and recognition of landmark image collections using iconic scene graphs. In *ECCV*, pages 427–440. Springer, 2008.

[63] Yunpeng Li, David J. Crandall, and Daniel P. Huttenlocher. Landmark classification in large-scale image collections. In *ICCV*, 2009.

[64] T. Lindeberg. Edge detection and ridge detection with automatic scale selection. *IJCV*, 30(2):117–154, 1998.

[65] T. Lindeberg. Feature detection with automatic scale selection. *IJCV*, 30(2):79–116, 1998.

[66] Lingqiao Liu, Lei Wang, and Xinwang Liu. In defense of soft-assignment coding. In *ICCV*, 2011.

[67] D.G. Lowe. Local feature view clustering for 3D object recognition. In *CVPR*, 2001.

[68] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[69] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, page 14. California, USA, 1967.

[70] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767, 2004.

[71] A. McCallum, K. Nigam, and L.H Ungar. Efficient clustering of high-dimensional data sets with application to reference matching. In *6Th ACM International Conference on Knowledge Discovery and Data Mining*, page 178, 2000.

[72] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.

[73] Andrej Mikulik, Michal Perdoch, Ondrej Chum, and Jiri Matas. Learning a fine vocabulary. In *ECCV*, 2010.

[74] F. Moosmann, E. Nowak, and F. Jurie. Randomized clustering forests for image classification. *PAMI*, 30:1632–1646, 2008.

[75] M. Muja and D.G Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *ICCV*, 2009.

[76] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *CVPR*, 2006.

[77] M. Norouzi and D. Fleet. Minimal loss hashing for compact binary codes. In *ICML*, 2011.

[78] M. Norouzi and D. Fleet. Cartesian k-means. In *CVPR*, 2013.

[79] Mohammad Norouzi, Ali Punjani, and David J Fleet. Fast search in Hamming space with multi-index hashing. In *CVPR*, 2012.

[80] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

[81] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

[82] L. Paulevé, H. Jégou, and L. Amsaleg. Locality sensitive hashing: a comparison of hash function types and querying mechanisms. *Pattern Recognition Letters*, 31(11):1348–1358, August 2010.

[83] Michal Perdoch, Ondrej Chum, and Jiri Matas. Efficient representation of local geometry for large scale object retrieval. In *CVPR*, 2009.

[84] F. Perronnin. Universal and adapted vocabularies for generic visual categorization. *PAMI*, 30(7):1243–1256, 2008.

[85] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007.

[86] James Philbin, Ondrej Chum, Josef Sivic, Michael Isard, and Andrew Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR*, 2008.

[87] T. Quack, B. Leibe, and L. Van Gool. World-scale mining of objects and events from community photo collections. In *CIVR*, pages 47–56, 2008.

[88] K. Rapantzikos, Y. Avrithis, and S. Kollias. Detecting regions from single scale edges. In *in Proceedings of International Workshop on Sign, Gesture and Activity (SGA'10), European Conference on Computer Vision (ECCV 2010)*, September 2010.

[89] D. Robertson and R. Cipolla. An image-based system for urban navigation. In *BMVC*, 2004.

[90] F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image sets, or how do i organize my holiday snaps. In *ECCV*, 2002.

[91] G. Schindler, M. Brown, and R. Szeliski. City-scale location recognition. In *CVPR*, 2007.

[92] Peter H Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

[93] X. Shen, Z. Lin, J. Brandt, S. Avidan, and Y. Wu. Object retrieval and localization with spatially-constrained similarity measure and k-nn re-ranking. In *Computer Vision and Pattern Recognition*, 2012.

[94] C. Silpa-Anan and R. Hartley. Optimised KD-trees for fast image descriptor matching. In *CVPR*, 2008.

[95] I. Simon, N. Snavely, and S.M Seitz. Scene summarization for online image collections. In *ICCV*, 2007.

[96] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, pages 1470–1477, 2003.

[97] A.W.M. Smeulders, M. Worring, S. Santini, and A. Gupta. Content-based image retrieval at the end of the early years. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(12):1349–1380, Dec 2000.

[98] N. Snavely, S.M. Seitz, and R. Szeliski. Photo tourism: Exploring photo collections in 3D. In *Computer Graphics and Interactive Techniques*, pages 835–846, 2006.

[99] N. Snavely, S.M. Seitz, and R. Szeliski. Skeletal graphs for efficient structure from motion. In *CVPR*, 2008.

[100] U. Steinhoff, D. Omercevic, R. Perko, B. Schiele, and A. Leonardis. How computer vision can help in outdoor positioning. In *European Conference on Ambient Intelligence*, 2007.

[101] M. Tipping and B. Sch□”olkopf. A kernel approach for vector quantization with guaranteed distortion bounds. In *Artificial Intelligence and Statistics*, pages 129–134, 2001.

[102] G. Tolias and Y. Avrithis. Speeded-up, relaxed spatial matching. In *in Proceedings of International Conference on Computer Vision (ICCV 2011)*, Barcelona, Spain, November 2011.

[103] G. Tolias and Y. Avrithis. Speeded-up, relaxed spatial matching. In *ICCV*, 2011.

[104] G. Tolias, Y. Kalantidis, and Y. Avrithis. Symcity: Feature selection by symmetry for large scale image retrieval. In *in Proceedings of ACM Multimedia (Full paper) (MM 2012)*, Nara, Japan, October 2012. ACM.

[105] G. Tolias, Y. Kalantidis, Y. Avrithis, and S. Kollias. Towards large-scale geometry indexing by feature selection. *Computer Vision and Image Understanding*, 120(3):31–45, March 2014.

[106] Panu Turcot and David G Lowe. Better matching with fewer features: The selection of useful features in large database recognition problems. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 2109–2116. IEEE, 2009.

[107] Tinne Tuytelaars and Cordelia Schmid. Vector quantizing feature space with a regular lattice. In *ICCV*, Oct 2007.

[108] J. C. Van Gemert, C. J. Veenman, A. W. M. Smeulders, and J. M. Geusebroek. Visual word ambiguity. *PAMI*, 2010.

[109] C. Varytimidis, K. Rapantzikos, and Y. Avrithis. W□sh: Weighted □-shapes for local feature detection. In *in Proceedings of European Conference on Computer Vision (ECCV 2012)*, Florence, Italy, October 2012.

[110] A. Vedaldi and S. Soatto. Quick shift and kernel methods for mode seeking. In *ECCV*, 2008.

[111] Y. Weiss, A. Torralba, and R. Fergus. Spectral hashing. In *NIPS*, 2008.

[112] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *ICCV*, 2005.

[113] Jianxin Wu and James M. Rehg. Beyond the euclidean distance: Creating effective visual codebooks using the histogram intersection kernel. In *ICCV*, 2009.

[114] Y. Xia, K. He, F. Wen, and J. Sun. Joint inverted indexing. In *ICCV*, 2013.

[115] W. Zhang and J. Kosecka. Image based localization in urban environments. In *International Symposium on 3D Data Processing, Visualization and Transmission*, 2006.

[116] Yimeng Zhang, Zhaoyin Jia, and Tsuhan Chen. Image retrieval with geometry-preserving visual phrases. In *Computer Vision and Pattern Recognition*, pages 809–816. IEEE, 2011.

[117] Yantao Zheng, Ming Zhao, Yang Song, Hartwig Adam, Ulrich Buddemeier, Alessandro Bissacco, Fernando Brucher, Tat-Seng Chua, and Hartmut Neven. Tour the world: Building a web-scale landmark recognition engine. In *CVPR*, 2009.