ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Εντοπισμός πολυτροπικών χαρακτηριστικών για οπτικοακουστική αναγνώριση ομιλίας

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Κωνσταντίνος Β. Παρδάλης

Επιβλέπων : Στέφανος Δ. Κόλλιας

Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2008

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Εντοπισμός πολυτροπικών χαρακτηριστικών για οπτικοακουστική αναγνώριση ομιλίας

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Κωνσταντίνος Β. Παρδάλης

Επιβλέπων : Στέφανος Δ. Κόλλιας

Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 23η Ιουλίου 2008

.....	.....	.....
Στέφανος Κόλλιας	Ανδρέας Σταφυλοπάτης	Παναγιώτης Τσανάκας

Αθήνα, Ιούλιος 2008

.....

Κωνσταντίνος Β. Παρδάλης

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΙΝΑΚΑΣ ΣΧΗΜΑΤΩΝ

ΠΙΝΑΚΑΣ ΠΙΝΑΚΩΝ

ΠΕΡΙΛΗΨΗ

Η χρήση της οπτικής πληροφορίας για την βελτίωση της κατανόησης της ανθρώπινης ομιλίας αποτελεί ένα πολύ ενεργό πεδίο τα τελευταία χρόνια. Ο άνθρωπος είναι ικανός να κατανοήσει την ομιλία σε πολύ ικανοποιητικό βαθμό από την οπτική πληροφορία μόνο, ενώ η αξία της έχει αποδειχθεί με πληθώρα πειραμάτων, τόσο για την περίπτωση του ανθρώπου όσο και για την μηχανή. Η παρούσα εργασία σκοπό έχει να ασχοληθεί με τον προσδιορισμό και την εξαγωγή των οπτικών αυτών χαρακτηριστικών τα οποία δίνουν την δυνατότητα της κατανόησης του ανθρώπινου λόγου μόνο με την χρήση της οπτικής πληροφορίας. Να εντοπισθούν τα όρια των τεχνολογιών που υπάρχουν στον συγκεκριμένο τομέα και οι πιθανές βελτιώσεις που μπορεί να προκύψουν. Ορίζεται μία συνολική αρχιτεκτονική η οποία μπορεί να οδηγήσει στην εξαγωγή των χαρακτηριστικών αυτών και μελετάται το κάθε τμήμα της ξεχωριστά.

ABSTRACT

In recent years there is an emerging scientific field that tries to exploit the visual information in order to enhance the ability of the machine to detect the human speech. The human being is capable of perceiving speech efficiently, only by using visual information, a technique which is known as lip reading. The value of visual information for the perception of human speech has been proved by many different experiments for both the human and the machine respectively. The present work tries to define and extract the visual characteristics that can lead to the perception of human speech by using only visual information. To find out the limitations of current state of the art in this area and propose improvements when this is possible. Also a software architecture is defined which is capable of extracting efficiently these characteristics.

ΚΕΦΑΛΑΙΟ 1ο.

ΕΙΣΑΓΩΓΗ – ΓΕΝΙΚΗ ΠΕΡΙΓΡΑΦΗ

Το πρώτο σύστημα αναγνώρισης ανθρώπινης ομιλίας με χρήση ταυτόχρονα ακουστικής και οπτικής πληροφορίας παρουσιάστηκε στο [10] το 1984. Από τότε μέχρι και σήμερα έχουν εκδοθεί δεκάδες άρθρα πάνω στο θέμα της χρήσης της οπτικής πληροφορίας για την αναγνώριση του ανθρώπινου λόγου. Η σημασία της οπτικής οδού στην κατανόηση της ομιλίας στον άνθρώπο παρουσιάζεται έντονα στο φαινόμενο McGurk το οποίο περιγράφηκε για πρώτη φορά στο [13]. Είναι λογικό λοιπόν η έρευνα να στραφεί στην χρήση της πληροφορίας η οποία μπορεί να εξαχθεί από το οπτικό σήμα αφού είναι φανερή η σημασία του στην ανθρώπινη κατανόηση του λόγου. Οι εφαρμογές οι οποίες έχουν προκύψει κατά καιρούς είναι πολλές και ποικίλουν μεταξύ τους. Σε κάθε περίπτωση πάντως τα αναφερθέντα οπτικοακουστικά συστήματα αναγνώρισης ομιλίας, υπερέχουν σε ρυθμούς αναγνώρισης αυτών τα οποία χρησιμοποιούν μόνο την ακουστική πληροφορία. Τα βασικά στοιχεία τα οποία συνθέτουν τον σχεδιασμό ενός οπτικοακουστικού συστήματος ομιλίας είναι:

- Σχεδιασμός και επιλογή των οπτικών χαρακτηριστικών και εξαγωγή αυτών
- Η επιλογή των μονάδων ομιλίας
- Κατηγοριοποίηση / ταξινόμηση
- Ενσωμάτωση οπτικών και ακουστικών αποτελεσμάτων

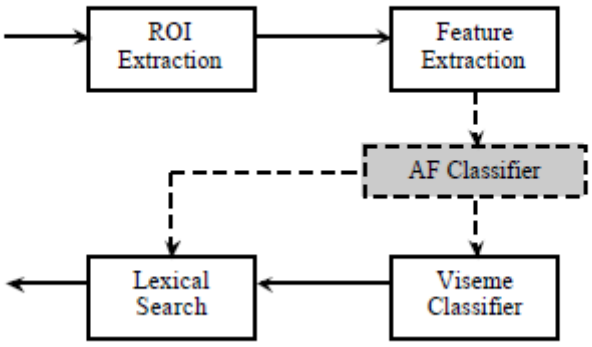
Το βασικό μειονέκτημα των συστημάτων αυτόματης αναγνώρισης ομιλίας είναι η ευαισθησία τους στον θόρυβο τόσο του περιβάλλοντος όσο και του διαύλου. Με την έννοια *δίαυλο* εννοούμε τον θόρυβο ο οποίος υπεισέρχεται στο σύστημα εξαιτίας της επεξεργασίας που κάνουμε ή των τμημάτων του συστήματος και ο οποίος δεν προέρχεται από το περιβάλλον. Πολλοί τρόποι έχουν προταθεί κατά καιρό για να αντιμετωπισθεί αυτό το πρόβλημα, όπως ειδικοί αλγόριθμοι επεξεργασίας ήχου, προσαρμοστικοί αλγόριθμοι μείωσης θορύβου και άλλοι. Η προσέγγιση που αναφέραμε παραπάνω, της ενσωμάτωσης και της οπτικής πληροφορίας στην διαδικασία της αναγνώρισης κερδίζει έδαφος τελευταία και για την αντιμετώπιση του προβλήματος του θορύβου.

Η χρήση της οπτικής πληροφορίας για την κατανόηση του λόγου, αν και ερευνάται έντονα τις τελευταίες δύο δεκαετίες, δεν έχει εισαχθεί ακόμα σε μεγάλο βαθμό στην επικοινωνία μεταξύ ανθρώπου και μηχανής. Υπάρχουν πολλοί λόγοι για τους οποίους συμβαίνει αυτό. Ένας βασικός λόγος είναι η αυξημένη ανάγκη τόσο σε επεξεργαστική ισχύ όσο και σε αποθηκευτικό χώρο που αυτές οι μέθοδοι απαιτούν. Άλλο ένα βασικό πρόβλημα το οποίο υπάρχει και το οποίο δεν είναι τόσο προφανές είναι η έλλειψη βάσεων δεδομένων οπτικοακουστικού υλικού με βάση τις οποίες μπορεί να γίνει η ανάπτυξη και ο έλεγχος τέτοιων συστημάτων. Ο πρώτος λόγος τείνει πλέον να εκλείψει μιας και υπάρχει πλέον διαθεσιμότητα μεγάλης υπολογιστικής ισχύς σε χαμηλό κόστος. Για τον δεύτερο λόγο έχουν γίνει πλέον κάποια βήματα και έχουν δημιουργηθεί βάσεις δεδομένων οι οποίες μπορούν να χρησιμοποιηθούν για την ανάπτυξη τέτοιων εφαρμογών. Θα πρέπει όμως να αναφέρουμε πως η λήψη των ομιλητών σε όλες σχεδόν τις περιπτώσεις γίνεται σε ελεγχόμενο περιβάλλον. Αυτό το γεγονός παρόλο που μπορεί να βοηθήσει σε μεγάλο βαθμό την έρευνα, καθιστά την διαδικασία δημιουργίας συστημάτων που μπορούν να λειτουργήσουν υπό φυσιολογικές συνθήκες πολύ δύσκολη.

Όπως προαναφέραμε η οπτική πληροφορία μπορεί να βοηθήσει πολύ στην αναγνώριση του λόγου κάτω από συνθήκες θορύβου. Αυτό δεν σημαίνει όμως πως η οπτική πληροφορία δεν είναι ευαίσθητη και αυτή στον θόρυβο, απλά ο θόρυβος ο οποίος επηρεάζει τα δύο κανάλια (οπτικό και ακουστικό) είναι διαφορετικής φύσης. Για παράδειγμα η χρήση χαμηλής ποιότητας μηχανημάτων λήψης μπορεί να εισάγει θόρυβο στην εικόνα. Μελέτες έχουν δείξει πως τα αποτελέσματα των συστημάτων οπτικής αναγνώρισης του λόγου, διαφέρουν σε μεγάλο βαθμό όταν το περιβάλλον λήψης μετακινηθεί από το ελεγχόμενο στο φυσικό, για παράδειγμα στον χώρο εργασίας ή μέσα σε ένα αυτοκίνητο. Αυτό το γεγονός έχει οδηγήσει τους ερευνητές να αναζητήσουν μεθόδους οι οποίες μπορούν να βελτιώσουν την διακριτική ικανότητα των αλγορίθμων οπτικής αναγνώρισης του λόγου κάτω από συνθήκες θορύβου.

1.1. Αρχιτεκτονική ενός συστήματος οπτικής αναγνώρισης του λόγου

Η αρχιτεκτονική ενός συστήματος οπτικής αναγνώρισης ομιλίας είναι τυπική και σε γενικές γραμμές η ίδια σε όλες τις περιπτώσεις που αναφέρονται στην βιβλιογραφία. Ένα τυπικό παράδειγμα όπως αυτό παρουσιάζεται στο [21] παρουσιάζεται στο **Σχήμα 1**.



Σχήμα 1: Τυπικό παράδειγμα αρχιτεκτονικής ενός συστήματος οπτικής αναγνώρισης του λόγου

Το πρώτο τμήμα του συστήματος συνήθως χωρίζεται σε επιμέρους τμήματα τα οποία είναι τα εξής. Αρχικά στο καρέ εντοπίζεται το κεφάλι του ομιλητή και στην συνέχεια εντοπίζονται συγκεκριμένα τα σημεία ενδιαφέροντος που στην δική μας περίπτωση είναι η περιοχή των χειλιών. Στην συνέχεια αφού όπως αναφέραμε και προηγουμένως έχουν εντοπισθεί τα χαρακτηριστικά τα οποία θα εξαχθούν, τα αποτελέσματα της επεξεργασίας από το προηγούμενο στάδιο, τροφοδοτούνται σε ένα σύστημα εξαγωγής χαρακτηριστικών. Στην συνέχεια αυτά τα χαρακτηριστικά συνθέτουν ένα διάνυσμα το οποίο τροφοδοτεί ένα σύστημα ταξινόμησης το οποίο και καλείται να προσδιορίσει την φωνητική μονάδα την οποία προφέρει ο ομιλητής. Στην περίπτωση που το σύστημα καλείται να κάνει χρήση και της ακουστικής πληροφορίας, ανά περίπτωση έχουν προταθεί διάφορες μέθοδοι για την ενσωμάτωση της με την οπτική πληροφορία. Συνήθως η αρχιτεκτονική ενός τέτοιου υβριδικού συστήματος έχει την ίδια δομή με αυτή του σχήματος 1.1 αλλά τα αποτελέσματα που προκύπτουν τροφοδοτούνται μαζί με τα αποτελέσματα ενός συστήματος ακουστικής αναγνώρισης σε έναν αλγόριθμο ο οποίος αναλαμβάνει να κάνει την ενσωμάτωση και να εξάγει το τελικό αποτέλεσμα.

1.2. Σκοπός – συμβολή της εργασίας

Σκοπός της παρούσας εργασίας είναι η μελέτη της διαδικασίας εξαγωγής χαρακτηριστικών από το οπτικό σήμα με σκοπό την αναγνώριση της ανθρώπινης ομιλίας. Παρόλο που δεν θα ασχοληθούμε καθόλου με το τελευταίο στάδιο της αρχιτεκτονικής ενός τέτοιου συστήματος, όπως αυτό παρουσιάζεται στο Σχήμα 1, θα ασχοληθούμε με όλα τα προηγούμενα. Αρχικά θα

επιχειρήσουμε τον εντοπισμό του προσώπου. Στην συνέχεια τον προσδιορισμό της περιοχής των χειλιών. Τέλος θα ορίσουμε τα χαρακτηριστικά τα οποία μας ενδιαφέρουν καθώς και την ακουστική μονάδα της ομιλίας και θα ορίσουμε έναν τρόπο εντοπισμού και αναπαράστασης των χαρακτηριστικών αυτών.

Τόσο το πρόσωπο όσο και τα χείλη δεν αποτελούν άμεσα χαρακτηριστικά τα οποία οδηγούν στην αναγνώριση του λόγου. Σε μία όμως πιο γενική θεώρηση μπορούμε να υποθέσουμε πως και τα δύο αυτά αποτελούν χαρακτηριστικά της εικόνας τα οποία οδηγούν στην αναγνώριση αυτή. Ταυτόχρονα η έλλειψη δεδομένων που να αποτελούνται μόνο από χείλη κάνει πολύ δύσκολη την ενασχόληση μόνο με τον εντοπισμό των απαραίτητων χαρακτηριστικών. Για τους παραπάνω λόγους θα γίνει εκτενής μελέτη όλων των προαναφερθέντων τμημάτων της αρχιτεκτονικής.

Σε κάθε περίπτωση γίνεται μελέτη και αναφορά όλων των σύγχρονων μεθόδων που έχουν προταθεί. Με βάση την βιβλιογραφία καθώς και τις απαιτήσεις που έχουμε ορίσει για το σύστημα μας, γίνεται επιλογή κάποιας από τις μεθόδους και στην συνέχεια ακολουθεί πειραματική επαλήθευση των αρχικών υποθέσεων που έχουμε κάνει, καθώς και των αποτελεσμάτων που αναφέρονται στην βιβλιογραφία. Βασικοί άξονες της εργασίας αυτής είναι ο ορισμός και η εν μέρη υλοποίηση ενός συστήματος το οποίο θα καταφέρνει να εξάγει χαρακτηριστικά για ένα σύστημα αυτόματης αναγνώρισης ομιλίας, αποδοτικά. Αυτό σημαίνει πως επιδιώκουμε υψηλούς ρυθμούς αναγνώρισης αλλά ταυτόχρονα απαιτούμε και από το σύστημα να λειτουργεί όσο το δυνατόν πιο κοντά σε πραγματικό χρόνο. Αυτοί οι δύο βασικοί άξονες οδηγούν την επιλογή την οποία κάνουμε για την μέθοδο σε κάθε περίπτωση αλλά και την αξιολόγηση των δεδομένων που προκύπτουν από την πειραματική επαλήθευση.

1.3. Διάρθρωση της εργασίας

Η διάρθρωση της εργασίας σε γενικές γραμμές ακολουθεί την δομή της αρχιτεκτονικής την οποία προαναφέραμε στο κεφάλαιο 1.2.

Στο δεύτερο κεφάλαιο ασχολούμαστε με το πρόβλημα του εντοπισμού του προσώπου. Το πρόβλημα αυτό από μόνο του αποτελεί πεδίο έντονης έρευνας ενώ τα χαρακτηριστικά προβλήματα που παρουσιάζει σε μεγάλο βαθμό εντοπίζονται και στα υπόλοιπα τμήματα του συστήματος. Για αυτό τον λόγο γίνεται μία εκτενέστατη μελέτη όλων των μεθόδων που έχουν

προταθεί για την αντιμετώπιση του προβλήματος αυτού και με βάση τα πειραματικά αποτελέσματα που προκύπτουν και τους άξονες που ορίσαμε στο κεφάλαιο 1.3 γίνεται η επιλογή της μεθόδου.

Στο τρίτο κεφάλαιο παρουσιάζεται η μεθοδολογία εντοπισμού της περιοχής των χειλιών. Παρουσιάζονται τα ιδιαίτερα προβλήματα που παρουσιάζει ο εντοπισμός αυτός καθώς και οι προτεινόμενες μέθοδοι από την βιβλιογραφία. Τέλος γίνεται παρουσίαση της μεθόδου που επιλέχθηκε, αναφορά στις λεπτομέρειες της υλοποίησης καθώς και τα πειραματικά αποτελέσματα.

Στο τέταρτο κεφάλαιο γίνεται ο ορισμός των χαρακτηριστικών τα οποία θα εντοπίσουμε με σκοπό να συνθέσουμε το διάνυσμα χαρακτηριστικών που θα οδηγήσει στην αναγνώριση της λέξης. Ταυτόχρονα ορίζεται η μονάδα ομιλίας και τέλος ορίζουμε την μεθοδολογία την οποία θα ακολουθήσουμε για τον εντοπισμό των χαρακτηριστικών και την σύνθεση του διανύσματος.

Στο πέμπτο κεφάλαιο γίνεται θεωρητική ανάπτυξη της μεθόδου ταξινόμησης των διανυσμάτων στήριξης. Οι μηχανές διανυσμάτων στήριξης χρησιμοποιούνται εκτενέστατα στην βιβλιογραφία αλλά και στην παρούσα εργασία και κρίθηκε σκόπιμη η αναφορά σε αυτές μιας και αυτή μπορεί να οδηγήσει τόσο στην κατανόηση της επιλογής όσο και τον παραμέτρων που χρησιμοποιήθηκαν κατά την διάρκεια της εργασίας. Τέλος πολλά από τα προβλήματα αλλά και τις λύσεις οι οποίες προέκυψαν δικαιολογούνται θεωρητικά με βάση αυτή την ανάπτυξη.

Τέλος στο έκτο κεφάλαιο γίνεται μία επισκόπηση των συμπερασμάτων και αποτελεσμάτων τα οποία προέκυψαν κατά την διάρκεια της εκπόνησης αυτής της εργασίας, καθώς και προτάσεις για μελλοντικές βελτιώσεις του συστήματος.

ΚΕΦΑΛΑΙΟ 2ο.

ΕΝΤΟΠΙΣΜΟΣ ΠΡΟΣΩΠΩΝ

Το πρώτο βήμα στην ανάγνωση χειλιών σε ένα αυτοματοποιημένο σύστημα, είναι ο εντοπισμός της περιοχής ενδιαφέροντος (ROI , Region Of Interest). Στην προκειμένη περίπτωση αυτή συνίσταται από την περιοχή των χειλιών κυρίως. Αυτό προϋποθέτει αρχικά τον εντοπισμό του προσώπου και εν συνεχεία τον εντοπισμό άλλων χαρακτηριστικών, όπως στην προκειμένη περίπτωση των χειλιών. Στο κεφάλαιο που ακολουθεί θα γίνει μία περίληψη των τεχνικών που υπάρχουν στην περιοχή του εντοπισμού του προσώπου. Θα δοθεί αρκετά μεγάλη βαρύτητα στο συγκεκριμένο τμήμα του συστήματος αφενός γιατί το συγκεκριμένο πρόβλημα αποτελεί ένα ιδιαίτερο πεδίο έρευνας, τόσο από πλευράς επεξεργασίας εικόνας όσο και από πλευράς μηχανικής εκμάθησης, αφετέρου οι μεθοδολογίες που θα παρουσιαστούν στην πλειοψηφία τους είναι ικανές να γενικευθούν έτσι ώστε να μην εντοπίζουν μόνο πρόσωπα αλλά και τα υπόλοιπα χαρακτηριστικά τα οποία είναι απαραίτητα για το σύστημα που περιγράφουμε.

Οι τεχνικές οι οποίες έχουν αναπτυχθεί έως τώρα στον εντοπισμό προσώπων μπορούν να χωριστούν σε δύο βασικές κατηγορίες. Τεχνικές οι οποίες βασίζονται στον εντοπισμό με βάση χαρακτηριστικά τα οποία εξάγονται από την εικόνα και τεχνικές οι οποίες βασίζονται στην επεξεργασία ενός κυλιόμενου πλαισίου πάνω στην εικόνα (sliding windows).

Οι τεχνικές οι οποίες βασίζονται στην εξαγωγή χαρακτηριστικών, ακολουθούν μία από κάτω προς τα πάνω προσέγγιση. Αρχικά εξάγονται από την εικόνα χαρακτηριστικά χαμηλού επιπέδου (για παράδειγμα γίνεται τμηματοποίηση της εικόνας). Στην συνέχεια ένα υψηλού επιπέδου σύστημα γνώσης αναλύει τα χαρακτηριστικά και συμπεραίνει το κατά πόσο τα χαρακτηριστικά αυτά ανήκουν σε πρόσωπο η όχι. Οι τεχνικές αυτές συνήθως λειτουργούν στο επίπεδο των εικονοστοιχείων (pixels), για παράδειγμα στο χρώμα του δέρματος, ή χρησιμοποιούν κάποια φίλτρα βασισμένα σε πυρήνες (για παράδειγμα τεχνικές εντοπισμού ακμών). Συνήθως επιλέγεται κάποια χαμηλού επιπέδου τεχνική η οποία μπορεί αφενός να υλοποιηθεί εύκολα, αφετέρου δεν προσδίδει στο όλο σύστημα υπολογιστικό φόρτο. Αυτή η απλότητα η οποία υπάρχει ως προς τους αλγορίθμους εξαγωγής χαρακτηριστικών έχει σαν βασικό πλεονέκτημα το

ότι αυτές οι μέθοδοι είναι ιδιαίτερα γρήγορες και σχετικά εύκολες ως προς την υλοποίηση τους. Βασικό τους μειονέκτημα όμως είναι πως η λειτουργία τους σε επίπεδο εικονοστοιχείου τις κάνει ιδιαίτερα ευαίσθητες στις εναλλαγές φωτισμού, θορύβου και στις περιπτώσεις όπου έχουμε μερική κάλυψη του προσώπου από άλλα αντικείμενα (occlusion).

Οι τεχνικές οι οποίες βασίζονται στην χρήση κάποιου κυλιόμενου πλαισίου χρησιμοποιούν αντίθετη προσέγγιση, από πάνω προς τα κάτω. Οι τεχνικές αυτές χρησιμοποιούν κάποιο πλαίσιο εντοπισμού, συγκεκριμένων διαστάσεων το οποίο κινείται πάνω στην εικόνα εισόδου. Στην συνέχεια ο αλγόριθμος αναλαμβάνει να συμπεράνει κατά πόσο το περιεχόμενο του πλαισίου εντοπισμού εμπεριέχει κάποιο πρόσωπο ή όχι. Αφού τελειώσει ο αλγόριθμος με την επεξεργασία κάποιου συγκεκριμένου τμήματος το οποίο έχει εξαχθεί με το πλαίσιο εντοπισμού, στην συνέχεια το πλαίσιο μετακινείται πάνω στην εικόνα και η διαδικασία επαναλαμβάνεται στο νέο τμήμα εικόνας το οποίο έχει προκύψει. Αυτή η διαδικασία συνεχίζεται έως ότου έχουμε διασχίσει όλη την εικόνα και έχουν εξετασθεί όλα τα δυνατά πλαίσια τα οποία μπορούν να προκύψουν. Με την χρήση όμως πλαισίου υπεισέρχεται στο σύστημα η αδυναμία εντοπισμού προσώπων διαφορετικής κλίμακας. Αυτό το πρόβλημα αντιμετωπίζεται σχετικά εύκολα με συγκεκριμένη μέθοδο. Η αρχική εικόνα εισόδου σχηματίζει μία πυραμίδα εικόνων διαφορετικών διαστάσεων, συνήθως με χρήση πυραμίδων Gauss οι οποίες εξετάζονται στο κεφάλαιο όπου θα γίνει αναφορά στην επεξεργασία εικόνας που υπάρχει στο σύστημα. Με αυτό τον τρόπο μπορούμε να είμαστε αρκετά σίγουροι πως θα εντοπίσουμε το πρόσωπο σε κάποιο επίπεδο της πυραμίδας και στην συνέχεια θα προβάλουμε τις διαστάσεις του στην αρχική εικόνα. Συνήθως οι τεχνικές που βασίζονται στην χρήση πλαισίων είναι πιο αποδοτικές αλλά απαιτούν και πολύ μεγαλύτερο χρόνο υπολογισμού.

Στην συνέχεια θα περιγράψουμε έναν αριθμό μεθόδων εντοπισμού προσώπων. Οι μέθοδοι αυτοί αποτελούν τις βασικές που έχουν προταθεί κατά καιρούς στην βιβλιογραφία με σκοπό στο τέλος να καταλήξουμε στην μέθοδο η οποία δίνει τα καλύτερα αποτελέσματα στην περίπτωση που μελετάμε έχοντας λάβει υπόψιν όλες τις παραμέτρους που συνθέτουν ένα σύστημα ανάγνωσης χειλιών.

2.1. Επισκόπηση του προβλήματος εντοπισμού προσώπου

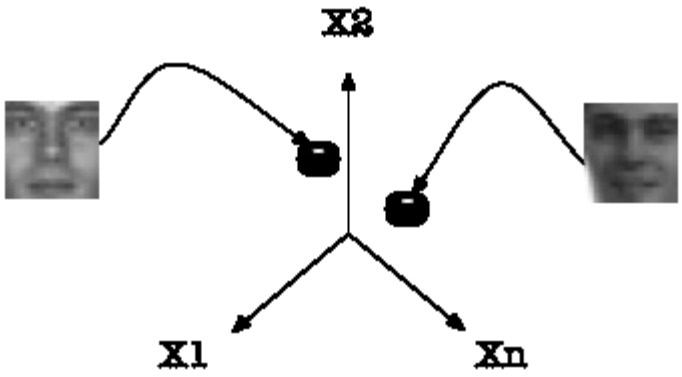
Πριν περιγράψουμε τους διάφορους αλγορίθμους εντοπισμού προσώπων σε εικόνες, είναι χρήσιμο να κάνουμε μία σύντομη εισαγωγή στα προβλήματα που παρουσιάζει αυτή η διαδικασία.

Το πρόβλημα εντοπισμού προσώπων προκύπτει τόσο από θεωρητικές ανάγκες όσο και από πρακτικές. Αρχικά λόγω της φύσης του προβλήματος και της δυσκολίας που παρουσιάζει αποτελεί ένα πολύ καλό τρόπο ελέγχου και αξιολόγησης μεθόδων και αλγορίθμων που αφορούν τον γενικότερο χώρο της υπολογιστικής όρασης. Επίσης μεγάλο πλήθος πρακτικών εφαρμογών απαιτούν την ικανότητα από την μηχανή να μπορεί να εντοπίζει αρχικά και στην συνέχεια να αναγνωρίζει πρόσωπα, όπως για παράδειγμα στην περίπτωση συστημάτων ασφαλείας και γενικότερα βιομετρικών εφαρμογών. Απτό παράδειγμα αποτελεί το πώς μπορούμε γενικά να αναγνωρίσουμε κάποιον όπου ουσιαστικά υπάρχουν τρεις διαφορετικές μέθοδοι. Αναγνώριση με βάση κάτι το οποίο κατέχει κάποιος, όπως για παράδειγμα μία ταυτότητα, με βάση κάτι το οποίο γνωρίζει, για παράδειγμα ένα password και τέλος κάτι το οποίο είναι ο ίδιος, δηλαδή τα ιδιαίτερα χαρακτηριστικά τα οποία αυτός κατέχει όπως για παράδειγμα τα δακτυλικά αποτυπώματα και το πρόσωπο του.

Αν αναλογιστούμε συγκεκριμένα το πρόβλημα της αναγνώρισης και ταυτοποίησης προσώπου, θα παρατηρήσουμε πως αποτελεί μία διαδικασία την οποία ο άνθρωπος μπορεί να την κάνει ιδιαίτερα εύκολα κάτω από σχεδόν οποιεσδήποτε συνθήκες. Προκύπτει όμως πως στην περίπτωση της μηχανής αυτή η διαδικασία είναι εξαιρετικά δύσκολη. Παρακάτω θα αναφέρουμε τους λόγους για τους οποίους συμβαίνει αυτό.

Ας θεωρήσουμε αρχικά την περίπτωση του προβλήματος ταυτοποίησης προσώπου, δηλαδή την διαδικασία κατά την οποία μας ενδιαφέρει να διακρίνουμε το κατά πόσο δύο πρόσωπα ανήκουν στο ίδιο άτομο η όχι. Αυτή η διαδικασία παρόλο που είναι διαφορετική από τον εντοπισμό προσώπων ουσιαστικά παρουσιάζει τις ίδιες δυσκολίες που καλούμαστε να αντιμετωπίσουμε και στην περίπτωση του εντοπισμού ενώ παρέχει ένα πιο απτό παράδειγμα για την κατανόηση τους. Μια αρχική προσέγγιση η οποία θα μπορούσαμε να εφαρμόσουμε και η οποία από πλευράς μαθηματικής διαίσθησης προκύπτει φυσικά, είναι να θεωρήσουμε ένα χώρο n διαστάσεων όπου n είναι ο αριθμός των εικονοστοιχείων τα οποία αποτελούν την εικόνα. Με αυτό τον τρόπο

μπορούμε να θεωρήσουμε το κάθε πρόσωπο σαν ένα σημείο σε έναν n-διάστατο χώρο όπου $x \in R^n$ και η τιμή του κάθε εικονοστοιχείου είναι μία συντεταγμένη του χ , όπως φαίνεται και στο **Σχήμα 2**. Αυτή η τεχνική έχει προταθεί στην βιβλιογραφία και συγκεκριμένα με χρήση ευκλείδειας μετρικής. Με αυτό τον τρόπο μπορούμε να συγκρίνουμε εικόνες μεταξύ τους παίρνοντας το μέτρο μεταξύ εικόνων σημείων και συγκρίνοντας το. Προφανώς αναμένουμε πως μεταξύ ιδίων προσώπων το μέτρο θα είναι περίπου μηδέν ενώ μεταξύ διαφορετικών προσώπων το μέτρο θα είναι διάφορο του μηδενός.

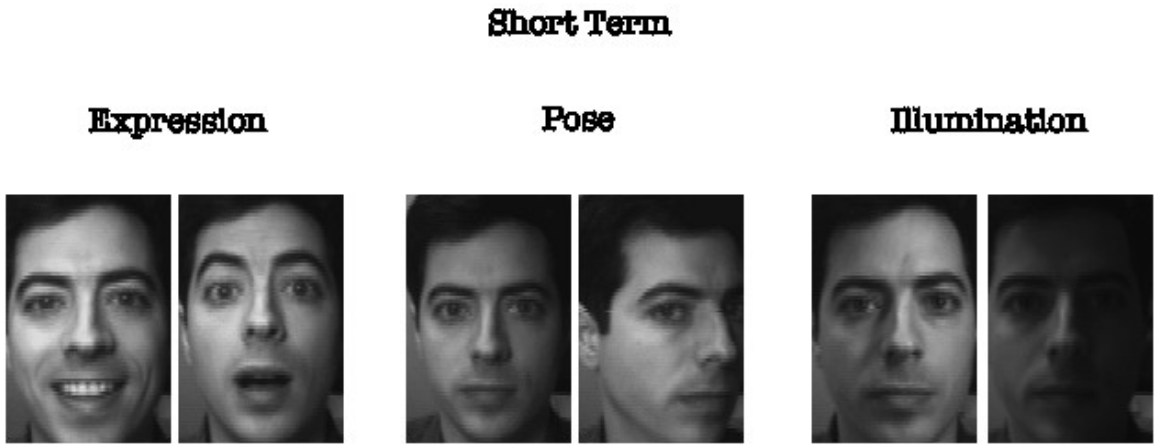


Σχήμα 2: Μεταφορά των εικόνων σε ένα n-διάστατο χώρο για σύγκριση τους με χρήση ευκλείδειου μέτρου.

Από τα παραπάνω μπορούμε να πούμε πως η ίδια λογική μπορεί να χρησιμοποιηθεί και στην περίπτωση που θέλουμε να εντοπίσουμε το πρόσωπο, κάνοντας όμως κάποιο διαφορετικό μετασχηματισμό έτσι ώστε να λάβουμε υπόψη τα ιδιαίτερα χαρακτηριστικά των προσώπων και να τα συγκρίνουμε με εικόνες οι οποίες δεν είναι πρόσωπα. Αντίστοιχα μπορούμε να χρησιμοποιήσουμε και διαφορετικές μετρικές. Τέτοια συστήματα έχουν παρουσιασθεί στην βιβλιογραφία και σημαντικό παράδειγμα είναι η χρήση των eigenfaces .

Παρόλα τα παραπάνω όμως στην πράξη παρατηρούμε πώς η διαδικασία αυτή δεν είναι ικανοποιητική. Ο κύριος λόγος που συμβαίνει αυτό, είναι η ευμεταβλησία που παρουσιάζουν οι εικόνες και η οποία μπορεί να διακριθεί σε δύο βασικές κατηγορίες. Η πρώτη κατηγορία είναι αυτή στην οποία η μεταβλητότητα είναι προσωρινή, σε αυτή ανήκουν η πόζα η έκφραση και ο φωτισμός. Στην δεύτερη κατηγορία έχουμε μεταβλητότητα η οποία δεν είναι προσωρινή, για παράδειγμα η ύπαρξη τριχών στο πρόσωπο, η ηλικία, το makeup και πολλά άλλα. Εστιάζουμε το

ενδιαφέρον μας κυρίως στην πρώτη κατηγορία την οποία θα ορίσουμε και μαθηματικά και για την οποία παραδείγματα παραθέτουμε στο **Σχήμα 3**.



Σχήμα 3: Παραδείγματα προσωρινής μεταβλητότητας στην περίπτωση εικόνων που περιέχουν πρόσωπα.

Στην περίπτωση λοιπόν που θέλουμε να μπορούμε να αναγνωρίσουμε πρόσωπα τα οποία βρίσκονται υπό διαφορετικές συνθήκες σαν αυτές που παρουσιάστηκαν παραπάνω θα πρέπει να αναζητήσουμε μία συνάρτηση $f(\cdot)$, τέτοια ώστε για ένα σημείο x στον n-διάστατο χώρο που περιγράψαμε παραπάνω και το οποίο αποτελεί μία εικόνα, τέτοια ώστε για $x_1, x_2, ..., x_n$, εικόνες οι οποίες δείχνουν το ίδιο πρόσωπο αλλά υπό διαφορετικές συνθήκες όπως αυτές που αναφέρθηκαν παραπάνω, να ισχύει, $f(x_1) = f(x_2) = ... = f(x_n) = a$, ενώ για ένα άλλο πρόσωπο y η γενικά διαφορετικό αντικείμενο να ισχύει $f(y_1) = f(y_2) = ... = f(y_n) = b$, όπου $a \neq b$.

Αυτό το οποίο διατυπώσαμε παραπάνω ουσιαστικά σημαίνει πως αναζητούμε την ύπαρξη σταθερών τέτοιων ώστε αν υπάρχει μεταβολή στον φωτισμό ή γεωμετρικές μεταβολές, να παραμένουν αμετάβλητες. Όπως έχει αποδειχθεί τόσο για την περίπτωση του φωτισμού όσο και για την περίπτωση των γεωμετρικών μεταβολών (πόζα και έκφραση) τέτοιες σταθερές δεν υπάρχουν για τρισδιάστατα αντικείμενα.

2.2. Χρωματική Σταθερότητα

Η χρωματική σταθερότητα αποτελεί ένα παράδειγμα υποκειμενικής σταθερότητας καθώς και βασικό χαρακτηριστικό της ανθρώπινης χρωματικής αντίληψης, το οποίο εγγυάται πως το χρώμα το οποίο αντιλαμβανόμαστε παραμένει σχετικά σταθερό υπό μεταβλητές συνθήκες φωτισμού. Για παράδειγμα ένα μήλο, συνεχίζουμε να το αντιλαμβανόμαστε ως κόκκινο αντικείμενο, είτε το βλέπουμε το μεσημέρι όπου το φως είναι λευκό, είτε το απόγευμα όπου το φως είναι κόκκινο. Αυτό το χαρακτηριστικό μας δίνει την δυνατότητα να αναγνωρίζουμε αντικείμενα. Μερικά είδη εκτός του ανθρώπου, όπως για παράδειγμα οι πίθηκοι και τα χρυσόψαρα έχουν αντίστοιχους μηχανισμούς χρωματικής σταθερότητας. Είναι πολύ πιθανό πώς όλα τα ζώα τα οποία κατέχουν χρωματική όραση παρουσιάζουν και χρωματική σταθερότητα.

Η χρωματική σταθερότητα λειτουργεί μόνο στην περίπτωση όπου ο προσπίπτον φωτισμός περιέχει μεγάλο εύρος από μήκη κύματος. Τα διαφορετικά κωνικά κύτταρα του ματιού αντιδρούν σε διαφορετικά μήκη κύματος του φωτός που ανακλάτε από κάθε αντικείμενο μίας σκηνής. Από αυτή την πληροφορία, το οπτικό σύστημα προσπαθεί να διακρίνει κατά προσέγγιση την σύνθεση του φωτός το οποίο φωτίζει την σκηνή. Αυτόν τον φωτισμό το οπτικό σύστημα τον παραβλέπει με σκοπό να αντιληφθεί το πραγματικό χρώμα του αντικειμένου, δηλαδή τα μήκη κύματος που πραγματικά αντανακλά το αντικείμενο. Αυτά τα μήκη κύματος καθορίζουν σε μεγάλο βαθμό το χρώμα που αντιλαμβάνεται το οπτικό σύστημα. Ο ακριβής αλγόριθμος που χρησιμοποιείται στην παραπάνω λειτουργία παραμένει άγνωστος.

Η περιγραφή του παραπάνω φαινόμενου έγινε για πρώτη φορά το 1971 από τον Edwin Land, ο οποίος σχημάτισε μία θεωρία με το όνομα retinex theory για να το εξηγήσει. Η λέξη retinex προκύπτει από τις λατινικές λέξεις retina και cortex, για να δώσει έμφαση στο ότι στον μηχανισμό αυτό συμμετέχουν τόσο το μάτι όσο και ο εγκέφαλος.

Αν αναλογιστούμε αυτά τα οποία αναφέραμε παραπάνω για τις δυσκολίες που παρουσιάζει ο εντοπισμός προσώπων από μία μηχανή, εύκολα μπορούμε να καταλήξουμε στο συμπέρασμα πως η χρωματική σταθερότητα είναι ένα ιδιαίτερα επιθυμητό χαρακτηριστικό το οποίο θα θέλαμε να υπάρχει στους αλγορίθμους. Με αυτό τον τρόπο θα μπορούσαμε τουλάχιστον να αντιμετωπίσουμε τις δυσκολίες που προσθέτει η ύπαρξη της μεταβλητότητας φωτισμού στα πρόσωπα. Έχουν σχεδιασθεί διάφοροι αλγόριθμοι οι οποίοι προσομοιώνουν την χρωματική

σταθερότητα που παρουσιάζει το ανθρώπινο μάτι, ενώ σχεδόν σε όλες τις περιπτώσεις εφαρμογής αλγορίθμων για εντοπισμό αντικειμένων απαιτείται προ επεξεργασία των εικόνων ώστε να μειωθούν τα προβλήματα λόγω φωτισμού, όπως θα δούμε και παρακάτω. Τέλος η χρωματική σταθερότητα είναι ένα κλασσικό παράδειγμα το οποίο μας δείχνει την υπεροχή που παρουσιάζει η όραση των θηλαστικών έναντι της όρασης μηχανής που έχουμε πετύχει έως τώρα.

2.3. Εντοπισμός βασισμένος στο χρώμα του δέρματος

Θα ξεκινήσουμε την παρουσίαση των μεθόδων εντοπισμού προσώπων με την ίσως πιο δημοφιλή μέθοδο στον συγκεκριμένο χώρο, η οποία είναι ο εντοπισμός του προσώπου λαμβάνοντας υπόψη το χρώμα του δέρματος.

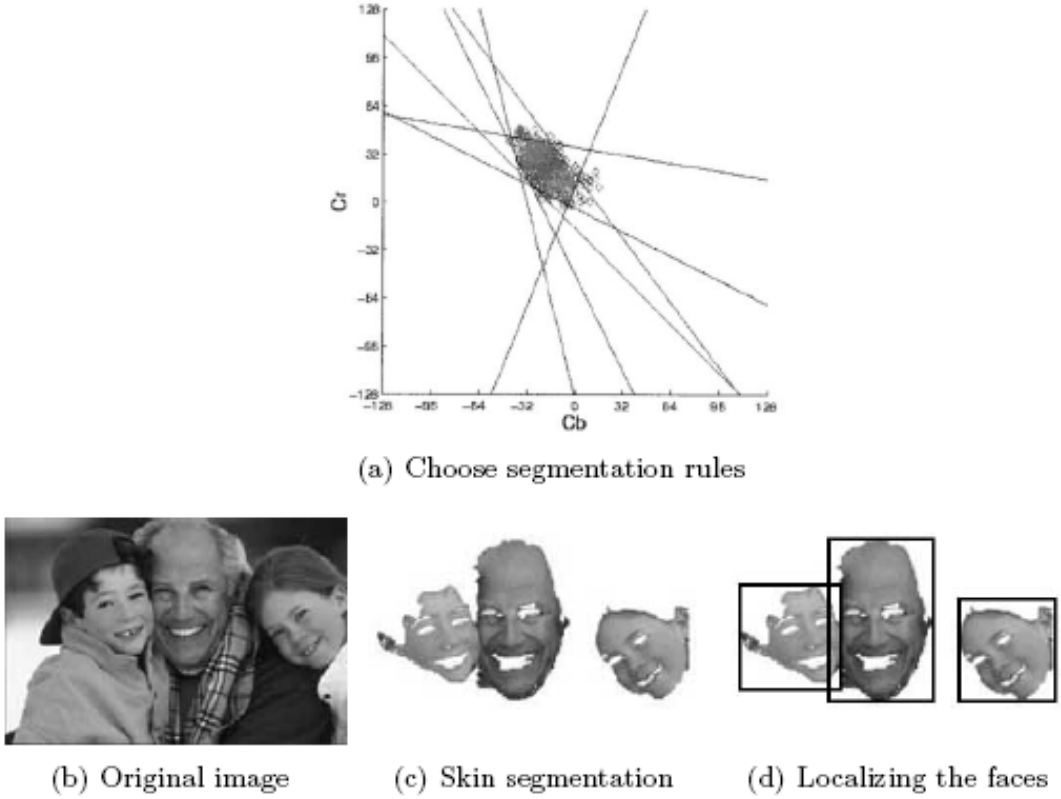
Οι μέθοδοι εντοπισμού με βάση το χρώμα του δέρματος έχουν γίνει ιδιαίτερα δημοφιλής τελευταία, κυρίως λόγω της απλότητας και την αποδοτικότητας τους σε σχέση με τους γεωμετρικούς μετασχηματισμούς που μπορεί να υποστεί το πρόσωπο. Επίσης πολύ σημαντικό είναι πως πλέον υπάρχει πρόσβαση σε υλικό video πολύ υψηλής χρωματικής ανάλυσης. Επίσης ενώ το χρώμα του δέρματος ανθρώπων διαφορετικής φυλής το αντιλαμβανόμαστε ως διαφορετικό, στην ουσία διαφέρει μόνο ως προς την ένταση του και όχι ως προς το χρώμα του. Αυτή η ιδιότητα του χρωματικού αναλλοίωτου του ανθρωπίνου δέρματος, μας δίνει την δυνατότητα να υλοποιήσουμε απλές και συνεπείς μεθόδους χρωματικής τμηματοποίησης. Ο εντοπισμός προσώπων με βάση το χρώμα του δέρματος μπορεί να χωριστεί σε τρία βασικά βήματα.

- Απόφαση ως προς τον χρωματικό χώρο που θα χρησιμοποιήσουμε.
- Δημιουργία δερματικού μοντέλου και τμηματοποίηση της εικόνας με βάση αυτό το μοντέλο.
- Εντοπισμός του προσώπου στην τμηματοποιημένη πλέον εικόνα.

Για τις συγκεκριμένες μεθόδους εντοπισμού έχουν προταθεί κατά καιρούς διάφοροι χρωματικοί χώροι. Μερικά δημοφιλή παραδείγματα είναι τα, RGB, κανονικοποιημένο RGB, YCbCr (YUV), HIS (Hue, Saturation, Intensity), καθώς και πολλά άλλα. Το ποιος χρωματικός χώρος είναι ο πλέον κατάλληλος αποτελεί ακόμα αντικείμενο έρευνας. Οι Shin et. al , έδειξαν πως η ικανότητα

διαχωρισμού μεταξύ δερματικού και όχι δερματικού χρώματος εξαρτάται από το χρωματικό μοντέλο που έχουμε επιλέξει. Χρησιμοποιώντας 4 διαφορετικές μετρικές διαχωρισμού σε 18 διαφορετικούς χρωματικούς χώρους, κατέληξαν στο ότι οι χώροι RGB και YCbCr απέδωσαν καλύτερα. Παρόλα αυτά τα αποτελέσματα τους έχουν αμφισβητηθεί από τους Vezhnevets et. al .

Το επόμενο βήμα είναι η δημιουργία του δερματικού μοντέλου με βάση τον χρωματικό χώρο που έχουμε επιλέξει. Αυτό το μοντέλο χρησιμοποιείται για τον διαχωρισμό των εικονοστοιχείων μεταξύ δέρματος και όχι δέρματος. Έχουν προταθεί πολλά δερματικά μοντέλα τα οποία κυμαίνονται από απλούς κανόνες τμηματοποίησης , έως πιο πολύπλοκα στατιστικά μοντέλα και προσαρμοστικές μεθόδους .



(a) Choose segmentation rules

(b) Original image (c) Skin segmentation (d) Localizing the faces

Σχήμα 4: Ένα τυπικό σύστημα εντοπισμού προσώπων με χρήση του δερματικού χρώματος. Αφού έχει γίνει επιλογή του χρωματικού χώρου, επιλέγονται οι κανόνες τμηματοποίησης του δέρματος(α). Στην συνέχεια η αρχική εικόνα (β) διαχωρίζεται με βάση αυτούς τους κανόνες (c). Στο τελευταίο στάδιο ένα υψηλού επιπέδου σύστημα γνώσης αναλύει την τμηματοποιημένη εικόνα και εντοπίζει τα πρόσωπα (d).

Όπως αναφέραμε παραπάνω ένας από τους βασικούς λόγους που η συγκεκριμένη μέθοδος προτιμάται είναι το αναλλοίωτο που παρουσιάζει ως προς τις γεωμετρικές μεταβολές που μπορεί να υποστεί το πρόσωπο. Δυστυχώς όμως αδυνατεί να παρουσιάσει χαρακτηριστικά χρωματικής σταθερότητας όπως αυτά αναφέρθηκαν προηγουμένως. Αυτό έχει ως αποτέλεσμα η απόδοση της να εξαρτάται σε πολύ μεγάλο βαθμό από τον φωτισμό της σκηνής. Για να αντιμετωπιστεί αυτό το πρόβλημα έχουν προταθεί διάφορες λύσεις οι οποίες όμως απαιτούν την ύπαρξη προσωρινής γνώσης ή να είναι γνωστά τα χαρακτηριστικά της κάμερας και των συνθηκών φωτισμού .

Το τελευταίο βήμα αποτελείται από κάποιο υψηλού επιπέδου συστήματος γνώσης το οποίο προσπαθεί να διακρίνει την περιοχή των προσώπων χρησιμοποιώντας την τμηματοποιημένη εικόνα. Σε αυτό το βήμα έχουν προταθεί κατά καιρούς πολλές διαφορετικές προσεγγίσεις και αναφέρουμε μερικές. Οι Singh et al αρχικά εντοπίζουν χαρακτηριστικά του προσώπου όπως τα μάτια και το στόμα από την εικόνα, βασισμένοι στην εικασία πως αυτά είναι περιοχές οι οποίες εμφανίζονται ως πιο σκούρες στην εικόνα. Οι Wang et al αρχικά χρησιμοποιούν κάποια μέθοδο εντοπισμού διαγραμμάτων (contours) στον δυαδικό χάρτη δέρματος που έχει δημιουργηθεί. Στην συνέχεια προσπαθούν να εντοπίσουν χαρακτηριστικά του προσώπου στο εσωτερικό του εντοπισμένου διαγράμματος και επιχειρούν να διακρίνουν τους άξονες συμμετρίας του προσώπου βασιζόμενοι στην θέση του στόματος.

2.3.1 Σχόλια

Όπως και σε όλα τα συστήματα τα οποία βασίζονται στην εξαγωγή χαρακτηριστικών, η απόδοση των μεθόδων με χρήση του δερματικού χρώματος βασίζεται κυρίως στην απόδοση της μεθόδου τμηματοποίησης. Εάν η απόδοση αυτής είναι άσχημη τότε το υψηλού επιπέδου σύστημα εντοπισμού έχει να φέρει σε πέρας πολύ δύσκολο έργο. Στο μεγαλύτερο τμήμα της βιβλιογραφίας που αφορά τις συγκεκριμένες τεχνικές, τα πρόσωπα που χρησιμοποιήθηκαν για τα πειράματα πάρθηκαν σε ελεγχόμενο περιβάλλον κυρίως σε σχέση με τον φωτισμό. Οι ρυθμοί εντοπισμού που αναφέρονται κυμαίνονται αρκετά υψηλά, ανάμεσα σε 80% και 95% .

2.4. Μέθοδος βασισμένη σε κατανομές

Οι Sung και Pongio ανέπτυξαν ένα σύστημα βασισμένο σε κατανομές για τον εντοπισμό προσώπων. Το σύστημα τους το οποίο ανήκει στην κατηγορία του κυλιομένου παραθύρου, αρχικά μοντελοποιεί εικόνες δείγματα μεγέθους 19×19 εικονοστοιχείων σε πολυδιάστατα διανύσματα δειγμάτων. Στην συνέχεια προσπαθούν να υποδιαιρέσουν τον δειγματικό χώρο σε υποκλάσεις. Για να προσεγγίσουν τις υποκλάσεις χρησιμοποιούν πολυδιάστατες συστοιχίες Gauss. Το σύστημα επιγραμματικά αποτελείται από τα παρακάτω βήματα.

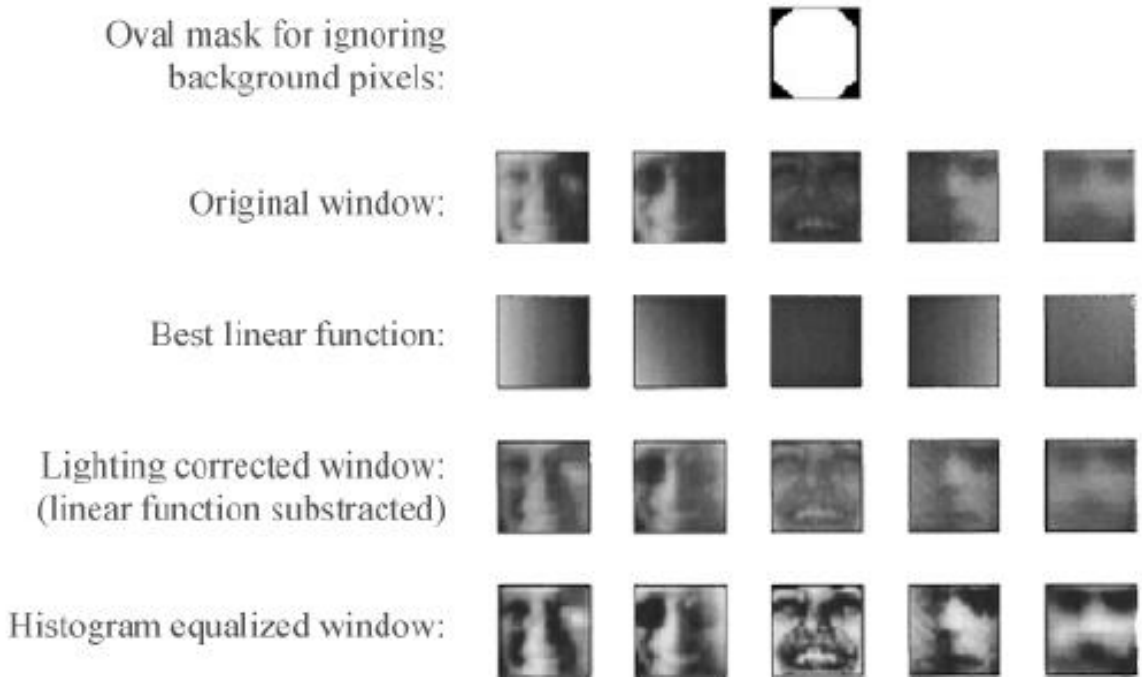
- Αρχικά η εικόνα στο παράθυρο εντοπισμού περνάει από επεξεργασία κατά την οποία μεταβάλλεται το μέγεθος της ώστε να είναι 19×19 και εφαρμόζονται πάνω της οι μέθοδοι που αναφέρονται στο **Σχήμα 5**. Αυτή η διαδικασία έχει σαν αποτέλεσμα να βελτιώνεται η ποιότητα της εικόνας ενώ ταυτόχρονα να μειώνονται οι διαστάσεις του διανύσματος εικόνα από $\Re^{361}$ σε $\Re^{283}$.

- Στην συνέχεια κατασκευάζεται ένα μοντέλο κατανομής από κανονικά πρόσωπα και όχι πρόσωπα χρησιμοποιώντας 12 πολυδιάστατες συστοιχίες Gauss. Οι συστοιχίες των 283 διαστάσεων κατασκευάζονται με χρήση ενός τροποποιημένου αλγορίθμου βασισμένου στον K-means algorithm ο οποίος υπολογίζει τις κεντροειδείς των συστοιχιών και τους πίνακες συνδιακύμανσης.

- Δεδομένης μίας νέας εικόνας, η απόσταση μεταξύ αυτής και κάθε συστοιχίας υπολογίζεται, αυτό δίνει ως αποτέλεσμα 12 αποστάσεις μεταξύ της εικόνας και των κέντρων των 12 συστοιχιών. Για κάθε μία από αυτές τις αποστάσεις υπολογίζονται δύο τιμές. Η πρώτη είναι το μέτρο Mahalanobis μεταξύ της εικόνας και του κέντρου της συστοιχίας σε έναν υποχώρο ο οποίος σχηματίζεται από τα 75 μεγαλύτερα ιδιοδιανύσματα της συστοιχίας. Η δεύτερη τιμή είναι η ευκλείδεια απόσταση μεταξύ της νέας εικόνας και της προβολής της στον υποχώρο, τελικά δημιουργείται ένα διάνυσμα μετρήσεων 24 διαστάσεων από την εικόνα.

- Τέλος ένα πολυεπίπεδο perceptron (MLP) χρησιμοποιείται για να κατηγοριοποιηθούν νέες εικόνες ως πρόσωπα η όχι πρόσωπα με χρήση των δυανυσμάτων 24 διαστάσεων που αναφέρθηκαν στο προηγούμενο βήμα.

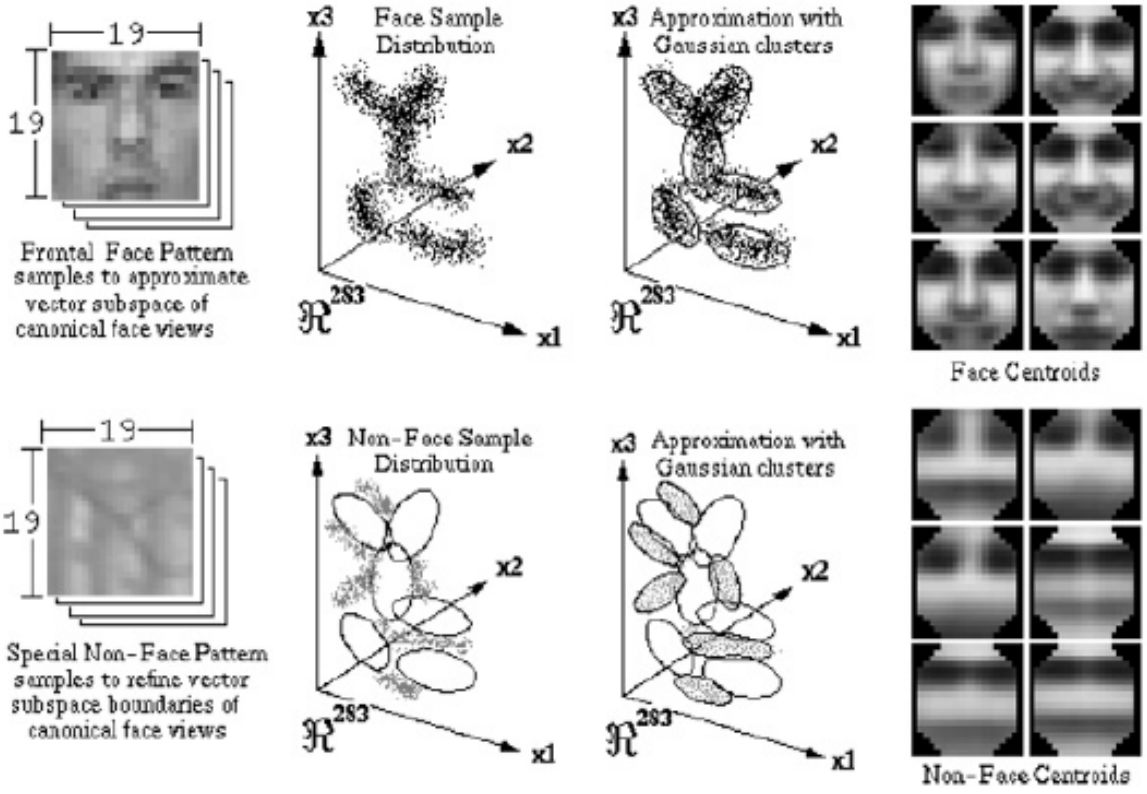
Για να ταυτοποιήσουμε ένα καινούργιο πρότυπο, η εικόνα αρχικά περνάει από μία προεργασία όπως περιγράφεται στο πρώτο βήμα της διαδικασίας και στην συνέχεια το διάνυσμα μετρήσεων εξάγεται από αυτή. Τελικά το MLP αναλαμβάνει να διαπιστώσει κατά πόσο το διάνυσμα μετρήσεων που έχει προκύψει ανήκει σε πρόσωπο η όχι.



Σχήμα 5: Τα βήματα προ επεξεργασίας τα οποία γίνονται στις εικόνες κατά το πρώτο βήμα της μεθόδου που παρουσιάστηκε παραπάνω. Αρχικά εφαρμόζεται μία οβάλ μάσκα στην εικόνα με σκοπό να αφαιρεθούν κάποια περιττά εικονοστοιχεία ενώ ταυτόχρονα να μειωθούν και οι διαστάσεις του διανύσματος εικόνας που προκύπτει. Στην συνέχεια αναζητείται ένα επίπεδο φωτεινότητας το οποίο να ταιριάζει καλύτερα και με την σειρά του αφαιρείται το επίπεδο αυτό από τα εικονοστοιχεία της εικόνας. Στο τέλος εφαρμόζεται ιστογραμμική κανονικοποίηση στην εικόνα με σκοπό να διορθωθούν τα διαφορετικά κέρδη τα οποία προέρχονται από την χρήση διαφορετικών καμερών και να βελτιωθεί η αντίθεση της εικόνας.

2.4.1 Αποτελέσματα

Οι Sung και Poggio δοκίμασαν το σύστημα τους σε δυο διαφορετικές βάσεις προσώπων. Αναφέρουν ρυθμούς εντοπισμού της τάξης του 96.3% με 3 εσφαλμένους εντοπισμούς στην πρώτη βάση, η οποία περιέχει 301 εμπρόσθιες και σχεδόν εμπρόσθιες όψεις προσώπων 71 διαφορετικών ανθρώπων. Στο δεύτερο σύνολο εικόνων (MIT υποσύνολο από την βάση MIT+CMU) αναφέρουν ποσοστό εντοπισμού της τάξης του 79.9% με 5 εσφαλμένους εντοπισμούς. Τέλος να αναφέρουμε πως δεν γίνεται καμία αναφορά στην ταχύτητα του συστήματος.

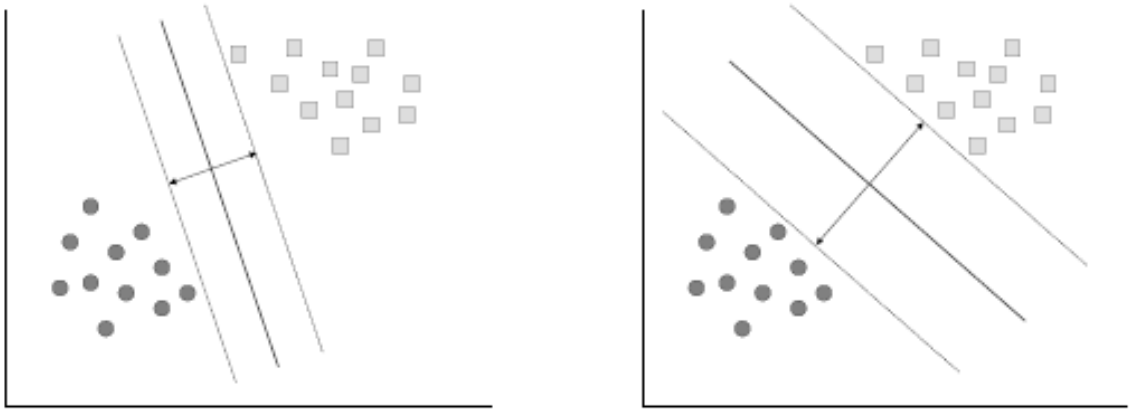


Σχήμα 6: Το βασισμένο σε κατανομές σύστημα το οποίο έχει προταθεί από τους Sung και Poggio . Η πρώτη γραμμή δείχνει μία εμπειρική κατανομή προτύπων προσώπων. Μοντελοποιούν αυτή την κατανομή με χρήση 6 πολυδιάστατων συστοιχιών Gauss των οποίων τα κέντρα φαίνονται στα δεξιά. Η κάτω γραμμή δείχνει την κατανομή των όχι προσώπων τα οποία με την σειρά τους μοντελοποιούνται με αντίστοιχες συστοιχίες. Συνολικά το μοντέλο αποτελείται από 12 συστοιχίες των 283 διαστάσεων.

2.5. Μηχανές διανυσμάτων στήριξης

Οι μηχανές διανυσμάτων στήριξης (Support Vector Machines, SVMs) αποτελούν έναν αλγόριθμο ταξινόμησης προτύπων, ο οποίος αναπτύχθηκε απο τους Vapnik et. Al. Οι περίσσοτεροι μέθοδοι κατηγοριοποίησης που βασίζονται στην εκπαίδευση μηχανής (machine learning), βασίζονται στην αρχή της ελαχιστοποίησης του σφάλματος στα δεδομένα εκπαίδευσης. Η αρχή αυτή ονομάζεται ελαχιστοποίηση εμπειρικού ρίσκου. Σε αντίθεση με την παραπάνω αρχή, οι μηχανές διανυσμάτων στήριξης λειτουργούν σε μία άλλη επαγωγική αρχή, η οποία αποκαλείται δομική ελαχιστοποίηση ρίσκου. Αυτή ελαχιστοποιεί κάποιο ανώτερο φράγμα στο σφάλμα γενικοποιήσης.

Για την κατηγοριοποίηση οι μηχανές διανυσμάτων στήριξης λειτουργούν αναζητώντας ένα υπερεπίπεδο στον χώρο των πιθανών εισόδων. Αυτό το υπερεπίπεδο επιχειρεί να διαχωρίσει τα θετικά παραδείγματα απο τα αρνητικά. Το υπερεπίπεδο αυτό επιλέγεται έτσι ώστε να έχει ως βασικό χαραχτηριστικό την μέγιστη απόσταση απο το κοντινότερο θετικό και αρνητικό παράδειγμα όπως φαίνεται και στο **Σχήμα 7**. Διαισθητικά με αυτό τον τρόπο η κατηγοριοποίηση γίνεται σωστά για δεδομένα ελέγχου τα οποία είναι κοντά, αλλά όχι ταυτόσημα με τα δεδομένα εκπαίδευσης.



(a) Small distance. (b) Maximum distance

Σχήμα 7: Διαχωριστικό υπέρ επίπεδο για το SVM . Στην περίπτωση (a) το υπέρ επίπεδο το οποίο επιλέχθηκε έχει μικρή απόσταση απο το κοντινότερο απο τα θετικά και αρνητικά παραδείγματα, για αυτό τον λόγο έχει χειρότερη ικανότητα γενίκευσης. Στην περίπτωση (b) το υπέρ επίπεδο έχει μέγιστη απόσταση μεταξύ των δύο κλάσεων με αποτέλεσμα να αναμένεται μεγαλύτερη ικανότητα γενίκευσης.

Οι μηχανές διανυσμάτων στήριξης χρησιμοποιήθηκαν για πρώτη φορά στο πρόβλημα του εντοπισμού προσώπων απο τους Osuna et. Al. . Εκπαίδευσαν την μηχανή διανυσμάτων στήριξης χρησιμοποιώντας μία βάση δεδομένων απο πρόσωπα και όχι-πρόσωπα μεγέθους 19x19 εικονοστοιχείων. Το προτεινόμενο σύστημα χρησιμοποιούσε βήματα προ επεξεργασίας ίδια με αυτά του συστήματος των Sung and Poggio όπως αυτά παρουσιάζονται στο **Σχήμα 5**. Με αυτό τον τρόπο επιτυγχάνεται η μείωση των διαστάσων του χώρου εισόδου και βελτιώνεται η ποιότητα των εικόνων. Επίσης έκαναν και χρήση της τεχνικής bootstrapping την οποία και θα αναφέρουμε παρακάτω.

2.5.1 Τεχνική BootStrapping

Στην περίπτωση του προβλήματος εντοπισμού προσώπων παρουσιάζεται μία βασική δυσκολία. Τα αρνητικά παραδείγματα είναι δύσκολο να χαραχτηριστούν ενώ ταυτόχρονα η κλάση των όχι – προσώπων είναι πολύ μεγαλύτερη απο αυτή των προσώπων. Για αυτό τον λόγο απαιτούνται πολύ περισσότερα παραδείγματα όχι – προσώπων παρα προσώπων για την σωστή

διαχωροποίηση των κλάσεων. Για να επιτύχουμε αυτό το αποτέλεσμα κάνουμε χρήση της τεχνικής bootstrapping. Αυτή η μέθοδος χρησιμοποιεί εικόνες οι οποίες δεν περιέχουν καθόλου πρόσωπα. Σε κάθε επανάληψη της εκπαίδευσης οι λάθος κατηγοριοποιήσεις σε αυτές τις εικόνες γίνονται μέρος του συνόλου παραδειγμάτων εκπαίδευσης για την επόμενη επανάληψη. Με αυτό τον τρόπο καταφέρνουμε να αποφύγουμε την χρήση τεράστιων συνόλων απο παραδείγματα που δεν είναι πρόσωπα κατα την εκπαίδευση, πολλά απο τα οποία μπορεί να μην αποτελούν καν σημαντικά παραδείγματα. Καλές περιπτώσεις πηγών μή προσώπων είναι εικόνες με τοπία, κτίρια, δένδρα κοκ. Κυρίως λόγω των πολλών και διαφορετικών προτύπων που περιέχουν .

2.5.2 Συμπεράσματα

Οι Osuna et. Al. δοκίμασαν το SVM σε δύο διαφορετικές βάσεις προσώπων. Η πρώτη περιέχει 313 εικόνες υψηλής ποιότητας με ένα πρόσωπο σε κάθε εικόνα. Το δεύτερο σύνολο περιέχει 23 εικόνες σε μεταβλητή ποιότητα με σύνολο 155 προσώπων. Για λόγους σύγκρισης χρησιμοποίησαν και το σύστημα των Sung και Poggio στις ίδιες βάσεις. Συνολικά 4 669 960 τμήματα που εξήχθησαν απο τις παραπάνω εικόνες δοκιμάστικαν απο το σύστημα για την πρώτη βάση, ενώ για την δεύτερη συνολικά δοκιμάστικαν 5 383 682 πλαίσια. Ο ρυθμός εντοπισμού και το πλήθος των εσφαλμένων θετικών εντοπισμών παρουσιάζεται στον **Πίνακας 1**. Η μηχανή διανυσμάτων στήριξης λειτούργησε καλύτερα αν παρατηρήσουμε το τμήμα του ρυθμού εντοπισμού του πίνακα, και ελάχιστα χειρότερα στην περίπτωση του πλήθους των εσφαλμένων θετικών εντοπισμών. Οι συγγραφείς αναφέρουν τέλος πως το δικό τους σύστημα ήταν 30 φορές πιό γρήγορο απο αυτό των Sung και Poggio.

	Database 1		Database 2	
	DR	FP	DR	FP
SVM	97.1%	4	74.2%	20
Sung *et. al.*	94.6%	2	74.2%	11

Πίνακας 1: Απόδοση του συστήματος με χρήση μηχανής διανυσμάτων στήριξης συγκριτικά με τα αποτελέσματα του συστήματος των Sung και Poggio.

2.6. Νευρωνικά Δίκτυα

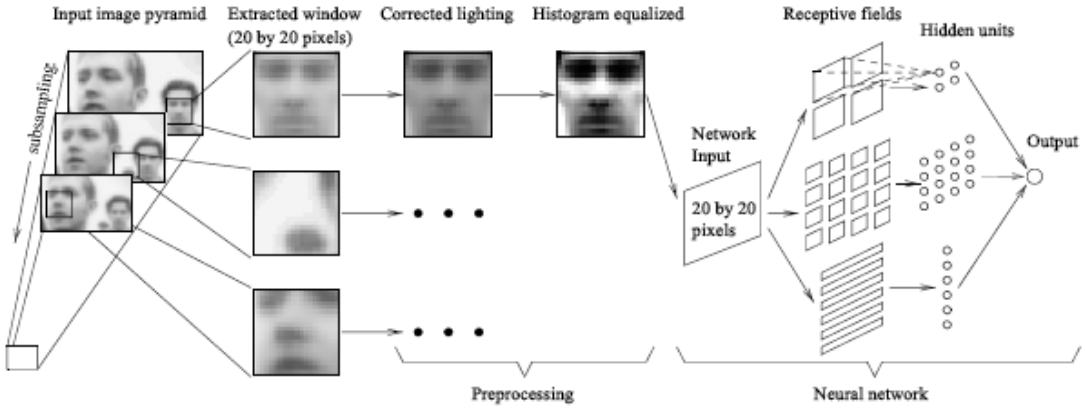
Το πρόβλημα του εντοπισμού προσώπων μπορεί να θεωρηθεί ως ένα δυαδικό πρόβλημα αναγνώρισης προτύπων. Για αυτό τον λόγο πολλές αρχιτεκτονικές νευρωνικών δικτύων έχουν προταθεί κατα καιρούς στην βιβλιογραφία. Ίσως η πιο σημαντική απο αυτές είναι η δουλεία που έγινε απο τους Rowley et. Al. . Η αρχιτεκτονική τους αποτελείται απο τρεία στάδια, την προ επεξεργασία, το νευρωνικό δίκτυο και τέλος την μετά – επεξεργασία.

Η μέθοδος προ – επεξεργασίας που χρησιμοποιήθηκε σε αυτή την εργασία είναι η ίδια με αυτή που αναφέραμε προηγουμένως απο τους Sung και Poggio και η οποία περιγράφεται στο **Σχήμα 5**. Οι πρό – επεξεργασμένες εικόνες διοχετεύονται στην συνέχεια στο νευρωνικό δίκτυο. Το δίκτυο αυτό ειναι σχεδιασμένο να επεξεργάζεται εικόνες διαστάσεων 20 x 20 εικονοστοιχείων, με αποτέλεσμα να έχουμε συνολικά 400 νευρώνες εισόδου. Υπάρχει ένα κρυφό στρώμα νευρώνων το οποίο αποτελείται απο 26 συνολικά νευρώνες. Οι 4 απο αυτούς επεξεργάζονται υποπεριοχές των 10 x 10 εικονοστοιχείων, 16 νευρώνες επεξεργάζονται υποπεριοχές των 5 x 5 και τέλος 6 επεξεργάζονται οριζόντιες λωρίδες διαστάσεων 20 x 5 εικονοστοιχείων όπως παρουσιάζεται και στο **Σχήμα 8**.

Η εκπαίδευση του συστήματος έγινε με χρήση 1050 παραδειγμάτων προσώπων διαφόρων μεγεθών, πυκνότητας, προσανατολισμού και θέσης. 1000 παραδείγματα εικόνων που δεν περιείχαν πρόσωπα χρησιμοποιήθηκαν επίσης στην εκπαίδευση και οι οποίες πάρθηκαν τυχαία απο διάφορες εικόνες που δεν περιείχαν καθόλου πρόσωπα. Το δίκτυο όταν του δινόταν ως είσοδος ένα άγνωστο απο το σύνολο εκπαίδευσης πρότυπο, επέστρεφε ως έξοδο μία τιμή η οποία ήταν μεταξύ -1 (όχι πρόσωπο) και 1 (πρόσωπο).

Το τελευταίο στάδιο της μετά – εκπαίδευσης μετρούσε τους εντοπισμούς σε μία μικρή περιοχή και στην περίπτωση που ξεπερνούσαν ένα συγκεκριμένο κατώφλι τότε θεωρούσε πιθανό το να υπάρχει σε αυτή την περιοχή πρόσωπο. Επίσης σε αυτό το τμήμα του συστήματος τα διάφορα τετράγωνα εντοπισμού που είχαν βρεθεί σε κάποια μικρή περιοχή και τα οποία το ένα υπερκάλυπτε το άλλο ενωνώντουσαν σε ένα κεντρικό τετράγωνο που περιείχε το πρόσωπο. Για να βελτιώσουν την απόδοση του συστήματος οι ερευνήτες εκπαίδευσαν διάφορα νευρωνικά δίκτυα και συνδίασαν τις εξόδους τους με κάποια απλά συστήματα διαιτησίας όπως για παράδειγμα με χρήση λογικών τελεστών (Και / Η) και ψήφου.

Ένα βασικό μειονέκτημα του συστήματος είναι πως περιορίζεται στον εντοπισμό μόνο εμπρόσθιων προσώπων , παρόλα αυτά οι συγγραφείς επέκτηναν το σύστημα τους για να καλύψουν και την περίπτωση προσώπων τα οποία εμφανίζονται υπό γωνία στο επίπεδο της φωτογραφίας.



Σχήμα 8: Το σύστημα εντοπισμού προσώπων με χρήση νευρωνικών δικτύων των Rowley et. Al.

2.6.1 Αποτελέσματα

Στην περίπτωση των εμπρόσθιων προσώπων οι Rowley et. Al. υποστηρίζουν πως πέτυχαν ρυθμούς αναγνώρισης μεταξύ 77.9% και 90.3% σε πρόσωπα που άνηκαν σε σύνολο 130 δοκιμαστικών εικόνων, επιτυγχάνοντας ταυτόχρονα ικανοποιητικό ρυθμό εσφαλμένων εντοπισμών. Ανάλογα την εφαρμογή υποστηρίζουν πως το σύστημα μπορεί να γίνει λιγότερο η περισσότερο ακριβές με την μετατροπή των κατωφλιών και των παραμέτρων του συστήματος διαιτησίας.

Το σύστημα δοκιμάστηκε σε μεγάλη ποικιλία εικόνων ως προς τα ποιοτικά τους χαρακτηριστικά, με πολλά πρόσωπα καθώς και με διάφορα φόντα. Μία γρήγορη έκδοση του συστήματος μπορεί να επεξεργασθεί μία εικόνα διαστάσεων 320 x 240 εικονοστοιχείων μεταξύ 2 και 4 δευτερολέπτων σε έναν σταθμό εργασίας SGI Indigo 2 R4400 των 200 MHz.

2.7. Ανιχνευτής προσώπων των Viola και Jones

Οι Paul Viola και Michael Jones στο [27] περιγράφουν μία πρωτοποριακή μέθοδο εντοπισμού αντικειμένων σε εικόνες, η οποία μπορεί να εφαρμοσθεί και στην περίπτωση του εντοπισμού προσώπων. Ο Αλγόριθμος τους βασίζεται στην χρήση κυλιώμενου πλαισίου όπως και άλλοι οι οποίοι αναφέρθηκαν προηγουμένως ενώ κάνει χρήση φίλτρων αντί να λειτουργεί στα εικονοστοιχεία απευθείας όπως λειτουργούσαν στις γενικές τους περιπτώσεις οι περισότεροι άλλοι αλγόριθμοι. Τα φίλτρα αυτά καλύπτουν ορθογώνιες περιοχές της υπό εξέταση εικόνας στο εσωτερικό του παραθύρου εντοπισμού, ενώ ταυτόχρονα ο υπολογισμός τους είναι εφικτός σε πολύ μικρούς χρόνους χωρίς ιδιαίτερες απαιτήσεις υπολογιστικής ισχύος, αυτό επιτυγχάνεται με μία ιδιαίτερη αναπαράσταση της εικόνας η οποία αποκαλείται ολοκληρωτική εικόνα (integral image).

Ο αριθμός των πιθανών φίλτρων που μπορούν να εφαρμοστούν σε μία εικόνα είναι πάρα πολύ μεγάλος. Για να μπορέσουμε να επιλέξουμε μόνο αυτά τα οποία είναι τα πλέον αποδοτικά, ο ανιχνευτής αντικειμένων εκπαιδεύεται με χρήση μίας τροποποιημένης μεθόδου εκπαίδευσης της κατηγορίας AdaBoost, μέθοδος η οποία πρωτοαναφέρθηκε απο τους Freund και Schapire . Ο συγκεκριμένος ενισχυτικός αλγόριθμος (boosting algorithm) επιλέγει ένα υποσύνολο απο ταξινομητές (classifiers) οι οποίοι αποκαλούντε ασθενείς ταξινομητές και οι οποίοι ουσιαστικά αποτελούνται απο ένα ένα απο τα φίλτρα τα οποία προαναφέραμε. Η επίλογη αυτή γίνεται από ένα μεγάλο σύνολο πιθανών ταξινομητών. Το βασικό χαρακτηριστικό είναι το ότι κανένας απο αυτούς τους ταξινομητές μόνος του δεν είναι ικανός να να αποδώσει μεγάλο βαθμό σωστής ταξινόμησης, παρόλα αυτά στην περίπτωση που αυτοί οι ταξινομητές συνδιασθούν μεταξύ τους, προκύπτει ένας ισχυρός ταξινομητής με πολύ καλά αποτελέσματα.

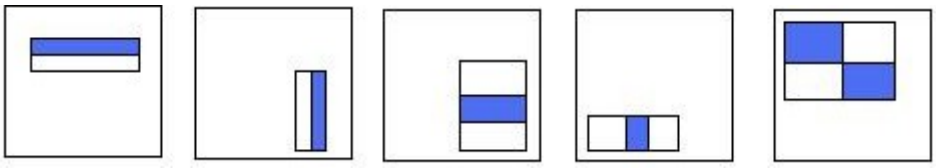
Για να επιτύχουν ακόμα μεγαλύτερη μείωση των ασθενών ταξινομητών ανά παράθυρο εντοπισμού, οι συγγραφείς προτείνουν την χρήση μίας διάταξης εντοπισμού βασισμένη στην προσοχή (attentional cascade). Αυτή η διάταξη συνδιάζει επιτυχώς πιό σύνθετους ταξινομητές στο σχήμα της. Με χρήση λοιπόν αυτής της προσέγγισης που βασίζεται σε στρώματα και η οποία χαρακτηρίζεται απο αυξημένη πολυπλοκότητα, είναι δυνατή η απόριψη παραθύρων εντοπισμού τα οποία δεν περιέχουν πρόσωπα πολύ γρήγορα, ενώ ταυτόχρονα να αποδίδεται μεγαλύτερη υπολογιστική ισχύς σε παράθυρα εντοπισμού τα οποία είναι πιό πιθανό να περιέχουν κάποιο πρόσωπο.

Στην συνέχεια θα αναφερθούμε με μεγαλύτερη λεπτομέρεια στην συγκεκριμένη μέθοδο μιάς και αποτελεί και την μέθοδο που τελικά επιλέχθηκε για την παρούσα διπλωματική εργασία, ενώ στο τέλος θα αναφερθούμε στα συμπεράσματα και στην σύγκριση μεταξύ της συγκεκριμένης μεθόδου και άλλων οι οποίες δοκιμάστικαν στα πλαίσια αυτής της εργασίας.

2.7.1 Φίλτρα

Οι Viola και Jones αποφάσισαν να χρησιμοποιήσουν φίλτρα έναντι μή επεξεργασμένων εικονοστοιχείων για τους παρακάτω λόγους. Τα φίλτρα τα οποία χρησιμοποιήθηκαν καλύπτουν μεγαλύτερη χωρική περιοχή το οποίο τους δίνει την δυνατότητα να εξάγουν πληροφορίες οι οποίες χαρακτηρίζουν μεγαλύτερες δομές,ακμές, γωνίες και γραμμές στο εσωτερικό του παραθύρου εντοπισμού. Επίσης τα φίλτρα αυτά μπορούν να υπολογισθούν πολύ γρήγορα αν γίνει χρήση της συγκεκριμένης αναπαράστασης της εικόνας που προ αναφέραμε. Οι συγγραφείς χρησιμοποίησαν τα πέντε απλά διαδικά φίλτρα τα οποία παρουσιάζονται στο **Σχήμα 9**. Αυτά τα φίλτρα εν μέρη βασίζονται στην εργασία του Παπαγεωργίου [7], ο οποίος χρησιμοποίησε συναρτήσεις με βάση Haar για τον εντοπισμό προσώπων και πεζών σε εικόνες [23]. Τα φίλτρα περιέχουν δύο, είτε τρείς η τέσσερις ορθογώνιες περιοχές. Αυτές οι περιοχές είναι ίσες σε μέγεθος και σχήμα και είναι τοποθετημένες είτε οριζόντια είτε κάθετα.

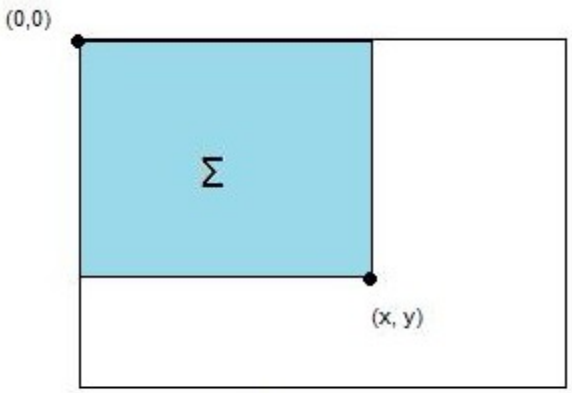
Στην περίπτωση που χρησιμοποιηθεί κάποιο παράθυρο εντοπισμού διαστάσεων 24 x 24, τότε υπάρχουν παραπάνω απο 180000 διαφορετικά ορθογώνια φίλτρα στο εσωτερικό του. Ο υπολογισμός όλων αυτών των φίλτρων ανά παράθυρο εντοπισμού θά ήταν μία υπερβολικά κοστοβόρα διαδικασία. Για αυτό τον λόγο μόνο ένα μικρό υποσύνολο αυτών των φίλτρων χρειάζεται να επιλεγούν για να δημιουργηθεί ένας ισχυρός ταξινομητής.



Σχήμα 9 Τα πέντε Haar Φίλτρα τα οποία χρησιμοποιήθηκαν απο τους Viola και Jones τοποθετημένα στο παράθυρο εντοπισμού. Για να υπολογισθεί ένα απο τα φίλτρα, το άθροισμα των εικονοστοιχείων στην σκιασμένη περιοχή αφαιρούνται απο το άθροισμα των εικονοστοιχείων της μή σκιασμένης περιοχής. Θα μπορούσαμε να κάνουμε και το αντιθέτο, δηλαδή να αφαιρέσουμε το άθροισμα της μή σκιασμένης περιοχής απο αυτό της σκιασμένης, τότε η μόνη διαφορά που θα υπήρχε είναι το πρόσημο του αποτελέσματος. Τα δύο πρώτα φίλτρα που παρουσιάζονται απο τα αριστερά προς τα δεξιά τείνουν στον εντοπισμό ακμών, οριζόντιων και καθέτων αντίστοιχα. Τα άλλα δύο φίλτρα τείνουν στον εντοπισμό γραμμών. Τέλος το τελευταίο φίλτρο τείνει στον εντοπισμό διαγωνίων γραμμών .

2.7.2 Ολοκληρωτική αναπαράσταση εικόνας (Integral Image)

Οι συγγραφείς προτείνουν την χρήση μίας ειδικής αναπαράστασης της εικόνας η οποία αποκαλείται, ολοκληρωτική αναπαράσταση εικόνας (Integral Image), για τον υπολογισμό των ορθογώνιων φίλτρων. Για την ακρίβεια η αναπαράσταση αυτή είναι αντίστοιχη του πίνακα αθροιστικής περιοχής (Summed Area Table SAT) ο οποίος χρησιμοποιείται στην περίπτωση της τεχνικής ταυτοποίησης υφής όπως πρώτοχρησιμοποιήθηκε απο τον Crow . Η μετονομασία της αναπαράστασης έγινε εσκεμένα για να μπορέσει να υπάρχει διάκριση ως προς την χρήση της, ταυτοποίηση υφής έναντι ανάλυσης εικόνας.



(a) Integral Image Value at point (x,y)

Σχήμα 10: Σχηματική αναπαράσταση του υπολογισμού της ολοκληρωτικής εικόνας για συγκεκριμένο σημείο.

Original Image

4 5 9
2 3 6

Integral Image

4 5 14
2 14 29

(b) Integral Image in numbers

Σχήμα 11: Παράδειγμα υπολογισμού ολοκληρωτικής εικόνας με συγκεκριμένες τιμές οικονοστοιχείων.

Η ολοκληρωτική τιμή μίας εικόνας στην περιοχή (x,y) ορίζεται ως το άθροισμα των τιμών όλων των εικονοστοιχείων απο πάνω και αριστερά του (x,y) όπως παρουσιάζεται και στο **Σχήμα 10**.

$$ii(x,y) = \sum_{j=0}^{x}\sum_{k=0}^{y} I(j,k) \quad (1.1)$$

Όπου το $ii(x,y)$ είναι η τιμή της ολοκληρωτικής εικόνας στο σημείο (x,y) και το I(x,y) είναι η τιμή της αρχικής εικόνας. Η παραπάνω εξίσωση μπορεί να ξαναγραφεί με χρήση των παρακάτω αναδρομικών σχέσεων.

$$r(x,y) = r(x,y-1) + I(x,y) \quad (1.2)$$

$$ii(x,y) = ii(x-1,y) + r(x,y) \quad (1.3)$$

Όπου το r(x,y) ονομάζεται συσωρευτικό άθροισμα στήλης, r(x,-1) = 0, ii(-1,y) = 0 και ii(x,-1) = 0. Με χρήση αυτών των αναδρομικών σχέσεων η ολοκληρωτική εικόνα μπορεί να υπολογισθεί με μόνο ένα πέρασμα της αρχικής εικόνας.

2.7.3 Υπολογισμός ορθογώνιου αθροίσματος εικονοστοιχείων.

Το άθροισμα των εικονοστοιχείων μίας ορθογώνιας περιοχής ορίζεται ώς,

$$pixelsum(x,y,w,h) = \sum_{j=x}^{x+w-1}\sum_{k=y}^{y+h-1} I(j,k)$$

Όπου w και h ειναι το μήκος και το πλάτος αντίστοιχα της οριζόμενης περιοχής στο σημείο (x,y) και I(j,k) είναι η τιμή της εικόνας στο σημείο (i,j). Στην συνέχεια θα δείξουμε πως η τιμή αυτή μπορεί να υπολογισθεί γρήγορα με χρήση της ολοκληρωτικής μορφής εικόνας. Η τιμή της ολοκληρωτικής μορφής εικόνας στο χαμηλώτερο δεξί σημείο μίας ορθογώνιας περιοχής μπορεί να ορισθεί ως εξής,

$$ii(x+w-1,y+h-1) = \sum_{j=o}^{x+w-1}\sum_{k=0}^{y+h-1} I(j,k)$$

Η παραπάνω εξίσωση μπορεί να μετατραπεί στους παρακάτω 4 όρους.

$$ii(x+w-1,y+h-1)=\sum_{j=o}^{x-1}\sum_{k=0}^{y-1}I(j,k)+\sum_{j=0}^{x-1}\sum_{k=y}^{y+h-1}I(j,k)+\sum_{j=x}^{x+w-1}\sum_{k=0}^{y-1}I(j,k)+\underbrace{\sum_{j=x}^{x+w-1}\sum_{k=y}^{y+h-1}I(j,k)}_{pixelsum}$$ Αν

παρατηρήσουμε τους παραπάνω 4 όρους θα διαπιστώσουμε πώς ο τελευταίος όρος ισούται με το άθροισμα των εικονοστοιχείων. Για να μπορέσουμε να πάρουμε αυτή την τιμή πρέπει να αφαιρέσουμε τους τρείς υπόλοιπους όρους. Μπορούμε να το επιτύχουμε αυτό με την χρήση των τρειών παρακάτω ολοκληρωτικών μορφών εικόνας όπως φαίνεται στο **Σχήμα 12**.

$$ii(x-1,y-1)=\sum_{j=0}^{x-1}\sum_{k=0}^{y-1}I(j,k)$$

$$ii(x+w-1,y-1)=\sum_{j=o}^{x+w-1}\sum_{k=0}^{y-1}I(j,k)$$

$$ii(x-1,y+h-1)=\sum_{j=0}^{x-1}\sum_{k=0}^{y+h-1}I(j,k)$$

Οι παραπάνω σχέσεις μπορούν να μετατραπούν στις παρακάτω ισοδύναμες μορφές,

$$ii(x-1,y-1)=\sum_{j=0}^{x-1}\sum_{k=0}^{y-1}I(j,k)$$

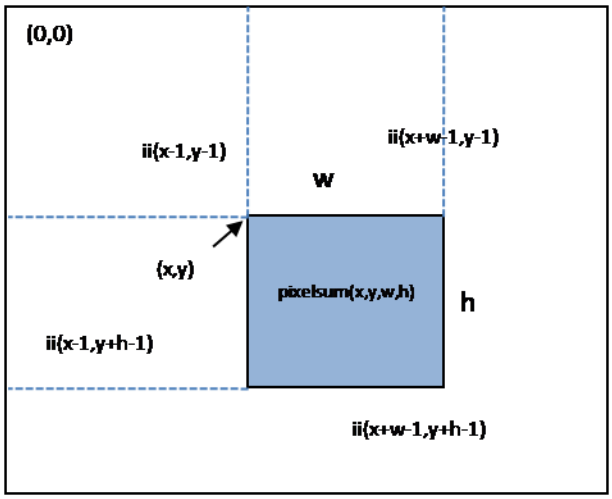
$$ii(x+w-1,y-1)=ii(x-1,y-1)+\sum_{j=x-1}^{x+w-1}\sum_{k=0}^{y-1}I(j,k)$$

$$ii(x-1,y+h-1)=ii(x-1,y-1)+\sum_{j=0}^{x-1}\sum_{k=y}^{y+h-1}I(j,k)$$

Αν χρησιμοποιήσουμε τις παραπάνω τρείς εξισώσεις μπορούμε εύκολα να καταλήξουμε στην παρακάτω σχέση,

$$pixelsum(x,y,w,h)=ii(x+w-1,y+h-1)+ii(x-1,y-1)-ii(x-1,y+h-1)-ii(x+w-1,y-1)$$

Συμπερένουμε λοιπόν πως με την χρήση της ολοκληρωτικής μορφής εικόνας κάθε ορθογώνιο άθροισμα εικονοστοιχείων μπορεί να υπολογιστεί με 4 αναζητήσεις, 2 αφαιρέσεις και μία πρόσθεση.



Σχήμα 12:. Υπολογισμός του αθροίσματος των εικονοστοιχείων με την χρήση ολοκληρωτικής αναπαράστασης εικόνας. Χρησιμοποιώντας 4 τιμές της ολοκληρωτικής αναπαράστασης ii(x+w-1,y+h-1), ii(x-1,y-1),ii(x-1,y+h-1) και ii(x+w-1,y-1) μπορούμε να υπολογίσουμε το άθροισμα των εικονοστοιχείων στην σκιασμένη ορθογώνια περιοχή.

2.7.4 Υπολογισμός των φίλτρων

Τα 5 φίλτρα τα οποία πρωτάθηκαν απο τους συγγραφείς αποτελούνται απο 2 ή περισσότερες ορθογώνιες περιοχές οι οποίες πρέπει να προστεθούν ή να αφαιρεθούν μεταξύ τους. Το ένα απο αυτά τα φίλτρα, το φίλτρο οριζόντιων ακμών παρουσιάζεται στο **Σχήμα 13**. Για να υπολογίσουμε το αποτέλεσμα H_{h_edge}της εφαρμογής του φίλτρου F_{h_edge}στην εικόνα I, θα πρέπει το άθροισμα των εικονοστοιχείων της σκιασμένης περιοχής να αφαιρεθούν απο αυτά της μή σκιασμένης περιοχής όπως φαίνεται στο **Σχήμα 13**

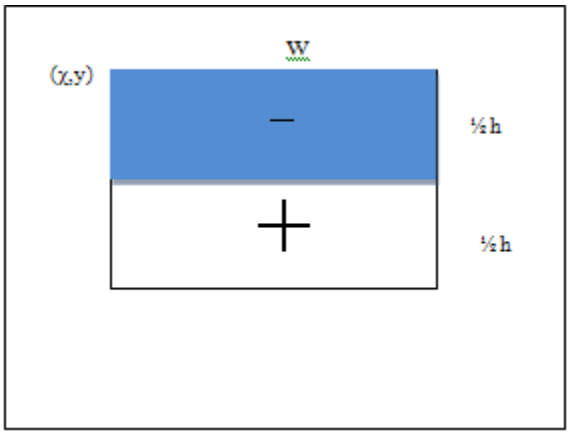
$$H_{h\_edge}(x,y)=\sum_{j=x}^{x+w-1}\sum_{k=y+\frac{1}{2}h}^{y+h-1}I(j,k)-\sum_{j=x}^{x+w-1}\sum_{k=y}^{y+\frac{1}{2}h-1}I(j,k)$$

Όπου w και h είναι το μήκος και πλάτος αντίστοιχα του φίλτρου. Όταν χρησιμοποιούμε την ολοκληρωτική μορφή εικόνας μπορούμε να χρησιμοποιήσουμε το άθροισμα των

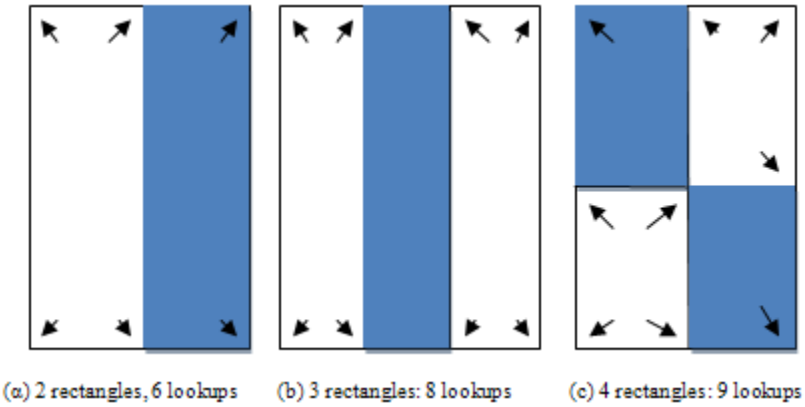
εικονοστοιχείων όπως αναφέραμε προηγουμένως και να ξαναγράψουμε την μορφή του φίλτρου ως εξής,

$$H_{h\_edge}(x,y) = pixelsum\left(x, y+\frac{1}{2}h, w, \frac{1}{2}h\right) - pixelsum\left(x, y, w, \frac{1}{2}h\right)$$

Με χρήση των αθροισμάτων εικονοστοιχείων το αποτέλεσμα της εφαρμογής του φίλτρου μπορεί εύκολα να υπολογισθεί. Εφόσον το φίλτρο αποτελείται απο δύο παράπλευρες ορθογώνιες περιοχές, υπάρχουν δύο ίσα σημεία της ολοκληρωτικής μορφής εικόνας και μόνο έξι αναζητήσεις απαιτούνται για τον υπολογισμό της εφαρμογής του φίλτρου. Οι αναζητήσεις για τα υπόλοιπα φίλτρα παρουσιάζονται στο **Σχήμα 14**. Με αυτή την μέθοδο τα αποτελέσματα της εφαρμογής των φίλτρων μπορεί να υπολογισθεί πολύ γρήγορα και σε σταθερό χρόνο, ανεξάρτητα απο το μέγεθος της περιοχής. Αντί να δημιουργούνται πυραμίδες της εικόνας για να μεταβάλουμε τις διαστάσεις της, όπως είδαμε σε προηγούμενους αλγορίθμους, οι Viola και Jones, προτάσουν την αντίστοιχη αλλαγή μεγέθους του παραθύρου αναζήτησης. Το παράθυρο αυτό περιέχει μόνο ένα απλό φίλτρο το οποίο μπορεί εύκολα να αλάξει διαστάσεις με χρήση συμπερασμού (extrapolation) του μεγέθους και της θέσης του. Με αυτό τον τρόπο επιτυγχάνουμε μεγάλη οικονομία υπολογισμών αφού σε κάθε άλλη περίπτωση θα έπρεπε να υπολογίσουμε πυραμίδες Gauss οι οποίες απαιτούν σύνθετους υπολογισμούς.



Σχήμα 13: Υπολογισμός του αποτελέσματος της εφαρμογής του φίλτρου H_{h_edge}



Σχήμα 14: Αριθμός αναζητήσεων ανά φίλτρο. Ένα φίλτρο με 2 ορθογώνιες περιοχές περιέχει 2 ίσα σημεία της ολοκληρωτικής μορφής εικόνας και για αυτό τον λόγο χρειάζεται 6 αναζητήσεις. Αντίστοιχα το φίλτρο με 6 ορθογώνιες περιοχές χρειάζεται 8 αναζητήσεις και τέλος το φίλτρο με 4 ορθογώνιες περιοχές χρειάζεται 9 αναζητήσεις.

2.7.5 Κανονικοποίηση εικόνας.

Οι Viola Και Jones στην εργασία τους κανονικοποιούν τις εικόνες σε μοναδιαία διακύμανση (variance) έτσι ώστε κατά την διάρκεια της εκπαίδευσης να ελαχιστοποιηθεί η επίδραση των διαφορετικών συνθηκών φωτισμού. Για την περίπτωση της κανονικοποίησης της εικόνας κατα την διάρκεια του εντοπισμού, πολλαπλασιάζουν εκ των υστέρων τα αποτελέσματα των φίλτρων με την σταθερή απόκλιση **σ** της εικόνας στο εσωτερικό του παραθύρου εντοπισμού. Για να υπολογίσουν την σταθερή απόκλιση των τιμών των εικονοστοιχείων χρησιμοποιήθηκε η παρακάτω σχέση,

$$\sigma = \sqrt{\mu^2 - \frac{1}{\mathrm{N}}\sum \chi^2} \quad (1.4)$$

Όπου **μ** είναι ο μέσος, **χ** η τιμή του εικονοστοιχείου και **N** ο συνολικός αριθμός των εικονοστοιχείων στο εσωτερικό του παραθύρου εντοπισμού. Ο μέσος μπορεί να υπολογισθεί με χρήση της ολοκληρωτικής μορφής εικόνας. Για να συμβεί αυτό οι Viola και Jones χρησιμοποίησαν μία τετραγωνική ολοκληρωτική μορφή εικόνας, η οποία είναι μία

ολοκληρωτική μορφή εικόνας μόνο με τετραγωνικές τιμές. Με αυτό τον τρόπο ο υπολογισμός του **σ** απαιτεί μόνο 8 αναζητήσεις και μερικές εντολές.

2.7.6 Επιλογή των φίλτρων με χρήση του AdaBoost

Όπως προαναφέραμε υπάρχουν πάνω απο 180 000 πιθανά φίλτρα τα οποία θα μπορούσαμε να υπολογίσουμε σε ένα μόνο απο τα παράθυρα εντοπισμού. Ο υπολογισμός όλων αυτών θα αποτελούνε μία πολύ δαπανηρή διαδικασία απο πλευράς χρόνου και επεξεργαστικής ισχύος. Για αυτό τον λόγο μόνο ένα μικρό υποσύνολο αυτών των φίλτρων χρειάζεται να επιλεχθεί έτσι ώστε να δημιουργηθεί ένας αποδοτικός ταξινομητής. Για αυτό το έργο οι συγγραφείς χρησιμοποίησαν έναν ενισχυτικό (boosting) αλγόριθμο.Ο αρχικός ενισχυτικός αλγόριθμος ο οποίος αποκαλύται και αλγόριθμος διακριτής ενίσχυσης (discrete boosting algorithm) προτάθηκε απο τους Freund και Schapire . Οι τελευταίοι ανέπτυξαν έναν αλγόριθμο ο οποίος ακολουθιακά ταιριάζει τους αποκαλούμενους ασθενείς ταξινομητές σε δείγματα με διαφορετικό βάρος του συνόλου δεδομένων. Οι ταξινομητές αυτοί όπως και προαναφέρθηκε, αποκαλούνται ασθενείς διότι ακόμα και ο καλύτερος απο αυτούς δεν θα επιτύγχανε καλά αποτελέσματα στην ταξινόμηση ολόκληρου του συνόλου δεδομένων. Σε κάθε κύκλο ενίσχυσης ο ασθενής ταξινομητής με το μικρότερο σφάλμα ταξινόμησης επιλέγεται απο το πλήθος των ασθενών ταξινομητών. Στην συνέχεια τα δείγματα τα οποία ταξινομήθηκαν εσφαλμένα απο τους ασθενείς ταξινομητές αποκτούν μεγαλύτερη βαρύτητα στον επόμενο κύκλο ενίσχυσης. Ο αλγόριθμος επαναλαμβάνεται πολλές φορές, προσθέτοντας μεγαλύτερη βαρύτητα στα δείγματα τα οποία είναι δύσκολο να ταξινομηθούν. Ο τελικός ταξινομητής , ο οποίος αποκαλείται ισχυρός ταξινομητής, αποτελείτα απο το ζυγισμένο (weighted) συνδιασμό των ασθενών ταξινομητών.

Οι Viola και Jones υιοθέτησαν τον αλγόριθμο της διακριτής ενίσχυσης (Discrete AdaBoost) περιορίζοντας το είδος του ασθενούς ταξινομητή με τον περιορισμό να αποτελείται αποκλείστικα και μόνο απο ένα φίλτρο το οποίο ξεχωρίζει με τον καλύτερο δυνατό τρόπο τα θετικά απο τα αρνητικά παραδείγματα. Για κάθε φίλτρο ένα αποδοτικό κατώφλι προσδιορίζεται έτσι ώστε να προκύπτει ο ελάχιστος αριθμός λάθος ταξινομήσεων. Ο ασθενής ταξινομητής αυτός ορίζεται ως εξής,

$$h_j(x) = \begin{cases} 1 & p_j f_j(x) < p_j \Theta_j \\ 0 & otherwise \end{cases}$$

Όπου το $f_j(x)$ είναι το αποτέλεσμα της εφαρμογής του φίλτρου στην εικόνα εισόδου χ, Θ_j είναι το κατώφλι, p_j είναι η ισοτιμία (parity) η οποία δείχνει την κατεύθυνση του συμβόλου ισότητας και χ είναι η εικόνα εισόδου. Ο αλγόριθμος εκμάθησης παρουσιάζεται στον **Πίνακας 2**.

- Με δεδομένα εικόνες παραδείγματα $(x_1, y_1), \ldots, (x_n, y_n)$ όπου $y_i = 0,1$ για αρνητικά και θετικά παραδείγματα αντίστοιχα.
- Αρχικοποίησε τα βάρη $w_{1,i} = \frac{1}{m}, \frac{1}{2l}$ για $y_i = 0,1$ αντίστοιχα, όπου m και l ο αριθμός των αρνητικών και θετικών αντίστοιχα παραδειγμάτων.
- Για $t = 1, \ldots, T$:
    1. Κανονικοποίησε τα βάρη,

$$w_{t,i} \leftarrow \frac{w_{t,i}}{\sum_{j=1}^{n} w_{t,j}}$$

Όπου w_t η κατανομή πιθανότητας.

2. Για κάθε φίλτρο, j , εκπαίδευσε τον ταξινομητή h_j ο οποίος είναι περιορισμένος να χρησιμοποιήσει μόνο ένα φίλτρο. Το σφάλμα υπολογίζεται σε σχέση με το w_t με ακόλουθη σχέση.

$$\varepsilon_j = \sum_i w_i \left| h_j(x_i) - y_i \right|$$

3. Επέλεξε τον ταξινομητή, h_t, με το μικρότερο σφάλμα ε_t.

4. Ανανεώσε τα βάρη με βάση την ακόλουθη σχέση,

$$w_{t+!,i} = w_{t,i} \beta_t^{1-|h_i(x_i)-y_i|}$$

Όπου $\beta_t = \frac{\varepsilon_t}{1-\varepsilon_t}$.

- Ο τελικός ταξινομητής είναι ο ακόλουθος,

$$h(x) = \begin{cases} 1 \sum_{t=1}^{T} a_t h_t(x) \geq \Theta \sum_{t=1}^{T} a_t \\ 0 otherwise \end{cases}$$

Όπου $a_t = -\log \beta_t$ και **θ** είναι μία τιμή εντός του πεδίου $[0 \ldots 1]$. Συνήθως το **θ** επιλέγεται να χει την τιμή $\frac{1}{2}$.

Πίνακας 2: Ο αλγόριθμος της διακριτής ενίσχυσης (Discrete AdaBoost). Οι Viola και Jones υιοθέτησαν την αρχική έκδοση του αλγορίθμου με τον περιορισμό όμως οι ασθενής ταξινομητές να χρησιμοποιούν αποκλειστικά και μόνο ένα φίλτρο. Τ ασθενείς ταξινομητές κατασκευάζονται με την μέθοδο που παρουσιάζεται παραπάνω. Όταν αυτοί έχουν δημιουργηθεί τότε συμετέχουν στην ζυγισμένη επιλογή του τελικού ισχυρού ταξινομητή.

2.7.7 Πλαίσιο επιλογής βασισμένο στην προσοχή (Attentional Cascade)

Υπάρχει ένα μεγάλο πλήθος πιθανός παραθύρων ελέγχου σε μία και μόνο φωτογραφία απο το οποίο η πλειοψηφία δεν περιέχει πρόσωπο. Αυτό συνεπάγεται πως θα ήταν καλύτερο αυτά τα παράθυρα να απορρίπτονται όσο το δυνατόν πιό γρήγορα. Για αυτό τον λόγο οι Viola και Jones επινόησαν την μέθοδο της επιλογής πλαισίου με βάση την προσοχή. Αυτό το πλαίσιο αποτελείται απο πολλά στάδια. Το πρώτο στάδιο αποτελείται απο έναν μικρό αριθμό απο ασθενής ταξινομητές, με αυτό τον τρόπο στο συγκεκριμένο στάδιο μπορούμε να κάνουμε εκτιμήσεις πολύ γρήγορα. Αυτά τα απλά στάδια εκπαιδεύονται με τέτοιο τρόπο ώστε να απορρίπτουν μεγάλο αριθμό απο αρνητικά παραδείγματα ενώ ταυτόχρονα να αποδέχονται όλα τα θετικά. Επόμενα στάδια στο πλαίσιο είναι πιό πολύπλοκα και χρησιμοποιούνται με σκοπό να ελαχιστοποιήσουν την παραγωγή θετικών σφαλμάτων (εικόνες που προκύπτουν ως πρόσωπα ενώ δεν είναι).

Η εικόνα στο πλαίσιο εντοπισμού κατατάσσεται ως πρόσωπο μόνο στην περίπτωση που αναγνωρίζεται ως τέτοιο απο όλα τα στάδια του πλαισίου που περιγράφουμε. Έτσι ένα μόνο αρνητικό αποτέλεσμα ελέγχου σε οποιοδήποτε στάδιο είναι αρκετό ώστε να απορριφθεί το πλαίσιο και να συνεχιστεί ο έλεγχος στο επόμενο.

Οι Viola και Jones εκπαίδευσαν έναν ισχυρό ταξινομητή ο οποίος περιείχε μόνο δύο ασθενής ταξινομητές στο πρώτο στάδιο του πλαισίου. Μετέτρεψαν το αρχικό κατώφλι του ισχυρού ταξινομητή, με σκοπό να επιτύχουν ρυθμό εντοπισμού 100%. Σε αυτό το κατώφλι ο ρυθμός εντοπισμού εσφαλμένων θετικών εντοπισμών ήταν στο 40%, αποδίδοντας ρυθμό απόρριψης ύψους 60% σε όλα τα πλαίσια εντοπισμού τα οποία δεν περιείχαν πρόσωπο. Πρακτικά αυτό σημαίνει πως 60% απο το σύνολο όλων των πιθανών πλαισίων απορίπτεται με χρήση μόνο δύο ασθενών ταξινομητών.

Η εκπαίδευση των σταδίων έγινε με χρήση μίας βάσης που περιείχε 4916 πρόσωπα. Τα αρνητικά παραδείγματα προέκυψαν απο 9500 φωτογραφίες οι οποίες δεν περιείχαν κανένα πρόσωπο. Σε κάθε στάδιο 10000 εικόνες χωρίς πρόσωπα δημιουργήθηκαν τυχαία απο το σύνολο τςν 9500 που προ αναφέραμε. Και σε αυτή την περίπτωση οι συγγραφείς χρησιμοποιήσαν την τεχνικη bootstrapping για την εκπαίδευση.

Ο τελικός ταξινομητής που προέκυψε αποτελούνταν απο 32 στάδια, με σύνολο 4297 φίλτρα. Συνέχισαν να προσθέτουν και άλλα στάδια εώς ότου ο εντοπισμός εσφαλμένων θετικών δειγμάτων να φτάσει στο μηδέν, ενώ ταυτόχρονα να διατηρείται υψηλός ρυθμός εντοπισμού.

2.7.8 Δυνατές βελτιώσεις του ταξινομητή

Για να μπορέσουμε να βελτιώσουμε την ταχύτητα του ταξινομητή εις βάρος όμως της ακρίβειας του μπορούμε να μεταβάλλουμε τρείς παραμέτρους.

1. Η αρχική κλίμακα η οποία είναι η κλίμακα του πλαισίου εντοπισμού κατά την εκκίνηση της διαδικασίας ταξινόμησης. Για παράδειγμα όταν το αρχικό μέγεθος του πλαισίου είναι 24 x 24 εικονοστοιχεία και η αρχική κλίμακα του είναι 2.0, τότε ο ταξινομητής θα ξεκινήσει χρησιμοποιώντας πλαίσιο διαστάσεων 48 x 48. Συνήθως η αρχική κλίμακα που θέτουμε είναι 1.0. Στην περίπτωση όμως που το το μέγεθος των προσώπων που αναμένουμε είναι μεγαλύτερο, τότε βοηθάει να αλάξουμε την αρχική κλίμακα του πλαισίου.

2. Όταν ο ταξινομητής τελειώσει την διάσχισει της εικόνας με κάποιο πλαίσιο συγκεκριμένου μεγέθους, συνεχίζει κλιμακώνοντας το μέγεθος του με βάση το βήμα κλιμάκωσης. Για παράδειγμα όταν το αρχικό πλαίσιο είναι 24 x 24 και το βήμα κλιμάκωσης είναι 1.25 τότε το επόμενο πλαίσιο θα έχει μέγεθος αυξημένο κατά 25%, δηλαδή 30 x 30 εικονοστοιχεία. Με βάση τα πειράματα των συγγραφέων βρέθηκε πως το 1.25 αποδίδει καλά ως βήμα κλιμάκωσης.

3. Το πλαίσιο μετακινείτε πάντα με βάση κάποιο προκαθορισμένο βήμα. Δηλαδή όταν τελειώσει ο έλεγχος ενός συγκεκριμένου τμήματος της εικόνας στην συνέχεια αλλάζει χωρική θέση κατά κάποια εικονοστοιχεία. Ο αριθμός των εικονοστοιχείων κατά των οποίων θα μετακινηθεί το πλαίσιο εξαρτάται απο την κλίμακα **s** καθώς και το βήμα θέσης **Δ**. Στην περίπτωση που η δεδομένη κλίμακα είναι s τότε το πλαίσιο θα μετακινηθεί κατα $[s\Delta]$ εικονοστοιχεία, στην επόμενη θέση, όπου το $[\ ]$ είναι ο τελεστής στρογγυλοποίησης. Όταν έχουμε $\Delta \geq 1$ παρατηρείται μικρή μείωση του ρυθμού εντοπισμού, ενώ ταυτόχρονα παρατηρείται μείωση των εσφαλμένων θετικών εντοπισμών.

2.7.9 Ομαδοποίηση

Ο τελικός εκπαιδευμένος ταξινομητής μετακινείται κατα μήκος ολόκληρης της εικόνας σε διάφορες κλίμακες και περιοχές. Επειδή ο ταξινομητής δεν χαρακτηρίζεται απο ευαισθήσια σε μικρές μεταβολές θέσης και κλίμακας, συνήθως προκύπτει ένα μεγάλο πλήθος απο εντοπισμούς οι οποίοι όμως βρίσκονται σε διαφορετικές θέσεις γύρω απο κάποιο πρόσωπο, όπως φαίνεται στο σχήμα 1.14. Για να μπορέσουν να συνδιάσουν όλους αυτούς τους εντοπισμούς οι συγγραφείς χρησιμοποίησαν έναν απλό αλγόριθμο ομαδοποίησης. Ο αλγόριθμος αυτός τοποθετεί τα διάφορα επικαλυπτόμενα πλαίσια εντοπισμού σε ένα μόνο. Αυτοί οι εντοπισμοί τοποθετούνται σε ένα πλαίσιο, μόνο στην περίπτωση που αλληλοκαλύπτονται τα όρια τους. Για κάθε σύνολο πλαισίων εντοπισμού που προέκυψαν, ο μέσος όρος του μεγέθους και της θέσης υπολογίζονται, με αποτέλεσμα να προκύπτει ένα ορθογώνιο για κάθε σύνολο επικαλυπτόμενων εντοπισμένων πλαισίων.

2.7.10 Αποτελέσματα

Οι Viola και Jones δοκίμασαν τον ταξινομητή τους στην βάση MIT + CMU η οποία περιέχει εμπρόσθιες εικόνες προσώπων και αναφέρουν αποτελέσματα τα οποία είναι συγκρίσιμα με αυτα των Rowley et. Al. και Sung και Poggio. Παρόλα αυτά όμως ο δικός τους ταξινομητής επιτυγχάνει τις αποδόσεις αυτές με λειτουργία η οποία είναι πολύ λιγότερο απαιτητική απο πλευράς υπολογιστικών πόρων, με άμεσο αποτέλεσμα να είναι πολύ πιό γρήγορος. Συγκεκριμένα ο ταξινομητής αυτός είναι περίπου 15 φορές πιό γρήγορος απο αυτόν των Rowley-Baluja-Kanade. Κατά μέσο όρο μόνο 8 ασθενείς ταξινομητές προέκυψαν απο τους 4297 για κάθε πλαίσιο εντοπισμού. Ο ανιχνευτής προσώπων λειτουργεί με ρυθμούς των 15 καρρέ ανά δευτερόλεπτο σε έναν υπολογιστή Pentium III σε συχνότητα των 700 MHz, όταν επεξεργάζεται εικόνες διαστάσεων 384 x 288 εικονοστοιχείων.

2.8. Συμπεράσματα

Όπως αναφέρθηκε και προηγουμένως, η τελική επιλογή του ταξινομητή για τον εντοπισμό των προσώπων ήταν αυτός που προτείνεται απο τους Viola και Jones. Αρχικά έγινε έντονος πειραματισμός με ταξινομητές οι οποίοι βασίζονται στην χρήση μηχανών διανυσμάτων στήριξης, αλλά εγκαταλήφθηκαν για τους παρακάτω λόγους.

Ένας απο τους βασικούς λόγους που μας οδήγησε στο να δοκιμάσουμε τους συγκεκριμένους ταξινομητές είναι το ότι τα τελευταία χρόνια υπάρχει πολύ έντονη έρευνα στον τομέα αυτό, ενώ απο την βιβλιογραφία που κυκλοφορεί φαίνεται πως έχουν την δυνατότα να αποδόσισουν εξίσου καλά με άλλες τεχνικές ταξινόμησης προτύπων. Ταυτόχρονα προσφέρουν μία μαθηματική θεωρία που τα περιγράφει ιδιαίτερα κάλα, βασισμένη στο πεδίο της στατιστικής εκμάθησης. Πέρα απο την εργασία των Osuna et. al., πολλοί άλλοι χρησιμοποίησαν μηχανές διανυσμάτων στήριξης, σε διάφορες παραλλαγές και αναφέρουν πολύ ικανοποιητικά αποτελέσματα . Ταυτόχρονα επειδή οι συγκεκριμένοι ταξινομητές έχουν χρησιμοποιηθεί επιτυχώς και στον εντοπισμό άλλων χαρακτηριστικών, θεωρήσαμε πως θα ήταν καλύτερο να έχουμε μία συγκεκριμένη αρχιτεκτονική ταξινόμησης για ολόκληρο το σύστημα.

Το βασικό πρόβλημα το οποίο προέκυψε ήταν η αδυναμία μας να αναπαράγουμε τα αποτελέσματα τα οποία αναφέρονται στις συγκεκριμένες δημοσιεύσεις. Σε καμία περίπτωση ακόμα και όταν χρησιμοποιήσαμε ακρίβως τα ίδια σύνολα απο δεδομένα εκπαίδευσης και δοκιμής με αυτά της βιβλιογραφία, δεν επιτύχαμε ρυθμό εντοπισμού που να μπορεί να θεωρηθεί πως δεν είναι απλά τυχαίο γεγονός. Συγκεκριμένα ανέξαρτητα απο την αρχιτεκτονική του SVM που χρησιμοποιούσαμε, σε κάθε περίπτωση παρατηρούνταν μεγάλος αριθμός απο διανύσματα στήριξης στο προκύπτον μοντέλο. Ταυτόχρονα είχαμε φαινόμενα υπέρ γενίκευσης προς την πλευρά του συνόλου των όχι προσώπων. Η μηχανή τελικά κατέληγε να ταξινομεί την πλειοψηφία των προσώπων ώς όχι πρόσωπα.

Σε όλα τα πειράματα έγιναν δοκιμές με το σύνολο των κλασσικών πυρήνων που υλοποιούνται απο τα συστήματα SVM που υπάρχουν διαθέσιμα . Τόσο με πολυονυμικούς πυρήνες που αποτελούν την βασική επιλογή της βιβλιογραφίας για την περίπτωση του εντοπισμού προσώπων, όσο και με τους υπόλοιπους διαθέσιμους (Radial Basis Functions, Sigmoid Functions, Gaussian Radial Basis Functions). Σε κάθε περίπτωση έγινε η αντίστοιχη προ επεξεργασία που προτείνεται απο τους συγγραφείς, ενώ χρησιμοποιηθήκαν και διάφορες αναπαραστάσεις για τις εικόνες (για παράδειγμα DCT μετασχηματισμός). Τέλος έγιναν και δοκιμές με χρήση μεθόδων μείωσης των διαστάσεων του χώρου εισόδου, συγκεκριμένα χρησιμοποιήθηκε η μέθοδος PCA (Principal Component Analysis). Σε κάθε περίπτωση η μεταβολή των αποτελεσμάτων ήταν πολύ μικρή, είτε θετική είτε αρνητική, και δεν διαπιστώθηκε κάποια σοβαρή βελτίωση.

Θα πρέπει εδώ να σημειώσουμε πως ένα βασικό μειονέκτημα που έχουν οι μηχανές διανυσμάτων στήριξης είναι πως δεν υπάρχει κάποια συγκεκριμένη θεωρία η οποία να μας οδηγεί στην σωστή επιλογή των παραμέτρων που ορίζουν τον πυρήνα του χώρου εισόδου. Αυτό αναφέρεται γενικά στην βιβλιογραφία. Ο βασικός τρόπος επιλογής που υπάρχει ουσιαστικά είναι το πείραμα. Μία μέθοδος η οποία χρησιμοποιείται ευρέως για τον εντοπισμό των παραμέτρων ανάλογα το πρόβλημα είναι η μέθοδος της δικτυωτής αναζήτησης (Grid Search). Η συγκεκριμένη μέθοδος υποστηρίζεται απο τα πακέτα λογισμικού τα οποία χρησιμοποιηθήκαν (libsvm, svmlight), αλλά δεν απέδωσαν τα επιθυμητά αποτελέσματα. Τέλος οι ίδιοι συγγραφείς των περισοτέρων μελετών δεν παρέχουν τις παραμέτρους του πυρήνα που χρησιμοποιήσαν, με αποτέλεσμα να δισχερένεται πολύ η αναπαραγωγή των αποτελεσμάτων.

Το βασικό συμπέρασμα το οποίο προέκυψε απο τον πειραματισμό με τα SVMs είναι πως η δυσκολία να δημιουργηθεί το σωστό υπέρ επίπεδο διαχωρισμού είναι κυρίως αποτέλεσμα της φύσης των δύο συνόλων (πρόσωπα και όχι πρόσωπα). Προφανώς το σύνολο των όχι προσώπων είναι πολύ μεγαλύτερο απο αύτο των προσώπων και ουσιαστικά μπορεί να περιέχει οτιδήποτε. Οι δύο αυτές κλάσεις επομένως δεν είναι ισορροπημένες. Σε αυτό το πρόβλημα έρχεται να βοηθήσει η μέθοδος του bootstrapping που προαναφέραμε. Στα πειράματα μας δεν χρησιμοποιηθήκε απο μας αλλα θεωρήσαμε πως τα σύνολα των δεδομένων που είχαμε στην κατοχή μας και τα οποία ήταν αποτέλεσμα των πειραμάτων της βιβλιογραφίας έιχαν προέρθει απο αυτόν τον τρόπο. Για τον λόγο αυτό αποφασίσαμε να μην χρησιμοποιήσουμε τους συγκεκριμένους ταξινομητές στην περίπτωση του εντοπισμού προσώπων, αλλά να γίνει χρήση τους παρακάτω, για παράδειγμα στον εντοπισμό των χειλιών. Με αυτό τον τρόπο θα μπορέσουμε να στηρίξουμε και το παραπάνω συμπέρασμα, αφού τα χείλη θα πρέπει να συγκριθούν με ένα πολύ πιο περιορισμένο σύνολο, αυτό που προκύπτει απο την τμηματοποίηση του υπόλοιπου προσώπου μόνο.

Το σύστημα το οποίο καλούμαστε να σχεδιάσουμε και να υλοποιήσουμε αποτελείται απο πολλά τμήματα τα οποία απαιτούν την χρήση κάποιου ταξινομητή. Για αυτό τον λόγο έχει πολύ μεγάλη σημασία η ταχύτητα του πέρα απο την ακρίβεια. Ο συγκεκριμένος ταξινομητής ο οποίος επιλέχθηκε εν τέλη υπερτερεί έντονα σε αυτό τον τομέα αφού μπορεί να λειτουργήσει ακόμα και σε πραγματικό χρόνο υπό συνθήκες. Αυτός ήταν ακόμα ένας σημαντικός παράγοντας που

οδήγησε στην τελική υιοθέτηση του απο το σύστημα, ειδικά απο την στιγμή που ο εντοπισμός του προσώπου δεν αποτελεί το σημαντικότερη τμήμα της εργασίας αυτής.

Ακόμα ένας σημαντικός λόγος είναι η η διαθεσιμότητα του συγκεκριμένου αλγορίθμου σε έτοιμη υλοποίηση. Αποτελεί μέρος του πακέτου OpenCV της Intel, το οποίο είναι πακέτο ανοιχτού λογισμικού. Με αυτό τον τρόπο είχαμε έτοιμη μία αποδοτική υλοποίηση του, κάτι το οποίο έκανε τον πειραματισμό με τον συγκεκριμένο ταξινόμητη πολύ πιό γρήγορο και αποδοτικό.

Τέλος η ίδια η φύση του ταξινομητή δίνει πολλές δυνατότητες έρευνας και βελτίωσης. Η χρήση της μεθόδου AdaBoost μας δίνει την ελευθερία να επιλέξουμε εάν επιθυμούμε διαφορετικούς ασθενείς ταξινομητές πέρα απο αυτούς τους οποίους πρωτείνουν οι συγγραφείς. Για παράδειγμα στην βιβλιογραφία έχουν προκύψει πολύ ενθαρυντικά αποτελέσματα απο την χρήση Gabor φίλτρων αλλά ακόμα και μηχανών διανυσμάτων στήριξης .

ΚΕΦΑΛΑΙΟ 3ο.

ΕΝΤΟΠΙΣΜΟΣ ΧΕΙΛΙΩΝ

Το επόμενο βήμα μετά τον εντοπισμό του προσώπου στο σύστημα, είναι ο προσδιορισμός της περιοχής των χειλιών. Το μέρος αυτό του συστήματος καλείται να εντοπίσει την περιοχή του προσώπου στην οποία βρίσκονται τα χείλη, η έξοδος του οποίου θα χρησιμοποιηθεί στην συνέχεια ως είσοδος στο σύστημα το οποίο θα επιχειρήσει τον προσδιορισμό των χαρακτηριστικών εκείνων τα οποία θα μας οδηγήσουν στην αναγνώριση της λέξης την οποία προφέρει ο ομιλητής. Αντίστοιχα με την περίπτωση του προσώπου και σε αυτή την περίπτωση οι μέθοδοι οι οποίοι μπορούν να χρησιμοποιηθούν μπορούν πάλι να κατηγοριοποιηθούν σε δύο βασικές κατηγορίες. Η πρώτη είναι αυτή η οποία βασίζεται σε χαμηλού επιπέδου χαρακτηριστικά τα οποία εξάγονται από την εικόνα, για παράδειγμα χρώμα ή περίγραμμα. Η δεύτερη βασίζεται πάλι στην επεξεργασία κυλιόμενου πλαισίου πάνω στην εικόνα. Στην συνέχεια θα παρουσιάσουμε χαρακτηριστικές περιπτώσεις μεθόδων οι οποίες έχουν προταθεί στην βιβλιογραφία για την επίλυση του προβλήματος εντοπισμού χειλιών ενώ στην συνέχεια θα καταλήξουμε στην μέθοδο η οποία τελικά επιλέχθηκε και θα γίνει εκτενέστερη ανάλυση και αναφορά σε αυτή.

3.1. Κατηγοριοποιηση των μεθοδων εντοπισμου χειλιων

Όπως αναφέραμε και παραπάνω ο εντοπισμός των χειλιών όπως και άλλων χαρακτηριστικών έχει πολλές ομοιότητες με τον εντοπισμό του προσώπου στην εικόνα. Οι βασικές μέθοδοι χωρίζονται σε δύο κατηγορίες, την ανάλυση από πάνω προς τα κάτω (top – down analysis) και την από κάτω προς τα πάνω (Bottom – up multiscale analysis). Στην συνέχεια θα αναφερθούμε επιγραμματικά στις χαρακτηριστικές περιπτώσεις και των δύο ομάδων και θα αναφέρουμε τα βασικά τους πλεονεκτήματα και μειονεκτήματα.

3.1.1 Αναλυση απο πανω προς τα κατω (Top-Down Analysis)

Η από πάνω προς τα κάτω ανάλυση, η όπως συνήθως αναφέρεται «ανάλυση βασισμένη σε μοντέλα», προϋποθέτει κάποιες από πριν υποθέσεις για το ποια χαρακτηριστικά της οπτικής

ομιλίας είναι σημαντικά. Συνήθως το πιο χρησιμοποιούμενο χαρακτηριστικό είναι το σχήμα των χειλιών όπως και παρουσιάζεται και σε μεγάλο πλήθος εργασιών ([29], [28], [16]). Ένας βασικός λόγος για τον οποίο συμβαίνει αυτό είναι γιατί το σχήμα των χειλιών είναι πάντα εμφανές και διαθέσιμο ειδικά στις περιπτώσεις όπου οι εικόνες που χρησιμοποιούνται έχουν ως περιεχόμενο πρόσωπα σε εμπρόσθια λήψη. Επίσης όπως θα φανεί και παρακάτω που θα αναφερθούμε στα χαρακτηριστικά της άρθρωσης, από το σύνολο τους ελάχιστα είναι αυτά τα οποία είναι εμφανή στην εικόνα και συνήθως το πιο άμεσο χαρακτηριστικό με τον λιγότερο θόρυβο είναι αυτό του σχήματος των χειλιών. Η μέθοδος που θα αναφέρουμε χρησιμοποιεί ενεργά μοντέλα σιλουέτας για να γίνει εφικτή η παρακολούθηση του εσωτερικού και εξωτερικού περιγράμματος των χειλιών.

Ένα ενεργό μοντέλο σιλουέτας είναι ουσιαστικά ένας επαναληπτικός αλγόριθμος ο οποίος έχει περιορισμούς που ορίζονται από κάποιο σχήμα. Οι περιορισμοί συνήθως προέρχονται από κάποιο στατιστικό μοντέλο σιλουέτας, το οποίο συνήθως αναφέρεται και ως μοντέλο κατανομής σημείων, το οποίο προέρχεται από τα στατιστικά κάποιου συνόλου εκπαίδευσης το οποίο έχει επιλεγεί με το χέρι. Το μοντέλο κατανομής σημείων περιγράφει έναν ελαχιστοποιημένο χώρο από έγκυρες σιλουέτες χειλιών. Τα σημεία αυτού του χώρου αποτελούν ομογενείς αναπαραστάσεις σιλουέτας χειλιών οι οποίες μπορούν να χρησιμοποιηθούν απευθείας ως χαρακτηριστικά εισόδου σε ένα σύστημα ανάγνωσης χειλιών.

Το μοντέλο κατανομής σημείων συνήθως δημιουργείται από κάποιο σύνολο εκπαίδευσης, στο οποίο συγκεκριμένα σημεία έχουν με το χέρι επιλεγεί. Αυτή η διαδικασία μπορεί και να αυτοματοποιηθεί όπως παρουσιάζεται και στο [2]. Κάθε σιλουέτα του συνόλου παραδειγμάτων αναπαρίσταται με βάση τις (x, y) συντεταγμένες των σημείων που όπως αναφέρθηκε προηγουμένως έχουν επιλεγεί με το χέρι, τα οποία έχουν το ίδιο νόημα σε όλα τα παραδείγματα του συνόλου παραδειγμάτων/εκπαίδευσης. Τα περιγράμματα του εσωτερικού και εξωτερικού χειλιού ορίζονται συνολικά από 44 σημεία, 24 σημεία για το εξωτερικό χείλος και 20 για το εσωτερικό. Τα σημεία αυτά χωρίζονται σε δύο κατηγορίες, αυτά τα οποία ο χρήστης μπορεί με βεβαιότητα να τοποθετήσει και καλούνται πρωτεύοντα. Ανάμεσα σε αυτά τα σημεία και σε ίσες αποστάσεις τοποθετούνται τα δευτερεύοντα σημεία. Για την ελαχιστοποίηση των λαθών η τοποθέτηση των δευτερευόντων σημείων ομαλοποιείται με χρήση παρεμβολής καμπύλων (spline interpolation).

Αν η νιοστή σιλουέτα είναι η,

$$\chi_v = (\chi_{v1}, y_{v1}, \chi_{v2}, y_{v2}, \dots, \chi_{v44}, y_{v44})^T$$

Τότε δύο όμοιες σιλουέτες χ_1και χ_2 ευθυγραμμίζονται με την ελαχιστοποίηση της παρακάτω σχέσης,

$$E = (\chi_1 - M(s,\theta)[\chi_2 - t])^T W(x_1 - M(s,\theta)[x_2] - t) \text{ (1)}$$

Όπου ο μετασχηματισμός της πόζας για την κλίμακα, s, την περιστροφή θ και η μεταφορά θέσης στο χ και $y(t_x, t_y)$ επίπεδο είναι,

$$M(s,\theta)\begin{bmatrix} x_{jk} \\ y_{jk} \end{bmatrix} = \begin{pmatrix} (scos\theta)x_{jk} - (ssin\theta)y_{jk} \\ (ssin\theta)x_{jk} + (scos\theta)y_{jk} \end{pmatrix} \text{ (2)}$$

$$t = (t_{x1}, t_{y1}, \dots, t_{xN}, t_{yN}) \text{ (3).}$$

Ο W είναι ένας διαγώνιος πίνακας βαρών για κάθε σημείο. Τα βάρη αυτά είναι αντιστρόφως ανάλογα της διακύμανσης του κάθε σημείου.

Για την ευθυγράμμιση του συνόλου εκπαίδευσης, μπορεί να χρησιμοποιηθεί ο αλγόριθμος που περιγράφεται στο [33]. Με δεδομένο το σύνολο των ευθυγραμμισμένων μοντέλων σιλουέτας, η μέση σιλουέτα, $\bar{x}_s$, μπορεί να υπολογισθεί. Στην συνέχεια οι άξονες αυτοί οι οποίοι περιγράφουν την μεγαλύτερη διακύμανση της μέσης σιλουέτας μπορούν να εντοπισθούν χρησιμοποιώντας τεχνικές όπως η ανάλυση πρωτογενών παραγόντων (PCA). Στην συνέχεια κάθε έγκυρη σιλουέτα μπορεί να προσεγγισθεί με την πρόσθεση μειωμένων υποσυνόλων, t, των καταστάσεων που περιγράφονται από το μοντέλο κατανομής σημείων, στην μέση σιλουέτα,

$$x_s = \bar{x}_s + P_s b_s \text{ (4)}$$

Όπου P_s είναι ο πίνακας των πρώτων t ιδιοτιμών,

$$P_s = (p_1, p_2, \dots, p_t)$$

Και b_s είναι ένα διάνυσμα από t βάρη, $b_s = (b_1, b_2, \dots, b_t)^T$. Καθώς τα ιδιοδιανύσματα είναι ορθογωνικά, οι παράμετροι σιλουέτας b_s, μπορούν να υπολογισθούν από ένα σύνολο σημείων παραδειγμάτων, χ_s , με βάση την παρακάτω σχέση,

$b_s = P_s^T (x_s - \overline{x_s})$. (5)

Τα παραπάνω δίνουν την δυνατότητα, έγκυρες σιλουέτες να παρασταθούν σε έναν συμπαγή στατιστικό χώρο σιλουέτας. Το πλήθος των καταστάσεων είναι πολύ μικρότερο από αυτό του αριθμού των σημείων τα οποία επιλέχθηκαν, διότι τα σημεία αυτά επιλέχθηκαν έτσι ώστε με σαφή τρόπο να καθορίζουν την σιλουέτα των χειλιών, κάτι το οποίο έχει ως αποτέλεσμα τα σημεία αυτά να χαρακτηρίζονται από υψηλή συσχέτιση μεταξύ τους. Επίσης δεν υπάρχουν προβλήματα κατά την εφαρμογή της ανάλυσης πρωτογενών παραγόντων, γιατί όλες οι μεταβλητές είναι τιμές είτε του χ είτε του y στο ορθογώνιο σύστημα αξόνων που ορίζει την εικόνα. Ο βαθμός του μοντέλου κατανομής σημείων επιλέγεται έτσι ώστε οι πρώτες t ιδιοτιμές του πίνακα συνδιακύμανσης να περιγράφει το 95% της συνολικής διακύμανσης.

Για να μπορέσουμε επαναληπτικά να ταιριάξουμε ένα μοντέλο κατανομής σημείων σε μία εικόνα, πρέπει να επιλεγεί κάποια συνάρτηση κόστους. Στην βιβλιογραφία πολύ συχνά χρησιμοποιείται ένα στατιστικό μοντέλο των συνδεδεμένων επιπέδων του γκρι, τα επίπεδα αυτά προκύπτουν από τα κάθετα διανύσματα σε κάθε σημείο κάποιου μοντέλου σιλουέτας ([15], [17], [16]). Αυτή η επιλογή δίνει την δυνατότητα στην ανάλυση πρωτογενών παραγόντων να αναπαραστήσει όλα τα κάθετα διανύσματα των επιπέδων του γκρι του μοντέλου σιλούετας με ένα μόνο στατιστικό μοντέλο, έτσι ώστε να λαμβάνονται υπόψη ο συσχετισμός μεταξύ των επιπέδων γκρι σε διαφορετικά σημεία. Με αυτό τον τρόπο κατασκευάζεται ένα μοντέλο κατανομής των επιπέδων του γκρι (GLDM) με βάση την παρακάτω σχέση,

$x_p = \overline{x_p} + P_p b_p$ (6)

Ο βαθμός του μοντέλου κατανομής των επιπέδων γκρι επιλέγεται πάλι έτσι ώστε οι πρώτες t ιδιοτιμές να περιγράφουν το 95% της συνολικής διακύμανσης.

Ο αρχικός αλγόριθμος του ASM, μοντελοποιούσε τα επίπεδα του γκρι για κάθε επιλεγμένο σημείο και στην συνέχεια ταίριαζε μία συγκεκριμένη εικόνα με το να υπολογίζει τις μεταβολές του μοντέλου για κάθε σημείο. Μία εναλλακτική μέθοδος είναι η χρήση του συνδυασμού παραμέτρων πόζας και σχήματος με σκοπό την δημιουργία ενός διανύσματος μεταβλητών το οποίο αποτελεί παράμετρο ενός απλοποιημένου αλγόριθμου downhill για την ελαχιστοποίηση. Εκτεταμένες αναφορές για τα παραπάνω υπάρχουν στα [17], [16], [32], [33] και [18].

Ένα βασικό πρόβλημα το οποίο μπορεί να προκύψει με τον παραπάνω αλγόριθμο είναι το ακόλουθο. Επειδή τα υποκείμενα τα οποία συμμετέχουν συνήθως στις εικόνες τις οποίες χρησιμοποιούνται έχουν μεγάλες διαφορές μεταξύ τους, το μοντέλο κατανομής επιπέδων του γκρι το οποίο προκύπτει, συνήθως αποτελείται από πολύ μεγάλο πλήθος καταστάσεων. Σε ένα τέτοιο χώρο πολλών διαστάσεων η συνάρτηση κόστους σπάνια έχει κάποιο ξεκάθαρο ελάχιστο και ο αλγόριθμος ελαχιστοποίησης συνήθως αποτυγχάνει να βρει την κατάλληλη πόζα και σχήμα των χειλιών του ομιλητή. Μία λύση ο οποία προτείνεται στην βιβλιογραφία [14], είναι η δημιουργία μοντέλων για κάθε ομιλητή ξεχωριστά με αποτέλεσμα την δημιουργία ενός συνόλου από μοντέλα κατανομής επιπέδων του γκρι. Αυτή η τακτική έχει δύο βασικά μειονεκτήματα, το πρώτο είναι πως πρέπει να υπάρχει a priori γνώση για το ποιος είναι ο ομιλητής έτσι ώστε να επιλεγεί το κατάλληλο μοντέλο, ενώ στην περίπτωση που επιθυμούμε την επέκταση της βάσης στην οποία εργαζόμαστε με νέους ομιλητές, θα πρέπει να κατασκευαστούν εξ αρχής μοντέλα για αυτούς.

Ο παραπάνω αλγόριθμος επιλέχθηκε να παρουσιαστεί αναλυτικά γιατί αποτελεί αντιπροσωπευτικό δείγμα των μεθόδων από πάνω προς τα κάτω (top-down analysis) οι οποίες εμφανίζονται στην βιβλιογραφία. Μία κλασσική επέκταση του μοντέλου που παρουσιάστηκε είναι τα μοντέλα ενεργής σιλουέτας (Active appearance model AAM). Αυτά τα μοντέλα όπως και τα ASM είναι στατιστικά μοντέλα τα οποία όμως εκτός από το να κωδικοποιούν μόνο την σιλουέτα, ταυτόχρονα κωδικοποιούν και τα επίπεδα του γκρι της εικόνας που ορίζει την σιλουέτα. Παραδείγματα αυτών των μοντέλων μπορούν να βρεθούν άφθονα στην βιβλιογραφία ([5], [3], [25]). Τέλος πρέπει να αναφέρουμε πως ένα βασικό θετικό χαρακτηριστικό αυτών των προσεγγίσεων είναι το ότι τα μοντέλα αυτά τα οποία προκύπτουν μπορούν να χρησιμοποιηθούν κατευθείαν ως είσοδοι ενός ταξινομητή ο οποίος θα επιχειρήσει τον εντοπισμό του φωνήματος,

για παράδειγμα κάποιο μαρκοβιανό μοντέλο, χωρίς να είναι απαραίτητη κάποια επεξεργασία ή ο μετασχηματισμός τους σε κάποιο άλλο διάνυσμα εισόδου.

Οι αλγόριθμοι που αναφέραμε παραπάνω δημιουργούν μοντέλα της σιλουέτας των χειλιών και προσπαθούν να τα ταυτίσουν με περιοχές του προσώπου έτσι ώστε να εντοπισθούν αφενός τα χείλη του ομιλητή αλλά αφετέρου ανάλογα το μοντέλο να γίνει και η επιλογή του φωνήματος που παράγεται εκείνη την στιγμή. Έχουν παρουσιασθεί αλγόριθμοι οι οποίοι βασίζονται στην σιλουέτα του προσώπου για να γίνει ο εντοπισμός του κέντρου των χειλιών και να δημιουργηθεί κάποιο περικλύων πολύγωνο που θα περιέχει τα χείλη του ομιλητή. Τα μοντέλα αυτά βασίζονται στο γεγονός του ότι η γεωμετρία του προσώπου παρόλο που διαφέρει από άνθρωπο σε άνθρωπο σε γενικές γραμμές είναι η ίδια. Όπως παρουσιάζεται στο [34], η χρήση μίας κάμερας η οποία λειτουργεί σχεδόν στο υπέρυθρο φάσμα δίνει την δυνατότητα του εύκολου εντοπισμού του κέντρου των ματιών. Ένα μοντέλο της γεωμετρίας του προσώπου έχει δημιουργηθεί χειροκίνητα από ένα σύνολο εικόνων. Με αυτό τον τρόπο μέσες τιμές για τις αποστάσεις μεταξύ των βασικών χαρακτηριστικών του προσώπου εξάγονται (π.χ. μάτια, μύτη, γωνίες στόματος). Οι τιμές αυτές χρησιμοποιούνται για να κανονικοποιηθεί κάποιο καινούργιο πρόσωπο το οποίο έχει εντοπισθεί σε αυτό το μοντέλο. Με τον τρόπο αυτό δημιουργείται ένα σύνολο από ευρύστικούς κανόνες οι οποίοι και με την χρήση της κανονικοποίησης του προσώπου στο πρότυπο που έχει δημιουργηθεί μπορεί να οδηγήσει στον εντοπισμό της περιοχής του προσώπου. Παρατηρούμε ότι ενώ η προσέγγιση είναι η ίδια, δηλαδή πάλι έχουμε την χρήση κάποιου μοντέλου σιλουέτας, σε αυτή την περίπτωση χρησιμοποιούνται γεωμετρικά χαρακτηριστικά και όχι στατιστικά μοντέλα. Παρόλα αυτά οι μεθοδολογίες που παρουσιάστηκαν παραπάνω για την προσέγγιση της σιλουέτας των χειλιών μπορεί εύκολα να εφαρμοσθεί και στην περίπτωση της σιλουέτας προσώπου και στην συνέχεια να εφαρμοσθούν οι ευριστικές μέθοδοι για την εύρεση του κέντρου του στόματος. Στην περίπτωση αυτή όμως αναφερόμαστε στο πρόβλημα του εντοπισμού προσώπου και στην συνέχεια της εφαρμογής γεωμετρικών μεθόδων για τον εντοπισμού των χειλιών, κάτι το οποίο είναι ακατάλληλο για την παρούσα αρχιτεκτονική την οποία παρουσιάζουμε, μιας και ο εντοπισμός του προσώπου αντιμετωπίζεται ξεχωριστά.

Ένα άλλο χαρακτηριστικό στο οποίο πρέπει να αναφερθούμε είναι η χρονική απόδοση αυτών των μεθόδων. Αναφέραμε πριν πως το ότι τα μοντέλα αυτά μπορούν να χρησιμοποιηθούν κατευθείαν ως είσοδος σε κάποιον ταξινομητή αποτελεί πλεονέκτημα των μεθόδων αυτών.

Ταυτόχρονα όμως θα πρέπει να σημειωθεί πως οι μέθοδοι αυτοί συνήθως κοστίζουν πολύ σε υπολογιστικούς πόρους με άμεσο αποτέλεσμα να απαιτείται πολύς χρόνος για την απαραίτητη επεξεργασία. Σε ένα σύστημα το οποίο επιθυμούμε να λειτουργεί σε πραγματικό χρόνο αυτό είναι ένα μη επιθυμητό χαρακτηριστικό. Για αυτό τον λόγο ανά περίπτωση προτιμούνται μέθοδοι οι οποίες εντοπίζουν σε κάποιο καρέ την επιθυμητή περιοχή και στην συνέχεια χρησιμοποιείται άλλες μέθοδοι για την παρακολούθηση της περιοχής αυτής (visual tracking), οι μέθοδοι αυτοί παρακολούθησης δίνουν την δυνατότητα της λειτουργίας του συστήματος σε πραγματικό χρόνο. Στην περίπτωση αποτυχίας του αλγορίθμου παρακολούθησης μπορεί να επαναληφθεί η εφαρμογή του αλγορίθμου εντοπισμού και να ξαναγίνει αρχικοποίηση στο νέο καρέ. Επειδή πολλές φορές ο εντοπισμός της αστοχίας του αλγορίθμου παρακολούθησης είναι μία δύσκολη διαδικασία, επιλέγεται κάποιο κατώφλι και γίνεται επανάληψη του αλγορίθμου εντοπισμού σε κάθε καρέ με βάση αυτό το κατώφλι.

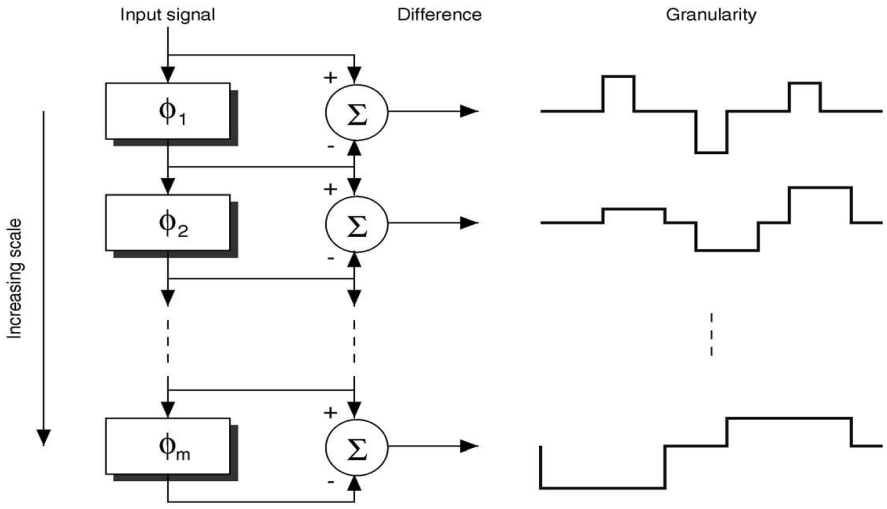
3.1.2 Ανάλυση από κάτω προς τα πάνω (Bottom-Up Analysis)

Η από κάτω προς τα πάνω ανάλυση λειτουργεί συνήθως στο επίπεδο των εικονοστοιχείων με σκοπό από τα χαρακτηριστικά που μπορούν να εξαχθούν από αυτά να προκύψει η περιοχή ενδιαφέροντος. Αυτές οι μέθοδοι επειδή λειτουργούν στο επίπεδο των εικονοστοιχείων έχουν την προοπτική να μειώσουν τα σφάλματα τα οποία προκύπτουν από την εφαρμογή των προσεγγίσεων που βασίζονται στην χρήση μοντέλων, κυρίως γιατί τα μοντέλα αυτά βασίζονται σε υποθέσεις για τα χαρακτηριστικά της εικόνας τα οποία μπορεί να είναι και εσφαλμένα.

Μία κλασσική εφαρμογή της ανάλυσης από κάτω προς τα πάνω, όπως περιγράφεται και στο [14], γίνεται με χρήση πολυεπίπεδης χωρικής ανάλυσης (Multiscale Spatial Analysis, MSA), η οποία βασίζεται στην χρήση φίλτρων τύπου κόσκινου (sieves). Η μέθοδος αυτή μπορεί να βελτιωθεί περισσότερο με την χρήση ανάλυσης πρωτογενών παραγόντων, η ανάλυση αυτή χρησιμοποιείται ευρύτατα στην αντιμετώπιση αντίστοιχων προβλημάτων και το σύνολο των χαρακτηριστικών που εξάγονται αναφέρονται συνήθως στην βιβλιογραφία ως eigenlips. Η ονομασία αυτή προέρχεται από μία αντίστοιχη προσέγγιση για τον εντοπισμό προσώπων από τους Turk and Pentland [24], οι οποίοι χρησιμοποίησαν τον όρο “eigenfaces”. Στην περίπτωση της παρούσας μεθόδου η ανάλυση πρωτογενών παραγόντων, χρησιμοποιείται για να γίνει το σύστημα πιο αποδοτικό. Για να γίνει αυτό τα χαρακτηριστικά προκύπτουν αφού πρώτα έχουν

μετασχηματιστεί σε έναν μη γραμμικό χώρο κλίμακας (nonlinear scale-space). Ο μετασχηματισμός αυτός γίνεται με την χρήση των φίλτρων που προαναφέραμε και έχει ως βασικό χαρακτηριστικό την αποδέσμευση της χωρικής πληροφορίας από την ένταση των εικονοστοιχείων.

Ένα κόσκινο είναι μία σειριακή δομή φίλτρων βασισμένη σε μαθηματικούς μορφολογικούς τελεστές, αυτά τα φίλτρα έχουν την δυνατότητα διαδοχικά να αφαιρούν χαρακτηριστικά από το σήμα εισόδου καθώς αυξάνουν διάσταση. Η διαδικασία αυτή αποτυπώνεται και στο **Σχήμα 15**. Σε κάθε στάδιο το στοιχείο φιλτραρίσματος, φ, αφαιρεί το μαθηματικό ακρότατο από αυτή μόνο την διάσταση. Το πρώτο στάδιο, φ_1, αφαιρεί το ακρότατο διάστασης 1 κοκ έως φτάσει στην μέγιστη διάσταση m που έχει καθοριστεί. Τα ακρότατα τα οποία έχουν αφαιρεθεί ονομάζονται «κόκκοι» (granules). Η διαδικασία αυτή είναι αναστρέψιμη και διατηρεί την αιτιακή σχέση διάστασης – χώρου. Για να μπορέσει να ορισθεί ένα κόσκινο σε μία εικόνα πρέπει να γίνει μία θεώρηση της εικόνας ως γράφος, θεωρώντας την ως ένα σύνολο από συνδεόμενα εικονοστοιχεία. Με αυτό τον τρόπο μπορούν να ορισθούν φίλτρα για την εφαρμογή μορφολογικών τελεστών όπως το opening, closing κοκ. Στην συνέχεια και αφού γίνει η εφαρμογή των φίλτρων στην εικόνα πρέπει να αφαιρεθούν τα περιττά στοιχεία από αυτή. Ένας τρόπος για να γίνει αυτό είναι με την χρήση ιστογραμμάτων διάστασης τα οποία προσεγγίζουν την κατανομή των οριζοντίων κόκκων στην εικόνα. Τα ιστογράμματα αυτά αφού εφαρμοσθούν στην εικόνα εξάγουν ένα πλήθος από χαρακτηριστικά ανάλογα και με τις αρχικές επιλογές που έχουμε κάνει. Συνήθως το πλήθος αυτό τον χαρακτηριστικών ορίζει ένα διάνυσμα το οποίο θα επιθυμούσαμε να είναι μικρότερο για τις ανάγκες της δημιουργίας στην συνέχεια ενός στατιστικού μοντέλου για τον εντοπισμό της περιοχής που επιθυμούμε. Σε αυτό το στάδιο μπορεί να γίνει χρήση της ανάλυσης πρωτογενών παραγόντων.



Σχήμα 15: Δομή των φίλτρων τύπου κόσκινου.

3.2. Περιγραφη της μεθοδου που επιλεχθηκε για τον εντοπισμο των χειλιων

Παραπάνω περιγράψαμε χαρακτηριστικές περιπτώσεις μεθόδων εντοπισμού της περιοχής των χειλιών όπως αυτές εμφανίζονται στην βιβλιογραφία, αφού αναφέραμε πρώτα τις δύο βασικές κατηγορίες στις οποίες αυτές ανήκουν. Σε αυτό το κομμάτι θα αναφερθούμε στην μέθοδο που επιλέξαμε στην εργασία αυτή για την περίπτωση του εντοπισμού των χειλιών.

Στην περίπτωση μας η μέθοδος που ακολουθήθηκε ανήκει στην πρώτη κατηγορία που προαναφέραμε, σε αυτή της από πάνω προς τα κάτω ανάλυσης. Ο αλγόριθμος ακολουθεί μία μέθοδο αντίστοιχη αυτής του εντοπισμού του προσώπου σε μία εικόνα. Η εικόνα σαρώνεται διαδοχικά με την χρήση ενός παραθύρου, με κάποιο προκαθορισμένο βήμα. Το περιεχόμενο του κάθε παραθύρου περνάει από κατάλληλη επεξεργασία και στην συνέχεια κάποιος ταξινομητής καλείται να αποφασίσει αν το περιεχόμενο αυτό ανήκει στο περίγραμμα χειλιών η όχι. Στην προκειμένη περίπτωση ο ταξινομητής που επιλέχθηκε είναι ένα SVM. Ο λόγοι για τους οποίους καταλήξαμε στην συγκεκριμένη μέθοδο και ταξινομητή είναι οι ακόλουθοι. Αρχικά ένας βασικός λόγος είναι η ομοιογένεια του συστήματος. Για καθαρά λόγους που σχετίζονται με την αρχιτεκτονική λογισμικού ένα σύστημα το οποίο αποτελείται από πολλά επί μέρους μέρη πρέπει να διακρίνεται από απλότητα και ομοιογένεια στην επιλογή των μερών αυτών. Ένα σύστημα το οποίο αποτελείται από πολλές διαφορετικές τεχνολογίες ταξινομητών είναι δύσκολο και στην

επέκταση του αλλά και στην υλοποίηση του. Για παράδειγμα, η διατήρηση σταθερού ταξινομητή μας δίνει την δυνατότητα να έχουμε μία σταθερή αναπαράσταση των δεδομένων χωρίς να είναι απαραίτητος ο μετασχηματισμός τους ανάλογα τις ανάγκες. Ένας άλλος λόγος είναι το ότι επιθυμούσαμε να διαπιστώσουμε την ορθότητα αυτών που αναφέραμε στο κεφάλαιο για τον εντοπισμό του προσώπου. Εκεί αναφέραμε πως η αποτυχία του συγκεκριμένου ταξινομητή οφείλεται σε μεγάλο βαθμό στην μορφή των δεδομένων και των κλάσεων που τα χαρακτηρίζουν. Στην περίπτωση που είχαμε θετικά αποτελέσματα στον εντοπισμό των χειλιών τότε θα μπορούσαμε να στηρίξουμε και με πειραματική επαλήθευση αυτή την υπόθεση. Τέλος στην βιβλιογραφία έχουν παρουσιασθεί πολύ καλά αποτελέσματα στην χρήση αυτών των ταξινομητών για τον εντοπισμό χειλιών [21] χωρίς να απαιτείται κάποια ιδιαίτερα πολύπλοκη επεξεργασία των δεδομένων εισόδου του ταξινομητή. Αυτό κρίνεται ιδιαίτερα απαραίτητο στην προκειμένη περίπτωση αφού ο σκοπός δεν είναι η κατασκευή ενός συστήματος εντοπισμού χειλιών, αλλά η περιγραφή και υλοποίηση ενός συστήματος για την ανάγνωση χειλιών.

3.2.1 Παρουσιαση δεδομενων

Για τις ανάγκες του τμήματος του εντοπισμού του χειλιών χρησιμοποιήθηκε ένα υποσύνολο από τα δεδομένα της οπτικοακουστικής βάσης CUAVE. Συγκεκριμένα για την εκπαίδευση των ταξινομητών επιλέχθηκε ένα σύνολο από τρεις διαφορετικούς ομιλητές, και οι τρείς ήταν άντρες, ενώ για λόγους ελέγχου της απόδοσης των ταξινομητών ο ένας από αυτούς επιλέχθηκε να έχει μούσι. Σε κάθε περίπτωση τα δεδομένα προέκυψαν από τα τμήματα του βίντεο όπου οι ομιλητές κοιτούσαν κάθετα την κάμερα, αφού το σύστημα το οποίο περιγράφουμε δεν λαμβάνει υπόψην του την περίπτωση λήψης υπό γωνία. Η λήψη των βίντεο έχει γίνει κάτω από ελεγχόμενες συνθήκες οι οποίες αναφέρονται από τον δημιουργό του βίντεο, ενώ οι διαστάσεις του καρέ είναι 720 επί 480 εικονοστοιχεία, το δείγμα του βίντεο είναι στα 24 bit, ενώ ο κωδικοποιητής που χρησιμοποιήθηκε είναι ο Cinepac. Περισσότερες πληροφορίες για την συγκεκριμένη βάση μπορούν να αναζητηθούν στο [9]. Χαρακτηριστικά καρέ από αυτά που χρησιμοποιήθηκαν παρουσιάζονται στο **Σχήμα 16**. Οι ομιλητές που χρησιμοποιήθηκαν είναι με βάση την ονοματολογία της βάσης, οι, s01m, s02m και s03m.

Σχήμα 16: Παραδείγματα ομιλητών που χρησιμοποιήθηκαν για την εκπαίδευση του ταξινομητή εντοπισμού χειλιών από την βάση CUAVE.

Όπως μπορούμε να παρατηρήσουμε και στα καρέ του **Σχήμα 16**, το φόντο του βίντεο είναι ομοιόμορφο και χρώματος ματζέντα, κάτι το οποίο δίνει την δυνατότητα της πιο γρήγορης και σωστής επεξεργασίας όπως θα φανεί και παρακάτω στην αναφορά για την επεξεργασία των δεδομένων.

3.2.2 Επεξεργασια δεδομενων

Όπως φαίνεται και στο διάγραμμα του συστήματος, η είσοδος του τμήματος εντοπισμού χειλιών παίρνει ως είσοδο τμήμα του αρχικού καρέ, το οποίο περιέχει το πρόσωπο του ομιλητή. Στο ολοκληρωμένο σύστημα αυτή η είσοδος προέρχεται από την έξοδο του συστήματος εντοπισμού προσώπου. Στην παρούσα φάση για την επεξεργασία των εικόνων και την εξαγωγή των προσώπων χρησιμοποιήθηκε η μέθοδος των Viola – Jones όπως αυτή αναφέρθηκε στο κεφάλαιο για τον εντοπισμό του προσώπου. Για καθέναν από τους ομιλητές που εμφανίζονται στο **Σχήμα 16**, έγινε η εξαγωγή του προσώπου από κάθε καρέ του βίντεο. Τα καινούργια καρέ τα οποία προέκυψαν εμφανίζονται δειγματοληπτικά στο **Σχήμα 17**. Για τις ανάγκες της ανάπτυξης του τμήματος εντοπισμού χειλιών δεν είναι απαραίτητη η χρήση κάποιου συστήματος εντοπισμού χειλιών. Το περιεχόμενο της βάσης και οι συνθήκες κάτω από τις οποίες έγινε η λήψη, δίνουν την δυνατότητα να χρησιμοποιηθεί κάποιος αλγόριθμος εντοπισμού, όπως για παράδειγμα ο αλγόριθμος των Lucas-Kanade. Αυτό μπορεί να εφαρμοσθεί διότι το φόντο είναι αυτό που περιγράψαμε παραπάνω. Επίσης στην περίπτωση αυτή θα μπορούσε να γίνει απλά αφαίρεση του φόντου, το οποίο όμως θα δημιουργούσε πρόβλημα στις διαστάσεις του διανύσματος εισόδου που θα δημιουργούσαμε παρακάτω, αφού είναι δύσκολο μία τέτοια μέθοδος να εγγυηθεί το μέγεθος της εξόδου της. Η ίδια επιφύλαξη υπάρχει και στην περίπτωση που θα χρησιμοποιηθεί κάποιος αλγόριθμος εντοπισμού, πάλι πρέπει να ληφθεί υπόψην ότι το διάνυσμα εισόδου, οπότε

και το προκύπτον καρέ θα πρέπει να έχει σταθερό μέγεθος. Σε ότι αναφορά το μέγεθος του διανύσματος εισόδου, πρέπει να αναφέρουμε ότι στην περίπτωση της βάσης που χρησιμοποιήθηκε η απόσταση του ομιλητή είναι σταθερή από την κάμερα, αυτό έχει ως αποτέλεσμα το περιβάλλον πολύγωνο το οποίο επιστρέφει το σύστημα εντοπισμού προσώπου να έχει σταθερό μέγεθος. Σε ένα πιο ρεαλιστικό σενάριο όπου η απόσταση του ομιλητή δεν είναι ούτε σταθερή αλλά ούτε και γνωστή από πριν, η χρήση των γκαουσιανών πυραμίδων της μεθόδου Viola – Jones, μας δίνει την δυνατότητα με μία προβολή στην διάσταση που επιθυμούμε να διασφαλίσουμε το μέγεθος του διανύσματος εισόδου του ταξινομητή. Πέρα από αυτή την λύση υπάρχει και η δυνατότητα εφαρμογής κάποιας μεθόδου μεταβολής των διαστάσεων της εικόνας όπως για παράδειγμα η γραμμική παρεμβολή (linear interpolation).

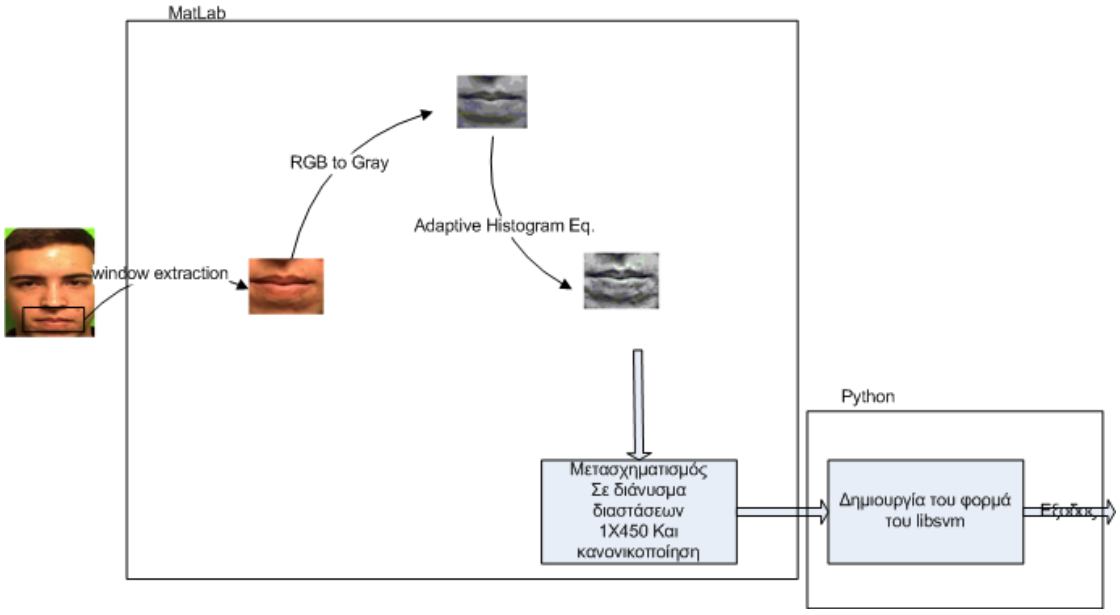
Σχήμα 17: Δείγμα της εξόδου του συστήματος εντοπισμού προσώπου για τους τρείς ομιλητές που επιλέχθηκαν για το σύστημα του εντοπισμού χειλιών.

Αφού πάρουμε τα αποτελέσματα του αλγορίθμου εντοπισμού προσώπου, το επόμενο βήμα είναι να δημιουργήσουμε τα σύνολα εκπαίδευσης και επαλήθευσης. Για να γίνει αυτό τμηματοποιήσαμε της εικόνες με βάση κάποιο προκαθορισμένο παράθυρο. Το μέγεθος του παραθύρου αυτού επιλέχθηκε να είναι 100 επί 75 εικονοστοιχεία. Το μέγεθος αυτό επιλέχθηκε έτσι ώστε στην συνέχεια που θα μετατρέψουμε το μέγεθος του παραθύρου σε μικρότερες διαστάσεις για να πάρουμε την είσοδο του ταξινομητή να μην υπάρξουν αλλοιώσεις στην εικόνα. Ταυτόχρονα το μέγεθος αυτό κρίθηκε ευρυστικά ικανοποιητικό ώστε η διάσχιση όλης της εικόνας να μην απαιτεί πολύ χρόνο ενώ ταυτόχρονα να μην υπάρχει περίπτωση να χάσουμε πληροφορία. Για να επιτευχθεί αυτό δημιουργήθηκε ένα πρόγραμμα με χρήση της βιβλιοθήκης openCV η οποία δεδομένου ενός καρέ, των διαστάσεων του παραθύρου καθώς και το βήμα μετακίνησης, να διασχίζει όλο το καρέ και να παράγει τα τμήματα που απαιτούνται. Ο κώδικας αυτός είναι απαραίτητος και για το σύνολο του συστήματος μιας και για να λειτουργήσει το σύστημα όπως αναφέραμε η διαδικασία που ακολουθείται από τον αλγόριθμο είναι να διασχίζει την εικόνα, να την τροφοδοτεί στον ταξινομητή και στην συνέχεια με βάση την απάντηση του να κρατάει το καρέ που απεικονίζει τα χείλη. Ο κώδικας αυτός όπως και ο υπόλοιπος που δημιουργήθηκε για την παρούσα εργασία εμφανίζεται στο αντίστοιχο παράρτημα. Στην συνέχεια χειροκίνητα τα καρέ κατηγοριοποιήθηκαν σε δύο κατηγορίες, η πρώτη είναι αυτή που περιέχει χείλη και η δεύτερη αυτή που περιέχει αρνητικά παραδείγματα. Ένα υποσύνολο των δεδομένων των δύο συνόλων παρουσιάζεται στο **Σχήμα 18**

Σχήμα 18: Παραδείγματα θετικών και αρνητικών παραδειγμάτων για την εκπαίδευση του SVM. Στην πρώτη γραμμή έχουμε θετικά και στην δεύτερη αρνητικά παραδείγματα.

Θα πρέπει εδώ να αναφέρουμε πως η διάσχιση της εικόνας δεν έγινε στο σύνολο του καρέ. Εκμεταλλευόμενοι την γνώση πως το καρέ εισόδου θα περιέχει κεντραρισμένο πρόσωπο, διασχίζουμε το καρέ από την μέση και κάτω, αφού η περιοχή που αναζητούμε να εντοπίσουμε βρίσκεται σε αυτό το τμήμα. Με αυτό τον τρόπο καταφέρνουμε να μειώσουμε τον χρόνο που απαιτείται για την επεξεργασία του καρέ.

Στην συνέχεια πρέπει τα δύο σύνολα εικόνων που έχουμε ετοιμάσει να τα επεξεργαστούμε κατάλληλα ώστε να αποτελέσουν κατάλληλη είσοδο για τον ταξινομητή. Για να γίνει αυτό κάναμε την επεξεργασία στο περιβάλλον matlab και ο κώδικας παρατίθεται στο παράρτημα. Το πρώτο βήμα της επεξεργασίας είναι η μετατροπή της έγχρωμης εικόνας που παίρνουμε σε μία εικόνα αποχρώσεων του γκρι. Το δεύτερο βήμα είναι η μεταβολή των διαστάσεων της εικόνας σε ένα μέγεθος το οποίο κρίνεται καταλληλότερο για τους ταξινομητές SVM που χρησιμοποιούμε. Οι νέα διάσταση των εικόνων εισόδου μετατρέπεται πλέον στα 25 επί 18 εικονοστοιχεία. Στην συνέχεια γίνεται εφαρμογή προσαρμοστικής κανονικοποίησης με ιστόγραμμα στην εικόνα. Ο λόγος για αυτό το βήμα είναι το ότι στην βιβλιογραφία έχει αποδειχθεί πως αυτή η διαδικασία βελτιώνει την διακριτική ικανότητα των ταξινομητών. Τέλος τα δεδομένα μετασχηματίζονται σε ένα διάνυσμα 1 επί 450 που είναι η κατάλληλη είσοδος για τον ταξινομητή. Τα δεδομένα που προκύπτουν κανονικοποιούνται στο διάστημα [-1,1] και αποθηκεύονται στον δίσκο ως δύο διαφορετικά αρχεία μαζί με τους δείκτες για το αν το κάθε διάνυσμα είναι θετικό η αρνητικό παράδειγμα. Στην συνέχεια τα δεδομένα αυτά περνάνε από μία τελική επεξεργασία με scripts της python έτσι ώστε να καταλήξουμε να έχουμε ένα αρχείο για τα θετικά και ένα για τα αρνητικά παραδείγματα, στο φορμά εισόδου που επιβάλει το λογισμικό libsvm. Στο **Σχήμα 19** παρουσιάζεται σχηματικά όλη η διαδικασία της προ επεξεργασίας των δεδομένων εκπαίδευσης.



Σχήμα 19: Συνοπτική διαγραμματική παρουσίαση της διαδικασίας προ-επεξεργασίας των δεδομένων για την εκπαίδευση του ταξινομητή.

3.2.3 Εκπαιδευση ταξινομητων για την αναγνωριση της περιοχης των χειλιων

Ως ταξινομητής, όπως προαναφέραμε επιλέχθηκε μία μηχανή διανυσμάτων στήριξης (SVM). Για τις ανάγκες της εφαρμογής που περιγράψαμε χρησιμοποιήσαμε το ανοιχτό πακέτο λογισμικού για μηχανές διανυσμάτων στήριξης που παρουσιάζεται στο [8]. Τα δεδομένα μετά την επεξεργασία που αναφέραμε στο προηγούμενο κεφάλαιο είναι έτοιμα για να εισαχθούν στον ταξινομητή για την εκπαίδευση, αφού έχουν τροποποιηθεί κατάλληλα με την επεξεργασία τους με χρήση κατάλληλων scripts της python.

Για την εκπαίδευση του ταξινομητή τα δεδομένα τα οποία δημιουργήθηκαν χωρίστηκαν σε δύο ομάδες. Ένα τμήμα θετικών παραδειγμάτων από κάθε ομιλητή καθώς και ένα σύνολο από αρνητικά παραδείγματα, αποτέλεσαν το δείγμα εισόδου για την εκπαίδευση. Τα υπόλοιπα θετικά και αρνητικά παραδείγματα, χρησιμοποιήθηκαν για την αξιολόγηση του ταξινομητή που προέκυπτε από την εκπαίδευση. Στον **Πίνακας 3**παρουσιάζονται τα πλήθη των δεδομένων που χρησιμοποιήθηκαν ανά περίπτωση. Το σύνολο παραδειγμάτων εκπαίδευσης αποτελείται από συνολικά 18960 θετικά και αρνητικά παραδείγματα. Για τον πρώτο ομιλητή δημιουργήθηκε ένα

σύνολο ελέγχου μεγέθους 879 θετικών παραδειγμάτων. Για τον δεύτερο ομιλητή δημιουργήθηκε ένα σύνολο ελέγχου μεγέθους 1789 θετικών παραδειγμάτων. Για τον τρίτο ομιλητή δημιουργήθηκε ένα σύνολο ελέγχου μεγέθους 879 θετικών παραδειγμάτων. Τέλος το σύνολο ελέγχου των αρνητικών παραδειγμάτων έχει μέγεθος 12548.

Δεδομένα εκπαίδευσης			**Δεδομένα ελέγχου**	**Πλήθος**
Θετικά	7094		Ομιλητής $1^{ος}$ (θετικά)	879
Αρνητικά	11865		Ομιλητής $2^{ος}$ (θετικά)	1789
			Ομιλητής $3^{ος}$ (θετικά)	879
			Αρνητικά (σύνολο)	12548

Πίνακας 3: Παρουσίαση του καταμερισμού δεδομένων εκπαίδευσης και ελέγχου των αποτελεσμάτων της εκπαίδευσης των μηχανών διανυσμάτων στήριξης.

Μετά την επιλογή των δεδομένων εκπαίδευσης και ελέγχου, καθώς και την προετοιμασία τους, το επόμενο βήμα είναι η εκπαίδευση του ταξινομητή ώστε στο τέλος να καταλήξουμε με ένα σύστημα το οποίο είναι ικανό να διαχωρίσει τις εικόνες μεταξύ των συνόλων που έχουμε ορίσει. Στην προκειμένη περίπτωση, μεταξύ εικόνων που περιέχουν χείλη και εικόνων που δεν περιέχουν χείλη. Για να γίνει αυτό πρέπει να ορισθούν και κατά περίπτωση να εντοπιστούν ένα σύνολο παραμέτρων που χαρακτηρίζουν τον ταξινομητή και τα οποία δίνουν τα βέλτιστα αποτελέσματα. Στην περίπτωση των μηχανών διανυσμάτων στήριξης το πρώτο βήμα στον ορισμό των παραμέτρων είναι η επιλογή του πυρήνα. Η σημασία και χρήση των πυρήνων είναι πολύ ουσιαστική και παρουσιάζονται στο κεφάλαιο στο οποίο περιγράφουμε της μηχανής διανυσμάτων στήριξης. Ο πυρήνας έχει παρατηρηθεί κυρίως πειραματικά στην βιβλιογραφία πως επηρεάζει σε μεγάλο βαθμό την απόδοση της μηχανής, ανάλογα με το είδος των δεδομένων προς ταξινόμηση. Για την ακρίβεια έχουν κατασκευαστεί πυρήνες για συγκεκριμένα

προβλήματα τα οποία αποδεικνύεται κιόλας πως αποδίδουν καλύτερα σε σχέση με τους υπόλοιπους τους. Στην περίπτωση την δική μας αποφασίσαμε να χρησιμοποιήσουμε πολυωνυμικούς πυρήνες. Ο λόγος που μας οδήγησε σε αυτή την επιλογή είναι πως στην βιβλιογραφία παρουσιάζονται οι συγκεκριμένοι πυρήνες πως αποδίδουν καλύτερα στην περίπτωση της ταξινόμησης προσώπων. Επειδή τα δεδομένα μας αποτελούν τμήματα προσώπου θεωρήσαμε πως ο συγκεκριμένος πυρήνας αποτελεί καλό υποψήφιο. Πρέπει εδώ να τονίσουμε πως πέρα από τις αναφορές στην βιβλιογραφία, δεν υπάρχει κάποια σαφής θεωρία η οποία να μπορεί να μας οδηγήσει στην σωστή επιλογή τόσο του πυρήνα όσο και των παραμέτρων του για κάποιο συγκεκριμένο πρόβλημα ταξινόμησης. Ο πειραματισμός και η δοκιμή μαζί με κάποιες ευριστικές μεθόδους αποτελούν τις καθιερωμένες μεθόδους για την παραπάνω διαδικασία.

Ο πολυωνυμικός πυρήνας, όπως αυτός ορίζεται και στο libsvm, το οποίο χρησιμοποιήθηκε για την συγκεκριμένη εφαρμογή, έχει την ακόλουθη μορφή,

$(\gamma u v + coef0)^d$ *(7)*

Όπου **γ** και **coef0** αποτελούν παραμέτρους του πυρήνα, d είναι ο βαθμός του πολυωνύμου και **u,v** είναι διανύσματα εισόδου. Οι παράμετροι γ, coef0 και d εξαρτώνται από την φύση του προβλήματος και πρέπει να επιλεγούν σωστά ώστε να έχουμε τα επιθυμητά αποτελέσματα. Δεν υπάρχει κάποιος συγκεκριμένος σαφής τρόπος ο οποίος να μπορεί να μας οδηγήσει στην επιλογή τους και ο βασικός τρόπος είναι ο πειραματισμός και η χρήση κάποιας μεθόδους grid search για την προσέγγιση των βέλτιστων τιμών.

Η βιβλιοθήκη libsvm, περιέχει και μία εφαρμογή υλοποιημένη σε python η οποία εφαρμόζει την μέθοδο grid search για τον εντοπισμό βέλτιστων παραμέτρων του πυρήνα με βάση τα συγκεκριμένα δεδομένα εισόδου τα οποία διαθέτουμε. Το πρόβλημα αυτής της εφαρμογής είναι πώς χρησιμοποιεί μόνο τον πυρήνα RBF (Radial Basis function). Ο λόγος που γίνεται χρήση μόνο αυτού του πυρήνα είναι το ότι απαιτείται ο προσδιορισμός μόνο δύο παραμέτρων εν αντίθεση με τον πολυωνυμικό πυρήνα. Αυτό κάνει τόσο την διαδικασία του grid search, όσο και την οπτικοποίηση των αποτελεσμάτων πιο εύκολη. Αυτό το οποίο αποφασίσαμε να κάνουμε στην συγκεκριμένη περίπτωση είναι να πραγματοποιήσουμε το grid search με τον πυρήνα που διαθέτει το λογισμικό που είχαμε στην κατοχή μας έτσι ώστε να βρούμε κάποιες ανεκτές τιμές για τις δύο παραμέτρους και στην συνέχεια με χρήση ενός δικού μας λογισμικού σε python να

χρησιμοποιήσουμε αυτές τις τιμές και να μεταβάλουμε μόνο τον βαθμό του πολυωνύμου με σκοπό να εντοπίσουμε μία περιοχή γύρω από αυτές τις τιμές με μεγαλύτερη ακρίβεια, όπου τα αποτελέσματα θα είναι ακόμα καλύτερα. Στους **Πίνακας 4** έως **Πίνακας 10** παρουσιάζονται ενδεικτικές τιμές των παραμέτρων καθώς και τα αντίστοιχα αποτελέσματα των ελέγχων που πραγματοποιήθηκαν στα δεδομένα με το προκύπτον μοντέλο.

Αποτελέσματα	
Ομιλητής $1^{ος}$ (θετικά)	57.1%
Ομιλητής $2^{ος}$ (θετικά)	83.2%
Ομιλητής $3^{ος}$ (θετικά)	100%
Αρνητικά (σύνολο)	88%

Πίνακας 4: Αποτελέσματα εκπαίδευσης ταξινομητή με παραμέτρους πολυωνυμικού πυρήνα, d = 2, g = 0.0625, c = 1, r = 1

Αποτελέσματα	
Ομιλητής $1^{ος}$ (θετικά)	57.7%
Ομιλητής $2^{ος}$ (θετικά)	83.4%
Ομιλητής $3^{ος}$ (θετικά)	100%
Αρνητικά (σύνολο)	86.15%

Πίνακας 5: Αποτελέσματα εκπαίδευσης ταξινομητή με παραμέτρους πολυωνυμικού πυρήνα, d = 3, g = 0.0625, c = 1, r = 1

Αποτελέσματα	
Ομιλητής $1^{ος}$ (θετικά)	58.13%
Ομιλητής $2^{ος}$ (θετικά)	83.9%
Ομιλητής $3^{ος}$ (θετικά)	100%
Αρνητικά (σύνολο)	84.66%

Πίνακας 6:Αποτελέσματα εκπαίδευσης ταξινομητή με παραμέτρους πολυωνυμικού πυρήνα, d = 4, g = 0.0625, c = 10, r = 1

Αποτελέσματα	
Ομιλητής $1^{ος}$ (θετικά)	58.248%
Ομιλητής $2^{ος}$ (θετικά)	84.12%
Ομιλητής $3^{ος}$ (θετικά)	100%
Αρνητικά (σύνολο)	83.1%

Πίνακας 7: Αποτελέσματα εκπαίδευσης ταξινομητή με παραμέτρους πολυωνυμικού πυρήνα, d = 4, g = 0.0625, c = 10, r = 0.05

Αποτελέσματα	
Ομιλητής $1^{ος}$ (θετικά)	58.248%
Ομιλητής $2^{ος}$ (θετικά)	84.12%
Ομιλητής $3^{ος}$ (θετικά)	100%
Αρνητικά (σύνολο)	83.8%

Πίνακας 8: Αποτελέσματα εκπαίδευσης ταξινομητή με παραμέτρους πολυωνυμικού πυρήνα, d = 4, g = 0.0625, c = 10, r = 0.005

Αποτελέσματα	
Ομιλητής $1^{ος}$ (θετικά)	58.93%
Ομιλητής $2^{ος}$ (θετικά)	91.89%
Ομιλητής $3^{ος}$ (θετικά)	100%
Αρνητικά (σύνολο)	71.358%

Πίνακας 9:Αποτελέσματα εκπαίδευσης ταξινομητή με παραμέτρους πολυωνυμικού πυρήνα, d = 4, g = 0.000625, c = 100, r = 0.0005

Αποτελέσματα	
Ομιλητής 1ος (θετικά)	79.86%
Ομιλητής 2ος (θετικά)	93.9%
Ομιλητής 3ος (θετικά)	100%
Αρνητικά (σύνολο)	90.05%

Πίνακας 10:Αποτελέσματα εκπαίδευσης ταξινομητή με παραμέτρους πολυωνυμικού πυρήνα, d = 5, g = 0.000625, c = 100, r = 0.0005

Όπως μπορούμε να παρατηρήσουμε τον καλύτερο ρυθμό αναγνώρισης τον επιτυγχάνουμε με τις παραμέτρους που παρουσιάζονται στον **Πίνακας 10**. Ενδιαφέρον παρουσιάζει το γεγονός ότι στην περίπτωση του 3ου ομιλητή έχουμε σε κάθε περίπτωση το καλύτερο ποσοστό αναγνώρισης το οποίο μάλιστα παραμένει σταθερό ανεξάρτητα από τις τιμές των παραμέτρων. Αυτό μπορεί να δικαιολογηθεί, γιατί ο συγκεκριμένος ομιλητής είναι ο μόνος ο οποίος έχει μούσι, κάτι το οποίο κάνει το σύνολο παραδειγμάτων του να είναι σαφώς πιο διαχωρίσιμο από τα υπόλοιπα. Εν αντίθεση με τον πρώτο ομιλητή ο οποίος επιτυγχάνει πάντα το χειρότερο ρυθμό αναγνώρισης, κάτι το οποίο οφείλεται στο ότι χρωματικά τα χείλη του είναι όμοια σε σχέση με το υπόλοιπο δέρμα του. Πρέπει εδώ να αναφέρουμε πως στα δεδομένα ελέγχου σε κάθε περίπτωση είχαμε θετικά παραδείγματα μόνο ενός ομιλητή, ενώ τα αρνητικά παραδείγματα εμπεριείχαν παραδείγματα από όλους τους ομιλητές. Αυτό αποτελεί έναν καλό έλεγχο της ικανότητας γενίκευσης που έχει ο ταξινομητής μας.

Σε αυτό το σημείο πρέπει να αναφερθούμε και στην διάκριση μεταξύ των λαθών που κάνει ο ταξινομητής και στις επιπτώσεις που αυτά έχουν στο σύστημα μας. Έχουμε δύο κατηγορίες σφαλμάτων, η πρώτη κατηγορία εμπεριέχει τα χείλη τα οποία λανθασμένα έχουν κατηγοριοποιηθεί ως καρέ τα οποία δεν εμπεριέχουν χείλος, και η δεύτερη κατηγορία εμπεριέχει τα καρέ τα οποία δεν περιέχουν χείλη και τα οποία έχουν κατηγοριοποιηθεί ως καρέ με χείλη. Η πρώτη κατηγορία είναι πιο επιθυμητή από την δεύτερη. Ο λόγος για αυτό είναι το ότι αν ένα καρέ ταξινομηθεί πως δεν περιέχει χείλη, το σύστημα θα επαναλάβει τον έλεγχο στο επόμενο

και είναι αρκετά πιθανό ότι θα εντοπίσει χείλη, ενώ ταυτόχρονα τα υπόλοιπα μέρη του συστήματος δεν θα τροφοδοτηθούν με δεδομένα. Με αυτό τον τρόπο μπορούμε να εξασφαλίσουμε την ποιότητα των δεδομένων που θα τροφοδοτήσουν τα υπόλοιπα μέρη του συστήματος. Στην περίπτωση όμως της δεύτερης κατηγορίας σφαλμάτων, το υπόλοιπο σύστημα θα τροφοδοτηθεί με δεδομένα τα οποία δεν εμπεριέχουν χείλη και αυτό θα έχει ως αποτέλεσμα την είσοδο μη επιθυμητού θορύβου στα υπόλοιπα τμήματα του συστήματος και μάλιστα με τρόπο που δεν είναι εύκολος ο εντοπισμός του σημείου στο οποίο έχει γίνει το σφάλμα μιας και τα υπόλοιπα τμήματα εμπεριέχουν πάλι ταξινομητές. Για αυτό τον λόγο επιθυμούμε γενικά όχι μόνο να έχουμε υψηλό ρυθμό αναγνώρισης χειλιών αλλά και υψηλό ρυθμό επιτυχούς αναγνώρισης των αρνητικών παραδειγμάτων. Για όλους τους παραπάνω λόγους κρίνεται πως τα αποτελέσματα του μοντέλου που παρουσιάζεται στον **Πίνακας 10**είναι αρκετά ικανοποιητικό για τις ανάγκες του συστήματος μας.

ΚΕΦΑΛΑΙΟ 4ο.

ΕΝΤΟΠΙΣΜΟΣ ΚΑΙ ΕΞΑΓΩΓΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

Το επόμενο βήμα στο σύστημα μας, είναι ο εντοπισμός και η εξαγωγή των χαρακτηριστικών εκείνων, τα οποία αν τροφοδοτηθούν σε ένα κατάλληλο ταξινομητή θα μπορέσει να προκύψει η λέξη την οποία προφέρει ο ομιλητής. Αυτό το οποίο επιχειρούμε να κάνουμε είναι να δημιουργήσουμε ένα διάνυσμα, το οποίο αποτελείται απο αυτά τα χαρακτηριστικά και το οποίο θα αποτελέσει την είσοδο του ταξινομητή. Η πιο συνηθισμένη διαδικασία η οποία συναντάται στην βιβλιογραφία είναι η απευθείας χρήση της εικόνας των χειλιών ως είσοδος σε έναν ταξινομητή, για παράδειγμα ένα νευρονικό δίκτυο, και η ταύτηση του με κάποιο viseme. Πολλές μέθοδοι έχουν προταθεί για την βελτίωση της απόδοσης των συστημάτων που βασίζονται στην ταύτηση των visemes, όπως για παράδειγμα η εξαγωγή του περιγράμματος των χειλιών ή η εξαγωγή χρωματικών συνιστοσών. Για τα visemes καθώς και για μεθόδους βελτιστοποίησης, καθώς και για την μέθοδο που επιλέξαμε θα αναφερθούμε στην συνέχεια.

4.1. Περιγραφη μεθοδων εξαγωγης οπτικων χαρακτηριστικων

Οι μέθοδοι εξαγωγής των οπτικών χαρακτηριστικών χωρίζονται στις ίδιες κατηγορίες με αυτές που αναφέραμε, τόσο για τον εντοπισμό του προσώπου, όσο και για τον προσδιορισμό της

περιοχής των χειλιών. Συγκεκριμένα όπως αυτές συνήθως αναφέρονται στην βιβλιογραφία είναι οι, μέθοδοι οι οποίες βασίζονται στην εμφάνιση, μέθοδοι οι οποίοι βασίζονται στο σχήμα και τέλος μέθοδοι οι οποίοι συνδιάζουν και τις δύο παραπάνω κατηγορίες.

Οι μέθοδοι που βασίζονται στην εμφάνιση, επεξεργάζονται την ένταση και την χρωματική πληροφορία η οποία εμπεριέχεται σε μία περιοχή ενδιαφέροντος, συνήθως αυτή η περιοχή είναι το καρέ το οποίο περιέχει τα εντοπισμένα χείλη και σε κάποιες περιπτώσεις και τα μάγουλα. Οι διαστάσεις του διανύσματος χαρακτηριστικών το οποίο προκύπτει συνήθως ελαχιστοποιείται με κάποια στατιστική μέθοδο, όπως είναι η μέθοδος πρωτογενών συνιστωσών, η κάποια άλλη γραμμική ή όχι μέθοδος. Μερικά παραδείγματα της «από κάτω προς τα πάνω» προσέγγισης που αναφέραμε είναι τα ακόλουθα:

- Η μέθοδος που παρουσιάζεται στο [22] χρησιμοποιεί μόνο επίπεδα του γκρι
- Ανάλυση πρωτογενών συνιστωσών στην πυκνότητα των εικονοστοιχείων, όπως αυτή παρουσιάζεται στο [4]
- Κωδικοποίηση της κίνησης μεταξύ συνεχόμενων καρέ [20]
- Χρήση ακμών [1]
- Χρήση φίλτρων, όπως αυτά παρουσιάστηκαν και στην περίπτωση του εντοπισμού της περιοχής ενδιαφέροντος στο προηγούμενο κεφάλαιο [14].

Σε αντίθεση με τις μεθόδους που περιγράψαμε παραπάνω, η χρήση μεθόδων που βασίζονται στο σχήμα χρησιμοποιούν ένα «από πάνω προς τα κάτω» μοντέλο. Οι παράμετροι του μοντέλου τροποποιούνται με βάση τα χαρακτηριστικά της εικόνας και χρησιμοποιούντε ως οπτικά χαρακτηριστικά. Χαρακτηριστικά παραδείγματα αυτής της προσέγγισης χρησιμοποιούν κυρίως γεωμετρικά χαρακτηριστικά της εικόνας, όπως για παράδειγμα το ύψος και το πλάτος των χειλιών [1]. Περιγραφείς fourier του περιγράμματος των χειλιών [30]. Γενικά, το περίγραμμα των χειλιών απο μόνο του στερείται της απαραίτητης διακριτικής ικανότητας, για αυτό τον λόγο πολύ συχνά συνδιάζεται μαζί με χαρακτηριστικά της εμφάνισης. Έχει δειχθεί πως ο συνδιασμός της εμφάνισης με χαρακτηριστικά του σχήματος, βελτιώνει αισθητά την απόδοση τέτοιων συστημάτων [14]. Το αποτέλεσμα της σύνθεσης των δύο μεθόδων είναι η δημιουργία ενός

ενεργού μοντέλου εμφάνισης, αναφορές σε αυτό έχουν γίνει και στο κεφάλαιο για τον προσδιορισμό της περιοχής των χειλιών.

4.2. Μοντελοποιηση της μοναδας ομιλιας

Παραδοσιακά, η ομιλία θεωρείται πως αποτελείται απο μία ακολουθία απο συνεχόμενες βασικές μονάδες οι οποίες αποκαλούντε φωνήματα. Αυτή η θεωρία είναι σε αρμονία με την αρχική θεωρία της παραγωγικής φωνολογίας [26]. Η αγγλική γλώσσα, η οποία χρησιμοποιήθηκε και στην παρούσα εργασία, αποτελείται από 50 φωνητικές μονάδες.

4.2.1 Visemes

Το Viseme, αποτελεί την βασική μονάδα ομιλίας στο οπτικό πεδίο. Αποτελεί το αντίστοιχο του φωνήματος, το οποίο ορίζεται για το ακουστικό πεδίο [6]. Παρόλο που δεν υπάρχει απόλυτη ταύτηση των δύο όρων, πλέον χρησιμοποιείται ο όρος με την παραπάνω έννοια. Ο λόγος που δεν υπάρχει απόλυτη ταύτηση είναι διότι το φώνημα, αποτελεί το ελάχιστο διακριτό δομικό στοιχείο της ομιλίας, κάτι το οποίο δεν ισχύει απόλυτα για το viseme. Ένα συγκεκριμένο Viseme ορίζει τις συγκεκριμένες κινήσεις του προσώπου και του στώματος τα οποία προκύπτουν κατά την παραγωγή ενός φωνήματος. Αυτό έχει σαν αποτέλεσμα πολλά φωνήματα να έχουν το ίδιο viseme, κάτι το οποίο στερεί απο το τελευταίο την διακριτότητα που προαναφέραμε.

Όπως προαναφέραμε τα visemes και τα φωνήματα δεν έχουν μία ένα προς ένα αντιστοιχία. Πολύ συχνά πολλά φωνήματα μοιράζονται το ίδιο viseme. Αντίστοιχα υπάρχουν πολλά φωνήματα τα οποία ακουστικά και μόνο είναι δύσκολο να διακριθούν, ενώ οπτικά μέσω των visemes η διάκριση και αναγνώριση τους είναι πολύ πιό εύκολη. Ένα παράδειγμα το οποίο αποδίδει τα παραπάνω, είναι το γεγονός πως οταν δύο άνθρωποι μιλάνε στο τηλέφωνο υπάρχουν πολύ περισσότερα σφάλματα απο ότι οταν μιλάνε απευθείας. Για αυτό το λόγο πολλοί γλωσσολόγοι έχουν οδηγηθεί στο συμπέρασμα, πως η ομιλία και η κατανόηση της ως διαδικασία εξαρτάται τόσο απο την ακουστική οδό, όσο και απο την οπτική. Στην περίπτωση που μία απο τις δύο οδούς απουσιάζει, τότε η διαδικασία της κατανόησης της ομιλίας παύει να είναι ομαλή. Ένα κλασσικό πείραμα το οποίο αποδεικνύει τα παραπάνω είναι αυτό το οποίο παρουσιάζεται στο [13]. Μία τυπική αντιστοιχεία viseme σε φωνήματα παρουσιάζεται στον **Πίνακας 11**.

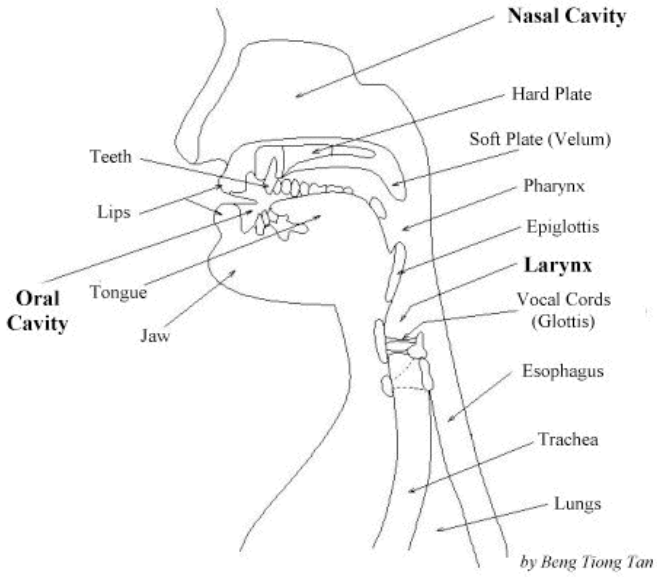
Viseme Index	Corresponding Phonemes
1	ax ih iy dx
2	ah aa
3	ae eh ay ey hh
4	aw uh u wow ao w oy
5	el l
6	er axr r
7	y
8	b p
9	bcl pcl m em
10	sz epi tcl dcl n en
11	ch jh sh zh
12	t d th dh g k
13	f v
14	gcl kcl ng

Πίνακας 11: Παρουσίαση ενδεικτικής αντιστοίχησης μεταξύ visemes και φωνημάτων.

4.3. Χαρακτηριστικα της αρθρωσης

Τα τελευταία χρόνια μία ανταγωνιστική θεωρία της μή γραμμικής φωνολογίας έχει αρχίσει να ελκύει το ενδιαφέρον της κοινότητας που ασχολείται με την αυτόματη αναγνώριση ομιλίας. Η θεωρία αυτή βασίζεται στην γνώση του μηχανισμού παραγωγής της ανθρώπινης ομιλίας, αντιμετοπίζει την ομιλία ως τον συνδιασμό πολλαπλών πηγών κρυφών χαρακτηριστικών άρθρωσης [31]. Ένα διάγραμμα το οποίο παρουσιάζει τον μηχανισμό παραγωγής της ανθρώπινης ομιλίας παρουσιάζεται στο **Σχήμα 20**. Η φωνητική οδός, η οποία αποτελεί το βάσικο όργανο παραγωγής της ομιλίας, αποτελείται απο τον φάρυγγα, την στωματική κοιλότητα και την ρινική κοιλότητα. Η γλωττίς, το ιστίον, η γλώσσα, τα χείλη και το σαγόνι, αποτελούν μηχανισμούς άρθρωσης. Η διαδικασία της μεταβολής του σχήματος της φωνητικής οδούς με σκοπό την παραγωγή διαφορετικών ήχων, ονομάζεται άρθρωση [11]. Με βάση τα παραπάνω, κάθε φώνημα μπορεί να ορισθεί με βάση κάποιο πλήθος χαρακτηριστικών άρθρωσης, για παράδειγμα η θέση του κορμού της γλώσσας, η θέση της άκρης της γλώσσας, η χρήση των δοντιών και άλλα.

Ένα απο τα βασικά πλεονεκτήματα της αναπαράστασης της ομιλίας ως τον συνδιασμό ενός πλήθους πηγών άρθρωσης, είναι η δυνατότητα της μοντελοποίησης κάθε πηγής (χαρακτηριστικού) ανεξάρτητα απο τα υπόλοιπα. Ακόμα είναι δυνατό να επιτραπεί αυτές οι πηγές να αποσυγχρονιστούν μεταξύ τους χωρίς να χάσουμε διακριτική ικανότητα. Για παράδειγμα, έχει παρατηρηθεί πως η αυθόρμητη ομιλία, είναι δύσκολο να μοντελοποιηθεί με βάση τα φωνήματα μόνα τους και η κατανόηση της αποτελεί βασική πρόκληση για τα σύγχρονα συστήματα αυτόματης κατανόησης ομιλίας. Τα μοντέλα προφοράς τα οποία έχουν προκύψει απο χρήση χαρακτηριστικών άρθρωσης, έχει αποδειχθεί πως καταφέρνουν σε πολύ μεγαλύτερο βαθμό να μοντελοποιήσουν την ποικιλία προφοράς που εμφανίζεται στην αυθόρμητη ομιλία [19]. Ένα άλλο βασικό πλεονέκτημα αυτής της προσέγγισης είναι πως τα μοντέλα αυτά που προκύπτουν καταφέρνουν να ανταπεξέλθουν σε πολύ μεγαλύτερο βαθμό σε περιβάλλον με έντονο θόρυβο.



Σχήμα 20: Μηχανισμός παραγωγής της ομιλίας στον άνθρωπο.

Δίνεται λοιπόν η δυνατότητα της μοντελοποίησης της οπτικής ομιλίας, με τρόπο ο οποίος είναι πιο συνεπής ως προς την πραγματική διαδικασία παραγωγής της ομιλίας στον άνθρωπο. Η μοντελοποίηση αυτή εν μέρη είναι επιρεασμένη απο την χρήση των σημείων άρθρωσης που προ αναφέραμε. Βέβαια στην παρούσα εργασία λειτουργούμε μόνο με οπτική πληροφορία και για αυτό τον λόγο μπορεί να γίνει χρήση μόνο ορατών απο το μάτι σημείων άρθρωσης. Έχοντας λοιπόν εντοπίσει την περιοχή ενδιαφέροντος που εμπεριέχει τα χείλη, μπορούμε να λάβουμε πληροφορίες για την θέση και την σχετική παραμετροποίηση των χειλιών, της γλώσσας, των δοντιών και της γνάθου. Τα υπόλοιπα σημεία άρθρωσης δεν είναι ορατά κάτω απο φυσιολογικές συνθήκες λήψης. Ταυτόχρονα, σε συνδιασμό με τα στατικά σημεία άρθρωσης, το βίντεο εμπεριέχει και δυναμικά σημεία άρθρωσης, για παράδειγμα το κλείσιμο των χειλιών και το άνοιγμα τους, το άν η γλώσσα εκτίνεται η μαζεύεται σε σχέση με τα δόντια, αν το κάτω χείλος ακουμπά τα πάνω δόντια κοκ.

Η μέθοδος της οπτικής μοντελοποίησης των σημείων άρθρωσης μιάζει σε μεγάλο βαθμό με την αντίστοιχη ακουστική μοντελοποίηση. Ουσιαστικά προτείνεται η μοντελοποίηση της οπτικής ομιλίας ως ένα σύνολο πολλαπλών πηγών, οπτικών γλωσσολογικών χαρακτηριστικών, εν αντιθέση με την μοντελοποίηση που θεωρεί μόνο μία πηγή από visemes. Οι περισσότερες απο τις διαδικασίες άρθρωσης και οι αντίστοιχες μεταβολές που παρατηρούντε βρίσκοντε σε άμεση αντιστοίχηση με το σύνολο χαρακτηριστικών που παρουσιάζεται στο μοντέλο προφοράς στο

[19]. Για παράδειγμα, το οπτικό χαρακτηριστικό των χειλιών που κλείνουν ή ανήγουν, ταυτίζεται με το χαρακτηριστικό LIP-OPEN. Για αυτό τον λόγο μπορούμε να θεωρήσουμε πως ένα σύστημα αναγνώρισης ομιλίας που βασίζεται στα χαρακτηριστικά άρθρωσης μπορεί να κάνει χρήση του ίδιου συνόλου χαρακτηριστικών με αυτά που αναφέροντε στο [19]. Παρόλα αυτά, λόγω της συμπληρωματικής σχέσης που έχουν μεταξύ τους η οπτική και ακουστική οδός, κάποια χαρακτηριστικά μπορούν να αποδοθούν και να εντοπισθούν καλύτερα απο την ηχητική ή οπτική αντίστοιχα πηγή, ειδικά σε περιπτώσεις όπου εμφανίζεται θόρυβος. Για παράδειγμα, είναι γνωστό απο μελέτες, πως η παρουσία ηχητικού θορύβου επηρεάζει τον εντοπισμό του σημείου άρθρωσης περισότερο απο τον προσδιορισμο του φωνισμού [12]. Από την άλλη ο προσδιορισμός της θέσης γίνεται πολύ πιο εύκολα στο οπτικό πεδίο κάτω απο συνθήκες παρουσίας θορύβου.

Το παρακάτω παράδειγμα θα μας βοηθήσει να αντιληφθούμε τις διαφορές μεταξύ ενός συστήματος αναγνώρισης visemes και ενός συστήματος που βασίζεται στον εντοπισμό των σημείων άρθρωσης. Αν θεωρήσουμε πως θέλουμε να μοντελοποιήσουμε το αγγλικό φώνημα /m/ σε δύο διαφορετικά φωνητικά περιεχόμενα, στην λέξη romantic και στην λέξη academic. Στο σχήμα 4.2 εμφανίζεται η εικόνα του σχήματος των χειλιών κατα την παραγωγή του φωνήματος στις δύο διαφορετικές λέξεις. Τα δύο αυτά παραδείγματα θεωρούντε πως ανήκουν στην ίδια κλάση viseme και πως έχουν την ίδια τιμή του χαρακτηριστικού ανοιχτό/κλειστό, συγκεκριμένα τελείως κλειστό. Παρόλα αυτά όμως όπως φένεται και στο σχήμα, η εμφάνιση των χειλιών είναι διαφορετική, στην περίπτωση της δεύτερης λέξης η απόσταση μεταξύ των γωνιών των χειλιών είναι κατά 25% πιό πλατιά. Αυτό το στοιχείο μπορεί να μας οδηγήσει στο συμπέρασμα της ύπαρξης πληροφορίας που σχετίζεται με συγκεκριμένο περιεχόμενο (εμφάνιση του ίδιου φωνήματος αλλά διαφορετική λέξη). Συγκεκριμένα το φώνημα /ow/ που προηγείται στην περίπτωση της λέξης romantic αναγκάζει το /m/ να να ανήκει στην κατηγορία rounded ενώ στην περίπτωση της λέξης όπου προηγείται το φώνημα /eh/ αυτό δεν ισχύει. Για αυτό τον λόγο αν μοντελοποιήσουμε το άνοιγμα και το σχήμα των χειλιών ως δύο διαφορετικά χαρακτηριστικά θα έχουμε την δυνατότητα να ανακτήσουμε περισσότερη πληροφορία απο το να μοντελοποιήσουμε απλά το /m/ ως viseme. Μία εναλλακτική λύση θα ήταν η χρήση μεγαλύτερου μήκους μονάδων ομιλίας, για παράδειγμα bi-visemes ή tri-visimes, παρόλα αυτά όμως αυτό θα οδηγούσε σε άλλα προβλήματα όπως για παράδειγμα η μείωση των διαθέσιμων δεδομένων εκπαίδευσης ανά κλάσης καθώς και την αύξηση των παραμέτρων του μοντέλου.

Θα πρέπει σε αυτό το σημείο να αναφέρουμε πως η χρήση της μεθόδου εξαγωγής χαρακτηριστικών άρθρωσης διαφέρει απο την περίπτωση της εξαγωγής γεωμετρικών χαρακτηριστικών. Η δεύτερη περίπτωση προυποθέτει την διαίρεση της εικόνας ή την ταύτηση κάποιου μοντέλου περιγράμματος στην εικόνα των χειλιών και βασίζεται κυρίως στην χρήση αλγορίθμων επεξεργασίας εικόνας. Η χρήση όμως των χαρακτηριστικών άρθρωσης προυποθέτει έναν ταξινομητή και την προεργασία που θα κάναμε και στην περίπτωση που εντοπίζαμε visemes. Η βασική διαφορά όμως είναι πως ο ταξινομητής, ταξινομεί τα δεδομένα εισόδου με βάση διάφορα χαρακτηριστικά των σημείων άρθρωσης, όπως για παράδειγμα το άνοιγμα κοκ. Ταυτόχρονα η συγκεκριμένη προσέγγιση δίνει την δυνατότητα τα δεδομένα που θα αποτελέσουν είσοδο στον ταξινομητή, να εξαχθούν με διαφορετικό τρόπο ανάλογα με το τί επιχειρούμε να εντοπίσουμε. Για παράδειγμα στην περίπτωση που θέλουμε να εντοπίσουμε την εμφάνιση ή οχι των δοντιών μπορούμε να περάσουμε πληροφορία χρώματος, ενώ στην περίπτωση που θέλουμε να δούμε την μεταβολή του σχήματος των χειλιών να γίνει χρήση δεδομένων απο τον υπολογισμό της οπτικής ροής σε διαδοχικά καρέ.

Λόγω της αποδομητικής φύσης της προσέγγισης με χρήση των χαρακτηριστικών άρθρωσης, έχουμε τα ακόλουθα πλεονεκτήματα στην περίπτωση της οπτικής αναγνώρισης του λόγου. Αρχικά, συνδιάζει πολλές και διαφορετικές πηγές πληροφορίας που σχετίζοντε με την παραγωγή του λόγου, η οποία πληροφορία προέρχεται απο την χρήση πολλών παράλληλων ταξινομητών. Για αυτό τον λόγο μπορούμε να αποφύγουμε το γεγονός του ότι κάποια χαρακτηριστικά μπορεί να είναι πιό δύσκολο να εντοπιστούν απο άλλα σε σχέση με τις εκάστοτε συνθήκες και τον θόρυβο που επικρατεί. Μπορεί να γίνει χρήση κάποιου βαθμού εμπιστοσύνης στα αποτελέσματα της ταξινόμησης για κάθε χαρακτηριστικό με τέτοιο τρόπο ώστε στην τελική ταξινόμηση το καθένα να έχει το δικό του βάρος. Τέλος πρέπει να αναφέρουμε πως έχουμε και το πλεονέκτημα της παραγωγής μεγαλύτερων συνόλων δεδομένων εκπαίδευσης και ελέγχου κάτι το οποίο διευκολύνει την διαδικασία αυτή.

Σχήμα 21: Διαφορά στο σχήμα των χειλιών κατα την προφορά του ίδιου φωνήματος /m/ σε δύο διαφορετικές λέξεις, “romantic” στα αριστερά και “academic” στα δεξιά.

3.1. Αντιστοιχηση φωνηματων και χαρακτηριστικων αρθρωσης

Για να μπορέσουμε να δημιουργήσουμε τα σύνολα δεδομένων εκπαίδευσης, πρέπει πρώτα να αντιστοιχήσουμε τα φωνήματα απο τα οποία αποτελείται το λεξικό προς αναγνώριση με τα διάφορα οπτικά χαρακτηριστικά άρθρωσης τα οποία θα κλειθούμε να εντοπίσουμε. Για να γίνει αυτό θα καταφύγουμε στο σύστημα IPA το οποίο και θα μας δώσει τις απαραίτητες πληροφορίες που χρειαζόμαστε. Στην συνέχεια αφού μπορέσουμε να εντοπίσουμε τα ιδιαίτερα χαρακτηριστικά άρθρωσης για το κάθε φώνημα μπορούμε να μεταφερθούμε και να κάνουμε και την κατάλληλη αντιστοίχηση με τα visemes αν το επιθυμούμε.

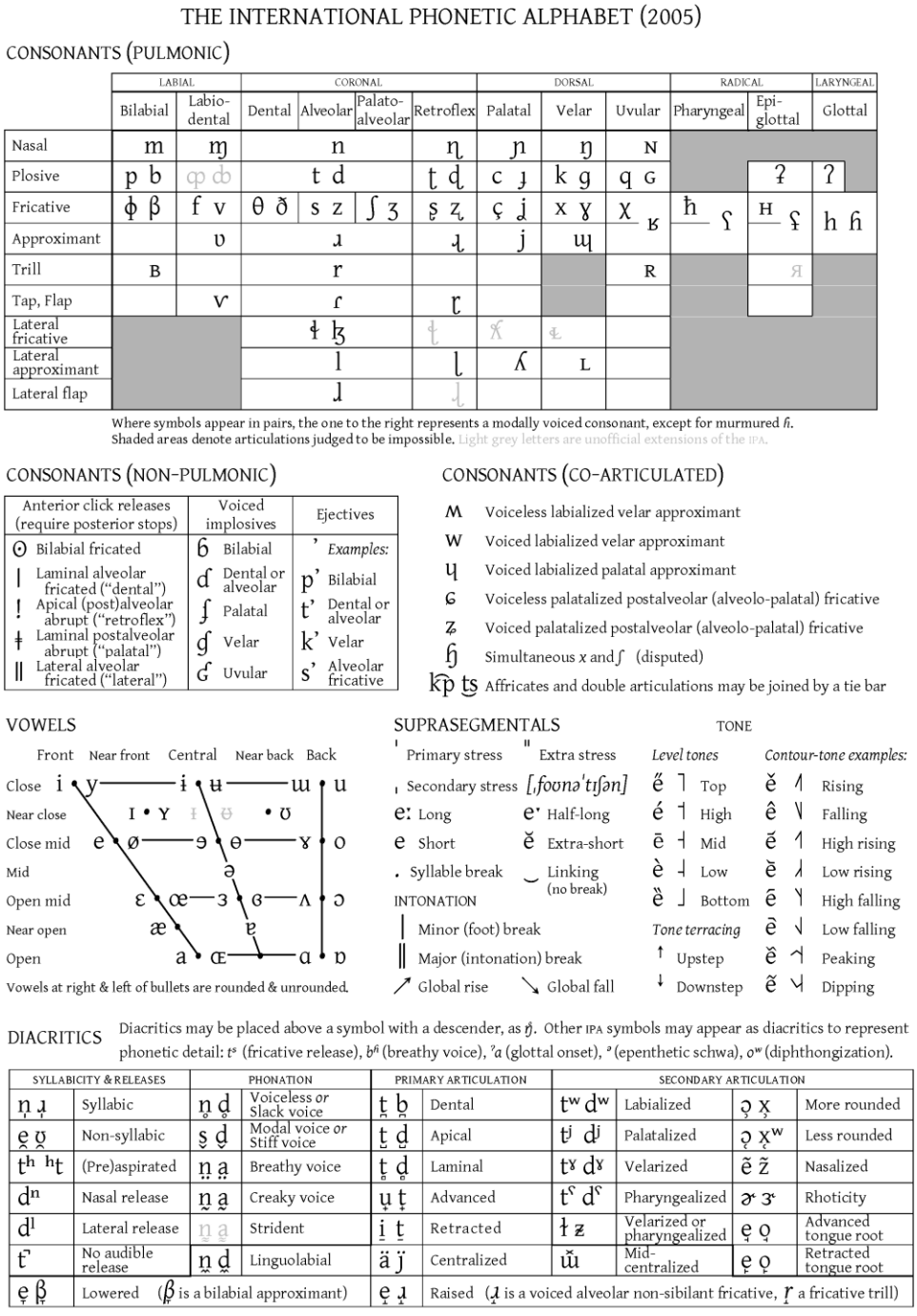
4.3.1 Το συστημα IPA

Το σύστημα IPA ή αλλιώς International Phonetic Alphabet, είναι ενα σύστημα φωνητικής παράστασης βασισμένο στο λατινικό αλφάβητο. Δημιουργήθηκε απο την διεθνή φωνητική ένωση ως ένας δεδομένος τρόπος των ήχων που συνθέτουν την ομιλούμενη γλώσσα. Το σύστημα αυτό χρησιμοποιείται απο γλωσσολόγους, θεραπευτές, δασκάλους και άλλες ιδικότητες. Η αναγραφή της προφοράς των λέξεων για παράδειγμα στα λεξικά βασίζεται σε αυτό το σύστημα.

Το σύστημα αυτό είναι σχεδιασμένο με τέτοιο τρόπο ώστε μόνο τα στοιχεία αυτά της ομιλίας τα οποία μπορούν να διακριθούν κατά την προφορά του λόγου να αναφέροντε. Δηλαδή, τα φωνήματα, η εκφορά και η διάκριση μεταξύ των λέξεων και συλλαβών. Για διάφορα άλλα χαρακτηριστικά του λόγου υπάρχουν διάφορες επεκτάσεις αυτού του συστήματος διαθέσιμες.

Η βασική αρχή του συστήματος είναι η διάθεση ενός συμβόλου για κάθε ξεχωριστή ηχητική μονάδα ή τμήμα του λόγου. Αυτό σημαίνει πως δεν γίνεται χρήση συνδιασμών γραμμάτων για

να αποδοθεί ένας συγκεκριμένος ήχος, ή απλών γραμμάτων για να αποδοθεί κάποιος σύνθετος ήχος. Επίσης δεν υπάρχουν γράμματα τα οποία να εξαρτάται ο ήχος τους απο το περιεχόμενο το οποίο αναπαριστούν. Τέλος το σύστημα δεν έχει ξεχωριστά γράμματα για δύο ήχους για τους οποίους δεν υπάρχει κάποια γνωστή γλώσσα η οποία να κάνει διάκριση μεταξύ τους. Η τελευταία αυτή ιδιότητα ονομάζεται και επιλεκτικότητα. Ανάμεσα στα σύμβολα του συστήματος IPA υπάρχουν 107 τα οποία αναπαριστούν σύμφωνα και φωνήεντα, 31 τα οποία είναι διακριτικά για να χαρακτηρίσουν ακόμα περισότερο τα παραπάνω και 19 σύμβολα των οποίων η χρήση είναι για να ορισθεί το μήκος, ο τόνος και η ο τονισμός. Στο **Σχήμα 22**παρουσιάζεται ένας ενδεικτικός πίνακας των συμβόλων του IPA.



Σχήμα 22: Πίνακας στον οποίο παρουσιάζονται τα σύμβολα του συστήματος IPA.

4.3.2 Αντιστοιχηση λεξεων του συστηματος με το συστημα IPA

Όπως έχουμε ήδη αναφέρει, το σύστημα μας καλείται να αναγνωρίσει ένα περιορισμένο σύνολο λέξεων της αγγλικής γλώσσας, το οποίο αποτελείται απο τις λέξεις των αριθμών μηδέν εώς εννέα. Για να γίνει η αντιστοίχηση με τα σύμβολα του IPA, έγινε χρήση κάποιου διαδικτυακού λεξικού το οποίο είναι σε θέση να δώσει την ακολουθία φωνητικών συμβόλων για δεδομένες λέξεις. Με βάση αυτό το λογισμικό καταλήξαμε στον **Πίνακας 12** στον οποίο αναφέρεται η κάθε λέξη με την αντίστοιχη ακολουθία φωνητικών συμβόλων κατά IPA.

Λέξη	**Σύμβολα IPA**
Zero	/'ziroU/
One	/wΛn/
Two	/tu:/
Three	/θri:/
Four	/fo:r/
Five	/faiv/
Six	/siks/
Seven	/’sev ən/
Eight	/eit/
Nine	/nain/

Πίνακας 12: Λέξεις του λεξικού του συστήματος με τις αντίστοιχες ακολουθίες φωνητικών συμβόλων κατά IPA.

Εφόσον έχουμε βρεί τις ακολουθίες φωνητικών συμβόλων για κάθε μία απο τις λέξεις του λεξικού μας, καλούμαστε να κατηγοριοποιήσουμε αυτά τα σύμβολα με βάση το φώνημα στο οποίο αντιστοιχούν και στην συνέχεια να προκύψουν απο αυτά τα χαρακτηριστικά της

άρθρωσης. Στον **Πίνακας 13** αναφέρονται τα διακριτά σύμβολα και ο αντίστοιχος φωνητικός χαρακτηρισμός τους.

w→ approximant Labial-velar
n→Nasal Alveolar nasal
t→ Plosive Alveolar
u →close back rounded vowel
Θ→fricative Dental
r→Trill Alveolar
I → close front unrounded vowel
f→fricative voiceless LabioDental
o→ Close-mid back rounded vowel
a→ open back unrounded vowel
v→fricative voiced labiodental
s→Fricative voiceless alveolar
k→Plosive Voiceless
e→Close-mid front unrounded vowel
ə→mid central unrounded
n→ Nasal Alveolar
a→open front unrounded vowel

Πίνακας 13: Σύμβολα και ο αντίστοιχος φωνητικός χαρακτηρισμός τους.

Για να μπορέσουμε να κατανοήσουμε τους χαρακτηρισμούς του **Πίνακας 13** θα πρέπει να κάνουμε μία σύντομη αναφορά στην επιστήμη της φωνολογίας απο όπου και οι παραπάνω πληροφορίες προέρχοντε. Η φωνολογία είναι η επιστήμη η οποία ασχολείται με την διαδικασία της φυσικής παραγωγής του ήχου κατά την προφορά του ανθρώπινου λόγου. Ασχολείται με τις φυσικές ιδιότητες των φωνημάτων, ώς μονάδες ήχου της ομιλίας, με την διαδικασία της φυσιολογικής παραγωγής τους, την ακουστική κατανόηση καθώς και την νευροφυσιολογική αντίληψη της ομιλίας. Η φωνητική επιστήμη χωρίζεται σε τρείς βασικές κατηγοριές. Η πρώτη κατηγορία ασχολείται με την διαδικασία της άρθρωσης (articulatory phonetics), δηλαδή με την θέση, την κίνηση και το σχήμα των σημείων της άρθρωσης, όπως αυτά αναφέρθηκαν παραπάνω, έτσι ώστε να προκύψει η παραγωγή του λόγου. Η δεύτερη κατηγορία ασχολείται με ακουστική του λόγου. Κυρίως με τις ηχητικές ιδιότητες απο την σκοπιά της φυσικής που χαρακτηρίζουν την ανθρώπινη ομιλία. Τέλος η Τρίτη κατηγορία ασχολείται με την αντίληψη του λόγου, πώς ο ήχος λαμβάνεται απο το εσωτερικού του αυτιού και στην συνέχεια γίνεται η επεξεργασία και κατανόηση του απο τον εγκέφαλο. Στην δική μας περίπτωση με ενδιαφέρει η πρώτη κατηγορία και με βάση τα αποτελέσματα αυτής μπορούμε να βγάλουμε χρήσιμα συμπεράσματα τα οποία θα μας οδηγήσουν στο να εντοπίσουμε τα χαρακτηριστικά αυτά τα οποία είναι ορατά και μπορούμε να τα εντοπίσουμε κατά την προφορά του λόγου.

Με βάση λοιπόν την φωνολογία της άρθρωσης, τα φωνήματα ταξινομούντε με βάση κάποια βασικά χαρακτηριστικά. Τα χαρακτηριστικά αυτά είναι τα ακόλουθα. Ο τρόπος άρθρωσης (manner of articulation), δηλαδή ποιά φυσική διαδικασία είναι αυτή η οποία παράγει τον ήχο. Για παράδειγμα ο εμποδισμός της ροής του αέρα στην φωνητική οδό. Το δεύτερη χαρακτηριστικό είναι η θέση της άρθρωσης (place of articulation), δηλαδή ποιά τμήματα της φωνητικής οδού συμμετέχουν στην παραγωγή του ήχου. Στην συνέχεια έχουμε τον φωνισμό του φωνήματος (phonation), για παράδειγμα υπάρχει ο χαρακτηρισμός voiceless που σημαίνει πως ο ήχος παράγεται χωρις την συμμετοχή των φωνητικών χορδών. Τέλος ένα ακόμα βασικό χαρακτηριστικό είναι ο μηχανισμός της ροής του αέρα κατά την παραγωγή του ήχου. Πέρα απο αυτά τα παραπάνω χαρακτηριστικά υπάρχουν και άλλα τόσο για να επιτευχθεί ο χαρακτηρισμός κάποιων ιδιαίτερων φωνημάτων, όσο και για να είναι δυνατός ακόμα καλύτερος χαρακτηρισμός της διαδικασίας παραγωγής του λόγου.

Στην περίπτωση μας αυτό το οποίο μας απασχολεί κυρίως είναι το δεύτερο χαρακτηριστικό, δηλαδή η θέση της άρθρωσης. Με βάση αυτό το χαρακτηριστικό μπορούμε να συμπεράνουμε αρχικά το κατα πόσο μπορούμε ή όχι να εντοπίσουμε το χαρακτηριστικό αυτό της άρθρωσης αποκλειστικά και μόνο απο την οπτική πληροφορία. Τα υπόλοιπα χαρακτηριστικά αν και χρήσιμα στην πλοιοψηφία τους δεν μπορούν να μας παρέχουν πληροφορία η οποία να μπορεί να εντοπισθεί οπτικά. Σε ένα σύστημα όμως το οποίο εκμεταλεύεται και την ηχητική πληροφορία θα μπορούσε να γίνει χρήση τους.

Τα παραπάνω που αναφέραμε ισχύουν κυρίως στην περίπτωση των συμφώνων. Στην περίπτωση των φονηέντων τα βασικά χαρακτηριστικά τους είναι, το ύψος, η θέση της γλώσσας και σχέση με το πίσω μέρος του στώματος (vowel backness), και η στρογγυλότητα του φωνήεντος (vowel roundness).

Με βάση λοιπόν όλα τα παραπάνω μπορούμε να εξάγουμε τις παρακάτω πληροφορίες για τα φωνήματα του **Πίνακας 12** σε συνδιασμό με τον χαρακτηρισμό που αναφέρουμε στον **Πίνακας 13**. Παρακάτω θα αναφέρουμε κυρίως την θέση της άρθρωσης και σε όποιες περιπτώσεις απαιτείται και κάποιο άλλο θα αναφέρεται.

- **w→ approximant Labial-velar**. Θέση άρθρωσης, labialized velar, τα χείλη είναι στρογγυλεμένα και το πίσω μέρος της γλώσσας ακουμπάει στο ιστίον.
- **n→Nasal Alveolar**. Θέση άρθρωσης, alveolar, η άκρη της γλώσσας ακουμπάει την φατνιακή προεξοχή.
- **t→ Plosive Alveolar**. Θέση άρθρωσης, alveolar, η άκρη της γλώσσας ακουμπάει την φατνιακή προεξοχή.
- **u →close back rounded vowel**. Η γλώσσα βρίσκεται όσο το δυνατόν πιό κοντά στο πάνω μέρος του στώματος με τέτοιο τρόπο ώστε να μη δημιουργείται κάποιο εμπόδιο το οποίο θα μπορούσε να παράγει σύμφωνο. Η γλώσσα βρίσκεται όσο το δυνατόν πιό πίσω στο στώμα. Τα χείλη είναι στρογγυλεμένα και προεξέχουν.
- **Θ→fricative Dental**. Θέση άρθρωσης, οδοντική, Η γλώσσα ακουμπάει τα δόντια για την παραγωγή του ήχου. Μπορεί να ακουμπάει τα πάνω, τα κάτω είται και τα δύο.

- **r→Trill Alveolar**. Θέση άρθρωσης, alveolar, η άκρη της γλώσσας ακουμπάει την φατνιακή προεξοχή.

- **I → close front unrounded vowel**. Η γλώσσα βρίσκεται όσο το δυνατόν πιό κοντά στο πάνω μέρος του στώματος με τέτοιο τρόπο ώστε να μη δημιουργείται κάποιο εμπόδιο το οποίο θα μπορούσε να παράγει σύμφωνο. Η γλώσσα βρίσκεται όσο το δυνατόν πιο μπροστά στο στώμα. Τα χείλη δεν είναι στρογγυλεμένα αλλά αφείνονται να χαλαρώσουν κατα μήκος.

- **f→fricative voiceless LabioDental**. Θέση άρθρωσης, labiodentals, η άρθρωση γίνεται με χρήση του κάτω χείλους και των άνω δοντιών.

- **o→ Close-mid back rounded vowel**. Η γλώσσα βρίσκεται ανάμεσα μεταξύ του να χαρακτηριστεί close και mid vowel. Η γλώσσα βρίσκεται όσο το δυνατόν πιό πίσω στο στώμα. Τα χείλη είναι στρογγυλεμένα και προεξέχουν.

- **a→ open back unrounded vowel**. Η γλώσσα βρίσκεται όσο το δυνατόν πιό μακριά απο το πάνω μέρος του στώματος. Η γλώσσα βρίσκεται όσο το δυνατόν πιό πίσω στο στώμα. Τα χείλη δεν είναι στρογγυλεμένα αλλά αφείνονται να χαλαρώσουν κατα μήκος.

- **v→fricative voiced labiodental**. Θέση άρθρωσης, labiodentals, η άρθρωση γίνεται με χρήση του κάτω χείλους και των άνω δοντιών.

- **s→Fricative voiceless alveolar**. Θέση άρθρωσης, alveolar, η άκρη της γλώσσας ακουμπάει την φατνιακή προεξοχή.

- **k→Plosive Voiceless**. Θέση άρθρωσης, alveolar, η άκρη της γλώσσας ακουμπάει την φατνιακή προεξοχή.

- **e→Close-mid front unrounded vowel**. Η γλώσσα βρίσκεται ανάμεσα μεταξύ του να χαρακτηριστεί close και mid vowel. Η γλώσσα βρίσκεται όσο το δυνατόν πιο μπροστά στο στώμα. Τα χείλη δεν είναι στρογγυλεμένα αλλά αφείνονται να χαλαρώσουν κατα μήκος.

- **ə→mid central unrounded**. Το ύψος θεωρείται μέσο, δηλαδή το φωνήεν είναι ανάμεσα στο να χαρακτηριστεί κλειστό και ανοιχτό. Η γλώσσα είναι τοποθετημένη έτσι ώστε το φωνήεν να είναι ανάμεσα στο να χαρακτηρισθεί front και back. Τα χείλη δεν είναι στρογγυλεμένα αλλά αφείνονται να χαλαρώσουν κατα μήκος.

- **n→ Nasal Alveolar**. Θέση άρθρωσης, alveolar, η άκρη της γλώσσας ακουμπάει την φατνιακή προεξοχή.

- **a→open front unrounded vowel**. Η γλώσσα βρίσκεται όσο το δυνατόν πιό μακριά απο το πάνω μέρος του στώματος. Η γλώσσα βρίσκεται όσο το δυνατόν πιό πίσω στο στώμα. Τα χείλη δεν είναι στρογγυλεμένα αλλά αφείνονται να χαλαρώσουν κατα μήκος.

4.4. Διαδικασια αναγνωρισης χαρακτηριστικων και παραγωγη του διανυσματος χαρακτηριστικων.

Στις προηγούμενες παραγράφους παρουσιάσαμε την μέθοδο την οποία θα ακολουθήσουμε για να εντοπίσουμε τα χαρακτηριστικά της άρθρωσης τα οποία συμμετέχουν στην παραγωγή των λέξεων του λεξικού μας. Στην συνέχεια αφού έχουμε πλέον εντοπίσει ποιά χαρακτηριστικά θα επιχειρήσουμε να εντοπίσουμε μένει να υλοποιήσουμε κάποιον ταξινομητή ο οποίος θα πραγματοποιήσει αυτή την διαδικασία. Για την ταξινόμηση των χαρακτηριστικών στην παρούσα εργασία προτιμήθηκε η χρήση μηχανών διανυσμάτων στήριξης. Ο λόγος που έγινε αυτό είναι γιατί ήδη σε άλλα τμήματα της αρχιτεκτονικής του συστήματος έχουν χρησιμοποιηθεί, οπότε υπάρχει συνοχή ως προς τις τεχνολογίες που χρησιμοποιούντε, ενώ ταυτόχρονα στην βιβλιογραφία έχουν παρουσιασθεί πολύ καλά αποτελέσματα απο την χρήση τους στο συγκεκριμένο πρόβλημα ταξινόμησης [21]. Στην προκειμένη περίπτωση θα κάνουμε χρήση ενός συνόλου απο ταξινομητές διανυσμάτων στήριξης. Ο καθένας θα εκπαιδευθεί στην εντοπισμό ενός χαρακτηριστικού άρθρωσης και το αποτέλεσμα όλων θα συνθέσει το τελικό διάνυσμα το οποίο θα τροφοδοτήσει το επόμενο στάδιο. Η έξοδος του κάθε ταξινομητή θα είναι η απάντηση της ύπαρξης ή οχι ενός συγκεκριμένου χαρακτηριστικού, με τον τρόπο αυτό θα καταλήξουμε να έχουμε ένα διάνυσμα το οποίο θα αποτελείται απο Boolean τιμές και το επόμενο στάδιο του συστήματος θα καλείται απο την χρονική ακολουθία αυτών των διανυσμάτων να εντοπίσει την λέξη η οποία προφέρεται. Για παράδειγμα ένας ταξινομητής θα εντοπίζει το κατα πόσο τα χείλη

είναι στρογγυλεμένα ή όχι ενώ κάποιος άλλος το κατα πόσο το κάτω χείλος ακουμπάει ή όχι τα πάνώ δόντια κοκ, για κάθε ένα απο τα χαρακτηριστικά που αναφέρονται στην παράγραφο 4.3.2 και τα οποία μπορούν να εντοπισθούν.

Πρέπει να παρατηρήσουμε την αντιστοίχηση που κάναμε στο 4.3.2 και να επισημάνουμε τα εξής. Τα σύμφωνα γενικά παρουσιάζουν ιδιαίτερο πρόβλημα ώς προς την αναγνώριση τους για τους ακόλουθους λόγους. Στις περίσσοτερες των περιπτώσεων η παραγωγή των συμφώνων χρησιμοποιεί το ίδιο σημείο άρθρωσης ενώ διαφοροποιούντε άλλοι παράγοντες οι οποίοι δεν έιναι εμφανείς απο την οπτική πληροφορία και μόνο. Ταυτόχρονα σε πολλές περιπτώσεις το σημείο της άρθρωσης αυτό καθ'αυτό δεν είναι εμφανές. Πρέπει όμως εδώ να επισημάνουμε πως σε ένα σύστημα ολοκληρωμένης αναγνώρισης ομιλίας όπου λαμβάνεται υπόψην και το ηχητικό σήμα μπορεί να γίνει χρήση αυτής της πληροφορίας και να εντοπισθούν χαρακτηριστικά όπως για παράδειγμα ο τρόπος της άρθρωσης κάτι το οποίο θα έκανε τα σύμφωνα να είναι πιό διακριτά μεταξύ τους. Ιδιαίτερα η χρήση ταυτόχρονα των δύο οδών (οπτικής και ακουστικής) σε συνδιασμό με τα χαρακτηριστικά άρθρωσης που μπορεί η καθεμία να προσφέρει θα μπορούσε ενδεχομένων να οφελήσει τον ρυθμό αναγνώρισης. Τα φωνήεντα απο την άλλη πλευρά εξαιτίας της μεγάλης συμμετοχής των χειλιών στην παραγωγή τους είναι πιό εύκολα να εντοπισθούν και να ταξινομηθούν. Τέλος πέρα απο τα χαρακτηριστικά τα οποία μπορούμε να εξάγουμε μόνο σε αυτό το στάδιο, βασιζόμαστε σε πολύ μεγάλο βαθμό τόσο στην διαχωρισημότητα που έχουν οι λέξεις του λεξικού μας όσο και της ικανότητας του ταξινομητή στο επόμενο στάδιο ώστε να βελτιωθεί κατά πολύ ο ρυθμός αναγνώρισης.

ΚΕΦΑΛΑΙΟ 5ο.

ΜΗΧΑΝΕΣ ΔΙΑΝΥΣΜΑΤΩΝ ΣΤΗΡΙΞΗΣ

5.1. Εισαγωγή

Οι μηχανές διανυσμάτων στήριξης αποτελούν μία νέα μέθοδο για την ταξινόμηση προτύπων και δεδομένων. Ανήκουν στην κατηγορία των αλγορίθμων επιβλεπόμενης εκμάθησης και εκτός της χρήσης τους για ταξινόμηση μπορούν να προσεγγίσουν και συναρτήσεις (Regression). Ανήκουν σε μία κατηγορία γενικευμένων γραμμικών ταξινομητών, ενώ μπορούν να θεωρηθούν ως μία ειδική περίπτωση της κανονικοποίησης Tikhonov. Το βασικό τους χαρακτηριστικό που τα διαφοροποιεί απο άλλους αλγορίθμους ταξινόμησης της ίδιας κατηγορίας, όπως για παράδειγμα τα νευρωνικά δίκτυα, είναι το γεγονός του ότι οι συγκεκριμένες μηχανές ελαχιστοποιούν ταυτόχρονα το εμπειρικό λάθος ταξινόμησης (empirical classification error), ενώ ταυτόχρονα μεγιστοποιούν την απόσταση διαχωρισμού(geometric margin). Για αυτό τον λόγο αποκαλούνται και ταξινομητές μεγίστου περιθωρίου (Maximum Margin Classifier).

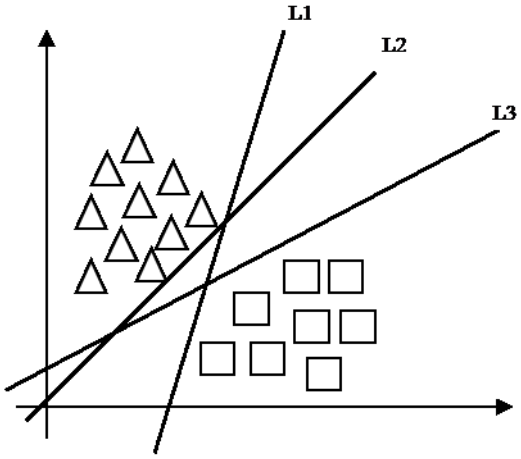
Ένα απλό παράδειγμα που μπορούμε να δόσουμε για την κατανόηση του προβλήματος ταξινόμησης με χρήση των μηχανών διανυσμάτων στήριξης και της κατανόησης της χρήση της απόστασης διαχωρισμού, είναι το ακόλουθο. Ας θεωρήσουμε μία επιφάνεια δύο διαστάσεων με σημεία πάνω της. Αυτά αναπαριστάνουν τις πόλεις μεταξύ δύο χωρών, επιθυμούμε να σχεδιάσουμε μία γραμμή, τέτοια ώστε οι πόλεις να διαχωρίζονται μεταξύ τους ανάλογα με την χώρα στην οποία ανήκουν, εύκολα κάποιος μπορεί να φανταστεί πως μία τέτοια γραμμή είναι τα συνορά των χωρών.

Το παραπάνω παράδειγμα είναι ιδιαίτερα απλοικό και μας δίνει να καταλάβουμε διαισθητικά την λειτουργία της απόστασης διαχωρισμού. Βέβαια στην περίπτωση των μηχανών διανυσμάτων στήριξης δεν περιοριζόμαστε σε σημεία που ανήκουν στον δισδιάστατο χώρο. Κάθε δεδομένο το οποίο μπορεί να αναπαρασθεί ως διάνυσμα n διαστάσεων, μπορούμε να το τοποθετήσουμε σε έναν χώρο n διαστάσεων και να το διαχωρίσουμε με μία καμπύλη n-1 διαστάσεων, όπως στην

περίπτωση του παραδείγματος ο διαχωρισμός έγινε με μία μονοδιάστατη γραμμή σε ένα επίπεδο δύο διαστάσεων. Παραδείγματα τέτοιων δεδομένων, αποτελούν οι φωτογραφίες ή τα δεδομένα φωνής. Μία φωτογραφία nxm διαστάσεων μπορεί να αναπαρασταθεί ως ένα διάνυμα 1x(nxm) και να τοποθετηθεί σε έναν χώρο R^d όπου $d = nxm$. Παρατηρούμε δηλαδή πως η εικόνα μετατρέπεται σε ένα σημείο σε αυτό τον χώρο.

5.2. Κίνητρο

Συχνά καλούμαστε να ταξινομήσουμε δεδομένα τα οποία δεν ανήκουν σε ένα επίπεδο δύο διαστάσεων, δηλαδή δεν είναι σημεία του $\Re^2$, αλλά να είναι πολυδιάστατα δεδομένα και να ανήκουν στο $\Re^p$ (στατιστική αναπαράσταση) ή στο $\Re^n$ (αναπαράσταση πληροφορικής). Αυτό το οποίο μας ενδιαφέρει είναι το κατα πόσο μπορούμε να διαχωρίσουμε αυτά τα σημεία με χρήση ενός υπερ-επιπέδου n-1 διαστάσεων. Αυτό αποτελεί μία τυπική περίπτωση γραμμικού ταξινομητή. Υπάρχουν πολλοί ταξινομητές οι οποίοι ικανοποιούν την παραπάνω απαίτηση, αλλά επιπλέων απαιτούμε εκτός απο την εύρεση αυτού του υπερ επιπέδου, να ικανοποιεί και την συνθήκη του ότι αυτό θα είναι το μέγιστο, δηλαδή θα επιτυγχάνει τον μέγιστο δυνατό διαχωρισμό μεταξύ των σημείων δεδομένων. Αν αυτό το υπέρ επίπεδο υπάρχει τότε μας ενδιαφέρει να μπορούμε να το εντοπίσουμε. Αυτό το υπερ επίπεδο ονομάζεται και επίπεδο μέγιστου διαχωρισμού. Ενώ ο ταξινομήτης που το καταφέρνει να το εντοπίσει ονομάζεται ταξινομητής μεγίστου περιθωρίου. Τα παραπάνω παρουσιάζονται σχηματικά στο **Σχήμα 23**.



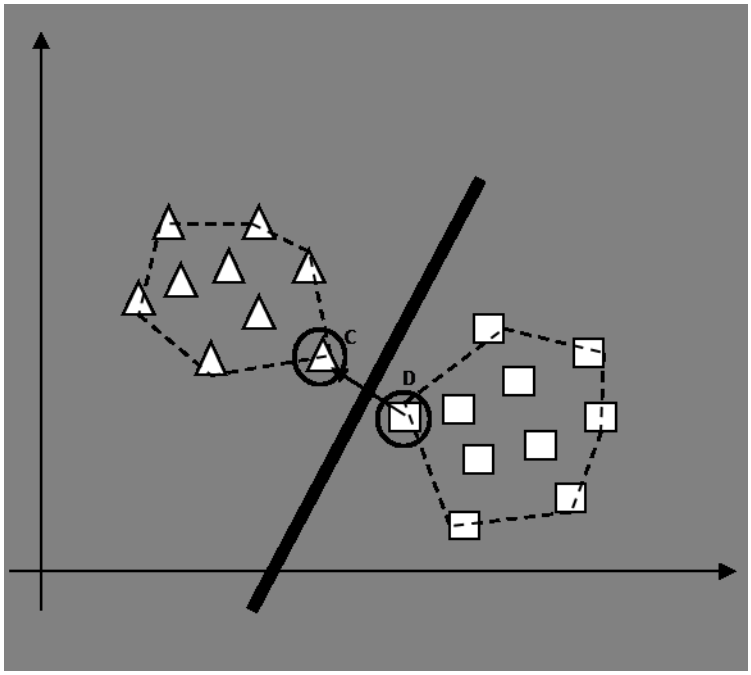
Σχήμα 23: Υπάρχουν πολλοί γραμμικοί ταξινομητές(υπέρ επίπεδα) οι οποίοι διαχωρίζουν τα σημεία στο επίπεδο αλλά μόνο ένας απο αυτούς επιτυγχάνει μέγιστο διαχωρισμό.

5.3. Βασικά χαρακτηριστικά των μηχανών διανυσμάτων στήριξης

Για να εμβαθύνουμε περισσότερο στην έννοια των επιπέδων διαχορισμού που αποτελούν βασικό χαρακτηριστικό των μηχανών διανυσμάτων στήριξης, ας θεωρήσουμε το διαδικό πρόβλημα ταξινόμησης με σημεία x_i όπου $(i = 1 \ldots m)$ και τα οποία έχουν ετικέτες $y_i = \pm 1$. Καθένα απο αυτά τα σημεία αναπαρίστανται σε έναν χώρο d διαστάσεων ο οποίος καλείται χώρος εισόδου ή χώρος χαρακτηριστικών. Έστω πως η συνάρτηση κατηγοριοποίησης (ή ταξινόμησης) είναι η $f(x) = sign(w \cdot x - b)$. Το διάνυσμα w καθορίζει την κατεύθυνση ενός διακεκριμένου επιπέδου (discriminant plane). Το αδιάστατο μέγεθος b καθορίζει την θέση του επιπέδου σε σχέση με την αρχή των αξόνων. Ας θεωρήσουμε πως τα δεδομένα είναι γραμμικών διαχωρίσημα μεταξύ τους (όπως στην περίπτωση του σχήματος 1.1). Τότε υπάρχει επίπεδο το οποίο ταξινομεί σωστά τις δύο κλάσεις, δηλαδή διαχωρίζει μεταξύ τους τα σημεία στο επίπεδο. Αν παρατηρήσουμε το **Σχήμα 23** θα δούμε πως τέτοια επίπεδα (στην προκειμένη περίπτωση επίπεδα μίας διάστασης – ευθείες γραμμές) υπάρχουν άπειρα. Αν λειτουργήσουμε λίγο διαισθητικά θα μπορέσουμε σχετικά εύκολα να καταλήξουμε στο συμπέρασμα πως το επίπεδο που επιθυμούμε είναι αυτό με την μέγιστη απόσταση απο τις δύο κλάσεις. Με αυτό τον τρόπο μπορούμε φυσικά να

περιμένουμε καλύτερη ικανότητα γενικοποίησης σε νέα δεδομένα, αφού οι κλάσεις είναι καλύτερα διαχωρισμένες μεταξύ τους.

Το θέμα που προκύπτει τώρα είναι το πως μπορούμε να κατασκευάσουμε αυτό το επίπεδο που αναφέραμε παραπάνω. Αρχικά θα ακολουθήσουμε μία γεωμετρική προσέγγιση. Θα εξετάσουμε τα κυρτά πολύγονα (convex hulls) τα οποία ορίζονται απο τα δεδομένα των κλάσεων μας. Κυρτό πολύγονο ενός συνόλου **Q** σημείων είναι το μικρότερο κυρτό πολύγονο P το οποίο έχει την ιδιότητα τα σημεία του **Q** βρίσκονται είται στο σύνορο του είται στο εσωτερικό του. Αφού λοιπόν βρούμε τα πολύγονα αυτά, τότε μπορούμε να εντοπίσουμε τα σημεία των δύο πολυγώνων τα οποία βρίσκονται πίο κοντά. Αν τώρα κατασκευάσουμε το επίπεδο το οποίο διχοτομεί αυτά τα δύο σημεία (w = d-c), τότε ο ταξινομητής που θα προκύψει απο το συγκεκριμένο επίπεδο αναμένουμε να είναι αποδοτικός. Τα παραπάνω φαίνονται στο **Σχήμα 24**.



Σχήμα 24: Τα κυρτά πολύγονα που κατασκευάζονται για τα δύο σύνολα των αντίστοιχων κλάσεων. Τα σημεία C και D αποτελούν τα σημεία των περιφερειών των δύο πολυγώνων που βρίσκονται πιό κοντά μεταξύ τους. Το επίπεδο το οποίο κατασκευάζεται με την προυπόθεση να διχοτομεί την απόσταση μεταξύ των δύο αυτών σημείων, αναμένουμε να έχει την καλύτερη απόδοση διαχωρισμού.

Τα κοντινότερα σημεία των κυρτών πολυγώνων μπορούν να προκύψουν απο την επίλυση του παρακάτω τετραγωνικού προβλήματος.

$$\min_a \quad \frac{1}{2}\|c-d\|^2$$

$$c=\sum_{y_i \in Class1} a_i x_i \quad d=\sum_{y_i \in Class-1} a_i x_i$$

(1)

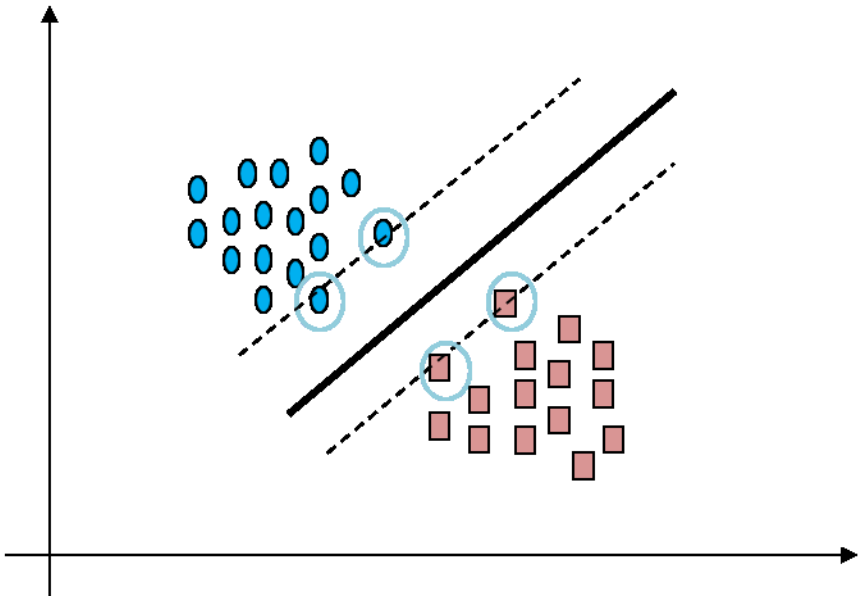
$$\sum_{y_i \in Class1} a_i = 1 \quad \sum_{y_i \in Class-1} a_i = 1$$

$$a_i \geq 0 \quad i=1,\ldots,m$$

Υπάρχουν πολλοί αλγόριθμοι για την επίλυση της παραπάνω τετραγωνικής μορφής που ορίζεται στην σχέση (1), ενώ έχουν αναπτυχθεί μέθοδοι πλέον οι οποίοι λαμβάνουν υπόψην τους και τα ιδιαίτερα χαρακτηριστικά των μηχανών διανυσμάτων στήριξης. Πρέπει τέλος να τονίσουμε πως η επίλυση εξαρτάται μόνο απο ορισμένα σημεία της περιφέρειας του κυρτού πολυγόνου και όχι απο όλα τα σημεία.

Μία εναλλακτική προσέγγιση στο παραπάνω πρόβλημα είναι το να μεγιστοποιήσουμε το διάστημα μεταξύ δύο παραλλήλων υποστηριχτικών επιπέδων, έναντι του υπολογισμού του επιπέδου που διχοτομεί την ευθεία μεταξύ των δύο σημείων. Ένα επίπεδο λέμε πως υποστηρίζει ένα σύνολο, όταν όλα τα σημεία του συνόλου αυτού βρίσκονται απο την μία πλευρά του επιπέδου. Για τα σημεία που ανήκουν στην κατηγορία με ετικέτα **+1** θα θέλαμε να υπάρχουν **w** και **b** τέτοια ώστε $w \cdot x_i > b$ ή $w \cdot x_i - b > 0$, ανάλογα με την ετικέτα της κλάσης. Αν υποθέσουμε πως η μικρότερη τιμή του $|w \cdot x_i - b|$ είναι **κ** τότε $w \cdot x_i - b > k$. Η παράμετρος στο εσωτερικό της συνάρτηση απόφασης είναι αναλλοίωτο ως προς τις θετικές μετατροπές στην διάσταση της, οπότε μπορούμε να ορίσουμε το παρακάτω $w \cdot x_i - b \geq 1$. Αντίστοιχα για τα σημεία της κλάσσης με ετικέτα -1 απαιτούμε $w \cdot x_i - b \leq -1$. Για να βρούμε το επίπεδο με την μεγαλύτερη απόσταση μεταξύ των δύο συνόλων, μπορούμε απλά να μεγιστοποιήσουμε την απόσταση ή περιθώριο

μεταξύ των επιπέδων στήριξης για κάθε ένα απο τα δύο σύνολα. Τα επίπεδα αυτά μετακινούνται το καθένα προς την πλευρά του αντίστοιχου συνόλου εώς ότου ένα μικρό υποσύνολο απο σημεία τα οποία καλούνται διανύσματα στήριξης να ταυτιστούν με αυτά. Τα παραπάνω φαίνονται στο **Σχήμα 25**.



Σχήμα 25: Τα επίπεδα υποστηρίξης εμφανίζονται με διακεκομένες γραμμές, ενώ τα διανύσματα στήριξης είναι κυκλωμένα. Με αυτό τον τρόπο βρίσκουμε το επίπεδο το οποίο μεγιστοποιεί το περιθώριο οπότε και αναμένουμε να έχει τα καλύτερα δυνατά αποτελέσματα διαχοροποίησης μεταξύ των δύο κλάσεων.

Η απόσταση ή περιθώριο μεταξύ των επιπέδων στήριξης, $w \cdot x = b + 1$ και $w \cdot x = b - 1$, είναι $\gamma = \frac{2}{\|w\|_2}$. Επομένως η ελαχιστοποίηση του περιθωρίου είναι αντίστοιχη με την ελαχιστοποίηση της ποσότητας $\frac{\|w\|_2}{2}$ στο ακόλουθο πρόβλημα τετραγωνική μορφής.

$$\min_{w,b} \quad \frac{1}{2}\|w\|^2$$

$$w \cdot x_i \geq b+1 \qquad y_i \in Class1$$
$$w \cdot x_i \geq b-1 \qquad y_i \in Class-1 \tag{2}$$

Πρέπει εδώ να παρατηρήσουμε πως η λύση απο τις δύο παραπάνω διαφορετικές προσεγγίσεις είναι η ίδια. Και στις δύο περιπτώσεις τα σημεία – διανύσματα τα οποία συμμετέχουν στην λύση είναι τα ίδια και αναφέρονται ως διανύσματα στήριξης. Η παρατήρηση αυτή δεν είναι τυχαία, είναι αποτέλεσμα της ιδέας του δυισμού, η οποία αποτελεί βασικό στοιχείο του μαθηματικού προγραμματισμού.

Η ιδιότητα του δυισμού και του μέγιστου περιθωρίου, μαζί με αυτή των πυρήνων είναι τα 3 βασικά στοιχεία τα οποία χαρακτηρίζουν την λειτουργία των μηχανών υποστήριξης διανυσμάτων, και αυτές οι οποίες μας οδηγούν στην κατανόηση της λειτουργίας τους. Η ιδιότητα του δυισμού είναι πολύ χρήσιμή γιατι μας δίνει την δυνατότητα να επιλύσουμε ένα πρόβλημα ισοδύναμα σε μία διαφορετική μορφή η οποία μπορεί να είναι πιο επιθυμητή για διάφορους λόγους, όπως για παράδειγμα η πολυπλοκότητα των περιορισμών ή των πράξεων που απαιτούνται.

5.4. Η περίπτωση των μη γραμμικά διαχωρίσημων συνόλων και η χρήση των πυρήνων.

Εώς τώρα είχαμε υποθέσει πως τα σύνολα των δεδομένων – σημείων ήταν γραμμικών διαχωρίσημα μεταξύ τους. Στην περίπτωση όμως που η συνθήκη αυτή δεν ισχύει τότε η μέθοδος με τα κυρτά πολύγωνα που παρουσιάσαμε παραπάνω δεν μπορεί να εφαρμοσθεί όπως φαίνεται και στο **Σχήμα 26**. Αυτό συμβαίνει γιατί σε αυτή την περίπτωση τα πολύγωνα αυτά θα τέμνονται μεταξύ τους. Πρέπει να παρατηρήσουμε όμως πως όπως φαίνεται και στο **Σχήμα 26**αν αφαιρεθεί το σημείο που δημιουργεί το πρόβλημα, τότε το πρόβλημα μας πάλι επιλύεται με την προηγούμενη μέθοδο. Για αυτό τον λόγο πρέπει να μπορούμε να καθορίσουμε την βαρύτητα που έχει το κάθε σημείο. Αυτό μπορεί να γίνει αν κάνουμε χρήση του ελαχιστοποιημένου κυρτού πολυγώνου στην θέση του πολυγώνου που είχαμε ορίσει νωρίτερα. Αυτό επιτυγχάνεται εισάγωντας τους παρακάτω περιορισμούς για την επιρροή του κάθε σημείου θέτωντας ένα

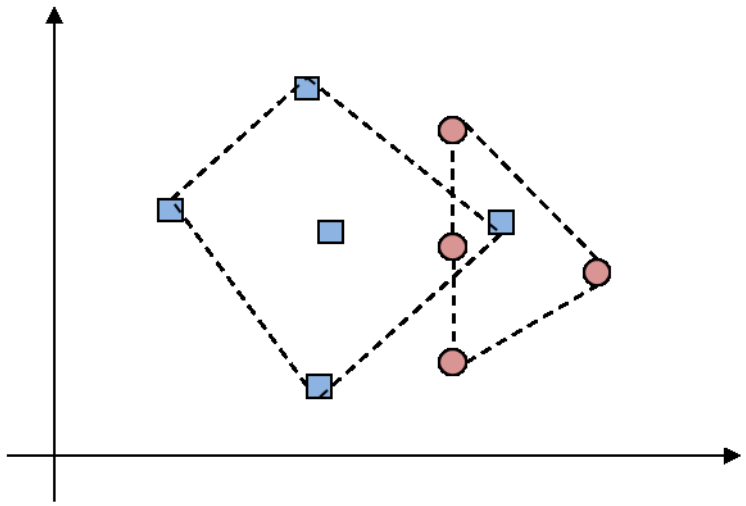
ανώτατο όριο D<1 στον πολλαπλασιαστή για αυτό το σημείο, τυπικά το ελαχιστοποιημένο κυρτό πολύγωνο ορίζεται ως εξής.

$$d = \sum_{y_i \in Class1} a_i x_i$$

$$\sum_{y_i \in Class1} a_i = 1 \qquad (4)$$

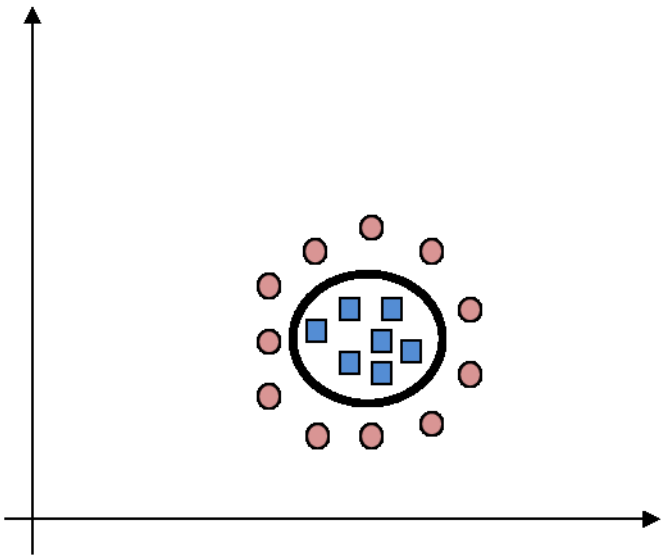
$$0 \le a_i \le D$$

Για D αρκετά μικρό τα ελαχιστοποιημένα κυρτά πολύγωνα δεν θα τέμνονται. Με τον παραπάνω τρόπο μπορούμε να τροποποιήσουμε τις τετραγωνικές μορφές για όλα τα δυικά προβλήματα που προκύπτουν και τελικά να μπορέσουμε να επιλύσουμε τα προβλήματα και να έχουμε διαχωρίσημες κλάσεις.



Σχήμα 26: Τα παραπάνω σύνολα δεν είναι γραμμικώς διαχωρίσημα όπως φαίνεται και στο σχήμα. Τα κυρτά πολύγωνα τέμνονται μεταξύ τους, αλλά η αφαίρεση του ενός τετραγώνου θα μπορούσε να μας οδηγήσει στην δημιουργία πολυγώνων τα οποία να μην τέμνονται. Αυτή την ιδέα εκμεταλευόμαστε ώστε να προκύψουν τελικά διαχωρίσημα σύνολα εισάγωντας την ιδέα του άνω φράγματος στην επιρροή του κάθε σημείου των συνόλων.

Στην έως τώρα προσέγγιση μας, αντιμετωπίσαμε την περίπτωση του γραμμικού διαχωρισμού, τόσο για τις περιπτώσεις που οι κλάσεις ήταν εξαρχής γραμμικά διαχωρίσημες, όσο και για αυτές που οι κλάσεις δεν ηταν γραμμικά διαχωρίσημες. Η βασική αρχή των μηχανών διανυσμάτων στήριξης είναι η κατασκευή του επιπέδου μέγιστου διαχωρισμού. Αυτό είναι ισοδύναμο με το δυικό πρόβλημα της εύρεσης των δύο κοντινότερων σημείων στα ελαχιστοποιημένα κυρτά πολύγωνα της κάθε κλάσης . Με την παραπάνω προσέγγιση οι μηχανές αυτές είναι δυνατό να κατασκευάσουν γραμμικές συναρτήσεις ταξινόμησης με καλές ιδιότητες γενίκευσης, τόσο θεωρητικά όσο και πρακτικά και για δεδομένα που βρίσκονται σε χώρους μεγάλων διαστάσεων. Υπάρχουν όμως περιπτώσεις που ο γραμμικός διαχωρισμός δεν «ταιριάζει» με τα δεδομένα τα οποία έχουμε ως έισοδο, και τα αποτελέσματα της ταξινόμησης με αυτές τις μηχανές να μην είναι ικανοποιητικά, όπως φαίνεται και στο **Σχήμα 27**. Για την αντιμετώπιση τέτοιων περιπτώσεων ,εισάγουμε την τελευταία βασική έννοια των μηχανών διανυσμάτων στήριξης που είναι οι χρήση των πυρήνων.



Σχήμα 27: Περίπτωση προβλήματος όπου γραμμική επιφάνεια διαχωρισμού δεν είναι ικανή να διαχωρίζει τις κλάσεις. Συγκεκριμένα όπως φαίνεται και απο το συγκεκριμένο σχήμα η καταλληλότερη μορφή διαχωριστικής επιφάνειας είναι τετραγωνική μορφής.

Αν αναλογιστούμε την περίπτωση του **Σχήμα 27**, θα προσέξουμε πως καμία γραμμική επιφάνεια διαχωρισμού δεν είναι ικανή να διαχωρίσει τις δύο κλάσεις. Μάλιστα όπως φαίνεται απο τον κύκλο που έχει σχηματιστεί η μορφή της επιφάνειας που διαχωρίζει ικανοποιητικά τις δύο κλάσεις είναι τετραγωνικής μορφής. Μία κλασσική μέθοδος για την μετατροπή ενός αλγορίθμου γραμμικής ταξινόμησης σε έναν αλγόριθμο μή γραμμική ταξινόμησης, είναι το να προσθέσουμε ιδιότητες στα δεδομένα, οι οποίες είναι μη γραμμικές συναρτήσεις των αρχικών δεδομένων. Με αυτό τον τρόπο υπαρκτοί αλγόριθμοι γραμμικής ταξινόμησης μπορούν να εφαμοστούν στο σύνολο δεδομένων που προέκυψε απο την επέκταση του αρχικού συνόλου στο χώρο χαραχτηριστικών (feature space) παράγωντας μη γραμμικές συναρτήσεις στον αρχικό χώρο εισόδου. Για το συγκεκριμένο παράδειγμα του **Σχήμα 27**, μπορούμε να κατασκευάσουμε μια τετραγωνική διαχωριστική επιφάνεια σε έναν διανυσματικό χώρο δύο διαστάσεων με ιδιότητες r, s, με το να προβάλουμε τον αρχικό δισδιάστατο χώρο εισόδου $[r,s]$, στον χώρο χαρακτηριστικών 5 διαστάσεων $[r,s,rs,r^2,s^2]$ και κατασκευάζοντας μία γραμμική διαχωριστική επιφάνεια σε αυτό τον χώρο. Πιό συγκεκριμένα ορίζουμε: $\Theta(x):\Re^2 \to \Re^5$ τότε $x=[r,s]$, $w\cdot x = w_1 r + w_2 s \to \Theta(x) = [r,s,rs,r^2,s^2]$ και $w\cdot\Theta(x) = w_1 r + w_2 s + w_3 rs + w_4 r^2 + w_5 s^2$. Με βάση τα προηγούμε η συνάρτηση ταξινόμησης τελικά παίρνει την μορφή $f(x) = sign(w\cdot\Theta(x) - b) = sign(w_1 r + w_2 s + w_3 rs + w_4 r^2 + w_5 s^2 - b)$. Η μορφή αυτή είναι γραμμική στον χώρο χαρακτηριστικών πέντε διαστάσεων αλλά είναι τετραγωνική στον δισδιάστατο χώρο εισόδου.

Για την περίπτωση πολυδιάστατων δεδομένων εισόδου, η παραπάνω μέθοδος έχει δύο πιθανά προβλήματα τα οποία μπορούν να προκύψουν. Τα δύο αυτά προβλήματα προκύπτουν απο το γεγονός του ότι οι διαστάσεις του χώρου χαρακτηριστικών αυξάνονται εκθετικά. Το πρώτο πρόβλημα που προκύπτει είναι η υπέρ γενίκευση. Οι μηχανές διανυσμάτων στήριξης μπορούν να αντιμετωπίσουν το συγκεκριμένο πρόβλημα διότι βασίζονται στην κατασκευή του μεγίστου περιθωρίου, επομένως αν γίνει κατάλληλη επιλογή της παραμέτρου **C**, έχουμε την δυνατότητα να το αποφύγουμε. Το δεύτερο πρόβλημα που προκύπτει είναι το ότι ο υπολογισμός της $\Theta(x)$ δεν είναι πρακτικός. Για το θέμα της αύξησης των διαστάσεων θα αναφερθούμε εκτενέστερα παρακάτω μιάς και αποτελεί σημαντικό θέμα για σχεδόν όλες τις περιπτώσεις μεθόδων

ταξινόμησης. Για το δεύτερο πρόβλημα οι μηχανές διανυσμάτων στήριξης βασίζονται στην χρήση των πυρήνων για να ανταπεξέλθουν.

Αν ερευνήσουμε την επιρροή που έχει η $\Theta(x)$ στην επίλυση κάποιου απο τα προβλήματα τετραγωνικής μορφής που αναφέραμε προηγουμένως θα παρατηρήσουμε τα εξής. Έστω $\Theta(x): \Re^n \to \Re^{n'}$ όπου $n' >> n$, τότε έχουμε να ελαχιστοποιήσουμε την ποσότητα:

$$\min_a \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j a_i a_j \Theta(x_i)\cdot\Theta(x_j) - \sum_{i=1}^{l} a_i$$

Με τους παρακάτω περιορισμούς:

$$\sum_{i=1}^{l} y_i a_i = 0$$

και

$$C \geq a \geq 0 \qquad i = 1,\ldots,m$$

Από τα παραπάνω μπορούμε να παρατηρήσουμε πως τα δεδομένα που προβάλουμε προκύπτουν ως εσωτερικό γινόμενο στην συνάρτηση που καλούμαστε να ελαχιστοποιήσουμε παραπάνω.Αυτό το οποίο κάνουμε στην περίπτωση αυτή ειναι να χρησιμοποιήσουμε ένα μαθηματικό τρικ γνωστό ως τρικ των πυρήνων (kernel trick) στο οποίο γίνεται χρήση πυρήνων Hilbert – Schmidt. Για να γίνει χρήση του παραπάνω τρικ στηριζόμαστε στο θεώρημα του Mercer το οποίο περιγράφεται ακριβώς στο παράρτημα. Με βάση λοιπόν αυτό το θεώρημα γνωρίζουμε πως για μία συγκεκριμένη προβολή **Θ** και για δύο σημεία u και v, το εσωτερικό γινόμενο των σημείων που προκύπτουν απο την προβολή μπορεί να προκύψει με χρήση των συναρτήσεων πυρήνων χωρίς να γνωρίζουμε συγκεκριμένα την δεδομένη προβολή, δηλαδή $\Theta(u)\cdot\Theta(v) \equiv K(u,v)$. Υπάρχουν πολλοί διαφορετικοί πυρήνες και σχεδιάζονται συνεχώς καινούργιοι για διάφορες περιπτώσεις προβλημάτων. Για παράδειγμα έχει παρατηρηθεί πως οι πολυωνυμικοί πυρήνες αποδίδουν καλύτερα σε περιπτώσεις ταξινόμησης προτύπων που σχετίζονται με εικόνες, ενώ άλλοι πυρήνες αποδίδουν καλύτερα στις περιπτώσεις δεδομένων που σχετίζονται με κείμενο. Μερικοί χαρακτηριστικοί πυρήνες είναι οι ακόλουθοι.

Θ(u)	K(u,v)
Degree d polynomial	$(u \cdot v + 1)^d$
Radial Basis Function Machine	$\exp\left(-\frac{\|u-v\|^2}{2\sigma}\right)$
Two-Layer Neural Network	sigmoid(η($u \cdot v$)+c)

Με βάση τα παραπάνω κανείς παρατηρεί πως για να μεταφερθούμε απο έναν γραμμικό ταξινομητή σε έναν μή γραμμικό, αρκεί στην συνάρτηση προς ελαχιστοποίηση του τετραγωνικού προβλήματος που προκύπτει να αντικαταστήσουμε το αρχικό εσωτερικό γινόμενο με τον υπολογισμό κάποιου πυρήνα. Με αυτό τον τρόπο αλλάζοντας πυρήνες μπορούμε να πάρουμε διαφορετικούς μή γραμμικούς ταξινομητές. Καμία αλλαγή στον αλγόριθμο δεν απαιτείται και όλα τα επιθυμητά χαρακτηριστικά που έχουν οι γραμμικές μηχανές διανυσμάτων στήριξης παραμένουν. Έτσι μπορούμε να εκπαιδεύσουμε μία πολυωνυμική μηχανή ή ένα σιγμοιδές νευρωνικό δίκτυο, με χρήση αποδοτικών αλγορίθμων οι οποίοι δεν έχουν πρόβληματα με τοπικά ελάχιστα.

5.5. Συμπεράσματα

Όπως παρατηρούμε απο όλα τα παραπάνω που αναφέρθηκαν, οι μηχανές διανυσμάτων στήριξης αποτελούν μηχανές ταξινόμησης με κάποια πολύ καλα χαρακτηριστικά. Έχουν ένα πολύ σαφές και συγκεκριμένο θεωρητικό υπόβαθρο ενώ ταυτόχρονα έχωντας προσελκύσει το ενδιαφέρον των ερευνητών για αυτό τον λόγο, υπάρχουν πλέον στην βιβλιογραφία και πολύ καλά πρακτικά αποτελέσματα. Λόγω της κατασκευής τους δίνουν την δυνατότητα να επιλυθούν βασικά προβλήματα που εντοπίζονται σε περιπτώσεις άλλων μεθόδων, όπως για παράδειγμα η αντιμετώπιση των τοπικών ελαχίστων. Ταυτόχρονα με την χρήση των πυρήνων έχουμε την δυνατότητα να δημιουργήσουμε εύκολα μή γραμμικές μηχανές και μάλιστα με την δυνατότητα της παραγωγής χώρων χαρακτηριστικών κατάλληλους ανά περίπτωση δεδομένων εισόδου.

Παρόλα αυτά υπάρχουν διάφορα προβλήματα που προκύπτουν στην χρήση τους και τα οποία πολλές φορές ειδικά για τον αρχάριο δυσκολεύουν πάρα πολύ την διαδικασία εξαγωγής ικανοποιητικών αποτελεσμάτων, παρόλο που με τις σωστές επιλογές και διαδικασίες έχει αποδειχθεί πειραματικά πως η απόδοση των μηχανών αυτών είναι ισάξια ή και υπερέχει απο άλλες πιο γνωστές. Βασικό παράδειγμα προβλήματος στην χρήση των μηχανών διανυσμάτων στήριξης είναι η επιλογή των παραμέτρων του πυρήνα. Για παράδειγμα στην περίπτωση της χρήσης πολυωνυμικού πυρήνα, πέρα της παραμέτρου C που πρέπει να επιλεγεί σωστά, έχουμε και τις παραμέτρους d για τον βαθμό και άλλων 2 παραμέτρων που τον ορίζουν. Συνολικά δηλαδή έχουμε 4 παραμέτρους οι οποίες πρέπει να εκτιμηθούν σωστά ώστε να έχουμε ικανοποιητικά αποτελέσματα. Δεν υπάρχει σαφές θεωρητικό πλαίσιο για την επιλογή αυτών των τιμών, ενώ οι εξαντλητικές μέθοδοι αναζήτησης τους είναι ιδιαίτερα χρονοβόρες. Τέλος διάφορες ευριστικές μέθοδοι που έχουν προταθεί κατα καιρούς έχουν ιδιαίτερη πολυπλοκότητα τόσο ως προς την θεωρητική τους θεμελίωση όσο και ως προς την πρακτική τους υλοποίηση.

Ένα άλλο χαρακτηριστικό το οποίο επηρεάζει πολύ έντονα την απόδοση των μηχανών διανυσμάτων στήριξης είναι η μορφή των δεδομένων στον χώρο εισόδου. Όπως φένεται και στα διάφορα αποτελέσματα που προέκυψαν και στα πειράματα της συγκεκριμένης εργασίας, ακόμα και μικρές αλλαγές στην αναπαράσταση των δεδομένων εισόδου μπορεί να καταστήσει την προηγούμενη επιλογή των παραμέτρων άχρηστη, ακόμα και αν αυτή απέδιδε με την χρήση κάποιας άλλης αναπαράστασης. Είναι χαρακτηριστικό το παράδειγμα που θα παρουσιαστεί και παρακάτω, όπου ενώ με απλή εισαγωγή εικόνων χωρίς καμία προ επεξεργασία καταφέραμε να επιτύχουμε πολύ καλούς ρυθμούς αναγνώρισης για την περίπτωση του εντοπισμού χειλιών, ακόμα και η μικρότερη παρέμβαση στα δεδομένα αυτα(για παράδειγμα διόρθωση του φωτισμού), κατέστησε την προηγούμενη επιλογή των παραμέτρων του πολυωνυμικού πυρήνα μή αποδοτική. Είναι εμφανές απο όλες τις εργασίες που έχουν γίνει πάνω στην χρήση των μηχανών διανυσμάτων στήριξης πως η σωστή αναπαράσταση των δεδομένων εισόδου, ανάλογα με το προς μελέτη πρόβλημα, παίζει πολύ σημαντικό ρόλο στην απόδοση τους. Για παράδειγμα ενώ έχουν επιτευθεί πολύ καλοί ρυθμοί αναγνώρισης στην περίπτωση προσώπων, με την χρήση μόνο των εικονοστοιχείων ώς είσοδο, έχει παρατηρηθεί πως τα αποτελέσματα είναι πολύ καλύτερα στην περίπτωση που χρησιμοποιηθεί άλλη αναπαράσταση για τις εικόνες. Για παράδειγμα DCT μετασχηματισμός ή Gabor φίλτρα .

Τέλος αναφέραμε πως σε γενικές γραμμές οι μηχανές διανυσμάτων στήριξης επιτυγχάνουν να αντιμετωπίσουν το πρόβλημα των πολλών διαστάσεων του χώρου χαρακτηριστικών. Παρόλο που επιτυγχάνουν να το αντιμετωπίσουν καλύτερα απο πολλές άλλες μηχανές, παραμένουν ευάισθητες ώς προς αυτό το πρόβλημα το οποίο αναφέρεται στην βιβλιογραφία και ώς «κατάρα των διαστάσεων» (curse of dimensionality).

ΚΕΦΑΛΑΙΟ 6ο.

ΣΥΜΠΕΡΑΣΜΑΤΑ

Ένα πρώτο γενικό συμπέρασμα το οποίο προκύπτει άμεσα μετά την εκπόνηση της συγκεκριμένης εργασίας, είναι η ιδιαιτερότητα που προκύπτει σε κάθε τμήμα της αρχιτεκτονικής. Παρόλη την προσπάθεια του να διατηρηθούν συγκεκριμένες μεθοδολογίες σε όλα τα τμήματα, αυτό δεν ήταν τελικά δυνατό. Οι ιδιαιτερότητες του κάθε τμήματος καθώς και οι απαιτήσεις που προέκυπταν με βάση τους βασικούς άξονες που είχαμε ορίσει αρχικά μας οδήγησαν να κάνουμε αλλαγές ακόμα και σε βασικές επιλογές, όπως για παράδειγμα αυτή του ταξινομητή. Σε κάθε περίπτωση πάντως αφού λάβαμε υπόψην τα ιδιαίτερα χαρακτηριστικά τα οποία προέκυπταν, καταφέραμε σε γενικές γραμμές να είμαστε συνεπείς στους βασικούς άξονες τους οποίους αναφέραμε και στην εισαγωγή, δηλαδή στον υψηλό βαθμό αναγνώρισης και ταυτόχρονα το κάθε τμήμα της αρχιτεκτονικής να λειτουργεί όσο το δυνατόν πιό κοντά σε πραγματικό χρόνο γίνεται. Ένα άλλο βασικό συμπέρασμα το οποίο προέκυψε, έχει να κάνει με την δομή της αρχιτεκτονικής αυτής καθ'αυτής. Σε ένα σύστημα που επικρατεί μία σειριακή δομή, όπου το κάθε τμήμα τροφοδοτεί με αποτελέσματα κάποιο επόμενο, η απόδοση και ο ρυθμός αναγνώρισης του κάθε τμήματος ξεχωριστά επιρεάζει σε μεγάλο βαθμό την τελική απόδοση. Αυτό συμβαίνει γιατί αν για παράδειγμα στο πρώτο τμήμα, του εντοπισμού του προσώπου, γίνει λάθος, όλο το σύστημα θα εξάγει λάθος συμπεράσματα σε κάθε επόμενο βήμα, ενώ ταυτόχρονα ο εντοπισμός του σημείου στο οποίο έγινε το σφάλμα είναι σχεδόν αδύνατη μιας και σφάλματα μπορούν να υπάρξουν σε κάθε ένα απο αυτά. Για αυτό τον λόγο η απόδοση και ο καλός βαθμός αναγνώρισης σε κάθε τμήμα θεωρείται ζωτικής σημασίας για ένα τέτοιο σύστημα .Στην συνέχεια παραθέτουμε ιδιαίτερα συμπεράσματα και παρατηρήσεις για κάθε τμήμα ξεχωριστά.

6.1. Συμπερασματα για τον εντοπισμο του προσωπου

Η περίπτωση του εντοπισμού του προσώπου μας απασχόλησε ιδιαίτερα. Η βασική αιτία για αυτό είναι όπως αναφέραμε και στο αντίστοιχο κεφάλαιο, το ότι αφενώς το συγκεκριμένο τμήμα αποτελεί πεδίο έρευνας απο μόνο του, αφετέρου τα προβλήματα τα οποία προκύπτουν συναντόντε και σε πολλές άλλες περιπτώσεις αναγνώρισης προτύπων.

Αρχικά επιλέξαμε να χρησιμοποιήσουμε μία σύγχρονη μέθοδο για τον εντοπισμό του προσώπου η οποία βασίζεται στην χρήση διανυσμάτων στήριξης. Η επιλογή αυτή έγινε, γιατί αφενώς στην βιβλιογραφία παρουσιάζοντε πολύ καλά αποτελέσματα, αφετέρου γιατί επιθυμούσαμε να διατηρήσουμε ένα συγκεκριμένο είδος ταξινομητή σε όλα τα τμήματα της αρχιτεκτονικής. Οι μηχανές διανυσμάτων στήριξης επιλέχθηκαν γιατι εκτός των άλλων έχουν ένα πολύ καλά ορισμένο μαθηματικό υπόβαθρο, ενώ ταυτόχρονα μας δίνουν την δυνατότητα να τις μετατρέψουμε σε ισοδύναμες με νευρωνικά δίκτυα αν αυτό ήταν σκόπιμο. Η μετατροπή αυτή βασίζεται σε συγκεκριμένη επιλογή πυρήνα και είναι ιδιαίτερα εύκολη να γίνει.

Τελικά η επιλογή αυτή κρίθηκε ατυχής. Αφενώς δεν επιτύχαμε να αναπαράγουμε τα αποτελέσματα που βρήκαμε στην βιβλιογραφία, αφετέρου η διαδικασία εκπαίδευσης μίας τέτοιας μηχανής ταξινόμησης με τον όγκο τον δεδομένων που απαιτεί το συγκεκριμένο πρόβλημα έκανε πάρα πολύ δύσκολη την όλη διαδικασία. Ταυτόχρονα παρατηρήθηκε πως ο χρόνος απόκρισης δεν ήταν ικανοποιητικός. Ο βασικός λόγος που τα αποτελέσματα που προέκυπταν απο την ταξινόμηση ήταν ιδιαιτέρως άσχημα είναι πως η μηχανή γενίκευε σε κάθε περίπτωση υπέρ της κλάσης των μή προσώπων. Αυτό συμβαίνει γιατί οι δύο κλάσεις, πρόσωπο και μή πρόσωπο διαφέρουν πολύ ως προς το δυνατό μέγεθος παραδειγμάτων. Ουσιαστικά απο την μία πλευρά έχουμε το πρόσωπο το οποίο είναι κάτι πολύ συγκεκριμένο απο πλευράς σχήματος, χρώματος κοκ, και απο την άλλη έχουμε το σύνολο όλων των άλλων πιθανών προτύπων εισόδου. Στην βιβλιογραφία προτίνεται ώς μέθοδος για την αντιμετόπιση αυτού το προβλήματος η διαδικασία του BootStraping. Στην παρούσα εργασία όμως, επειδή αυτή η διαδικασία είναι ιδιαιτέρως χρονοβόρα ενώ ταυτόχρονα ο εντοπισμός του προσώπου είναι μόνο ένα τμήμα της, δεν έγινε κάποια δοκιμή η υλοποίηση του BootStraping.

Τελικά καταλήξαμε στην επιλογή της μεθόδου εντοπισμού προσώπου που περιγράψαμε στο αντίστοιχο κεφάλαιο και η οποία βασίζεται στην χρήση της μεθόδου AdaBoost. Τόσο τα

αποτελέσματα των πειραμάτων, όσο και η δυνατότητα της μεθόδου αυτής να λειτουργεί σε σχεδόν πραγματικό χρόνο, την καθιστά ιδανική για τις ανάγκες μας. Τέλος πρέπει να αναφέρουμε πως βασικό πλεονέκτημα αυτής της μεθόδου είναι και η ύπαρξη πολλώς έτοιμων υλοποιήσεων σε διάφορες πλατφόρμες, με αποτέλεσμα να είναι έυκολη τόσο η πειραματική επαλήθευση της, όσο και η ενσωμάτωση της σε ένα σύστημα.

6.2. Συμπερασματα για τον εντοπισμο της περιοχης των χειλιων

Ένα βασικό συμπέρασμα που προκέκυψε απο την διαδικασία εντοπισμού της περιοχής των χειλιών είναι η επαλήθευση της ικασίας την οποία κάναμε για τους λόγους αποτυχίας της εφαρμογής των μηχανών διανυσμάτων στήριξης στην περίπτωση του εντοπισμού του προσώπου. Στην περίπτωση των χειλιών τα πρότυπα τα οποία καλείται το σύστημα να κατηγοριοποιήση ως χείλη και μή χείλη είναι σαφώς πιο περιορισμένα. Αυτό συμβαίνει γιατι τα πρότυπα αυτά προέρχοντε απο μία περιορισμένη περιοχή του προσώπου. Για τον λόγο αυτό τα αποτελέσματα που προέκυψαν κατά την εφαρμογή των μηχανών διανυσμάτων στήριξης στον εντοπισμό της περιοχής των χειλιών κρίθηκε ιδιαίτερα ικανοποιητική, όπως αυτά στηρίζονται και απο τις τιμές των πειραμάτων που παρουσιάζονται στο αντίστοιχο κεφάλαιο.

Ταυτόχρονα έγινε φανερό το πόσο επηρεάζει την χρονική απόκριση μίας τέτοιας μηχανής ταξινόμησης το μέγεθος του διανύσματος εισόδου. Στην περίπτωση των χειλιών το διάνυσμα είναι μικρότερο απο αυτό που προκύπτει στην περίπτωση του προσώπου με αποτέλεσμα η μηχανή να αποκρίνεται πολύ πιο γρήγορα προσεγγίζοντας έτσι πραγματικό χρόνο. Τα αποτελέσματα τα οποία προέκυψαν εκτός απο ικανοποιητικά για τις ανάγκες μας κρίθηκαν και ανταγωνιστικά αυτών της βιβλιογραφίας τα οποία παρουσιάστηκαν στο αντίστοιχο κεφάλαιο.

6.3. Συμπερασματα για την εξαγωγη και τον εντοπισμο των χαρακτηριστικων

Για τον ορισμό και την εξαγωγή των χαρακτηριστικών επιλέχθηκε η χρήση της περιγραφής του λόγου με βάση τα χαρακτηριστικά της άρθρωσης. Ένα βασικό πλεονέκτημα αυτής της μεθόδου είναι πως βασίζεται σε μία σαφώς ορισμένη επιστήμη, της φωνολογίας. Αποτέλεσμα αυτού είναι πως υπάρχει μία πληθώρα πληροφοριών για τους μηχανισμούς παραγωγής των ήχων που συνθέτουν τον λόγο. Ταυτόχρονα η ύπαρξη συστημάτων όπως το IPA δίνουν μία συστηματική μέθοδο για την αποδόμιση των λέξεων σε φωνητικές μονάδες και τον χαρακτηρισμό αυτών.

Ταυτόχρονα απο την βιβλιογραφία προκύπτουν πολύ ενθαρυντικά αποτελέσματα απο την χρήση των σημείων άρθρωσης. Παρά τα ενθαρυντικά συμπεράσματα όμως η μέθοδος αυτή είναι πολύ καινούργια και εμφανίζει κάποια προβλήματα. Ένα βασικό πρόβλημα σχετίζεται με αυτή καθ'αυτή την επιστήμη της φωνολογίας. Οι όροι καθώς και οι περιγραφές των διαδικασιών προσεγγίζουν σε μεγάλο βαθμό την ιατρική επιστήμη, με αποτέλεσμα να είναι δύσκολο για κάποιον που δεν είναι γνώστης του συγκεκριμένου πεδίου να εκμεταλευτεί την γνώση που υπάρχει. Σε πολλές περιπτώσεις τα χαρακτηριστικά της άρθρωσης που προκύπτουν εμφανίζοντε σε παραπάνω απο ένα φώνημα, η συγκεκριμένη συμπεριφορά είναι ιδιαίτερα εμφάνης στην παραγωγή των συμφώνων. Αυτό έχει σαν αποτέλεσμα να υπάρχει ιδιαίτερη δυσκολία στην διάκριση των φωνημάτων ειδικά όταν γίνεται χρήση μόνο της οπτικής πληροφορίας. Στην περίπτωση όμως των φωνηέντων ο ορισμός που δίνεται είναι επαρκής για τον σαφή καθορισμό και κατ'επέκταση εντοπισμό τους. Πρέπει εδώ να τονίσουμε πως το συγκεκριμένο πρόβλημα υπάρχει και στην περίπτωση της χρήσης άλλων μεθόδων όπως είναι τα visemes, αφού περισσότερα του ενός φωνήματος ανήκουν στην ίδια κατηγορία viseme.

Η μέθοδος που χρησιμοποιήθηκε για την ταξινόμηση των σημείων άρθρωσης και την παραγωγή του διανύσματος είναι πάλι οι μηχανές διανυσμάτων στήριξης. Τα χαρακτηριστικά άρθρωσης δίνουν την δυνατότητα να προκύψει ένα εμπλουτισμένο σύνολο δεδομένων εκπαίδευσης για τις μηχανες διανυσμάτων στήριξης και αυτό αποτελεί ακόμα ένα πλεονέκτημα τους. Ταυτόχρονα το περιορισμένο πλήθος των διαφορετικών προτύπων που προκύπτει δίνουν την δυνατότητα στους ταξινομητές να αποδόσουν σε ικανοποιητικό βαθμό και αρκετά γρήγορα. Το συμπέρασμα αυτό προκύπτει τόσο απο πειραματική επαλήθευση που έγινε κατά την διάρκεια της εργασίας αυτής, όσο και απο την διαθέσιμη βιβλιογραφία.

6.4. Προτασεις για μελλοντικες βελτιωσεις

Στην περίπτωση του εντοπισμού των προσώπων λίγα έχουν να προταθούν. Η ίδια η διαδικασία αποτελεί απο μόνη της πεδίο έρευνας, υπάρχει έντονος πλούτος απο διαφορετικές μεθόδους που λειτουργούν καλά ανά περίπτωση, ενώ καινούργιες μέθοδοι προκύπτουν συνεχώς. Μία πρόταση είναι να γίνει προσπάθεια να εφαρμοσθούν οι μηχανές διανυσμάτων στήριξης και σε αυτό το πρόβλημα. Ο λόγος για αυτό προκύπτει κυρίως απο ανάγκη που σχετίζεται με τον σχεδιασμό ενός πληροφοριακού συστήματος και όχι τόσο απο ερευνητικές ανάγκες. Σε ένα σύστημα

προσπαθούμε να έχουμε όσο δυνατόν μεγαλύτερη συνοχή απο πλευράς μεθόδων και τεχνικών που χρησιμοποιούμε. Ο λόγος που μας οδηγεί σε αυτό είναι το ότι προκύπτει ένα σύστημα το οποίο είναι πιο εύκολα επεκτάσιμο και προβλέψιμο.

Στην περίπτωση του εντοπισμού των χειλιών μία βασική πρόταση είναι η έρευνα και εφαρμογή μεθόδων προ επεξεργασίας στα δεδομένα που τροφοδοτούντε στις μηχανές ταξινόμησης. Ένα θέμα το οποίο δεν αγγίξαμε σε αυτή την εργασία είναι η ελαχτιστοποίηση των διαστάσεων του διανύσματος εισόδου των μηχανών ταξινόμησης. Η ελαχτιστοποίηση των διαστάσεων δίνει την δυνατότητα τόσο του να επιτύχουμε καλύτερους ρυθμούς αναγνώρισης αφαιρόντας περιττή πληροφορία, όσο και ταχύτερης αναγνώρισης, αφού όπως αναφέραμε το μέγεθος του διανύσματος επιρεάζει σε μεγάλο βαθμό την απόκριση της μηχανής ταξινόμησης. Ταυτόχρονα έχει φανεί απο αναφορές στην βιβλιογραφία πως διάφορες τεχνικές επεξεργασίας εικόνας που δίνουν την δυνατότητα εξομάλυνσης της εικόνας βοηθάνε στην απόδοση του ταξινομητή. Τέτοιες μέθοδοι δεν χρησιμοποιήθηκαν εκτενώς στην εργασία αυτή αλλα η μελλοντική ενσωμάτωση τους δεν θα αποτελούσε ιδιαίτερο πρόβλημα.

Στην εξαγωγή των χαρακτηριστικών υπάρχουν τρείς βασικές προτάσεις. Η πρώτη αφορα την χρήση ασαφούς λογικής για την περιγραφή των χαρακτηριστικών αυτών. Χαρακτηριστικά όπως το άνοιγμα των χειλιών και η σχετική θέση της γλώσσας είναι ιδανικά για να περιγραφούν με χρήση ασαφών συνόλων και η χρήση τους θα μπορούσε να αυξήσει την διακριτική ικανότητα της μετέπειτα ταξινόμησης. Η δεύτερη αφορά την συστηματική ταξινόμηση των πληροφοριών που υπάρχουν απο την επιστήμη της φωνολογίας. Οι πληροφορίες αυτές είναι πολλές σε πλήθος αλλά σαφώς ορισμένες και κατανοητές απο κάποιον γνώστη του αντικειμένου. Αυτό μας οδηγεί στο συμπέρασμα πως η χρήση μεθόδων που παρέχει το σημασιολογικό δίκτυο θα ήταν ιδανικές για την μηχανιστική αποτύπωση των κανόνων και πληροφορίων που παρέχει η επιστήμη αυτή, ενώ η ύπαρξη συστημάτων όπως το IPA κάνει την παραγωγή μίας οντολογίας για τα φωνήματα ακόμα πιό εύκολη. Μία σωστά ορισμένη οντολογία βασισμένη στις πληροφορίες που παρέχει η φωνολογία, θα μπορούσε να παρέχει την αυτόματη αποδόμισει των λέξεων σε φωνήματα και την ανάσυρση όλων αυτών των πληροφοριών οι οποίες θα οδηγούσαν στον χαρακτηρισμό και την παραγωγή του διανύσματος χαρακτηριστικών. Αυτό θα μπορούσε να βοηθήσει ιδιαίτερα στην έρευνα, αφού η παραπάνω διαδικασία είναι ιδιαίτερα χρονοβόρα και επιρεπής σε λάθη ειδικά απο κάποιον μή ειδικό. Ταυτόχρονα θα μπορούσε να βοηθήσει και σε ένα πραγματικό σύστημα

αφού η διαδικασία εκπαίδευσης των ταξινομητών θα μπορούσε να αυτοματοποιηθεί ενώ ταυτόχρονα θα ήταν δυνατή και η χρήση των κανόνων και την αξιολόγηση των αποτελεσμάτων της ταξινόμησης. Η τελευταία πρόταση αφορά την ενσωμάτωση του συστήματος οπτικής αναγνώρισης ομιλίας, με ένα σύστημα ακουστικής αναγνώρισης το οποίο να βασίζεται και αυτό στα χαρακτηριστικά άρθρωσης. Όπως αναφέραμε και στο αντίστοιχο κεφάλαιο υπάρχουν χαρακτηριστικά τα οποία ορίζονται στην φωνολογική ανάλυση των μονάδων του λόγου τα οποία μπορούν να εντοπισθούν ακουστικά αλλά όχι οπτικά. Αυτό θα οδηγούσε σε ένα σύστημα με πολύ καλύτερους ρυθμούς αναγνώρισης.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] A. Adjoudani and C. Benoit, On the integration of auditory and visual parameters in HMM-based ASR, in Speechreading by Humans and Machines, D. G. Stork and M. E. Hennecke, Eds. Berlin, Germany: Springer, pp. 461–471, 1996.

[2] A. Hill and C.J. Taylor, Automatic Landmark Generation for Point Distribution Models, Proc. British Machine Vision Conf., pp. 429-438, 1994.

[3] C. Bregler and S.M. Omohundro, Learning Visual Models for Lipreading, Computational Imaging and Vision, chapter 13, vol. 9, pp. 301-320, 1997.

[4] C. Bregler and Y. Konig, Eigenlips for Robust Speech Recognition, In Proc. ICASSP, 1994.

[5] C. Bregler, H. Hild, S. Manke, and A. Waibel, Improving Connected Letter Recognition by Lipreading, Proc. Int'l Conf. Acoustics, Speech and Signal Processing, vol. 1, pp. 557-560, 1993.

[6] C. Fisher, *Confusions among visually perceived consonants.* Journal of Speech and Hearing Research, 11(4):796–804, 1968.

[7] C. Papageorgiou, M. Oren, T. Poggio, A general framework for object detection, International Conference on Computer Vision, 1998.

[8] Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[9] E. Patterson, S. Gurbuz, Z. Tufekci, J.N. Gowdy, CUAVE: a new audio-visual database for multimodal human-computer interface research, Department of Electrical and Computer Engineering, Clemson University, Clemson, SC 29634, USA

[10] E. Petajan, Automatic lip-reading to enhance speech recognition, In Proc. Global Telecomm. Conf., pp. 265–272, Atlanta, GA, 1984.

[11] G. Fant, Acoustic Theory of Speech Production, Netherlands: Mouton and Co., 1960.

[12] G. Miller and P. Nicely, An Analysis of Perceptual Confusions among some English Consonants, J. Acoustical Society America, vol. 27, no. 2, pp. 338-352, 1955.

[13] H. McGurk, J. MacDonald, *Hearing lips and seeing voices.* Nature, 746–748, 1976.

[14] I. Matthews, T. F. Cootes, J. A. Bangham, S. C. and R. Harvey, Extraction of Visual Features for Lipreading, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 2, 2002, pp. 198–213.

[15] J. Haslam, C.J. Taylor, and T.F. Cootes, A Probabilistic Fitness Measure for Deformable Template Models, Proc. British Machine Vision Conf., pp. 33-42, 1994.

[16] J. Luettin and N.A. Thacker, Speech-reading Using Probabilistic Models, Computer Vision and Image Understanding, vol. 65, no. 2, pp. 163-178, Feb. 1997.

[17] J. Luettin, Visual Speech and Speaker Recognition, PhD thesis, Univ. of Sheffield, May 1997.

[18] J. Nelder and R. Mead, A Simplex Method for Function Minimization, Computing J., vol. 7, no. 4, pp. 308-313, 1965.

[19] K. Livescu and J. Glass, Feature-based Pronunciation Modeling for Speech Recognition, In Proc. HLT/NAACL, Boston, 2004.

[20] K. Mase and A. Pentland, Automatic Lipreading by optical flow analysis, Systems and Computers in Japan, vol. 22, no. 6, pp. 67-76, 1991.

[21] K. Saenko, T. Darrell, J. Glass, Articulatory Features for Robust Visual Speech Recognition, MIT Computer Science and Artificial Intelligence Laboratory 32 Vassar Street Cambridge, Massachusetts, USA

[22] M. Gordan, C. Kotropoulos, and I. Pitas, A support vector machine based dynamic network for visual speech recognition applications, EURASIP J. Appl. Signal Processing, vol. 2002, no. 11, pp. 1248–1259, 2002.

[23] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, T. Poggio, Pedestrian detection using wavelet templates, Proc. Computer Vision and Pattern Recognition, 1997.

[24] M. Turk and A. Pentland, Eigenfaces for Recognition, Journal of Cognitive Neuroscience, Volume 3, Number 1, MIT 1991.

[25] N. Brooke and S.D. Scott, PCA Image Coding Schemes and Visual Speech Intelligibility, Proc. Inst. of Acoustics, vol. 16, no. 5, pp. 123-129, 1994.

[26] N. Chomsky and M. Halle, The Sound Pattern of English, Harper and Row, New York, 1968.

[27] P. Viola and M. J. Jones, Robust Real-Time Face Detection, International Journal of Computer Vision, Springer, Volume 57, Number 2 / May, 2004.

[28] R. Kaucic and A. Blake, Accurate, Real-Time, Unadorned Lip Tracking, Proc Sixth Int'l Conf. Computer Vision, 1998.

[29] S. Basu, N. Oliver, and A. Pentland, 3D Modeling and Tracking of Human Lip Motions, Proc. Int'l Conf. Computer Vision, 1998.

[30] S. Gurbuz, Z. Tufekci, E. Patterson, and J. Gowdy, Application of affine-invariant fourier descriptors to lipreading for audio-visual speech recognition, in Proc. Int. Conf. Acoust., Speech, Signal Processing, pp. 177–180, Salt Lake City, UT, 2001.

[31] S. King, T. Stephenson, S. Isard, P. Taylor and A. Strachan, Speech recognition via phonetically featured syllables, In Proc. ICSLP, Sydney, 1998.

[32] T. Cootes, A. Hill, C.J. Taylor, and J. Haslam, The Use of Active Shape Models for Locating Structures in Medical Images, Image and Vision Computing, vol. 12, no. 6, pp. 355-366, 1994.

[33] T. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham, Active Shape Models—Their Training and Application, Computer Vision and Image Understanding, vol. 61, no. 1, pp. 38-59, Jan. 1995.

[34] T. Gernoth, R. Kricke, R.R. Grigat, Mouth Localization for Appearance-based Lip Motion Analysis, WSEAS Transactions on Signal Processing, vol. 3, no. 3, pp. 275-281, 2007.