Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων

τομεας τεχνολογιας πληροφορικης και υπολογιστων

εργαστηριο επεξεργασιας εικονας βιντεο και πολυμεσων

Ημερολογιοποίηση Ομιλητών με Βάση την Οπτική Πληροφορία

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Νικόλαου Σαραφιανού

Επιβλέπων: Στέφανος Κόλλιας
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2013

Εθνικο Μετσοβιο Πολυτεχνειο
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ
ΕΡΓΑΣΤΗΡΙΟ ΕΠΕΞΕΡΓΑΣΙΑΣ ΕΙΚΟΝΑΣ ΒΙΝΤΕΟ ΚΑΙ ΠΟΛΥΜΕΣΩΝ

Ημερολογιοποίηση Ομιλητών με Βάση την Οπτική Πληροφορία

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Νικόλαου Σαραφιανού

Επιβλέπων : Στέφανος Κόλλιας
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή επιτροπή την 11η Ιουλίου 2013.

.....	.....	.....
Στέφανος Κόλλιας	Γεώργιος Στάμου	Θεόδωρος Γιαννακόπουλος
Καθηγητής Ε.Μ.Π.	Λέκτορας Ε.Μ.Π.	Ερευνητής ,"Δημόκριτος"

Αθήνα, Ιούλιος 2013

.....
Νικόλαος Σαραφιανός
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Ηλεκτρονικών Υπολογιστών

Περίληψη

Η παρούσα διπλωματική εργασία μελετά το πρόβλημα της Ημερολογιοποίησης Ομιλητών με βάση την οπτική πληροφορία που εξάγεται από ένα βίντεο. Η Ημερολογιοποίηση Ομιλητών αποτελεί ένα ιδιαίτερα ενεργό πεδίο της Αναγνώρισης Προτύπων εξαιτίας της ολοένα και αυξανόμενης ανάγκης εξαγωγής και επεξεργασίας των πληροφοριών που περιέχονται στα πολυμέσα. Βρίσκει εφαρμογές σε ανίχνευση πνευματικών δικαιωμάτων, σε επιστημονικούς κλάδους που ασχολούνται με αυτόματη ανάλυση συμπεριφοράς, ενώ ταυτόχρονα, είναι μια πολύ σημαντική διαδικασία για ανάκτηση πληροφοριών με εφαρμογές σε επιστημονικά πεδία όπως η προσαρμογή των ομιλητών για αυτόματη ανίχνευση φωνής.

Συγκεκριμένα η εργασία μας επικεντρώνεται στην εξαγωγή οπτικών χαρακτηριστικών τα οποία να είναι ικανά για διαχωρισμό ομιλητών και σε συνδυασμό με μεθόδους που δημιουργούν ένα πιο αντιπροσωπευτικό χώρο χαρακτηριστικών, ομαδοποιούμε τα χαρακτηριστικά ώστε να δοθεί απάντηση στο αρχικό ερώτημα του "Ποιος μίλησε και πότε".

Αρχικά δίνεται έμφαση σε μεθόδους χωρισμού ενός βίντεο σε μικρότερα τμήματα που ονομάζονται shots. Αφού αναφέρουμε τις βασικές μεθόδους που χρησιμοποιούνται στη βιβλιογραφία, συγκρίνουμε τα αποτελέσματα και επισημαίνουμε τη συμβολή του χωρισμού ενός βίντεο σε επιμέρους shots στην Ημερολογιοποίηση Ομιλητών. Στη συνέχεια, αφού περιγράψουμε με λεπτομέρεια όλα τα στάδια της μεθόδου ανίχνευσης προσώπου των Viola & Jones, ερευνάμε τεχνικές εξαγωγής χαρακτηριστικών από αυτό.

Επιδιώκουμε στη συνέχεια τη μείωση των διαστάσεων του αρχικού χώρου των παραπάνω χαρακτηριστικών και συνεπώς μελετήσαμε και υλοποιήσαμε τεχνικές μείωσης των διαστάσεων σε ένα μικρότερο χώρο. Η κυριότερη μέθοδος με την οποία ασχοληθήκαμε ονομάζεται FLsD και δεδομένου ότι εκμεταλλεύεται τα πλεονεκτήματα υπάρχοντων μεθόδων μείωσης των διαστάσεων επιτυγχάνει πολύ καλύτερα αποτελέσματα. Επιπλέον στο χώρο όπου έχουν μειωθεί πια οι διαστάσεις επιχειρούμε να ομαδοποιήσουμε τα τελικά μας δεδομένα σε ομάδες οι οποίες θα αντιστοιχούν σε ομιλητές. Η αξιολόγηση όλων των παραπάνω τεχνικών και μεθόδων γίνεται μέσω πειραμάτων με τη βοήθεια των οποίων μας δίνεται η δυνατότητα να οπτικοποιήσουμε τα αποτελέσματα μας και να εξάγουμε συμπεράσματα για την απόδοση της μεθόδου Ημερολογιοποίησης Ομιλητών που προτείνουμε. Επιπλέον επισημαίνουμε τα περιθώρια βελτίωσης που υπάρχουν στην εν λόγω μέθοδο με στόχο να προσφέρουμε πολλαπλές κατευθύνσεις για μελλοντική εργασία.

Λέξεις-κλειδιά: αναγνώριση προτύπων, ημερολογιοποίηση ομιλητών, ανίχνευση αλλαγής shot, ανίχνευση προσώπου, ανίχνευση δέρματος, εξαγωγή χαρακτηριστικών με τη χρήση Gabor κυματιδίων, μείωση των διαστάσεων, ομαδοποίηση, ανίχνευση κίνησης των χειλιών

Abstract

The objective of this thesis is visual-based speaker diarization in videos. Speaker diarization is a notably active field of pattern recognition due to the increasing need for extraction and processing of information contained in multimedia. Speaker diarization is applied in copyright detection and in scientific fields that deal with automatic behavior analysis. It is also a significant procedure for information retrieval with applications in scientific fields such as speaker adaptation for automatic voice detection.

Specifically, our work focuses, in particular, on the extraction of speaker discriminant visual characteristics and in collaboration with dimensionality reduction methods that create a more representative feature space, we cluster our features in order to answer the initial question "Who spoke when".

Firstly we give emphasis in methods for video segmentation methods that result in shorter video segments called shots. Once we have mentioned the basic state-of-the-art methods, we compare the results and point out the contribution of video segmentation in shots in the speaker diarization problem. After describing with detail all the steps of the face detection method of Viola & Jones, we investigate face feature extraction techniques.

Moreover, we focus our attention on reducing the dimensions of the initial space of the aforementioned features and as a result we studied and implemented dimensionality reduction techniques. The main method being used in this thesis is called FLsD which takes advantage of the benefits of the existing dimensionality reduction methods and achieves far better results. In the reduced feature space, we applied some clustering methods in order to gather the features in groups, each one of which will correspond to a speaker. The evaluation of all the above techniques is performed through experiments in order to visualize the results and to draw conclusions about the performance of our speaker diarization method. Finally, we point out margin improvements of the current method, aiming to provide several directions for future work

Keywords: pattern recognition, speaker diarization, shot change detection, face detection, skin detection, feature extraction with Gabor wavelets, dimensionality reduction, clustering, lip movement detection

Ευχαριστίες

Η εμπειρία της παρακολούθησης των μαθημάτων της Τεχνολογίας και Ανάλυσης Εικόνων και Βίντεο και των Σημάτων και Συστημάτων που διδάσκονται στη σχολή μας από τον καθηγητή Σ. Κόλλια, υπήρξε καθοριστική όχι μόνο για την επιλογή ενός θέματος διπλωματικής εργασίας από τις παραπάνω περιοχές αλλά και γενικά για τη διαμόρφωση των ακαδημαϊκών μου ενδιαφερόντων. Θα ήθελα να τον ευχαριστήσω θερμά τόσο για την εμπιστοσύνη του που μου έδειξε για την ανάθεση αυτής διπλωματικής εργασίας όσο και για την προθυμία του να με βοηθήσει όποτε υπήρξε ανάγκη.

Καθοριστική υπήρξε καθ' όλη τη διάρκεια της περιόδου εκπόνησης αυτής της εργασίας στο Εργαστήριο Υπολογιστικής Νοημοσύνης στο Δημόκριτο η συνεργασία με το μεταδιδακτορικό ερευνητή Θεόδωρο Γιαννακόπουλο. Η βοήθεια και η καθοδήγηση του σε όλα τα στάδια εκπόνησης της διπλωματικής εργασίας ήταν εξαιρετικά σημαντικές. Για όλα τα παραπάνω καθώς και για την αμείωτη διάθεση του για ανταλλαγή ιδεών και συνεργασία θέλω να του εκφράσω τις ιδιαίτερες ευχαριστίες μου. Επιπλέον θα ήθελα να ευχαριστήσω τόσο τον μεταδιδακτορικό ερευνητή Σέργιο Πετρίδη για την πολύτιμη καθοδήγηση του τον τελευταίο ένα χρόνο όσο και τον κ. Περαντώνη για την εμπιστοσύνη που μου έδειξε. Το γεγονός ότι και οι τρεις κάθε φορά που χρειάστηκα τη βοήθεια τους ήταν εκεί πρόθυμοι να συμβάλλουν παρά την πληθώρα των υποχρεώσεων που είχαν, είναι κάτι που δε θα μπορούσα να αγνοήσω και τους ευχαριστώ θερμά για αυτό.

Τέλος θα ήθελα να ευχαριστήσω την οικογένεια μου για την συνεχή υποστήριξη που συνεχίζουν να προσφέρουν όλα αυτά τα χρόνια ενθαρρύνοντας κάθε βήμα προς την πραγμάτωση των στόχων μου, τη Β. Καλουδά για την στήριξη της τον τελευταίο χρόνο καθώς και τους φίλους μου για την μεγάλη συμπαράσταση που μου έδωσαν κατά τη φοίτηση μου στο Πολυτεχνείο.

Περιεχόμενα

Σελίδα

Κατάλογος Σχημάτων

Κατάλογος Αλγορίθμων

Σελίδα

Κεφάλαιο 1

Εισαγωγή

1.1 Γενικά για την Αναγνώριση Προτύπων

Ως ανθρώπινα όντα έχουμε την ιδιότητα να δεχόμαστε πληροφορίες από το περιβάλλον μέσω των αισθήσεων. Χρησιμοποιώντας μια σειρά από γενικές ιδέες ή πρότυπα που έχουμε μάθει για τα αντικείμενα, σε συνδυασμό με πληροφορίες από τις αισθήσεις μας και τη γνωστική ικανότητα της αναγνώρισης μπορούμε, για παράδειγμα, να αναγνωρίσουμε ένα χαρακτήρα του αλφάβητου, να διαχωρίσουμε ένα αρσενικό από ένα θηλυκό πρόσωπο ή να αναγνωρίσουμε ένα γνωστό πρόσωπο όταν ακούμε τη φωνή του στο τηλέφωνο. Τα προηγούμενα παραδείγματα και γενικότερα όλες οι διαδικασίες αναγνώρισης περιλαμβάνουν μια ταξινόμηση ή ένα προσδιορισμό των αντικειμένων, ανθρώπων ή γεγονότων. Στη συνέχεια γίνεται η λήψη μιας απόφασης και αν απαιτείται διενεργείται μια ενέργεια όπως για παράδειγμα η απόρριψη των κομματιών που προσδιορίστηκαν ως λανθασμένα ή κατεστραμμένα σε μια γραμμή παραγωγής. Η πολυπλοκότητα και η επαναληπτικότητα αυτών των ταξινομήσεων, η ανάγκη αύξησης της αξιοπιστίας και της αντικειμενικότητας των αποφάσεων οδήγησαν στην ανάπτυξη αλγορίθμων ικανών να αποσπούν τις πιο σημαντικές πληροφορίες από το περιβάλλον και να τις αναπαριστούν μαθηματικά. Με αυτό τον τρόπο αναπαράστασης δημιουργείται η δυνατότητα να σχηματιστεί μια έννοια κλάσης ή κατηγορίας ενώ ταυτόχρονα να αναγνωρίζονται και να ταξινομούνται τα αντικείμενα αυτόματα.

Η Αναγνώριση Προτύπων συνεπώς, είναι ο επιστημονικός κλάδος στόχος του οποίου είναι, δοθέντων κάποιων αντικειμένων στην είσοδο να εφαρμόσει σε αυτά μια σειρά από ενέργειες οι οποίες βασίζονται στην κατηγορία του προτύπου με απώτερο σκοπό την ταξινόμηση των αντικειμένων αυτών σε μια κατηγορία ή κλάση. Ανάλογα με την εφαρμογή τα αντικείμενα αυτά μπορεί να είναι εικόνες, ηχητικά κομμάτια, γραφικές παραστάσεις σημάτων ή οποιοδήποτε είδος από μετρήσεις που πρέπει να ταξινομηθεί.

Το πρόβλημα της αναζήτησης προτύπων σε δεδομένα είναι θεμελιώδες και έχει ένα μακρύ παρελθόν. Στο δυτικό κόσμο, τα θεμέλια για την Αναγνώριση Προτύπων φτάνουν μέχρι τον Πλάτωνα και στη συνέχεια τον Αριστοτέλη ο οποίος έκανε το διαχωρισμό μιας ουσιώδους ιδιότητας (που θα είναι κοινή για όλα τα μέλη μια κλάσης) και της τυχαίας ιδιότητας (που μπορεί να διαφέρει μεταξύ των μελών μιας κλάσης). Στον ανατολικό κόσμο ο Bodhidharma έδειχνε στους μαθητές του αντικείμενα και ζητούσε να μάθει τι είναι, παρουσιάζοντας με αυτό τον τρόπο την ταυτότητα των αντικειμένων και τη φύση της ταξινόμησης και της απόφασης.

Το πόσο ευρύ είναι το πεδίο στο οποίο εκτείνονται οι εφαρμογές της Αναγνώρισης Προτύπων γίνεται κατανοητό από τον αριθμό των κλάδων της επιστήμης που είτε βασίζονται σε αυτή σε μεγάλο βαθμό είτε επεκτείνουν τις βάσεις της. Χαρακτηριστικά παραδείγματα κλάδων που ανήκουν στην πρώτη κατηγορία είναι :

- Η Όραση Υπολογιστών (Computer Vision) ασχολείται με την περιγραφή με μια ή περισσότερες εικόνες του κόσμου που βλέπουμε και επιδιώκει να ανακατασκευάσει ορισμένες ιδιότητες τους όπως το σχήμα, την υφή, τη φωτεινότητα, καθώς και το πως κατανέμεται το χρώμα. Η Όραση Υπολογιστών βρίσκει εφαρμογή σε επιστημονικά και ερευνητικά πεδία όπως η επεξεργασία ιατρικών εικόνων (Medical Imaging), η βέλτιστη αναγνώριση χαρακτήρων (Optical Character Recognition), η παρακολούθηση με τη χρήση καμερών (Camera Surveillance) κ.α.

- Η Αναγνώριση Φωνής (Speech Recognition) έχει ως απώτερο σκοπό τη μετατροπή ενός ηχητικού σήματος που περιέχει ομιλία σε κείμενο. Χωρίζεται σε διάφορες υποκατηγορίες ανάλογα με

το αν υπάρχει εξάρτηση από τον ομιλητή ή όχι (Speaker Dependent vs Speaker Independent) με το αν ο λόγος είναι συνεχής ή διακριτός (Continuous Speech vs Isolated-word) κ.α.

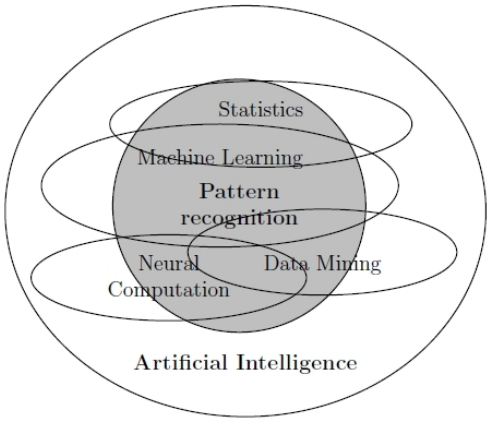
- Το Data Mining. Η μη αυτόματη επεξεργασία και ερμηνεία του τεράστιου αριθμού δεδομένων που υπάρχουν για ένα συγκεκριμένο πρόβλημα είναι μια αργή, υπολογιστικά ακριβή και υποκειμενική διαδικασία που καθιστά το χειρισμό τέτοιων δομών μη πρακτικό. Ο κλάδος του Data Mining στοχεύει στο να ανακαλύπτει πρότυπα σε μεγάλες βάσεις δεδομένων έτσι ώστε να εξάγονται αρχικά οι χρήσιμες πληροφορίες από αυτά τις οποίες στη συνέχεια μετατρέπει σε μια πιο εύχρηστη και κατανοητή στο χρήστη δομή.

Ενώ αντίστοιχα κλάδοι που ανήκουν στη δεύτερη κατηγορία είναι:

- Η Στατιστική η οποία είναι ο μαθηματικός τομέας που ασχολείται με τη συλλογή την οργάνωση και την ερμηνεία αριθμητικών δεδομένων.

- Η Μηχανική Εκμάθηση (Machine Learning) είναι η επιστήμη που έχει ως στόχο να μπορούν οι υπολογιστές να μάθουν και να δρουν ανάλογα χωρίς να είναι ρητά προγραμματισμένοι για μια τέτοια ενέργεια.

- Τα Νευρωνικά Δίκτυα (Neural Networks) τα οποία χρησιμοποιούνται για τη μοντελοποίηση σύνθετων σχέσεων μεταξύ εισόδων και εξόδων καθώς και για να βρουν πρότυπα σε δεδομένα. Βρίσκουν εφαρμογή σε αναγνώριση φωνής, αναγνώριση αντικειμένων σε ανάκτηση εικόνων κ.α.



Σχήμα 1: Αναγνώριση προτύπων και επιστημονικοί κλάδοι που σχετίζονται με αυτή

Τέλος πρέπει να επισημανθεί πως παρόλο που τεχνικές Αναγνώρισης Προτύπων έχουν εφαρμοστεί πρακτικώς σε κάθε επιστημονικό τομέα, η λήψη της κατάλληλης απόφασης και η ταξινόμηση των αντικειμένων με βάση αυτή, παραμένει ακόμα και σήμερα ένα ιδιαίτερα πολύπλοκο και απαιτητικό πρόβλημα. Τα κυριότερα προβλήματα που προσπαθεί να επιλύσει αποτελούν πρόκληση για περαιτέρω έρευνα και ανάπτυξη καινοτόμων προσεγγίσεων αποτελεσματικής αντιμετώπισης τους.

1.2 Ανάλυση Ομιλίας σε Ηχητικά και Οπτικά Σήματα

Κατά τη διάρκεια της δεκαετίας του 90 ο κλάδος της ανάλυσης πολυμεσικών δεδομένων (multimedia content analysis) κυριαρχείτο κυρίως από έρευνες που σχετίζονταν με την - βασισμένη στο περιεχόμενο - ανάκτηση πληροφοριών από εικόνες και βίντεο. Το κίνητρο πίσω από αυτές τις έρευνες βασιζόταν στο γεγονός ότι οι παραδοσιακές τεχνικές ανάκτησης που χρησιμοποιούσαν λέξεις κλειδιά, δεν ήταν πια εφαρμόσιμες σε εικόνες και βίντεο.

Ο κύριος λόγος για αυτό έχει να κάνει με το ότι το προαπαιτούμενο για να εφαρμοστούν τεχνικές αναζήτησης που να βασίζονται σε λέξεις κλειδιά, είναι η ύπαρξη μίας περιεκτικής περιγραφής του περιεχομένου για κάθε εικόνα ή βίντεο που είναι αποθηκευμένα στη βάση δεδομένων. Λαμβάνοντας υπόψη την ανάπτυξη τεχνικών Όρασης Υπολογιστών και Αναγνώρισης Προτύπων, συμπεραίνουμε πως είναι αδύνατο να μπορούν να εξαχθούν αυτόματα από υπολογιστές τέτοιες περιγραφές περιεχομένου.

Ένας ακόμη παράγοντας είναι πως το μη αυτόματο annotation περιεχομένων που εμπεριέχονται σε εικόνες ή βίντεο είναι υπερβολικά χρονοβόρο και ταυτόχρονα απαγορευτικό σε κόστος. Το γεγονός αυτό μας οδηγεί στην εφαρμογή του μόνο όταν κρίνεται απαραίτητο.

Τέλος λαμβάνοντας υπόψη ότι υπάρχουν πολλοί διαφορετικοί τρόποι για annotation της ίδιας εικόνας ή του βίντεο, γίνεται ξεκάθαρο πως όταν αυτό δεν γίνεται αυτόματα εισάγεται μεγάλη υποκειμενικότητα στο πρόβλημα, κάνοντας την αναζήτηση περιεχομένου με βάση τις λέξεις κλειδιά ακόμα πιο δύσκολη.

Απεναντίας, οι τεχνικές ανάκτησης που βασίζονται στο περιεχόμενο προσπαθούν να δίνουν τη δυνατότητα στο χρήστη να ανακτά τις επιθυμητές εικόνες ή τα βίντεο χρησιμοποιώντας τις ομοιότητες μεταξύ κάποιων χαμηλού επιπέδου χαρακτηριστικών όπως το χρώμα, η υφή, το σχήμα, η κίνηση κ.α. Η παραδοχή που γίνεται σε αυτές τις περιπτώσεις είναι ότι εικόνες που οπτικά φαίνονται όμοιες αποτελούνται από παρόμοια χαρακτηριστικά τα οποία μπορούν να ποσοτικοποιηθούν και να αξιολογηθούν από κατάλληλες μετρικές. Την τελευταία δεκαετία, έχουν γίνει μεγάλες προσπάθειες σε πολλά θεμελιώδη προβλήματα όπως τα εξαγόμενα χαρακτηριστικά, οι μετρικές ομοιότητας, η συνάφεια του feedback κ.α. Παρόλα αυτά η επιτυχία τέτοιου είδους συστημάτων είναι περιορισμένη κυρίως λόγω των χαμηλών επιδόσεων που παρουσιάζουν. Ένα χαρακτηριστικό παράδειγμα είναι αν δοκιμάσει κανείς να χρησιμοποιήσει σαν query ένα κόκκινο αυτοκίνητο, που θα έχει ως αποτέλεσμα οι περισσότερες εικόνες που θα επιστρέψουν τα συστήματα αυτά να περιέχουν αντικείμενα άσχετα με κόκκινα αυτοκίνητα. Ο λόγος που συμβαίνει αυτό είναι ότι υπάρχουν μεγάλα σημασιολογικά κενά μεταξύ των χαμηλού επιπέδου χαρακτηριστικών που χρησιμοποιούνται από τα συστήματα ανάκτησης που βασίζονται στο περιεχόμενο και των υψηλού επιπέδου σημασιολογιών που εμφανίζονται στις query εικόνες ή βίντεο. Παράλληλα θα πρέπει να επισημανθεί πως οι χρήστες τείνουν να κρίνουν την ομοιότητα μεταξύ δύο εικόνων βασιζόμενοι περισσότερο στη σημασιολογία και λιγότερο στην εμφάνιση ορισμένων χαρακτηριστικών σε αυτές όπως το χρώμα και η υφή. Συνεπώς ένα συμπέρασμα που μπορεί να εξαχθεί, είναι πως το κλειδί για την επιτυχία ενός συστήματος ανάκτησης πληροφοριών από μια εικόνα ή ένα βίντεο που βασίζεται στο περιεχόμενο, έγκειται στο σε τι βαθμό μπορεί να γεφυρωθεί ή να μειωθεί αυτό το σημασιολογικό κενό.

Ένας ευθύς αλλά αποτελεσματικός τρόπος για να γεφυρωθούν τα σημασιολογικά κενά είναι να εμβαθύνουμε την ανάλυση και την κατανόηση των περιεχομένων μιας εικόνας. Ενώ η κατανόηση των περιεχομένων γενικών εικόνων είναι ακόμα ανέφικτη, η αναγνώριση συγκεκριμένων κλάσεων αντικειμένων ή γεγονότων κάτω από συγκεκριμένες συνθήκες είναι εντός των δυνατοτήτων μας. Συμπερασματικά, πολλά ερευνητικά πεδία μπορούν να επωφεληθούν από την πολυμεσική ανάλυση που βασίζεται στο περιεχόμενο. Ο Giannakopoulos [24] περιγράφει κάποιες γενικές κατηγορίες τέτοιων εφαρμογών οι οποίες συνοψίζονται σε:

- Αναζήτηση και Ανάκτηση. Μεγάλες βάσεις πολυμεσικών δεδομένων ή συλλογές αρχείων είναι δυνατό να περιέχουν χιλιάδες πολυμεσικά αρχεία. Τέτοια παραδείγματα είναι βιβλιοθήκες από ταινίες και βίντεο, ψηφιακές μουσικές συλλογές ή αρχεία εικόνων. Είναι λοιπόν προφανές ότι η πρόσβαση και περιήγηση σε τέτοιες βάσεις αποτελεί ένα δύσκολο πρόβλημα.

- Ταξινόμηση(Classification). Αρκετές μέθοδοι έχουν επικεντρωθεί στην ταξινόμηση μίας εικόνας, ενός ηχητικού αρχείου ή ενός βίντεο σε προκαθορισμένες κατηγορίες. Για παράδειγμα στην ηχητική πληροφορία κάποιοι μέθοδοι ταξινόμησης σχετίζονται με την αναγνώριση του μουσικού είδους ενός τραγουδιού, με την αναγνώριση των μουσικών οργάνων σε ένα αρχείο ήχου κ.α. Αντίστοιχα για το βίντεο έχουν αναπτυχθεί μέθοδοι αναγνώρισης της τοποθεσίας στο background μιας εικόνας, ενός αθλητικού γεγονότος (αν πρόκειται για ποδόσφαιρο ή για μπάσκετ) για αναγνώριση συναισθήματος κ.α.

- Κατάτμηση (Segmentation). Είναι η διαδικασία εντοπισμού των τμημάτων ενός ηχητικού ή οπτικού σήματος τα οποία έχουν ακουστικά ή οπτικά ομοιογενές περιεχόμενο. Το κριτήριο ομοιογένειας εξαρτάται από το τι αποτέλεσμα επιδιώκουμε να εξάγουμε από μια τέτοια προσέγγιση. Οι μέθοδοι κατάτμησης σε βίντεο έχουν επικεντρωθεί στην αναγνώριση ενός shot (shot detection ή shot boundary detection). Το shot είναι ένα από τα πιο βασικά στοιχεία ενός βίντεο και συνεπώς η αναγνώριση τους μέσα σε αυτά είναι θεμελιώδης για την αποτελεσματική κατάτμηση τους. Στις περισσότερες περιπτώσεις, τα κριτήρια που χρησιμοποιούν οι μέθοδοι για αναγνώριση των shots βασίζονται στο χρώμα ή στην κίνηση μεταξύ διαδοχικών frames ενός βίντεο.

- Abstraction Στόχος της είναι η αναπαράσταση του πολυμεσικού περιεχομένου σε ένα πιο κλειστό τρόπο. Όσον αφορά το βίντεο στόχος είναι αρχικά να εξάγει στατικές εικόνες (ονομάζονται key frames) τα οποία αναπαριστούν ολόκληρο το περιεχόμενο του βίντεο και στη συνέχεια να δημιουργήσει ένα αντιπροσωπευτικό βίντεο μικρότερης διάρκειας (storyboard) . Στην πρώτη περίπτωση η διαδικασία ονομάζεται video summarization ενώ στη δεύτερη video skimming.

Συνοψίζοντας, οι παραδοσιακές προσεγγίσεις όπου οι άνθρωποι πρέπει να ανακαλύψουν τη βασική γνώση και να την κωδικοποιήσουν σε ένα σετ από προγραμματιστικούς κανόνες, εκτός από υπερβολικά δαπανηρές, είναι και σε μεγάλο βαθμό ανίκανες να αναλύσουν αποτελεσματικά το πολυμεσικό περιεχόμενο καθώς η γνώση που απαιτείται για την αναγνώριση υψηλού επιπέδου εννοιών ή γεγονότων μπορεί να είναι αρκετά περίπλοκη, ασαφής και δύσκολη να καθοριστεί. Με την ολοένα και αυξανόμενη πολυπλοκότητα και μεταβλητότητα των πολυμεσικών δεδομένων, οι τεχνικές εκμάθησης μηχανών έχουν γίνει το πιο ισχυρό εργαλείο τόσο για την ανάλυση των περιεχομένων όσο και για τη βελτίωση της τεχνητής νοημοσύνης που σχετίζεται με τέτοιου είδους αντικείμενα.

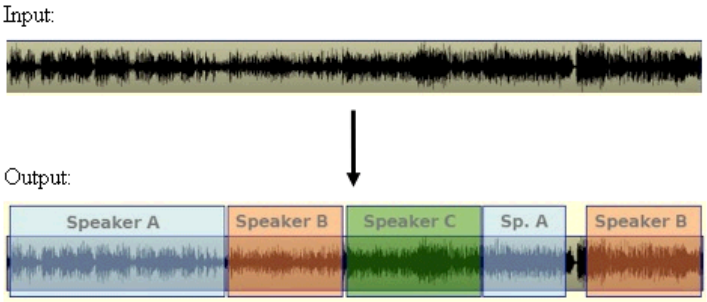
1.3 Το πρόβλημα της Ημερολογιοποίησης Ομιλητών σε Βίντεο

Τα τελευταία 10 χρόνια, τόσο το διαδίκτυο όσο και οι εφαρμογές του, έχουν αλλάξει και αναπτυχθεί σημαντικά, κυρίως λόγω της αύξησης των διαθέσιμων πόρων. Όσον αφορά τα πολυμέσα, η πιο εντυπωσιακή εξέλιξη είναι η συνεχώς αυξανόμενη χρήση ιστοσελίδων που προσφέρουν βίντεο (video sharing websites). Ταυτόχρονα όμως, δημιουργείται η ανάγκη για εφαρμογή αποτελεσματικής αναζήτησης, κατηγοριοποίησης (indexing) και πρόσβασης σε πληροφορίες σχετικές με αυτά τα αρχεία.

Παρόλο που η εξαγωγή των λέξεων που εμπεριέχονται σε ένα ηχητικό ή οπτικό σήμα, χρησιμοποιώντας τεχνολογίες αναγνώρισης φωνής, αποτελεί μια καλή βάση για την επίτευξη των παραπάνω στόχων, είναι σύνηθες τα αποτελέσματα που προκύπτουν να είναι δυσανάγνωστα και να μην έχουν αποσπάσει όλες τις πληροφορίες που εμπεριέχονταν στο σήμα. Συνεπώς, χρειάζονται επιπλέον τεχνολογίες έτσι ώστε να γίνει η κατάλληλη εξαγωγή των μεταδεδομένων (meta-data) εκείνων τα οποία θα εμπλουτίσουν τα αποτελέσματα μετατρέποντας τα σε πιο ευανάγνωστα ενώ παράλληλα θα παρέχουν περιεχόμενο και πληροφορίες πέρα από μια απλή αλληλουχία λέξεων. Χαρακτηριστικά παραδείγματα τέτοιων μεταδεδομένων είναι η εναλλαγή ομιλητών (speaker turns) ή τα σημεία στα οποία τελειώνει μια πρόταση και ξεκινάει μια άλλη (sentence boundaries).

Η κατάτμηση ενός βίντεο με βάση τους ομιλητές (speaker segmentation) καθώς και η ομαδοποίησή τους χωρίς επίβλεψη (speaker clustering) αποτελούν εργαλεία που διευκολύνουν το χειρισμό δεδομένων σε μεγάλα ηχητικά ή οπτικά αρχεία. Το speaker segmentation στοχεύει στο να διαχωρίσει ένα ηχητικό ή οπτικό σήμα σε ομοιογενή τμήματα έτσι ώστε κάθε τμήμα να περιέχει ιδανικά ένα ομιλητή. Η ομαδοποίηση των ομιλητών χωρίς επίβλεψη βασιζόμενη σε χαρακτηριστικά της φωνής του ομιλητή χρησιμοποιείται με στόχο την αναγνώριση όλων των τμημάτων ομιλίας που ανήκουν στον ίδιο ομιλητή και την απόδοση μιας μοναδικής ταμπέλας (label) σε αυτόν. Παρόλο που ο διαχωρισμός αυτός θα μπορούσε να πραγματοποιείται σε προηγούμενο στάδιο από την ομαδοποίηση, κάτι τέτοιο δε συνηθίζεται καθώς τα σφάλματα που θα προέκυπταν από το segmentation θα μείωναν την απόδοση του clustering. Εναλλακτικά οι 2 παραπάνω μέθοδοι μπορούν να βελτιστοποιηθούν από κοινού. Η κατάτμηση ενός βίντεο με βάση τους ομιλητές ακολουθούμενη από speaker clustering ονομάζεται Ημερολογιοποίηση Ομιλητών (Speaker Diarization).

Η Ημερολογιοποίηση Ομιλητών αποτελεί ένα ιδιαίτερα ενεργό πεδίο της Αναγνώρισης Προτύπων και έχει κερδίσει έντονο ερευνητικό ενδιαφέρον τα τελευταία χρόνια εξαιτίας της ολοένα και αυξανόμενης ανάγκης εξαγωγής και επεξεργασίας των πληροφοριών που περιέχονται στα πολυμέσα. Δοσμένου ενός ηχητικού (audio) ή/και ενός οπτικού (video) σήματος η Ημερολογιοποίηση Ομιλητών ορίζεται ως η διαδικασία της αυτόματης κατάτμησης ενός ομοιογενούς σήματος σε μικρότερες περιοχές με ομοιογένεια ως προς τον ομιλητή έχοντας ως στόχο να απαντήσει αυτόματα στην ερώτηση "Ποιος μίλησε και πότε". Μια τέτοια προσέγγιση περιλαμβάνει αναγνώριση ομιλίας ή όχι (Πότε υπάρχει ομιλία) σε συνδυασμό με ανίχνευση και ανάλυση των επικαλύψεων ομιλητών (Ποιος ομιλητής επικαλύπτει ποιόν).



Σχήμα 2: Παράδειγμα ενός συστήματος Ημερολογιοποίησης Ομιλητών.

Το να γνωρίζει κανείς πότε ένας ομιλητής μιλάει σε ένα ηχητικό σήμα ή σε ένα βίντεο αποτελεί ένα χρήσιμο δεδομένο προς επεξεργασία για πολλούς σκοπούς. Η Ημερολογιοποίηση Ομιλητών έχει χρησιμοποιηθεί για ανίχνευση πνευματικών δικαιωμάτων, καθώς και σε επιστημονικούς κλάδους που

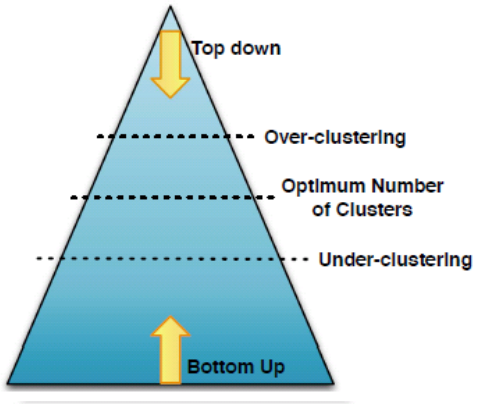
ασχολούνται με αυτόματη ανάλυση συμπεριφοράς. Παράλληλα είναι μια πολύ σημαντική διαδικασία για ανάκτηση πληροφοριών και έχει πολλές εφαρμογές σε επιστημονικά πεδία όπως η προσαρμογή των ομιλητών για αυτόματη ανίχνευση φωνής. Στον τομέα του εμπλουτισμού κειμένων που εξάγονται αυτόματα από ηχητικά ή οπτικά σήματα, η ημερολογιοποίηση ομιλητών μπορεί να χρησιμοποιηθεί τόσο σαν μια ανεξάρτητη εφαρμογή που αποδίδει περιοχές ομιλητών σε ένα ηχητικό ή οπτικό σήμα, όσο και σαν ένα στάδιο προ-επεξεργασίας για ανίχνευση φωνής.

Εν συνεχεία των παραπάνω εφαρμογών πρέπει να επισημανθεί πως η Ημερολογιοποίηση Ομιλητών βρίσκει μεγάλη εφαρμογή σε δελτία ειδήσεων, σε ηχογραφημένες τηλεφωνικές κλήσεις καθώς και σε συναντήσεις-συζητήσεις μεταξύ ανθρώπων σε ένα κλειστό χώρο (meetings). Τόσο στα δελτία ειδήσεων όσο και στα meetings συναντάμε σύνθετη αλληλεπίδραση μεταξύ των ομιλητών ο καθένας από τους οποίους παρουσιάζει μια διαφορετική συμπεριφορά. Επομένως οι ερευνητές από διάφορα ερευνητικά πεδία όπως η Συμπεριφοριστική Ψυχολογία, η Αλληλεπίδραση Ανθρώπου Υπολογιστή, η Όραση Υπολογιστών και η Επεξεργασία Σήματος έχουν εστιάσει την προσοχή τους στην ανάλυση τους. Λαμβάνοντας υπόψη ότι τα γεγονότα αυτά βιντεοσκοπούνται και ηχογραφούνται γίνεται ορατή η ανάγκη που έχει δημιουργηθεί για συστήματα τα οποία θα μπορούν αυτόματα να αναλύουν τέτοιου είδους γεγονότα.

Τα τελευταία χρόνια η επιστημονική κοινότητα έχει ασχοληθεί σε βάθος με το πρόβλημα της Ημερολογιοποίησης Ομιλητών με το ενδιαφέρον να επικεντρώνεται κυρίως σε ερευνητικά προγράμματα. Από την πρώιμη δουλειά με τηλεφωνικά δεδομένα, τα δελτία ειδήσεων έγιναν το επίκεντρο του ενδιαφέροντος στα τέλη της δεκαετίας του 90 και στις αρχές της επομένης καθώς η Ημερολογιοποίηση Ομιλητών εστίαζε την προσοχή της στο αυτόματο annotation τηλεοπτικών και ραδιοφωνικών εκπομπών. Στη συνέχεια υπήρξε μια ραγδαία ανάπτυξη του αντικειμένου που οδήγησε σε πολυμεσικές τεχνολογίες που αποσκοπούσαν στη βελτίωση της επικοινωνίας (από απόσταση) μεταξύ των ανθρώπων.

Η ανταπόκριση αυτών των πολυμεσικών τεχνολογιών θα πρέπει να ανταποκρίνεται στις ολοένα και αυξανόμενες απαιτήσεις και συνεπώς προτιμάται συχνά να χρησιμοποιούνται περισσότερες από μια πηγές από ροή δεδομένων (όπως η ακουστική πληροφορία, η οπτική, το κείμενο που μπορεί να υπάρχει σε μορφή διαφανειών σε μια παρουσίαση κ.α.) Με αυτό τον τρόπο γίνεται πιο εύκολη η αποσύνθεση του περιεχομένου ενός meeting σε πιο απλές μορφές ώστε να προσδιορισθεί όσο το δυνατόν καλύτερα η δομή του. Η Ημερολογιοποίηση Ομιλητών διαδραματίζει ένα σημαντικό ρόλο στην ανάλυση των δεδομένων αυτών, καθώς επιτρέπει σε περιεχόμενο τέτοιας μορφής να κατανεμηθεί σε εναλλαγές ομιλητών. Στη συνέχεια μπορούμε να προσθέσουμε σε αυτές γλωσσικό περιεχόμενο και άλλα μεταδεδομένα (όπως κυρίαρχους ομιλητές, το βαθμό της αλληλεπίδρασης μεταξύ των ομιλητών ή τα συναισθήματα).

Θα πρέπει να τονισθεί πως γίνεται επανειλημμένως αναφορά στο σενάριο ενός meeting το οποίο συχνά αναφέρεται και σαν "Πλήρες ως προς την Αναγνώριση Ομιλίας" καθώς αποτελεί ένα σενάριο όπου εμφανίζονται όλων των ειδών τα προβλήματα που μπορεί να εμφανιστούν σε οποιαδήποτε αναγνώριση ομιλίας. Τα meetings συνεπώς θέτουν ένα μεγάλο αριθμό από νέες προκλήσεις στην Ημερολογιοποίηση Ομιλητών οι οποίες δεν σχετίζονται άμεσα με παλαιότερες έρευνες. Τα περισσότερα state-of-the-art συστήματα Ημερολογιοποίησης Ομιλητών προσεγγίζουν το πρόβλημα με δυο διαφορετικούς τρόπους. Η πρώτη προσέγγιση είναι η bottom-up ενώ η δεύτερη η top-down και παρουσιάζονται στο παρακάτω σχήμα.



Σχήμα 3: Δύο Διαφορετικές Προσεγγίσεις των συστημάτων Ημερολογιοποίησης Ομιλητών(επανεκτύπωση από [2]

Η top-down που αποτελεί την πιο διαδεδομένη από τις δύο αρχικοποιείται με πολύ λίγες ομάδες(συνήθως μια) ενώ αντίθετα η bottom-up αρχικοποιείται με πολλές (συνήθως περισσότερες από τους ομιλητές που περιμένουμε να βρούμε). Ο στόχος και στις δύο περιπτώσεις είναι να υπάρξει σύγκλιση μετά από κάποιες επαναλήψεις στο βέλτιστο αριθμό προσθέτοντας ή ενώνοντας αντίστοιχα ομάδες. Αν ο τελικός αριθμός είναι μεγαλύτερος από το βέλτιστο τότε το σύστημα θεωρείται πως κάνει under-clustering ενώ όταν είναι χαμηλότερος κάνει over-clustering. Οι δύο αυτές προσεγγίσεις βασίζονται στα Κρυφά Μαρκοβιανά Μοντέλα (HMMs) όπου κάθε κατάσταση είναι ένα Gaussian Mixture Model (GMM) και αντιστοιχεί σε ένα ομιλητή.

Ο χωρισμός ενός ηχητικού σήματος σε ομογενή τμήματα αποτελεί βασικό στοιχείο της Ημερολογιοποίησης ομιλητών με βάση τη ηχητική πληροφορία. Η κλασσική προσέγγιση για την κατάτμηση αυτή, κάνει έλεγχο μιας υπόθεσης χρησιμοποιώντας τα ηχητικά σήματα σε δύο κυλιόμενα και πιθανότατα επικαλυπτόμενα, διαδοχικά παράθυρα. Για κάθε σημείο όπου υπάρχει περίπτωση αλλαγής υπάρχουν δυο υποθέσεις. Η πρώτη είναι ότι και τα δύο τμήματα προέρχονται από τον ίδιο ομιλητή και επομένως αναπαριστώνται από ένα μοντέλο ενώ η δεύτερη θεωρεί πως υπάρχουν δύο διαφορετικοί ομιλητές και κατ' επέκταση δύο διαφορετικά μοντέλα. Στην πραγματικότητα, η εκτίμηση των μοντέλων γίνεται σε καθένα από τα παράθυρα λόγου ενώ παράλληλα χρησιμοποιούμε κάποια κριτήρια ώστε να αποφασιστεί αν αντιπροσωπεύουν καλύτερα ένα ή δύο ξεχωριστά μοντέλα με τη βοήθεια κάποιων εμπειρικών ή δυναμικών προσαρμοζόμενων κατωφλίων. Τέλος, τα χαρακτηριστικά που επιλέγονται κατά κόρον στη βιβλιογραφία ([23]) για την ακουστική πληροφορία είναι τα Mel-frequency cepstrum coefficients (MFCCs). Το mel-frequency cepstrum (MFC) αποτελεί μια αναπαράσταση της ενέργειας βραχέως χρόνου του φάσματος ενός ήχου και βασίζεται σε ένα γραμμικό ημιτονοειδή μετασχηματισμό ενός φάσματος λογαριθμικής δύναμης μια μη γραμμική mel κλίμακα συχνότητας. Τα MFCCs είναι οι συντελεστές οι οποίοι από κοινού φτιάχνουν ένα MFC.

Στην οπτική πληροφορία αντίστοιχα, αντί για χρονικά μετακινούμενα παράθυρα επιλέγεται ως πρωταρχική μονάδα το shot που αποτελεί μια σκηνή μέσα σε ένα βίντεο στην οποία δεν παρουσιάζεται ιδιαίτερη κίνηση μεταξύ των καρέ και είναι τραβηγμένη από μια σταθερή σχετικά κάμερα. Όσον αφορά την επιλογή της μεθόδου εξαγωγής χαρακτηριστικών, μια συχνή επιλογή που χρησιμοποιούνται είναι η διαφορά μεταξύ δύο διαφορετικών καρέ ώστε να εντοπισθούν περιοχές όπου υπάρχει πιθανότητα να βρίσκεται ο ομιλητής. Μια άλλη συχνή μέθοδος είναι η χρήση του SIFT με στόχο την εξαγωγή χαρακτηριστικών στις περιοχές του προσώπου ενώ στη συνέχεια υπολογιζόταν η από κοινού πληροφορία

στη μέση ακουστική ενέργεια και στη διακύμανση των pixels σε κλίμακα γκρι. Οι Knox & Friedland [30] εξήγαγαν χαρακτηριστικά με τη χρήση της οπτικής ροής ώστε να εντοπίζουν πως μεταβάλλονται τα καρέ με το χρόνο.

Δεδομένου όμως, ότι οι άνθρωποι αντιλαμβάνονται τον κόσμο γύρω τους μέσα από πολλαπλές αισθήσεις με ένα συνδυαστικό τρόπο καθίσταται αναγκαίο για τα αυτόματα συστήματα που επιθυμούν να πετύχουν μια αντίστοιχη αντίληψη των ανθρώπινων αλληλεπιδράσεων την υιοθέτηση μιας πολυμεσικής προσέγγισης. Για παράδειγμα, είναι γνωστό πως ο λόγος και οι κινήσεις παράγονται με ένα συζευγμένο τρόπο για να εκφράσουν την ίδια σκέψη. Επομένως, συμπεραίνουμε πως ένα άτομο παρουσιάζει περισσότερη κίνηση όταν μιλάει σε σχέση με όταν ακούει. Τέτοιου είδους πληροφορίες μπορούν να χρησιμοποιηθούν για να βοηθήσουν την διαδικασία της Ημερολογιοποίησης Ομιλητών και αυτό ακριβώς εκμεταλλεύτηκαν οι Knox & Friedland λαμβάνοντας την οπτική ροή σε όλο το καρέ.

Συνοψίζοντας, η Ημερολογιοποίηση ομιλητών στοχεύει στο να δώσει απάντηση στο ερώτημα του "Ποιος μίλησε πότε" χωρίς να λαμβάνει υπόψη της πληροφορίες που έχουν εξαχθεί εκ των προτέρων για την ταυτότητα των ομιλητών και των αριθμό τους, ενώ παράλληλα δεν κάνει ταυτοποίηση των ομιλητών (speaker identification). Η παρούσα διπλωματική εργασία μελετά το πρόβλημα της Ημερολογιοποίησης Ομιλητών αποκλειστικά από τη σκοπιά του video. Οι κυριότερες συνεισφορές της στο εν λόγω πρόβλημα συνοψίζονται στα ακόλουθα σημεία:

- Παρέχεται μια λεπτομερής επισκόπηση μεθόδων αλλαγής shot σε ένα βίντεο και τονίζεται η καθοριστική συμβολή τους στην Ημερολογιοποίηση Ομιλητών με βάση την οπτική πληροφορία.

- Εξετάζεται σε βάθος η μέθοδος ανίχνευσης προσώπου των Viola & Jones και στη συνέχεια ενσωματώνουμε την ανίχνευση δέρματος ώστε να απορριφθούν περιπτώσεις λανθασμένης ανίχνευσης. Παράλληλα μελετήσαμε τρόπους και μεθόδους εξαγωγής χαρακτηριστικών απο το πρόσωπο ώστε να δοθεί μια περιεκτική σε πληροφορία αναπαράσταση του προσώπου.

- Παρουσιάζονται μέθοδοι μείωσης των διαστάσεων του αρχικού χώρου ενώ ταυτόχρονα μελετάμε τρόπους ομαδοποίησης των δεδομένων σε ομάδες κάθε μία από τις οποίες αντιστοιχεί σε ένα πρόσωπο - ομιλητή.

- Διεξάγονται πειράματα Ημερολογιοποίησης Προσώπου συγκρίνοντας τα αποτελέσματα για πληθώρα διαφορετικών συνδυασμών των παραμέτρων που χρησιμοποιούνται. Αρχικά επιλέγουμε τη βέλτιστη επιλογή παραμέτρων που θα μπορούσε να χρησιμοποιηθεί για την κατασκευή ενός συστήματος που επιθυμεί να δώσει απάντηση στο ερώτημα "Ποιο πρόσωπο εμφανίζεται και πότε".

- Για τις παραμέτρους αυτές βρίσκουμε τόσο ποια μέθοδος μείωσης των διαστάσεων βελτιώνει καλύτερα τον αρχικό χώρο ενώ παράλληλα παρουσιάζουμε διαγράμματα που συγκρίνουν τη συμπεριφορά του αρχικού χώρου σε σχέση με τον μειωμένο όταν αυξάνεται ο αριθμός ομιλητών στο βίντεο.

- Σχεδιάζεται και υλοποιείται μια μέθοδος ανίχνευσης κίνησης των χειλιών ώστε να γίνει η μετάβαση στην Ημερολογιοποίηση Ομιλητών. Επιπλέον πραγματοποιήσαμε μια πληθώρα πειραμάτων Ημερολογιοποίησης Ομιλητών εξάγουμε τις βέλτιστες επιλογές για την κατασκευή ενός μελλοντικού τελικού συστήματος και σχολιάζουμε τα αποτελέσματα.

1.4 Διάρθρωση της Διπλωματικής Εργασίας

Στο **Κεφάλαιο** 2 παρουσιάζουμε μεθόδους ανίχνευσης αλλαγής shot σε ένα βίντεο. Γίνεται εκτενής αναφορά στις τρεις βασικότερες που υπάρχουν στη βιβλιογραφία και εντοπίζονται τα πλεονεκτήματα και τα μειονεκτήματα τους. Αυτές οι μετρικές εξάγονται από δύο διαδοχικά καρέ και συνοψίζονται αρχικά στην απόλυτη διαφορά μεταξύ των pixels, στην απόλυτη διαφορά μεταξύ των ιστογραμμάτων και σε μια στατιστική μετρική που είναι ο λόγος πιθανότητας. Τέλος γίνεται αναφορά, τόσο στην αναγκαιότητα χρήσης τεχνικών ανίχνευσης αλλαγής shot στην Ημερολογιοποίηση Ομιλητών όσο και στο που αποσκοπεί η χρήση τους.

Στο **Κεφάλαιο** 3 αναφερόμαστε στην ανίχνευση και τον εντοπισμό του προσώπου μέσα σε ένα καρέ. Αφού γίνει επισήμανση των εφαρμογών που βρίσκει καθώς και κάποιων δυσκολιών που μπορεί να εμφανίζονται, στη συνέχεια εξετάζουμε τις κυριότερες μεθόδους ανίχνευσης και εντοπισμού ενός προσώπου. Έπειτα περιγράφουμε με λεπτομέρεια τον αλγόριθμο ανίχνευσης προσώπου των Viola & Jones μαζί με όλα τα στάδια του, δηλαδή την ολοκληρωτική εικόνα, τα Haar χαρακτηριστικά, τον αλγόριθμο AdaBoost και τους κατευθυνόμενους ακολουθιακούς ταξινομητές. Επιπρόσθετα για να βελτιωθεί η ακρίβεια των αποτελεσμάτων της αναγνώρισης προσώπου εισάγουμε και την ανίχνευση δέρματος ώστε να απορρίψουμε πιθανόν λανθασμένα αποτελέσματα. Στο τελευταίο μέρος του κεφαλαίου, περιγράφουμε την πορεία που ακολουθήσαμε συνδέοντας την ανίχνευση προσώπου με το πρόβλημα της Ημερολογιοποίησης Ομιλητών.

Στο **Κεφάλαιο** 4 αναπτύσσεται λεπτομερώς η διαδικασία της εξαγωγής των χαρακτηριστικών. Αρχικά επισημαίνουμε κάποιες εφαρμογές που βρίσκει σε διάφορους τομείς της Αναγνώρισης Προτύπων και στη συνέχεια κάνουμε διαχωρισμό μεταξύ της εξαγωγής και της επιλογής χαρακτηριστικών. Επιπροσθέτως, γίνεται αναφορά σε διάφορες μεθόδους εξαγωγής χαρακτηριστικών από το πρόσωπο οι οποίες συνοψίζονται στα Gabor Κυματίδια, στο μετασχηματισμό SIFT και τα SURF χαρακτηριστικά. Τέλος περιγράφουμε τη διαδικασία που ακολουθήσαμε για να εξάγουμε χαρακτηριστικά στο πρόβλημα που καλούμαστε να αντιμετωπίσουμε.

Στο **Κεφάλαιο** 5 το πρώτο αντικείμενο με το οποίο ασχολούμαστε είναι η μείωση των διαστάσεων των χαρακτηριστικών δίνοντας ταυτόχρονα τον ορισμό του Curse of Dimensionality. Οι μέθοδοι στις οποίες επικεντρωθήκαμε και παρουσιάζονται στο παρόν κεφάλαιο, είναι η Ανάλυση σε Κύριες Συνιστώσες (PCA), και η LDA. Επίσης γίνεται μια σύντομη αναφορά στη μέθοδο της Τυχαίας Προβολής που χρησιμοποιείται ως ένα προγενέστερο στάδιο της μείωσης διαστάσεων με LDA. Επιπλέον αναφερόμαστε στην FLsD μέθοδο μείωσης των διαστάσεων που χρησιμοποιήσαμε κάνοντας αναφορά στα πλεονεκτήματά της έναντι της PCA και της LDA και περιγράφουμε τη λειτουργία της. Μετέπειτα, παρατίθενται αρχικά κάποιες πληροφορίες με σκοπό να τονισθεί η ανάγκη ακριβούς ομαδοποίησης των δεδομένων και στη συνέχεια, παρουσιάζουμε δύο μεθόδους ομαδοποίησης των τελικών χαρακτηριστικών. Η πρώτη είναι ο αλγόριθμος k-means και η δεύτερη ένας fuzzy τρόπος ομαδοποίησης που βασίζεται σε μια βελτιωμένη έκδοση των Gustafson-Kessel (GK). Επίσης, αναλύουμε τη Silhouette μετρική που είναι ένας τρόπος αξιολόγησης του πόσο καλά έχει γίνει η ομαδοποίηση και τέλος περιγράφουμε τη διαδικασία απόδοσης μιας μοναδικής ετικέτας σε κάθε shot.

Στο **Κεφάλαιο** 6 χρησιμοποιούμε όλες τις μεθόδους που έχουν περιγραφεί στα παραπάνω κεφάλαια και πραγματοποιούμε το πρώτο σετ πειραμάτων που σχετίζονται με την Ημερολογιοποίηση Προσώπων σε ένα dataset στο οποίο πραγματοποιήσαμε annotation με βάση την οπτική πληροφορία. Για την εκτέλεση των πειραμάτων λαμβάνουμε όλους τους δυνατούς συνδυασμούς μεθόδων εξαγωγής χαρακτηριστικών, μείωσης των διαστάσεων, ομαδοποίησης καθώς επίσης και αν είναι επίσης γνωστός εκ των προτέρων ο αριθμός των προσώπων-ομιλητών που εμφανίζονται στο βίντεο. Τέλος αξιολογούμε και ερμηνεύουμε τα αποτελέσματα που προέκυψαν και εντοπίζουμε εκείνο των συνδυασμό

μεθόδων που παρουσιάζει τα καλύτερα αποτελέσματα.

Στο **Κεφάλαιο** 7 γίνεται το πέρασμα από την Ημερολογιοποίηση Προσώπων στην Ημερολογιοποίηση Ομιλητών που είναι και το αρχικό μας πρόβλημα. Αυτό γίνεται ανιχνεύοντας αρχικά την περιοχή των χειλιών και στη συνέχεια εισάγουμε μια μετρική κίνησης τους ώστε να αποσαφηνιστεί αν το εικονιζόμενο πρόσωπο σε κάθε καρέ ομιλεί ή όχι. Παράλληλα πραγματοποιήσαμε το δεύτερο σετ πειραμάτων που αφορά την Ημερολογιοποίηση Ομιλητών και πάλι για όλες τις μεθόδους και αξιολογούμε τα αποτελέσματα που προέκυψαν. Επιπλέον εξάγουμε τα παραπάνω αποτελέσματα σε δύο ξεχωριστές περιπτώσεις ανάλογα με το αν θέλουμε να δώσουμε απάντηση για όλο το βίντεο ή μόνο για τα σημεία για τα οποία είμαστε σίγουροι.

Στο **Κεφάλαιο** 8 πραγματοποιείται μια αποτίμηση των συμπερασμάτων που εξάγονται από την παρούσα διπλωματική εργασία και παρουσιάζονται κάποιες προτάσεις και επεκτάσεις για μελλοντική έρευνα.

Στο **Παράρτημα Α** γίνεται μια συνοπτική επισήμανση τεχνικών Αναγνώρισης Προσώπου από σε ένα σετ δεδομένων. Οι τεχνικές που χρησιμοποιούνται είναι οι μετασχηματισμοί SIFT και SURF, τα Eigenfaces και τα Fisherfaces. Η επισήμανση αυτών των μεθόδων πραγματοποιείται στο συγκεκριμένο παράρτημα και όχι στο κυρίως κείμενο της διπλωματικής εργασίας καθώς σε αντίθεση με την Ανίχνευση Προσώπου, η Αναγνώριση Προσώπου δεν είναι άμεσα συσχετισμένη με την Ημερολογιοποίηση Ομιλητών.

Στο **Παράρτημα Β** περιγράφουμε με συντομία το dataset του Canal 9 με τη χρήση του οποίου εκτελέσθηκαν τα δύο σετ πειραμάτων Ημερολογιοποίησης Προσώπου και Ομιλητών.

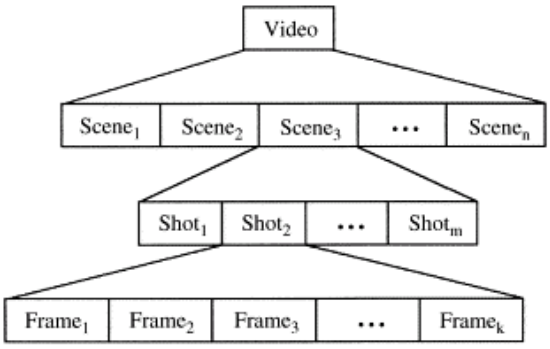
Κεφάλαιο 2

Ανίχνευση Αλλαγής Shot

2.1 Εισαγωγή

Στη βιβλιογραφία έχουν προταθεί αρκετές μέθοδοι οι οποίες στοχεύουν στο να συμπεριληφθεί ψηφιακό περιεχόμενο από βίντεο σε μια - προσβάσιμη από ένα δίκτυο δεδομένων - ψηφιακή βιβλιοθήκη. Το interface της πρόσβασης σε τέτοιου είδους περιεχόμενο θα πρέπει να επιτρέπει την περιήγηση και την αναζήτηση μέσα στο βίντεο. Δεδομένης της προσωρινά γραμμικής και ιδιαίτερα πυκνής σε δεδομένα, φύσης ενός ψηφιακού βίντεο, σε συνδυασμό με τους περιορισμούς που υφίστανται γύρω από το εύρος ζώνης του δικτύου, οδηγούμαστε στο συμπέρασμα πως η κατάτμηση ενός βίντεο που είναι αποθηκευμένο σε μια βάση δεδομένων αποτελεί το αναγκαίο πρώτο βήμα για τη δημιουργία του συγκεκριμένου interface. Η ανίχνευση των σημείων σε ένα βίντεο όπου συμβαίνει αλλαγή shot (Shot-change detection) είναι η διαδικασία εντοπισμού των αλλαγών στο περιεχόμενο του σκηνικού ενός βίντεο με απώτερο σκοπό την εξαγωγή εναλλακτικών αναπαραστάσεων του ίδιου βίντεο ώστε να διευκολύνεται η περιήγηση, η ανάκτηση κ.α. Επίσης υπάρχει η δυνατότητα να αναπαρασταθεί κάθε shot από ένα χαρακτηριστικό καρέ δημιουργώντας μια περιληπτική αναπαράσταση (storyboard) ενός βίντεο ή μιας ταινίας. Η διαδικασία αυτή μπορεί να εκτελεστεί σε βίντεο που είτε προβάλλονται ζωντανά είτε είναι αποθηκευμένα σε κάποια βάση δεδομένων δημιουργώντας μια αναπαράσταση που μπορεί να χρησιμοποιηθεί τόσο για τη δημιουργία μιας ενδεικτικής περίληψης ενός βίντεο όπως αναφέρθηκε παραπάνω όσο και για την γρήγορη εύρεση συγκεκριμένων σκηνών ενός βίντεο που μπορεί να θέλει να δει ο χρήστης.

Οι Gargi et al.[21] ορίζουν ένα shot ως μια ακολουθία από καρέ τα οποία είναι (ή φαίνεται να είναι) τραβηγμένα σε συνεχή χρόνο από την ίδια κάμερα. Ιδανικά, ένα shot μπορεί να εμπεριέχει περιστροφές στον οριζόντιο ή τον κάθετο άξονα μιας σταθερής κάμερας (panning - tilting) καθώς και περιπτώσεις όπου η κάμερα κάνει zooming.



Σχήμα 4: Η ιεραρχική δομή ενός βίντεο (επανεκτύπωση από το [31])

Στην πραγματικότητα όμως, οι αλγόριθμοι που ανιχνεύουν την αλλαγή shots σε βίντεο μπορεί να αντιδρούν σε σημαντικές μετακινήσεις τόσο αντικειμένων μέσα στην περιοχή λήψης της κάμερας, όσο και της ίδιας της κάμερας. Αλλαγές shot μπορεί να εμφανιστούν σε ένα βίντεο όταν έχουμε cuts, όταν ένα καρέ από ένα shot ακολουθείται από ένα καρέ ενός διαφορετικού shot καθώς και σε σταδιακές μεταβάσεις από μια εικόνα σε μια άλλη, υποκατηγορία των οποίων είναι τεχνικές όπως τα fade-ins, fade-outs. Ένα χαρακτηριστικό παράδειγμα αλλαγής σκηνής που χρησιμοποιείται κατά

κόρον από σκηνοθέτες κινηματογραφικών ταινιών, είναι η μετάβαση από fade-out σε μαύρο χρώμα ακολουθούμενη από fade-in για να υποδηλωθεί το πέρασμα του χρόνου ή την αλλαγή τοποθεσίας.

2.2 Μετρικές διαφοράς για Κατάτμηση του βίντεο

Η ανίχνευση των μεταβάσεων από μια σκηνή σε μια άλλη περιλαμβάνει την ποσοτικοποίηση της διαφοράς μεταξύ δύο διαδοχικών καρέ σε ένα βίντεο. Για να επιτευχθεί αυτό, πρέπει πρωτίστως να ορισθεί μια κατάλληλη μετρική αυτής της διαφοράς έτσι ώστε όταν ξεπερνάει ένα κατώφλι να καταγράφεται ο αριθμός του καρέ στο οποίο πραγματοποιείται αυτή η μετάβαση.

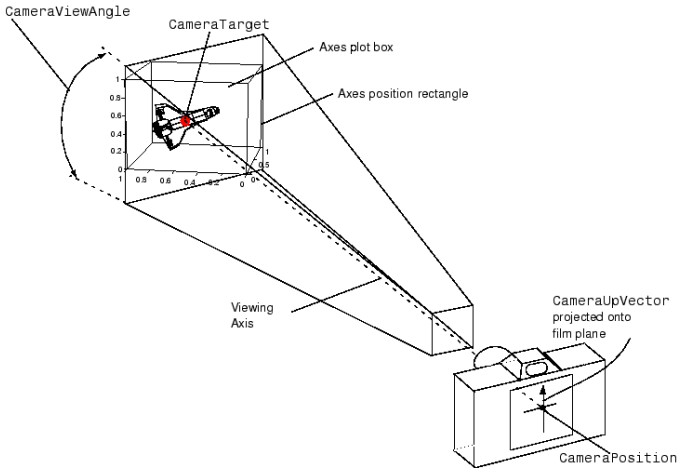
2.2.1 Διαφορά μεταξύ Ιστογραμμάτων

Μια μέθοδος που χρησιμοποιείται σε μεγάλο βαθμό στη βιβλιογραφία ([26],[37]) είναι η χρήση ιστογραμμάτων σε έγχρωμες εικόνες όπου υπολογίζεται η διαφορά μεταξύ των ιστογραμμάτων δύο διαδοχικών καρέ. Όταν η διαφορά αυτή λάβει κάποια μεγάλη τιμή τότε μαρκάρεται το συγκεκριμένο καρέ ως υποψήφιο για αλλαγή σκηνής. Η σύγκριση των δύο ιστογραμμάτων μπορεί να γίνει με πολλούς τρόπους όπως η απόλυτη διαφορά μεταξύ των αντίστοιχων bins (bin-to-bin histogram), η προσθήκη βαρών ανάλογα με το χρώμα και ο υπολογισμός της απόλυτης διαφοράς (δίνονται διαφορετικά βάρη στις κόκκινες πράσινες και μπλε διαφορές ιστογραμμάτων) και η χρήση της τομής των δύο ιστογραμμάτων η οποία δίνεται από τον παρακάτω ορισμό:

$$K_{\cap}(a,b) = \frac{\sum_{i=1}^{N} min(a_i, b_i)}{N}$$

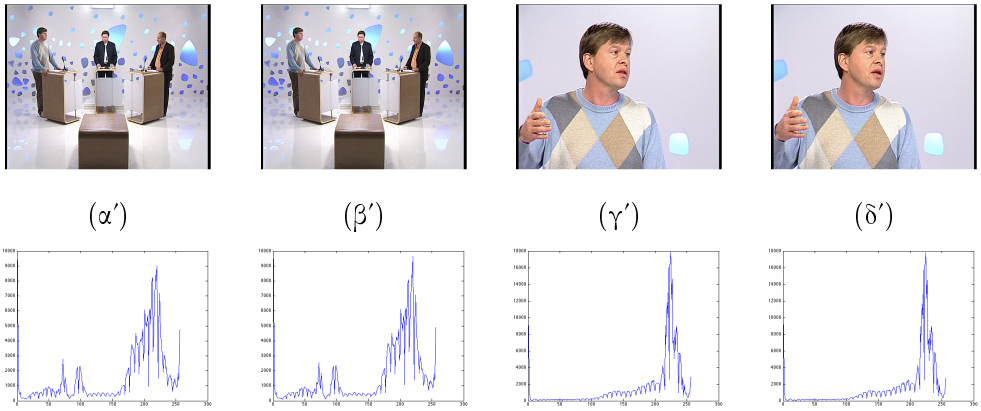
όπου a και b τα δύο ιστογράμματα με N bins το καθένα.

Οι Swain & Ballard [44], σημειώνουν πως η ιδιότητα των ιστογραμμάτων να παραμένουν αμετάβλητα σε μετακινήσεις η περιστροφές γύρω από τον άξονα που συνδέει το κέντρο του φακού της κάμερας με το κέντρο του αντικειμένου(viewing axis) καθώς και το ότι μεταβάλλονται ελάχιστα όταν αλλάζει η γωνία λήψης και η κλίση, είναι οι βασικοί παράγοντες για τους οποίους επιλέγονται σαν μετρική της διαφοράς μεταξύ δύο διαδοχικών καρέ. Για καλύτερη κατανόηση των προαναφερθέντων όρων δίνεται το παρακάτω σχήμα:



Σχήμα 5: Κάμερα και άξονες

Η ιδέα πίσω από αυτό τον αλγόριθμο είναι ότι δύο καρέ τα οποία έχουν σταθερό φόντο και αντικείμενα που δε μετακινούνται θα παρουσιάσουν μικρή διαφορά στα αντίστοιχα ιστογράμματα τους. Παράλληλα, η διαφορά ιστογραμμάτων δεν είναι ευαίσθητη σε μετακινήσεις αντικειμένων μέσα στην εικόνα, επειδή αγνοεί τις χωρικές αλλαγές που συμβαίνουν μέσα σε ένα καρέ. Από την άλλη όμως, υπάρχει το ενδεχόμενο δύο εικόνες να έχουν ίδια ιστογράμματα αλλά εντελώς διαφορετικό περιεχόμενο, γεγονός που θα οδηγήσει τον αλγόριθμο σε εσφαλμένα συμπεράσματα. Παρόλα αυτά, η πιθανότητα εμφάνισης τέτοιων περιστατικών σε ένα βίντεο είναι αρκετά σπάνια στην πράξη και συνεπώς δεν επηρεάζει την απόδοση του.



(α′) (β′) (γ′) (δ′)

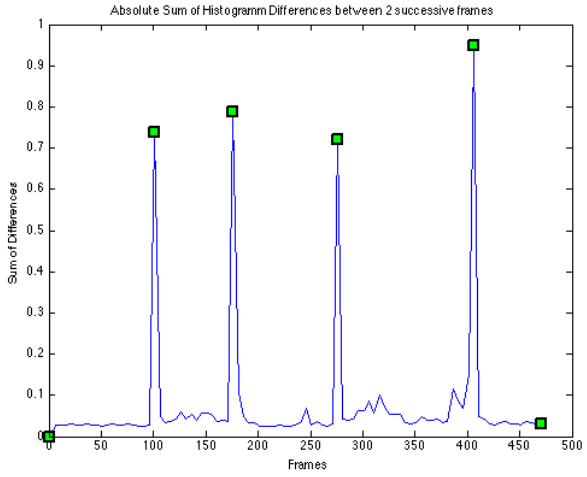
(ε′) (ϝ′) (ζ′) (η′)

Σχήμα 6: Διαδοχικά καρέ ενός βίντεο (α-δ) και τα αντίστοιχα ιστογράμματά τους (ε-η)

Οι Zhang et al. [52] ορίζουν $H_i(j)$ τις τιμές που επιστρέφει το ιστόγραμμα του i καρέ, όπου j είναι ένα από τα G πιθανά γκρι επίπεδα. (Ο αριθμός των bins των ιστογραμμάτων μπορεί να επιλεχθεί ανάλογά με τον επιθυμητό χρόνο υπολογισμού). Έτσι η διαφορά μεταξύ δυο διαδοχικών καρέ, αφού την κανονικοποιήσουμε ώστε να παίρνει τιμές από μηδέν έως ένα διαιρώντας με τις διαστάσεις κάθε καρέ M,N δίνεται από τον παρακάτω τύπο:

$$SD_i = \frac{\sum_{j=1}^{G} (|H_i(j) - H_{i+1}(j)|)}{M * N}$$

Αν η συνολική διαφορά SD_i είναι μεγαλύτερη από ένα κατώφλι τότε ορίζεται σε εκείνο το σημείο ένα σύνορο shot. Για το παράδειγμα των παραπάνω εικόνων (frames 99-102 του βίντεο) επιλέγοντας 256 bins για γκρι εικόνες και 2 για δυαδικές, ο παραπάνω τύπος εξήγαγε το παρακάτω σχήμα από το οποίο είναι ξεκάθαρο σε ποια ακριβώς καρέ του βίντεο έχουμε αλλαγή σκηνής όπου με πράσινο παρουσιάζεται η αρχή και το τέλος κάθε shot.



Σχήμα 7: Άθροισμα των απόλυτων διαφορών μεταξύ των αντίστοιχων bins μεταξύ δύο διαδοχικών ιστογραμμάτων

2.2.2 Διαφορά μεταξύ των pixels

Μια δεύτερη και πολύ απλή μέθοδος ανίχνευσης των συνόρων των shots σε ένα βίντεο είναι η σύγκριση των αντίστοιχων pixels σε δύο διαδοχικά καρέ με στόχο των εντοπισμό εκείνων τα οποία έχουν αλλάξει. Αυτή η προσέγγιση είναι γνωστή σαν διαφορά των pixels (pair-wise comparison). Στην απλούστερη περίπτωση των μονοχρωματικών εικόνων, ένα pixel θεωρείται ότι έχει αλλάξει αν η διαφορά μεταξύ των τιμών των εντάσεων σε δύο καρέ ξεπερνά ένα δοσμένο κατώφλι. Αυτή η μετρική μπορεί να παρουσιαστεί με τη μορφή μιας δυαδικής συνάρτησης $\Delta P_i(k,l)$ όπου (k,l) οι συντεταγμένες του pixel, i ο δείκτης που υποδηλώνει τον αριθμό του καρέ το οποίο συγκρίνεται με το επόμενο του για τον υπολογισμό της διαφοράς και $P_i(k,l)$ η τιμή της έντασης στο αντίστοιχο pixel. Συνεπώς δοσμένου ενός κατωφλίου T_1 η δυαδική συνάρτηση $\Delta P_i(k,l)$ ορίζεται ως εξής:

$$\Delta P_i(k,l) = \begin{cases} 1 & \text{αν } |P_i(k,l) - P_{i+1}(k,l)| \ > \ T_1 \\ 0 & \text{αλλιώς} \end{cases}$$

Με τη βοήθεια αυτής της μετρικής θεωρείται ότι υπάρχει ένα όριο ενός shot αν ο συνολικός αριθμός των pixels που έχει αλλάξει ξεπερνά ένα κατώφλι T_2. Επίσης για να μην εμφανίζονται πολύ μεγάλες τιμές και για να υπάρχει και ένα μέτρο σύγκρισης κανονικοποιείται και το αποτέλεσμα διαιρώντας με

το συνολικό αριθμό των pixels σε κάθε καρέ που για καρέ διαστάσεων M,N θα είναι $M \times N$. Ο τύπος της μετρικής της σύγκρισης μεταξύ των αντίστοιχων pixels σε δύο διαδοχικά καρέ δίνεται συνεπώς από την παρακάτω ανισότητα:

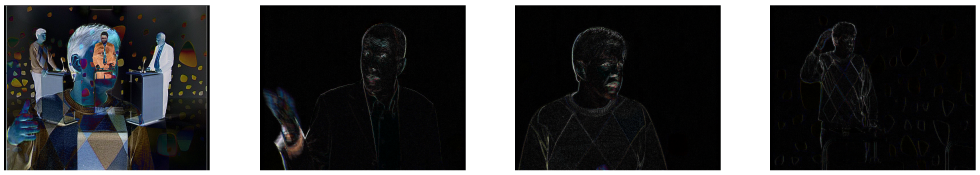
$$\frac{\sum_{k,l=1}^{M,N} \Delta P_i(k,l)}{M * N} * 100 \ > \ T_2$$

Η ίδια λογική μπορεί να επεκταθεί και σε έγχρωμες εικόνες υπολογίζοντας κάθε φορά τη μέση τιμή σε κάθε ένα από τα τρία κανάλια:

$$\frac{\sum_{k=1}^{M} \sum_{l=1}^{N} \sum_{dims=1}^{3} P_i(k,l)}{3 * M * N} \ > \ T_3$$

Ένα ουσιαστικό πρόβλημα που παρουσιάζει αυτή η μετρική είναι η ευαισθησία της στις κινήσεις τις κάμερας. Για παράδειγμα, όταν η κάμερα μετακινηθεί ελαφρά οριζόντια (camera panning) χωρίς όμως να αλλάξει το περιεχόμενο που τραβάει θα οδηγήσει στην αλλαγή ενός μεγάλου αριθμού από αντικείμενα τα οποία θα μετακινηθούν κατά την ίδια κατεύθυνση στα διαδοχικά καρέ που καλύπτουν τη μετακίνηση αυτή. Σε τέτοιες περιπτώσεις, η προσέγγιση αυτή θα οδηγήσει σε εσφαλμένα συμπεράσματα καθώς θα θεωρήσει πως τα περισσότερα pixels έχουν αλλάξει τιμή ενώ στην πραγματικότητα η κάμερα μετακινείται κατά λίγα pixels. Το φαινόμενο αυτό μπορεί να μειωθεί με τη χρήση φίλτρων εξομάλυνσης - αντικαθιστώντας δηλαδή πριν τη σύγκριση κάθε pixel με τη μέση τιμή των γειτονικών του απαλείφοντας ταυτόχρονα και θόρυβο από τις εικόνες.

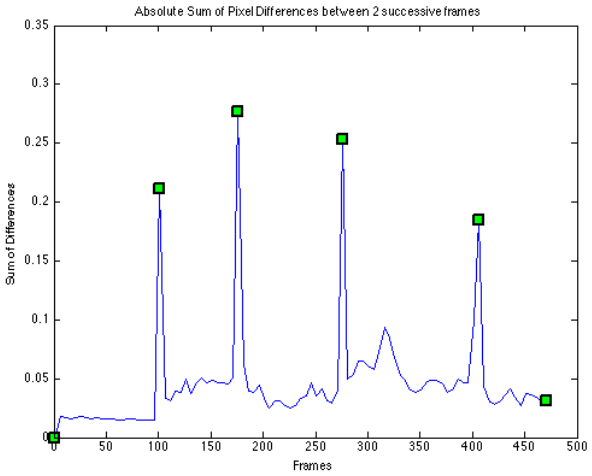
Για τα ίδια καρέ του βίντεο με πριν η απόλυτη τιμή της διαφοράς των καρέ στα σημεία όπου έχουμε αλλαγή shot θα είναι:



(α′) (β′) (γ′) (δ′)

Σχήμα 8: Απόλυτη τιμή της διαφοράς των καρέ στα σημεία όπου έχουμε αλλαγή shot

Στα σημεία όπου δεν λαμβάνει χώρα μια τέτοια αλλαγή οι αντίστοιχες εικόνες τις διαφοράς επιστρέφουν μαύρη εικόνα καθώς η διαφορά τους είναι περίπου ίση με μηδέν σε κάθε pixel. Η αντίστοιχη γραφική παράσταση για τα σύνορα των shots του βίντεο δίνεται παρακάτω



Σχήμα 9: Pair-wise comparison μεταξύ των pixels για ανίχνευση των συνόρων των shots σε ένα βίντεο

Συγκρίνοντας τα αποτελέσματα των δύο παραπάνω γραφικών συναρτήσεων φαίνεται πως η μέθοδος της διαφοράς ιστογραμμάτων αντιμετωπίζει καλύτερα τις διαφορές μεταξύ δύο καρέ που ανήκουν στο ίδιο shot αφού δεν επηρεάζεται από μικρές μεταβολές στην κίνηση που μπορεί να παρουσιάζονται. Παράλληλα στον παρακάτω πίνακα παρουσιάζεται ο χρόνος υπολογισμού των δύο αυτών μετρικών από όπου φαίνεται πως η μέθοδος διαφοράς ιστογραμμάτων είναι ταχύτερη κατά 5ms το καρέ που σημαίνει πως σε βίντεο διάρκειας μίας ώρας με 25 καρέ ανά δευτερόλεπτο εξοικονομείται υπολογιστικός χρόνος ίσος με 7.5 λεπτά.

Μετρική Διαφοράς	Χρόνος Υπολογισμού μιας Διαφοράς
Διαφορά Ιστογραμμάτων	35ms
Διαφορά μεταξύ pixels	40ms

Πίνακας 1: Μετρικές και αντίστοιχοι χρόνοι υπολογισμού για ανίχνευση αλλαγής shot

2.2.3 Λόγος Πιθανότητας

Για να γίνει η ανίχνευση των συνόρων πιο εύρωστη, αντί να πραγματοποιείται σύγκριση μεταξύ σκέτων pixels, μπορεί να γίνεται σύγκριση μεταξύ αντίστοιχων περιοχών σε δύο διαδοχικά καρέ χρησιμοποιώντας στατιστικές μετρικές δεύτερης τάξης των τιμών έντασης που εμφανίζουν. Οι Kasturi & Jain [29] αναφέρουν πως μια τέτοια μετρική για τη σύγκριση των αντίστοιχων περιοχών ονομάζεται λόγος πιθανότητας (likelihood ratio). Ας είναι, m_i και m_{i+1} οι μέσες τιμές έντασης για μια δοσμένη περιοχή σε δύο διαδοχικά καρέ, ενώ S_i και S_{i+1} οι αντίστοιχες διακυμάνσεις. Για δοσμένο κατώφλι t η παρακάτω σχέση υπολογίζει το λόγο πιθανότητας και ελέγχει αν περνάει ή όχι το κατώφλι:

$$\frac{(\frac{S_i+S_{i+1}}{2} + (\frac{m_i-m_{i+1}}{2})^2)^2}{S_i * S_{i+1}} > t$$

Τα σύνορα τώρα μπορούν να ανιχνευθούν αφού πρώτα χωρίσουμε το καρέ σε ένα σετ από περιοχές. Στη συνέχεια θεωρείται ότι υπάρχει σύνορο σε ένα shot όταν ο συνολικός αριθμός των περιοχών των οποίων το likelihood ratio ξεπερνά ένα κατώφλι είναι ικανοποιητικά μεγάλος (όπου το "ικανοποιητικά μεγάλος αριθμός" θα εξαρτάται από το πως έχει χωριστεί το καρέ σε υποπεριοχές). Το πλεονέκτημα που παρουσιάζουν τα blocks σε σχέση με τα pixels είναι πως ο λόγος πιθανότητας αυξάνει το επίπεδο ανοχής σε αργές και μικρές μετακινήσεις αντικειμένων μεταξύ των καρέ. Αυτή η αυξημένη ανοχή μειώνει την πιθανότητα αυτές οι μικρές μετακινήσεις να θεωρηθούν σαν αλλαγές shot.

Ένα πιθανό πρόβλημα του λόγου πιθανότητας σαν μετρική διαφοράς μεταξύ δύο διαδοχικών καρέ είναι πως υπάρχει περίπτωση δύο περιοχές που συγκρίνονται να έχουν ίδια μέση τιμή και διακύμανση αλλά τελείως διαφορετικές συναρτήσεις πυκνότητας πιθανότητας και συνεπώς να μην ανιχνευθεί καμία αλλαγή. Παρόλα αυτά κάτι τέτοιο είναι ιδιαίτερα σπάνιο και δεν είναι από μόνο του δυνατό να επηρεάσει την απόδοση της μετρικής.

2.3 Ανίχνευση Αλλαγής shot και Ημερολογιοποίηση Ομιλητών

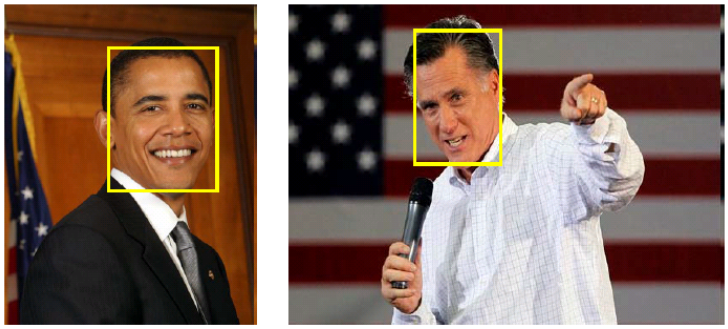
Ο χωρισμός ενός βίντεο σε επιμέρους shots αποτελεί ένα από τα βασικότερα και θεμελιώδη βήματα για την Ημερολογιοποίηση Ομιλητών που βασίζεται σε οπτικό περιεχόμενο που εξάγεται από βίντεο. Χωρίζοντας ένα βίντεο σε επιμέρους τμήματα το καθένα από τα οποία έχει την ιδιότητα ότι καλύπτεται από μια σχετικά σταθερή κάμερα μας δίνει τη δυνατότητα να θεωρήσουμε πως αν υπάρχει ένα πρόσωπο σε κάποιο καρέ ενός shot τότε θα είναι ίδιο για όλο το shot. Αντίστοιχα αν η κάμερα καλύπτει ένα γενικό πλάνο που εμπεριέχει όλα τα πρόσωπα - ομιλητές τότε αυτοί θα εμφανίζονται καθ' όλη τη διάρκεια του. Το παραπάνω σκεπτικό βασίζεται στο γεγονός πως αν ένας ομιλητής μιλάει σε ένα βίντεο στο τρέχων δευτερόλεπτο (και σε καθένα από τα 25 καρέ μέσα σε αυτό κατ' επέκταση), το πιθανότερο είναι να μιλάει και στο επόμενο δευτερόλεπτο. Επομένως χωρίζοντας ένα βίντεο σε επιμέρους τμήματα μέσα στα οποία οι αλλαγές που λαμβάνουν χώρα είναι από ασήμαντες μέχρι πολύ μικρές επιτυγχάνεται η επέκταση της παραπάνω παραδοχής, με τη διαφορά όμως ότι τώρα ο ομιλητής είναι σταθερός.

Όπως αναφέρεται και στη συνέχεια στο Κεφάλαιο 5 ο τελικός μας στόχος είναι να εξαχθούν κάποιες ταμπέλες (labels) οι οποίες θα λαμβάνουν ίδια τιμή για κάθε δευτερόλεπτο μέσα στο τρέχων shot. Με άλλα λόγια αν έχουμε ένα shot διάρκειας πέντε δευτερολέπτων στο οποίο μετά από κάποια επεξεργασία έχουμε αποδώσει μια ταμπέλα (έστω 1) στο πρόσωπο - ομιλητή που εμφανίζεται σε αυτό, τότε θα πρέπει να επιστραφεί ένας πίνακας διαστάσεων 1×5 με την τιμή 1. Γίνεται κατανοητό λοιπόν, πόσο ουσιώδης είναι ο χωρισμός σε επιμέρους shots καθώς επιτρέπει στην Ημερολογιοποίηση Ομιλητών να θεωρήσει ομοιογένεια ομιλητή για κάθε δευτερόλεπτο μέσα σε καθένα από αυτά.

Κεφάλαιο 3

Ανίχνευση Προσώπου και Δέρματος

3.1 Ανίχνευση Προσώπου - Εισαγωγή

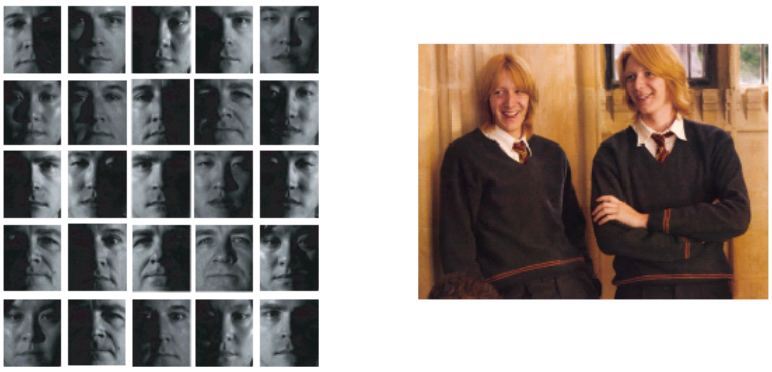


Σχήμα 10: Υπάρχουν καθόλου πρόσωπα στην εικόνα ; Αν ναι ποιοί είναι ;

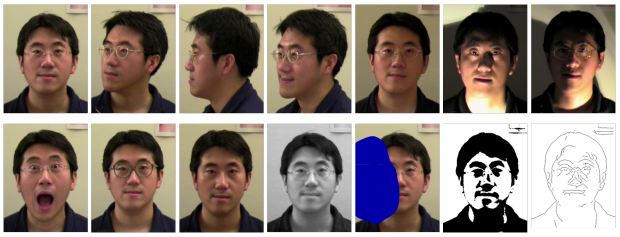
Στόχος της Ανίχνευσης Προσώπου είναι η αναγνώριση αρχικά της ύπαρξης ή μη ενός προσώπου και στη συνέχεια ο εντοπισμός της θέσης στην οποία βρίσκεται μέσα σε μια εικόνα ανεξάρτητα από παράγοντες όπως η θέση, η κλίμακα, η περιστροφή εντός της εικόνας, ο προσανατολισμός, η πόζα και ο φωτισμός.

Η ανίχνευση του προσώπου σε μια εικόνα είναι σημαντική καθώς αποτελεί το πρώτο βήμα για κάθε πλήρες αυτόματο σύστημα αναγνώρισης προσώπου. Χρησιμοποιείται σε πληθώρα εφαρμογών όπως σε συστήματα παρακολούθησης, σε συστήματα που επιτρέπουν την πρόσβαση σε χώρους, για παρακολούθηση προσώπου σε βίντεο (face tracking) κ.α.

Ταυτόχρονα όμως, η ανίχνευση προσώπου είναι μια αρκετά δύσκολη διαδικασία καθώς υπάρχουν πολλοί διαφορετικοί παράγοντες που επηρεάζουν τα αποτελέσματά της. Τέτοιοι παράγοντες είναι οι διαφορετικές συνθήκες φωτισμού, οι εναλλαγές στις εκφράσεις ή στην στάση του προσώπου, η ομοιότητα που είναι πιθανό να παρουσιάζεται σε πρόσωπα που έχουν ταξινομηθεί στην ίδια κλάση (inter-class similarity) όπως για παράδειγμα η παρακάτω φωτογραφία διδύμων αλλά και η διακύμανση εντός μίας κλάσης που μπορεί να οφείλονται σε μεταβολές της στάσης, του φωτισμού της έκφρασης, του χρώματος, αλλά και αξεσουάρ που μπορεί να φοράει το εικονιζόμενο πρόσωπο όπως γυαλιά, καπέλο κ.α.



Σχήμα 11: Εικόνες προσώπων κάτω από διαφορετικές συνθήκες φωτισμού καθώς και από δίδυμα αδέρφια



Σχήμα 12: Διαφορετικές πόζες του ίδιου προσώπου σε διαφορετικές συνθήκες

Τα κύρια ερευνητικά θέματα γύρω από την ανίχνευση προσώπου είναι πρωτίστως αναπαράστασης (Πως περιγράφεται ένα πρόσωπο;), δευτερευόντως κλίμακας (Πως αντιμετωπίζονται πρόσωπα διαφορετικών διαστάσεων;) αλλά και στρατηγικής αναζήτησης (Πως εντοπίζονται τα πρόσωπα;) καθώς επίσης ταχύτητας (Πως επιταχύνεται η διαδικασία;) και ακρίβειας (Πως γίνεται ακριβής εντοπισμός των προσώπων;).

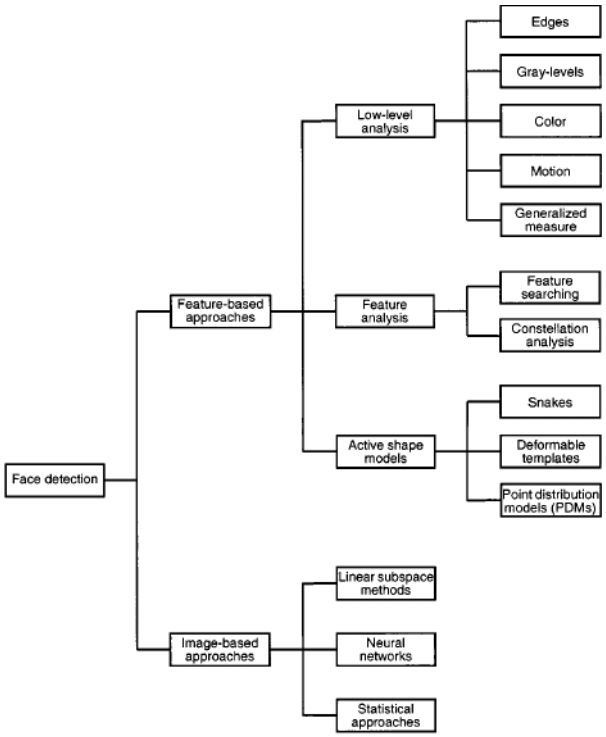
3.2 Μέθοδοι Ανίχνευσης - Εντοπισμού Προσώπων

Μια πρώτη μέθοδος ανίχνευσης και εντοπισμού προσώπων βασίζεται στη δημιουργία κανόνων που εξάγονται από την υπάρχουσα ανθρώπινη γνώση του τι συνιστά ένα τυπικό πρόσωπο. Για παράδειγμα μπορεί να θεωρηθεί πως το κέντρο ενός προσώπου παρουσιάζει ομοιομορφία ως προς τις τιμές της έντασης καθώς και ότι περιμένουμε να βρούμε ένα πρόσωπο με δύο μάτια συμμετρικά μεταξύ τους, μια μύτη και ένα στόμα και έτσι να καθοδηγήσουμε την αναζήτηση μέσα στην εικόνα. Τα θετικά μια τέτοιας προσέγγισης είναι πως είναι εύκολο να βρει κανείς απλούς κανόνες για να περιγράψει τα χαρακτηριστικά ενός προσώπου και τις σχέσεις που τα συνδέουν καθώς και ότι η μέθοδος αυτή δουλεύει καλά όταν επιθυμούμε να εντοπίσουμε την τοποθεσία ενός προσώπου σε μια εικόνα με απλό φόντο. Τα αρνητικά της είναι πως είναι δύσκολο να μεταφραστεί η ανθρώπινη γνώση σε κανόνες με ακρίβεια, με αποτέλεσμα είτε η μέθοδος αναγνώρισης προσώπου να αποτυγχάνει επειδή θέσαμε λεπτομερείς κανόνες, είτε να επιστρέφει ως πρόσωπα άσχετες εικόνες (false positives) γιατί οι κανόνες που τέθηκαν ήταν ελαστικοί.

Μια δεύτερη μέθοδος βασίζεται στην εύρεση δομικών χαρακτηριστικών τα οποία θα εμφανίζονται ανεξάρτητα από μεταβολές στην πόζα, στην οπτική γωνία και το φωτισμό. Η προσέγγιση αυτή ξεκινά με την ανίχνευση χαρακτηριστικών του προσώπου όπως τα μάτια, η μύτη και το στόμα και στη συνέχεια αναζητά επιπλέον χαρακτηριστικά του προσώπου όπως ακμές, σχήμα, υφή, χρώμα και ένταση στοχεύοντας στο να ανιχνεύσει ποια από αυτά τα χαρακτηριστικά είναι αμετάβλητα. Παρόλο που η χρήση αμετάβλητων χαρακτηριστικών συντελεί ένα μεγάλο προτέρημα για αυτή τη μέθοδο, συχνά αποτυγχάνει καθώς ο εντοπισμός των χαρακτηριστικών του προσώπου είναι αρκετά δύσκολο πρόβλημα όταν υπάρχει θόρυβος όσο και όταν το φόντο είναι πολύπλοκο.

Μια επιπλέον μεθοδολογία είναι η ανίχνευση προσώπου που βασίζεται σε ταίριασμα με τεμπλέτες. Εδώ αρχικά αποθηκεύονται αρκετά στάνταρ πρότυπα τα οποία περιγράφουν είτε ένα πρόσωπο σαν σύνολο είτε τα χαρακτηριστικά του προσώπου ξεχωριστά. Η διαδικασία αυτή γίνεται είτε με προκαθορισμένο τρόπο (βασιζόμενη σε ακμές ή περιοχές) ή με παραμορφώσιμο τρόπο (βασιζόμενη στα περιγράμματα χαρακτηριστικών του προσώπου). Για τις τεμπλέτες δεν χρησιμοποιείται εκμάθηση αλλά προγραμματίζονται με το χέρι ενώ στη συνέχεια χρησιμοποιώντας ως μετρική τη συσχέτιση εντοπίζονται τα πρόσωπα. Η ιδέα αυτή είναι απλή ως προς την υλοποίηση της, όμως πρωτίστως θα πρέπει η τεμπλέτες να αρχικοποιηθούν κοντά στα πρόσωπα μέσα στην εικόνα και δευτερευόντως είναι δύσκολο να δημιουργήσει κανείς τεμπλέτες για κάθε πιθανή πόζα του προσώπου.

Τέλος υπάρχουν και οι μέθοδοι που βασίζονται στην παρουσίαση (Appearance based methods). Τα μοντέλα με βάση αυτή την προσέγγιση εκπαιδεύονται από ένα σετ εικόνων εκπαίδευσης οι οποίες αποτυπώνουν αντιπροσωπευτικά την μεταβλητότητα της παρουσίασης ενός προσώπου. Αρχικά πραγματοποιείται η εκπαίδευση ενός ταξινομητή χρησιμοποιώντας τόσο πραγματικά παραδείγματα προσώπων όσο και εσφαλμένα. Οι περισσότερες state-of-the-art μέθοδοι ανήκουν σε αυτή την κατηγορία, η πιο γνωστή από τις οποίες είναι η μέθοδος ανίχνευσης προσώπου των Viola & Jones η οποία θα αναπτυχθεί παρακάτω.



Σχήμα 13: Προσεγγίσεις και μέθοδοι ανίχνευσης προσώπου (επανεκτύπωση από το [27])

3.3 Η Ανίχνευση Προσώπου των Viola & Jones

Ο αλγόριθμος ανίχνευσης προσώπου των Viola & Jones που παρουσιάσθηκε το 2001 [50], αποτελεί το πρώτο πλαίσιο για ανίχνευση αντικειμένων το οποίο παρουσίασε καλά ποσοστά ανίχνευσης σε πραγματικό χρόνο. Παρόλο που μπορεί να εκπαιδευτεί κατά τέτοιο τρόπο ώστε να ανιχνεύει μια ποικιλία από κλάσεις αντικειμένων, επικεντρώνεται στο πρόβλημα της ανίχνευσης προσώπου. Ο αλγόριθμος αυτός αφού χρησιμοποιήσει ένα διαφορετικό τρόπο αναπαράστασης της εικόνας (ολοκληρωτική εικόνα) στη συνέχεια κάνει χρήση Haar χαρακτηριστικών [51], του AdaBoost αλγορίθμου εκπαίδευσης και ενός κατευθυνόμενου ακολουθιακού ταξινομητή (Attentional Cascade Classifier).

3.3.1 Ολοκληρωτική Εικόνα

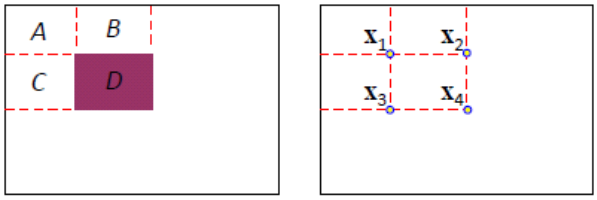
Τα χαρακτηριστικά σχήματος ορθογωνίου παραλληλογράμμου μπορούν να υπολογιστούν πολύ γρήγορα χρησιμοποιώντας μια ενδιάμεση αναπαράσταση της εικόνας που ονομάζεται ολοκληρωτική εικόνα. Η ολοκληρωτική εικόνα στο σημείο x,y περιέχει το άθροισμα όλων των pixels πάνω και αριστερά από το x,y ως εξής :

$$ii(x,y) = \sum_{\substack{x' \leq x \\ y' \leq y}} i(x',y')$$

όπου i(x,y) η πρωτότυπη εικόνα και ii(x,y) η ολοκληρωτική εικόνα. Χρησιμοποιώντας την παρακάτω επαναληπτική διαδικασία, είναι δυνατό να υπολογιστεί η ολοκληρωτική εικόνα με ένα πέρασμα πάνω από την πρωτότυπη:

$$s(x,y) = s(x,y-1) + i(x,y)$$
$$ii(x,y) = ii(x-1,y) + s(x,y)$$

με $s(x,y)$ να είναι το άθροισμα της σειράς, $s(x,-1) = 0$ και $ii(-1,y) = 0$.



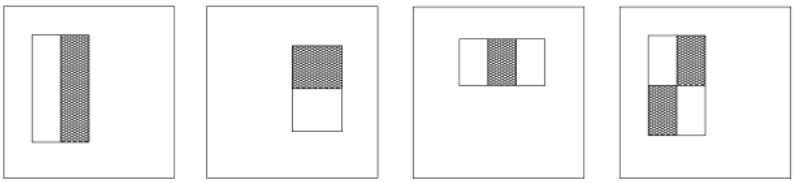
Σχήμα 14: Πρωτότυπη και ολοκληρωτική εικόνα

Χρησιμοποιώντας την ολοκληρωτική εικόνα το άθροισμα ενός ορθογωνίου σαν αυτό του D που φαίνεται στο παραπάνω σχήμα μπορεί να υπολογιστεί αφού γίνουν οι ακόλουθοι υπολογισμοί.
$ii(x_1) = A$
$ii(x_2) = A + B$
$ii(x_3) = B + C$
$ii(x_4) = A + B + C + D$
από όπου προκύπτει πως $D = ii(x_4) + ii(x_1) - ii(x_2) - ii(x_3)$. Το πλεονέκτημα αυτής της αναπαράστασής είναι πως ο υπολογιστικός χρόνος για οποιοδήποτε άθροισμα ορθογώνιων είναι σταθερό ενώ χρειάζονται το πολύ τρεις πράξεις για να πραγματοποιηθεί.

3.3.2 Haar Χαρακτηριστικά

Ο αλγόριθμος ανίχνευσης προσώπων των Viola & Jones ταξινομεί τις εικόνες βασιζόμενος στην τιμή κάποιων απλών χαρακτηριστικών και όχι των pixel απευθείας. Οι κυριότεροι λόγοι για αυτή την επιλογή είναι πως τα χαρακτηριστικά μπορούν να κωδικοποιήσουν την εξειδικευμένη υπάρχουσα γνώση, η οποία είναι δύσκολο να μαθευτεί όταν χρησιμοποιείται μια πεπερασμένη ποσότητα δεδομένων εκπαίδευσης. Ταυτόχρονα τα συστήματα που βασίζονται στα χαρακτηριστικά εμφανίζονται να λειτουργούν αρκετά πιο γρήγορα σε σχέση με τα αντίστοιχα που βασίζονται στα pixels.

Τα Haar χαρακτηριστικά τα οποία χρησιμοποιούνται για ανίχνευση προσώπου χωρίζονται σε 3 είδη. Το πρώτο αποτελείται από ένα ορθογώνιο παραλληλόγραμμο χωρισμένο σε δύο τμήματα στα οποία εξάγεται η τιμή της διαφοράς μεταξύ του αθροίσματος των pixel στη λευκή περιοχή από το άθροισμα των αντίστοιχων στη μαύρη περιοχή. Οι δύο περιοχές έχουν το ίδιο ακριβώς μέγεθος και σχήμα και είναι παρακείμενες είτε οριζόντια είτε κάθετα. Η 2η κατηγορία συνίσταται από ένα ορθογώνιο παραλληλόγραμμο αποτελούμενο από τρία τμήματα όπου υπολογίζουμε το άθροισμα των pixel μεταξύ των δύο εξωτερικών τμημάτων, αφαιρώντας στη συνέχεια αυτό του εσωτερικού στη μέση. Το τελευταίο είδος είναι το ορθογώνιο παραλληλόγραμμο με τέσσερα τμήματα όπου υπολογίζεται η διαφορά μεταξύ των διαγωνίων τμημάτων που φέρουν το ίδιο χρώμα. Αν θεωρηθεί πως η βασική ανάλυση του ανιχνευτή είναι 24×24 προκύπτει ότι πλήρες σετ των Haar χαρακτηριστικών είναι μεγαλύτερο από 180.000.



Σχήμα 15: Σχήματα των τριών ειδών των Haar χαρακτηριστικών

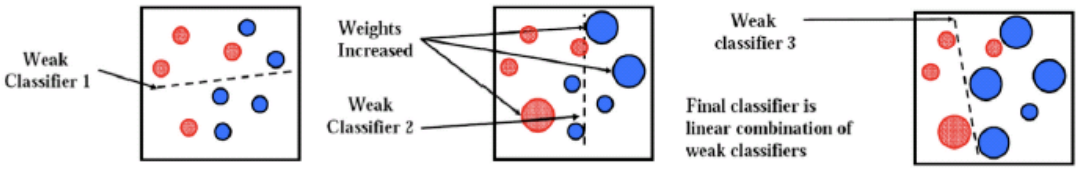
Η χρήση χαρακτηριστικών σχήματος ορθογωνίου παραλληλογράμμου θεωρείται θεμελιώδης συγκρινόμενη με εναλλακτικές μεθόδους όπως τα κατευθυνόμενα φίλτρα. Τα κατευθυνόμενα φίλτρα είναι μια εξαιρετική μέθοδος για λεπτομερή ανάλυση των συνόρων σε μια εικόνα, για συμπίεση εικόνας καθώς επίσης και για ανάλυση υφής. Αντιθέτως, τα χαρακτηριστικά σε σχήμα ορθογωνίου όπως τα Haar παρόλο που είναι ευαίσθητα στην παρουσία απλών δομών μέσα σε μια εικόνα όπως οι ακμές έχουν το πλεονέκτημα ότι έχουν αρκετά συμπαγή δομή. Σε αντίθεση με τα κατευθυνόμενα φίλτρα οι μόνες κατευθύνσεις τους είναι η κάθετη, η οριζόντια και η διαγώνια. Το σετ των Haar χαρακτηριστικών παρέχει όμως πλούσια αναπαράσταση της εικόνας υποστηρίζοντας με αυτό τον τρόπο την αποτελεσματική εκμάθηση.

3.3.3 Αλγόριθμος AdaBoost

Ο αλγόριθμος AdaBoost στην κανονική του μορφή χρησιμοποιείται για να ενισχύσει την επίδοση της ταξινόμησης ενός απλού (αδύναμου) αλγορίθμου. Οι Freund & Schapire [18] απέδειξαν ότι το λάθος που προκύπτει από την εκπαίδευση ενός ισχυρού ταξινομητή προσεγγίζει το μηδέν εκθετικά καθώς αυξάνεται ο αριθμός των εκτελέσεων του αλγορίθμου. Επειδή ο αριθμός των χαρακτηριστικών ξεπερνά τις 180.000 σε κάθε παράθυρο που παίρνουμε πάνω σε μια εικόνα και παρόλο που καθένα από αυτά τα χαρακτηριστικά μπορεί να υπολογιστεί πολύ αποδοτικά, το να υπολογιστεί το πλήρες σετ είναι απαγορευτικό από άποψη κόστους. Οι Viola & Jones κατέληξαν στο συμπέρασμα πως μόνο ένας πολύ μικρός αριθμός από αυτά τα χαρακτηριστικά μπορεί να συνδυαστεί ώστε να σχηματίσει έναν αποτελεσματικό ταξινομητή. Η κύρια πρόκληση είναι να βρεθούν αυτά τα χαρακτηριστικά. Για την επίτευξη αυτού του στόχου ο αδύναμος αλγόριθμος εκμάθησης σχεδιάζεται έτσι ώστε να επιλέγει εκείνο το χαρακτηριστικό σε σχήμα ορθογωνίου παραλληλογράμμου το οποίο διαχωρίζει καλύτερα τα θετικά από τα αρνητικά παραδείγματα. Για κάθε χαρακτηριστικό ο αδύναμος ταξινομητής επιλέγει τη συνάρτηση που αποδίδει το βέλτιστο κατώφλι έτσι ώστε να ταξινομείται λανθασμένα ο μικρότερος δυνατός αριθμός δειγμάτων. Ένας αδύναμος ταξινομητής $\mathrm{h}_j(x)$ αποτελείται από ένα χαρακτηριστικό f_j, ένα κατώφλι $\theta_j(x)$ και μια πιθανότητα p_j η οποία καθορίζει τη φορά του δείκτη της ανισότητας. Ο τύπος με τον οποίο ορίζεται είναι:

$$\mathrm{h}_j(x) = \begin{cases} 1 & \text{αν } \mathrm{p}_j \mathrm{f}_j(x) < \mathrm{p}_j \theta_j \\ 0 & \text{αλλιώς} \end{cases}$$

και φαίνεται πως πρόκειται για ιδιαίτερα απλούς ταξινομητές δεδομένου ότι η έξοδος τους λαμβάνει δυαδικές τιμές. Στη συνέχεια όταν η εκπαίδευση ολοκληρωθεί, εξάγεται ένας ισχυρός ταξινομητής ο οποίος λαμβάνει για είσοδο ένα νέο διάνυσμα x και το ταξινομεί χρησιμοποιώντας ένα άθροισμα με βάρη των αδύναμων ταξινομητών όπως φαίνεται και παρακάτω.



Σχήμα 16: Ο συνδυασμός των αδύναμων ταξινομητών με τα αντίστοιχα βάρη οδηγεί στον ισχυρό ταξινομητή

Αλγόριθμος 1 Αλγόριθμος Εκπαίδευσης AdaBoost

1: *Δίνονται εικόνες* $(x_1, y_1), \ldots, (x_n, y_n)$ *με* $y_i = 0, 1$ *για αρνητικά και θετικά παραδείγματα αντίστοιχα*

2: *Αρχικοποίησε τα βάρη* $w_{1,i} = \frac{1}{2m}, \frac{1}{2l}$ *για* $y_i = 0, 1$ *αντίστοιχα όπου* m *και* l *ο αριθμός των θετικών και αρνητικών περιπτώσεων αντίστοιχα*

3: *Για* $t = 1 \to T$

4: *Κανονικοποίησε τα βάρη* $w_{t,i} \leftarrow \frac{w_{t,i}}{\sum_{j=1}^{n} w_{t,j}}$

5: *Επίλεξε τον καλύτερο αδύναμο ταξινομητή λαμβάνοντας υπόψη το σφάλμα με βάρη:*

$$\varepsilon_t = \min_{f,p,\theta} = \sum_i w_i |h(x_i, f, p, \theta) - y_i|$$

6: *Όρισε* $h_t(x) = h(x, f_t, p_t, \theta_t)$ *με* f_t, p_t, θ_t *να είναι οι τιμές εκείνες για τις οποίες*

7: *ελαχιστοποιείται το σφάλμα* ε_t

8: *Ανανέωσε τα βάρη:*

$$w_{t+1,i} = w_{t,i} \beta_t^{1-e_i}$$

όπου $e_i = 0$ *αν το παράδειγμα* x_i *κατηγοριοποιηθεί σωστά ενώ σε αντίθετη περίπτωση*

9: $e_i = 1$ *ενώ το* $\beta_t = \frac{\varepsilon_t}{1 - \varepsilon_t}$

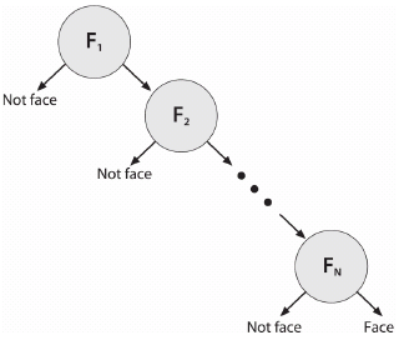
10: *Ο τελικός ισχυρός ταξινομητής θα είναι:*

$$C(x) = \begin{cases} 1 & \text{αν } \sum_{t=1}^{T} \alpha_t h_t(x) \geq \frac{1}{2} \sum_{t=1}^{T} \alpha_t \\ 0 & \text{αλλιώς} \end{cases}$$

όπου $\alpha_t = \log \frac{1}{\beta_t}$

3.3.4 Κατευθυνόμενοι Ακολουθιακοί Ταξινομητές

Το επόμενο βήμα είναι η κατασκευή κάποιων ακολουθιακών ταξινομητών οι οποίοι να επιτυγχάνουν αυξημένη απόδοση ανίχνευσης ενώ παράλληλα να μειώνουν δραστικά το υπολογιστικό κόστος. Το κλειδί πίσω από αυτό το σκεπτικό βασίζεται στο γεγονός ότι μπορούν να φτιαχτούν πιο μικροί και επομένως πιο αποτελεσματικοί boosted ταξινομητές οι οποίοι θα απορρίπτουν πολλά από τα αρνητικά υπο-παράθυρα ενώ ταυτόχρονα θα ανιχνεύουν σχεδόν όλες τις θετικές περιπτώσεις. Το κατώφλι σε τέτοιου είδους ταξινομητές μπορεί να προσαρμοστεί κατά τέτοιο τρόπο ώστε το false negative ποσοστό (περιπτώσεις όπου το επιστρεφόμενο αποτέλεσμα της μεθόδου είναι αρνητικό ενώ δε θα έπρεπε) να τείνει στο μηδέν. Επειδή μέσα σε μια εικόνα οι περισσότερες υπό-εικόνες αποτελούν περιπτώσεις στις οποίες δεν βρίσκεται πρόσωπο θα χρησιμοποιηθούν σαν πρώτο βήμα απλοί ταξινομητές οι οποίοι θα απορρίψουν την πλειονότητα των υποπαραθύρων σε όσο το δυνατόν αρχικά στάδια πριν γίνει η χρήση πιο δυνατών ταξινομητών οι οποίοι θα πετύχουν μικρότερα false positive ποσοστά. Η ανιχνευτική διαδικασία που πραγματοποιείται έχει τη μορφή ενός εκφυλισμένου δέντρου απόφασης. Ένα θετικό αποτέλεσμα από τον πρώτο (δυνατό) ταξινομητή ενεργοποιεί τη διαδικασία αξιολόγησης του δεύτερου ταξινομητή ο οποίος έχει ομοίως προσαρμοστεί ώστε να επιτυγχάνει υψηλά ποσοστά ανίχνευσης. Ένα αρνητικό αποτέλεσμα σε οποιοδήποτε στάδιο της διαδικασίας οδηγεί σε άμεση απόρριψη του συγκεκριμένου υπό-παραθύρου. Τα στάδια της ακολουθιακής διαδικασίας δημιουργούνται εκπαιδεύοντας ταξινομητές με τη χρήση του AdaBoost αλγορίθμου προσαρμόζοντας στη συνέχεια το κατώφλι ώστε να ελαχιστοποιούνται τα false negatives. Θα πρέπει να επισημανθεί πως όπως και στα δέντρα απόφασης, οι μεταγενέστεροι ταξινομητές στην ακολουθιακή διαδικασία έχουν εκπαιδευτεί χρησιμοποιώντας παραδείγματα τα οποία έχουν περάσει επιτυχώς όλα τα προηγούμενα στάδια. Συνεπώς, ο δεύτερος ταξινομητής αντιμετωπίζει ένα πιο δύσκολο έργο συγκριτικά με τον πρώτο αφού τα παραδείγματα τα οποία περνάνε το πρώτο στάδιο είναι δυσκολότερα από τα τυπικά παραδείγματα. Δοσμένου ενός ποσοστού ανίχνευσης (detection rate), οι ταξινομητές που βρίσκονται σε μετέπειτα στάδια εντός της ακολουθίας έχουν αντίστοιχα μεγαλύτερα false positive ποσοστά.

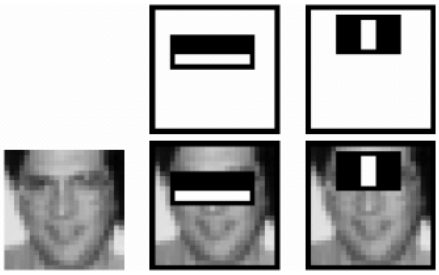


Σχήμα 17: Cascade ταξινομητής απόρριψης: κάθε κόμβος αναπαριστά ένα ισχυρό ταξινομητή που λειτουργεί κατά τέτοιο τρόπο ώστε σπάνια να χάνει ένα πραγματικό πρόσωπο ενώ παράλληλα θα απορρίπτει ένα πιθανότατα μικρό αριθμό από nonfaces

Η διαδικασία εκπαίδευσης αυτών των ακολουθιακών ταξινομητών έχει δύο είδη tradeoff. Στις περισσότερες περιπτώσεις οι ταξινομητές με τα περισσότερα χαρακτηριστικά επιτυγχάνουν μεγαλύτερα ποσοστά ανίχνευσης και μικρότερα false positive ποσοστά ενώ απαιτείται και περισσότερος υπολογιστικός χρόνος. Σαν αξίωμα, θα μπορούσε να οριστεί ένα πλαίσιο βελτιστοποίησης στο οποίο ο

αριθμός των σταδίων του ταξινομητή, ο αριθμός των χαρακτηριστικών σε κάθε στάδιο και το κατώφλι σε κάθε στάδιο, λαμβάνουν τέτοιες τιμές ώστε να ελαχιστοποιείται ο εκτιμώμενος αριθμός των χαρακτηριστικών προς αξιολόγηση. Στην πραγματικότητα όμως αυτό που συμβαίνει είναι σε κάθε στάδιο του ακολουθιακού ταξινομητή να μειώνεται το false positive ποσοστό καθώς και το ποσοστό ανίχνευσης. Αφού γίνει η επιλογή του ελάχιστου αριθμού στον οποίο θέλουμε να μειώσουμε τα false positives και της μέγιστης μείωσης που θέλουμε να επιτευχθεί με τη διαδικασία ανίχνευσης, εκπαιδεύεται κάθε στάδιο προσθέτοντας χαρακτηριστικά μέχρι τα ποσοστά ανίχνευσης και false positives να γίνουν ίσα.

Συνοψίζοντας ο αλγόριθμος ανίχνευσης προσώπου των Viola & Jones χρησιμοποιεί τα ακόλουθα βασικά συστατικά: τη χρήση της ολοκληρωτικής εικόνας για να γίνεται πιο αποδοτικά η συνέλιξη, τα Haar χαρακτηριστικά και τον AdaBoost αλγόριθμο τόσο για την επιλογή των χαρακτηριστικών όσο και για την εκμάθηση του ακολουθιακού ταξινομητή. Αποτελεί τον πιο διαδεδομένο αλγόριθμο ανίχνευσης προσώπου καθώς είναι γρήγορος, αρκετά εύρωστος και τρέχει και σε πραγματικό χρόνο. Στα αρνητικά του θα μπορούσε να επισημάνει κανείς πως απαιτεί πολύ χρόνο στο στάδιο της εκπαίδευσης.



Σχήμα 18: Εξαγόμενα Χαρακτηριστικά του ανιχνευτή των Viola & Jones. Το πρώτο και το δεύτερο χαρακτηριστικό επιλέγονται από τον AdaBoost. Το πρώτο χαρακτηριστικό μετράει τη διαφορά της έντασης μεταξύ της περιοχής των ματιών και της περιοχής του προσώπου που καλύπτεται από το πάνω μέρος των μάγουλων. Το χαρακτηριστικό αυτό βασίζεται στην παρατήρηση ότι η περιοχή των ματιών είναι συνήθως πιο σκοτεινή σε σχέση με αυτή στα μάγουλα. Το δεύτερο χαρακτηριστικό συγκρίνει τις εντάσεις στις περιοχές των ματιών με την ένταση της περιοχής όπου ξεκινάει η μύτη.

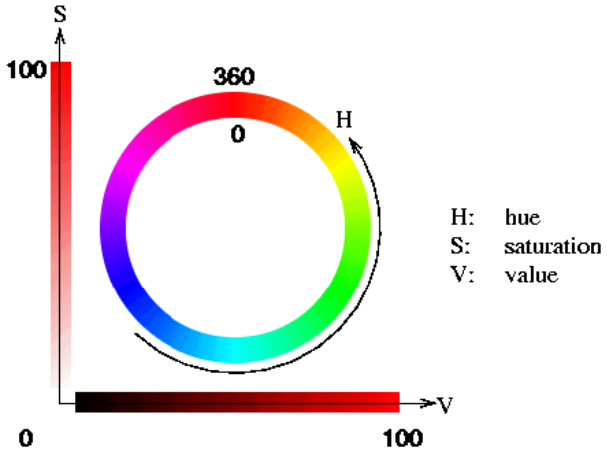
3.4 Ανίχνευση Δέρματος

Ανίχνευση δέρματος είναι η διαδικασία της εύρεσης των pixel σε χρώμα δέρματος σε μια εικόνα ή σε ένα βίντεο. Χρησιμοποιείται είτε σαν στάδιο προ-επεξεργασίας για να προσεγγιστούν οι περιοχές που είναι πιθανόν να υπάρχει ανθρώπινο πρόσωπο είτε μετά την ανίχνευση του προσώπου ώστε να απορριφτούν πιθανά λάθη που είναι δυνατόν να επιστρέψει ο αλγόριθμος. Ένας ανιχνευτής δέρματος συνήθως μετασχηματίζει ένα pixel σε ένα κατάλληλο χρωματικό χώρο και στη συνέχεια χρησιμοποιεί ένα ταξινομητή δέρματος ο οποίος αποδίδει μια ταμπέλα στο pixel ανάλογα με το αν περιέχει ή όχι δέρμα σχηματίζοντας εν τέλει ένα σύνορο που διαχωρίζει τις περιοχές όπου υπάρχει δέρμα από τις αντίστοιχες που περιέχουν άσχετη πληροφορία.

Οι Forsyth & Fleck [16] επεσήμαναν πως το χρώμα του ανθρώπινου δέρματος έχει ένα περιορισμένο εύρος από αποχρώσεις και δεν λαμβάνει μεγάλες τιμές κορεσμού αφού το δέρμα συνίσταται από

ένα συνδυασμό αίματος (κόκκινο) και μελανίνης(καφέ και κίτρινο). Συνεπώς, το χρώμα του ανθρώπινου δέρματος λαμβάνει ένα συγκεκριμένο εύρος τιμών μέσα σε ένα χρωματικό χώρο, όχι όμως το ίδιο για όλους τους χρωματικούς χώρους. Μια πληθώρα από χρωματικούς χώρους έχει χρησιμοποιηθεί με στόχο την εύρεση εκείνου του χώρου όπου το χρώμα του δέρματος δεν εξαρτάται από τις συνθήκες φωτισμού. Η επιλογή του χρωματικού χώρου επηρεάζει το σχήμα της κλάσης του δέρματος και κατ' επέκταση την διαδικασία ανίχνευσης.

Ο Maragos [34] αναφέρει πως η αντίληψη του χρώματος από ανθρώπους δημιουργεί τις εξής ψυχολογικές αισθήσεις χρώματος: απόχρωση (hue), κορεσμός (saturation), φωτεινότητα-ένταση (brightness-intensity). Η απόχρωση είναι η ιδιότητα του χρώματος που μεταβάλλεται πηγαίνοντας σε διαφορετικά κύρια χρώματα του φάσματος, π.χ. από κόκκινο σε πράσινο. Ο κορεσμός είναι η ιδιότητα του χρώματος που μεταβάλλεται με τον βαθμό απόχρωσης σε κάποιο χρώμα και σχετίζεται με την καθαρότητα του χρώματος. Η φωτεινότητα είναι η ιδιότητα του χρώματος που μεταβάλλεται πηγαίνοντας από το μαύρο στο λευκό.
Ο χρωματικός χώρος που επιλέξαμε είναι ο HSV (Hue, Saturation, Value όπου Value σημαίνει φωτεινότητα) ο οποίος αποτελεί ένα μη γραμμικό μετασχηματισμό του χώρου RGB και αντιστοιχεί καλύτερα στην ανθρώπινη αντίληψη της τοπολογίας του χώρου των χρωμάτων.



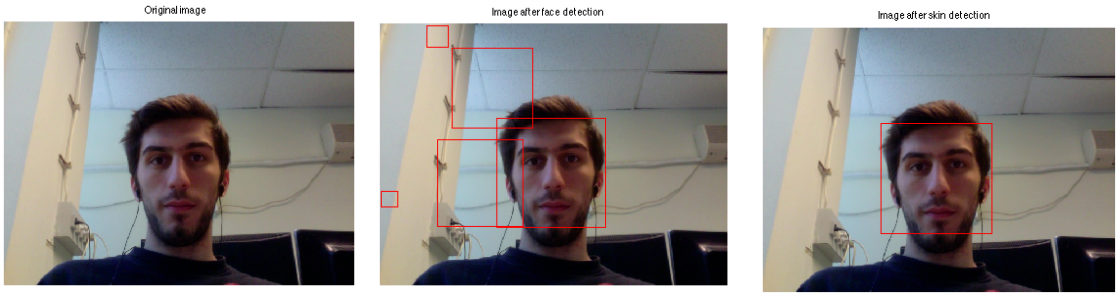
Σχήμα 19: Ο χρωματικός χώρος HSV

Αφού έχουμε κάνει ανίχνευση προσώπου και και αλλαγή του χρωματικού χώρου κρατάμε στη συνέχεια μόνο το πρώτο κανάλι της εικόνας που αντιστοιχεί στην απόχρωση και υπολογίζουμε τον παρακάτω λόγο :

$$Ratio = \frac{Number\ of\ pixels\ <\ Threshold}{Overall\ number\ of\ pixels}$$

Αν ο λόγος αυτός είναι μεγαλύτερος από ένα δεύτερο κατώφλι τότε θεωρούμε πως έχουμε πρόσωπο ενώ σε αντίθετη περίπτωση απορρίπτουμε το συγκεκριμένο bounding box του τρέχοντος frame και προχωράμε στο επόμενο όπως φαίνεται και στο παρακάτω σχήμα. Για την επιλογή των τιμών για τα δύο παραπάνω κατώφλια μετά από πειράματα επιλέξαμε να κρατάμε τα pixel που λαμβάνουν τιμή απόχρωσης μικρότερης ή ίσης του 10% ($Threshold = 10\%$) ενώ για την απόφαση του αν θα

κρατήσουμε ή όχι το πρόσωπο επελέγη να έχουμε $Ratio > 50\%$. Η βελτίωση που παρουσιάζει μια τέτοια προσέγγιση είναι εμφανής στην παρακάτω φωτογραφία από webcam



Σχήμα 20: Βελτίωση ανίχνευσης προσώπου με εφαρμογή ανίχνευσης δέρματος

3.5 Ανίχνευση Προσώπου και Ημερολογιοποίηση Ομιλητών

Οι Otsuka et al. [38] με απώτερο σκοπό την επέκταση των συμπερασμάτων τους σε meetings αναλύουν την πρόσωπο με πρόσωπο ομιλία, η οποία είναι ένας από τους πιο βασικούς τρόπους επικοινωνίας στην καθημερινή ζωή κατά την οποία η άνθρωποι ανταλλάσουν μηνύματα τόσο προφορικά μέσω διαλόγου όσο και με άλλες μορφές που σχετίζονται με τη συμπεριφορά του σώματος του ανθρώπου κατά τη διάρκεια της ομιλίας. Χαρακτηριστικά παραδείγματα μη προφορικών μηνυμάτων που ανταλλάσσονται κατά την ομιλία είναι το που κοιτάει ο ομιλητής, οι εκφράσεις του προσώπου, οι κινήσεις του κεφαλιού, οι κινήσεις των χεριών καθώς και η στάση του σώματος. Επομένως, για την καλύτερη κατανόηση των σκηνών ομιλίας χρησιμοποιούνται παρατηρήσεις και πληροφορίες που εξάγονται από την ευρύτερη συμπεριφορά ενός ομιλητή είτε αυτή είναι προφορική είτε όχι. Παράλληλα, τονίζουν πως η σημασία της μέτρησης των εκφράσεων του προσώπου προκύπτει από το γεγονός ότι είναι ένας εύλογος δείκτης του που κοιτάει και εστιάζει την προσοχή του ο κάθε ομιλητής.

Ένα από τα πρώτα και σημαντικότερα βήματα για να πραγματοποιηθεί η Ημερολογιοποίηση Ομιλητών είναι να κατασκευασθεί ένα διάνυσμα χαρακτηριστικών (feature vector) το οποίο θα αναπαριστά με τον καλύτερο δυνατό τρόπο την πληροφορία που θέλουμε να χρησιμοποιήσουμε από κάθε καρέ ενός βίντεο. Πολλές φορές η χρήση πληροφορίας από ολόκληρο το καρέ ενός βίντεο θα έκανε πιο εύκολη την περιγραφή του περιεχομένου του. Για παράδειγμα ένας εύκολος τρόπος διαχωρισμού των ανθρώπων που εμφανίζονται σε ένα βίντεο θα μπορούσε να είναι το τι χρώμα μπλούζα φοράνε, ή το τι φόντο υπάρχει πίσω τους όταν μιλάνε. Παρόλα αυτά, επειδή η χρήση όλου του καρέ σε ένα βίντεο θα κατέληγε σε ένα διάνυσμα χαρακτηριστικών με τεράστιες διαστάσεις το οποίο εκτός του ότι θα ήταν αδύνατο να το χειριστεί κανείς υπολογιστικά, θα περιείχε και αρκετή άσχετη πληροφορία, επιλέγεται να εξάγονται πληροφορίες από μια συγκεκριμένη περιοχή μέσα σε κάθε καρέ.

Στο πρόβλημα της Ημερολογιοποίησης Ομιλητών επιλέγεται η περιοχή αυτή να είναι εκείνη στην οποία εμφανίζεται πρόσωπο μέσα στην εικόνα. Εκτός των πλεονεκτημάτων αυτής της προσέγγισης που αναφέρθηκαν παραπάνω, εντός της περιοχής του προσώπου μπορεί να ανιχνευθεί επίσης η περιοχή στην οποία βρίσκεται το στόμα και η οποία θα χρησιμοποιηθεί στη συνέχεια έτσι ώστε να γίνει η μετάβαση από την Ημερολογιοποίηση Προσώπων (ποιο πρόσωπο εμφανίζεται και πότε μέσα στο βίντεο) στην Ημερολογιοποίηση Ομιλητών.

Για να επιταχυνθεί περισσότερο η διαδικασία ανίχνευσης προσώπου, μειώνουμε τις διαστάσεις της αρχικής εικόνας στο μισό και πραγματοποιούμε σε αυτή ανίχνευση προσώπου ώστε ο αλγόριθμος

των Viola & Jones να ψάχνει τόσο σε λιγότερα pixels όσο και σε μικρότερο αριθμό από κλίμακες για πιθανά πρόσωπα. Τέλος, αφού βρούμε τα bounding boxes των πιθανών προσώπων τα πολλαπλασιάζουμε επί δύο ώστε να επιστραφούν στις αρχικές διαστάσεις τις εικόνας. Στη συνέχεια αν έχει επιστραφεί πρόσωπο - δηλαδή το bounding box δεν είναι κενο - πραγματοποιούμε ανίχνευση δέρματος ώστε να επαληθευτεί το αποτέλεσμα και να αποθηκευτεί το πρόσωπο.

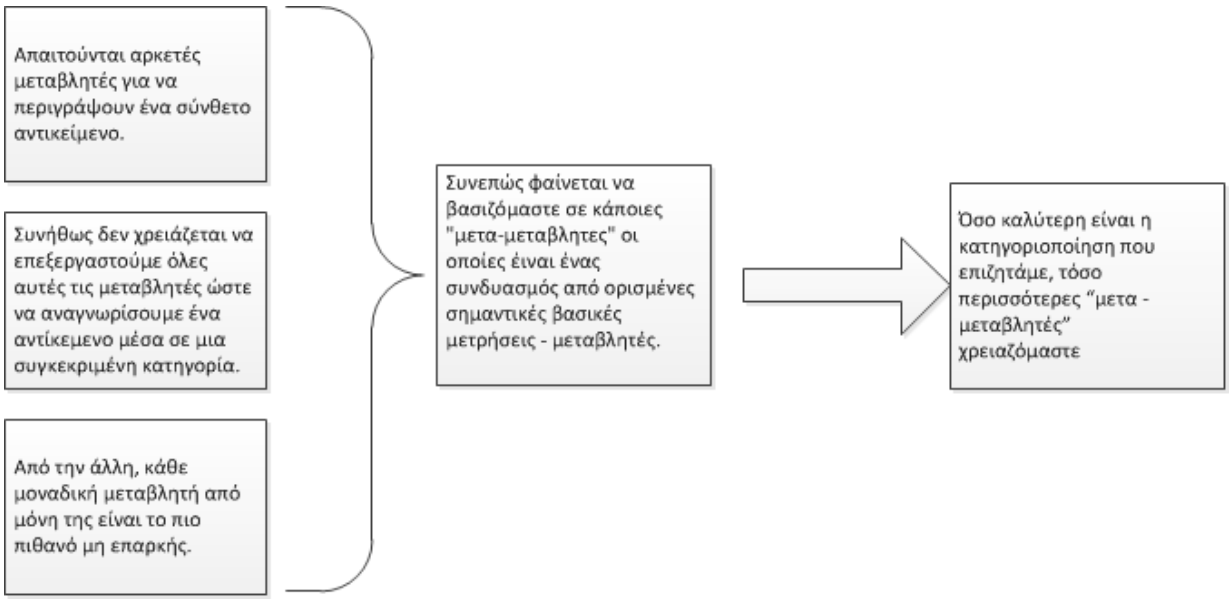
Κεφάλαιο 4

Εξαγωγή Χαρακτηριστικών

4.1 Εισαγωγή

Είτε συνειδητά είτε ασυνείδητα, ως άνθρωποι είμαστε πολύ καλοί στο να εξάγουμε χαρακτηριστικά - περιεκτικά σε πληροφορία - στις καθημερινές μας ζωές. Για παράδειγμα μπορούμε πολύ εύκολα να αναγνωρίσουμε το φύλο ενός ανθρώπου από απόσταση, χωρίς να εξετάσουμε καθόλου πιο λεπτομερή χαρακτηριστικά του. Αυτό συμβαίνει, επειδή γνωρίζουμε μια συγκεκριμένη μορφή που έχει καθένα από τα δύο φύλα π.χ. σχήμα του σώματος, μήκος μαλλιών ή ένα συνδυασμό αυτών των δύο. Μερικές φορές τυχαίνει να μην έχουμε δει ένα φιλικό πρόσωπο για πολλά χρόνια, όμως όταν τον δούμε ξανά ακόμα και αν έχει πάρει ή χάσει βάρος μπορούμε συνήθως ακόμα να τον αναγνωρίσουμε.

Είναι λοιπόν προφανές πως δεν χρειάζεται να επεξεργαστούμε κάθε χαρακτηριστικό ενός αντικειμένου ώστε να το αναγνωρίσουμε και να το κατηγοριοποιήσουμε καθώς έχουμε εκ των προτέρων αναγνωρίσει κάποια ιδιαίτερα - περιεκτικά σε πληροφορία - χαρακτηριστικά. Φαίνεται λοιπόν ότι ισχύουν κάποιοι γενικοί κανόνες που παρουσιάζονται στο ακόλουθο σχήμα:



Σχήμα 21: Γενικοί κανόνες για το είδος και τον αριθμό των εξαγόμενων χαρακτηριστικών

Αυτές οι “μετα-μεταβλητές” (meta variables) είναι ακριβώς τα περιεκτικά σε πληροφορία χαρακτηριστικά που αναζητάμε. Στην στατιστική εκμάθηση, η διαδικασία αναγνώρισης αυτών των μεταβλητών είναι γνωστή σαν εξαγωγή χαρακτηριστικών (feature extraction).

Ο Zhu [54] επισημαίνει πως υπάρχουν τουλάχιστον τρεις λόγοι για τους οποίους η εξαγωγή χαρακτηριστικών είναι ένα σημαντικό πρόβλημα στην προβλεπτική μοντελοποίηση και στη σύγχρονη ανάλυση δεδομένων. Ο πρώτος λόγος είναι η μείωση των διαστάσεων (Dimensionality reduction) στην οποία θα γίνει εκτενής αναφορά στο επόμενο κεφάλαιο. Περιληπτικά σε προβλήματα με πολύ μεγάλο αριθμό μεταβλητών, όλα τα προβλεπτικά μοντέλα δεν παρουσιάζουν την απαιτούμενη απόδοση

εξαιτίας του curse of dimensionality.[1] Είναι συνεπώς επιθυμητό να επιλέξουμε ένα αρκετά μικρότερο αριθμό από σχετικά και σημαντικά χαρακτηριστικά ώστε η εξαγωγή χαρακτηριστικών, να λειτουργήσει θετικά προς την κατεύθυνση της μείωσης των διαστάσεων του προβλήματος. Ταυτόχρονα τα μεγάλα σε διαστάσεις προβλήματα παρουσιάζουν και υπολογιστικά προβλήματα. Μερικές φορές δύο μεταβλητές μπορεί να είναι το ίδιο περιεκτικές σε πληροφορία, αλλά να είναι αρκετά συσχετισμένες μεταξύ τους γεγονός που συνήθως οδηγεί σε μη επιθυμητή συμπεριφορά κατά τον υπολογισμό τους. Χαρακτηριστικό παράδειγμα είναι το πρόβλημα της πολυσυγγραμικότητας (multicollinearity) στη μέθοδο των ελαχίστων τετραγώνων [15].

Δευτερευόντως, η εξαγωγή χαρακτηριστικών χρησιμοποιείται στον τομέα της Αυτόματης Διερευνητικής Ανάλυσης Δεδομένων (Automatic Exploratory Data Analysis). Σε πολλές κλασσικές εφαρμογές, χαρακτηριστικά περιεκτικά σε πληροφορία, συχνά επιλέγονται εκ των προτέρων από τους ερευνητές ώστε να φτιάξουν ένα μοντέλο. Όλο και πιο συχνά, σε σύγχρονες εφαρμογές του data-mining που υπάρχει μια αυξανόμενη ζήτηση για πλήρως αυτοματοποιημένα προβλεπτικά μοντέλα σε μορφή "μαύρου κουτιού", τα οποία θα έχουν από μόνα τους τη δυνατότητα της αναγνώρισης των σημαντικών χαρακτηριστικών. Η ανάγκη για τέτοια αυτοματοποιημένα συστήματα εμφανίζεται για δύο λόγους. Από τη μία, υπάρχουν ανάγκες οικονομικής φύσεως για επεξεργασία μεγάλου αριθμού δεδομένων σε ένα μικρό χρονικό διάστημα με μικρή ανθρώπινη επίβλεψη. Από την άλλη, υπάρχει το ενδεχόμενο, το είδος του προβλήματος καθώς και τα δεδομένα να είναι καινοφανή σε σημείο που να μην υπάρχουν καθόλου ειδικοί στο συγκεκριμένο κλάδο που να καταλαβαίνουν τα δεδομένα αρκετά καλά έτσι ώστε να μπορούν να διαλέξουν από αυτά τις πιο σημαντικές μεταβλητές πριν γίνει η ανάλυση τους. Κάτω από αυτές τις συνθήκες, η Automatic Exploratory Data Analysis αποτελεί το κλειδί του προβλήματος, καθώς αντί να βασιζόμαστε σε ιδέες που έχουν συλληφθεί εκ των προτέρων, υπάρχει τόσο η ανάγκη όσο και το ενδιαφέρον του να αφήσουμε τα δεδομένα να "μιλήσουν" από μόνα τους πέρα από την τυπική μοντελοποίηση.

Τέλος μια άλλη εφαρμογή της εξαγωγής χαρακτηριστικών είναι η οπτικοποίηση των δεδομένων (Data Visualization). Το ανθρώπινο μάτι έχει μια εκπληκτική ικανότητα στο να αναγνωρίζει συστηματικά πρότυπα στα δεδομένα. Ταυτόχρονα, είμαστε συνήθως ανίκανοι στο να συλλάβουμε δεδομένα τα οποία έχουν περισσότερες από τρεις διαστάσεις. Για να μεγιστοποιηθεί η χρήση, της ιδιαίτερα ανεπτυγμένης ανθρώπινης ικανότητας οπτικής αναγνώρισης, βάζουμε συνήθως τον άνθρωπο να αναγνωρίσει δύο ή τρία από τα πιο περιεκτικά σε πληροφορία χαρακτηριστικά στα δεδομένα, έτσι ώστε να μπορούμε να τα αναπαραστήσουμε γραφικά σε ένα χώρο μειωμένων διαστάσεων. Για να παραχθούν τέτοιες γραφικές απεικονίσεις, η εξαγωγή χαρακτηριστικών αποτελεί το καθοριστικό αναλυτικό βήμα προς αυτή την κατεύθυνση.

Σε κάθε μια από τις παραπάνω κατηγορίες εφαρμογών, η εξαγωγή χαρακτηριστικών δεν αποτελεί συνήθως το τελικό βήμα της ανάλυσης, αλλά είναι μια προσπάθεια για να διευκολυνθεί η υπολογιστική διαδικασία και η δημιουργία ενός μοντέλου ενώ, συχνά μπορεί να αποτελέσει ένα σημαντικό επιστημονικό πρόβλημα από μόνη της.

Για παράδειγμα στον ανθρώπινο εγκέφαλο πολλά διαφορετικά μέρη του είναι υπεύθυνα για να επιτελούν πολλές διαφορετικές διεργασίες. Στο πρότζεκτ της "χαρτογράφησης" του ανθρώπινου εγκεφάλου οι επιστήμονες επιδιώκουν την κατανόηση των διαφορετικών περιοχών εντός του εγκεφάλου και να συνδέσουν κάθε περιοχή με μια βασική διεργασία για την οποία είναι υπεύθυνη. Αυτό το επιτυγχάνουν συγκρίνοντας εικόνες του, τις οποίες έχουν λάβει όταν ένας άνθρωπος εκτελεί μια λειτουργία

[1] Το curse of dimensionality βασίζεται στο γεγονός ότι για ένα δεδομένο αριθμό δεδομένων υπάρχει ένας συγκεκριμένος αριθμός χαρακτηριστικών πάνω από τον οποίο η επίδοση του ταξινομητή πέφτει αντί να βελτιώνεται.

(ενεργή κατάσταση) και όταν είναι ακίνητος (ανενεργή κατάσταση). Οι ψηφιοποιημένες εικόνες αποτελούνται από χιλιάδες pixels καθένα από τα οποία αντιστοιχεί σε ένα σημείο στον εγκέφαλο και μπορεί να αντιμετωπισθεί σαν μια μεταβλητή. Αντί να φτιάξουν οι επιστήμονες ένα προβλεπτικό μοντέλο χρησιμοποιώντας όλες αυτές τις μεταβλητές για να διαχωρίσουν αυτές τις εικόνες στις δύο κατηγορίες ανάλογα με την κατάσταση (ενεργή ή ανενεργή), επιδιώκουν στο να ερευνήσουν ποιες περιοχές του ανθρώπινου εγκεφάλου ενεργοποιούνται όταν εκτελεί ο άνθρωπος μια διεργασία. Με άλλα λόγια, επικεντρώνουν την προσοχή τους στο να αναγνωρίσουν τις πιο σημαντικές μεταβλητές (pixels) οι οποίες για το συγκεκριμένο πρόβλημα διαφοροποιούν την ενεργή από την ανενεργή κατάσταση.

4.2 Εξαγωγή και Επιλογή Χαρακτηριστικών

Έχοντας ως στόχο να μειώσουμε τις αρχικές διαστάσεις του προβλήματος που καλούμαστε να αντιμετωπίσουμε, υπάρχουν δύο προσεγγίσεις για να το επιτύχουμε: η εξαγωγή χαρακτηριστικών και η επιλογή των χαρακτηριστικών (feature selection)[2]. Με την εξαγωγή χαρακτηριστικών επιδιώκεται η δημιουργία ενός υποσυνόλου από νέα χαρακτηριστικά, τα οποία είναι συνδυασμοί των ήδη υπαρχόντων.

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \xrightarrow{feature\ extraction} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix} = f(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix})$$

Δοσμένου ενός σετ χαρακτηριστικών $x_i \in R^N$ η εξαγωγή χαρακτηριστικών βρίσκει μια αντιστοίχιση $y = f(x) : R^N \to R^M$ με $M < N$ έτσι ώστε το μετασχηματισμένο διάνυσμα χαρακτηριστικών $y_i \in R^M$ να διατηρεί όσο το δυνατόν περισσότερη πληροφορία ή δομή του R^N. Μια βέλτιστη αντιστοίχιση $y = f(x)$ θα είναι αυτή που θα οδηγεί σε μηδενική αύξηση της ελάχιστης πιθανότητας λάθους.[3] Γενικότερα η βέλτιστη αντιστοίχιση $y = f(x)$ θα είναι μη γραμμική συνάρτηση, όμως δεν υπάρχει συστηματικός τρόπος για να παραχθούν μη γραμμικές συναρτήσεις. Ταυτόχρονα, η επιλογή ενός συγκεκριμένου υποσυνόλου εξαρτάται από το είδος του προβλήματος. Συνεπώς, η εξαγωγή χαρακτηριστικών περιορίζεται τις περισσότερες φορές σε γραμμικούς μετασχηματισμούς της μορφής: $y = Wx$. Οι μέθοδοι εξαγωγής χαρακτηριστικών που βρίσκουν εφαρμογή στο ανθρώπινο πρόσωπο θα αναπτυχθούν στο επόμενο υποκεφάλαιο.

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \xrightarrow{linear\ feature\ extraction} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix} = \begin{bmatrix} W_{11} & W_{12} \dots W_{1N} \\ W_{21} & W_{22} \dots W_{2N} \\ \vdots & \ddots \\ W_{M1} & W_{M2} \dots W_{MN} \end{bmatrix} * \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}$$

[2] Το συγκεκριμένο υποκεφάλαιο βασίστηκε κυρίως σε [40]

[3] Ένας κανόνας απόφασης του Bayes που εφαρμόζεται στον αρχικό χώρο R^N και στο μειωμένο χώρο R^M παρουσιάζει ίδια απόδοση κατά την ταξινόμηση.

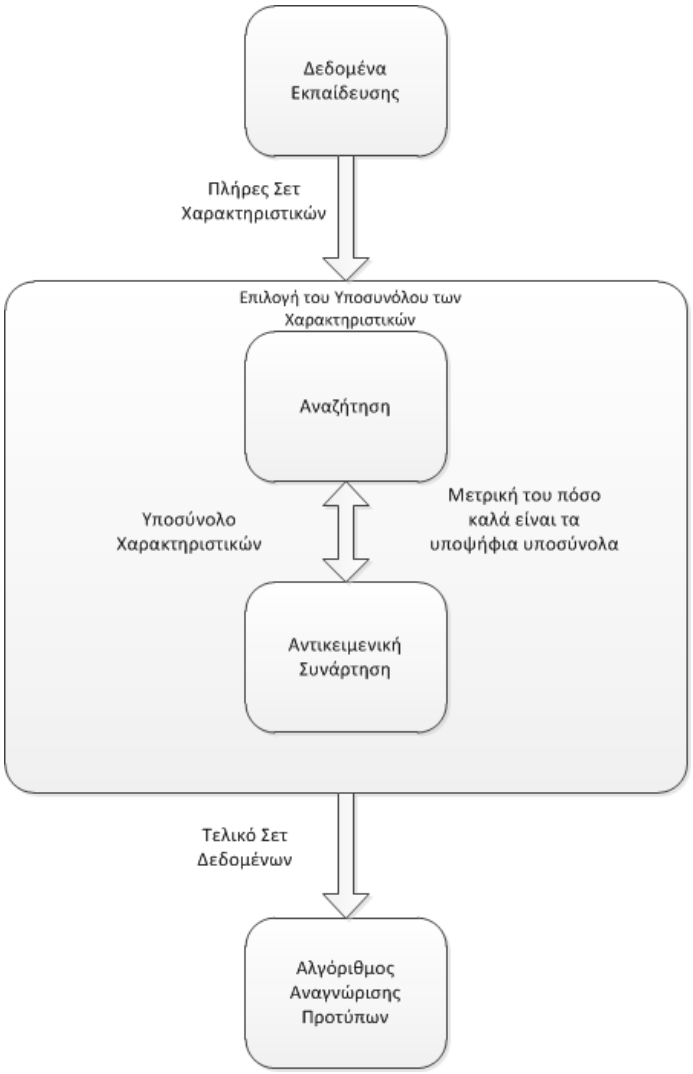
Η επιλογή χαρακτηριστικών απεναντίας, παρόλο που είναι μια ειδική περίπτωση της εξαγωγής χαρακτηριστικών, θεωρείται ως ξεχωριστή καθώς επιλέγει ένα υποσύνολο (τα πιο περιεκτικά σε πληροφορία) από όλα τα χαρακτηριστικά. Στοχεύει στο να βρει ένα υποσύνολο (στη βιβλιογραφία αναφέρεται και σαν subset selection [35])το οποίο να ελαχιστοποιεί κάποια συνάρτηση κόστους.

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \xrightarrow[feature\ selection]{} \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iM} \end{bmatrix}$$

Δοσμένου ενός σετ χαρακτηριστικών $x = \{x_i \mid i = 1...N\}$ η επιλογή των χαρακτηριστικών επιδιώκει στο να βρει το υποσύνολο $x_M = \{x_{i1}, x_{i1}, ..., x_{iM}\}$ με $M < N$ το οποίο να βελτιστοποιεί μια αντικειμενική συνάρτηση $J(Y)$. Η επιλογή της μεθόδου αυτής έναντι της εξαγωγής των χαρακτηριστικών προτιμάται μερικές φορές είτε γιατί μπορεί τα χαρακτηριστικά να είναι υπολογιστικά ακριβά είτε επειδή μπορεί να είναι πολύπλοκα στο να εξαχθούν. Ακόμα, μπορεί να θέλουμε να εξάγουμε κάποιους ουσιαστικούς κανόνες από τον ταξινομητή, καθώς όταν κάνουμε μετασχηματισμό ή προβολή χάνονται μονάδες μετρήσεις όπως πχ το μήκος ή το πλάτος. Τέλος, υπάρχουν περιπτώσεις όπου τα χαρακτηριστικά μπορεί να μην είναι αριθμητικά αλλά να είναι strings. Η υλοποίηση της μεθόδου της επιλογής των χαρακτηριστικών απαιτεί μια στρατηγική αναζήτησης ώστε να επιλεγούν τα υποψήφια υποσύνολα καθώς και μια αντικειμενική συνάρτηση ώστε να αξιολογήσει αυτούς τους υποψήφιους.

Όσον αφορά τη στρατηγική αναζήτησης, πρέπει να επισημανθεί πως η εξαντλητική αξιολόγηση των υποσυνόλων των χαρακτηριστικών περιλαμβάνει $\binom{N}{M}$ συνδυασμούς για μια δεδομένη τιμή του M και 2^N υπολογισμούς αν πρέπει να βελτιστοποιηθεί και το M. Γίνεται εμφανές πως ο αριθμός αυτών των συνδυασμών-υπολογισμών είναι ανέφικτος να πραγματοποιηθεί ακόμα και για κανονικές τιμές των M και N. Χρειάζεται επομένως, μια στρατηγική αναζήτησης η οποία θα καθοδηγεί την επιλογή του υποσυνόλου του χαρακτηριστικών όσο αυτή εξερευνά το χώρο όλων των δυνατών συνδυασμών των χαρακτηριστικών.

Η αντικειμενική συνάρτηση (objective function) αξιολογεί τα υποψήφια υποσύνολα και επιστρέφει μια μετρική του πόσο καλά είναι. Το feedback που μας επιστρέφει χρησιμοποιείται από τη στρατηγική αναζήτησης ώστε να επιλέξει τους νέους υποψήφιους. Μια απλή αντικειμενική συνάρτηση είναι το ποσοστό λάθους αντεπικύρωσης (cross-validation error rate).



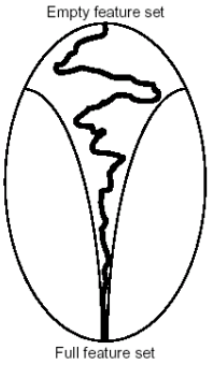
Σχήμα 22: Διάγραμμα Λειτουργίας της Αντικειμενικής Συνάρτησης στην Επιλογή Χαρακτηριστικών

Μια τυπική μέθοδος επιλογής χαρακτηριστικών είναι η ακολουθιακή προς τα εμπρός επιλογή (Sequential Forward Selection (SFS)) η οποία αποτελεί τον απλούστερο άπληστο αλγόριθμο αναζήτησης. Ξεκινώντας από ένα άδειο σετ, προσθέτει ακολουθιακά το χαρακτηριστικό x^+ που έχει ως αποτέλεσμα την υψηλότερη τιμή στην αντικειμενική συνάρτηση $J(Y_k + x^+)$ όταν συνδυάζεται με χαρακτηριστικά Y_k τα οποία έχουν ήδη επιλεγεί.

Αλγόριθμος 2 Αλγόριθμος ακολουθιακής προς τα εμπρός επιλογής χαρακτηριστικών

1: *Ξεκίνα με ένα άδειο σετ* $Y = \{\emptyset\}$.
2: *Επέλεξε το επόμενο καλύτερο χαρακτηριστικό* $x^+ = \arg\max_{x \in X - Y_k} J(Y_k + x)$.
3: *Ανανέωσε τα* $Y_{k+1} = Y_k + x$ *και* $k = k + 1$.
4: *Πήγαινε στο* 2.

Ο αλγόριθμος αυτός λειτουργεί καλύτερα όταν το βέλτιστο υποσύνολο περιέχει ένα μικρό αριθμό από χαρακτηριστικά. Όπως φαίνεται και στο παρακάτω σχήμα όταν η αναζήτηση γίνεται κοντά στο κενό σετ χαρακτηριστικών, μπορεί θεωρητικά να αξιολογηθεί ένας μεγάλος αριθμός από καταστάσεις, ενώ κοντά στο πλήρες σετ η περιοχή που εξετάζει ο αλγόριθμος είναι πολύ πιο στενή καθώς τα περισσότερα από τα χαρακτηριστικά έχουν ήδη επιλεγεί. Ο χώρος αναζήτησης παρουσιάζεται σε σχήμα έλλειψης ώστε να δοθεί έμφαση στο γεγονός ότι υπάρχουν λιγότερες καταστάσεις κοντά στο άδειο και στο γεμάτο σετ. Το βασικό μειονέκτημα του είναι ότι δεν έχει τη δυνατότητα να διαγράψει τα "παλιά" χαρακτηριστικά μετά την πρόσθεση άλλων χαρακτηριστικών.



Σχήμα 23: Αναζήτηση χαρακτηριστικών με τον αλγόριθμο Sequential Forward Selection

4.3 Μέθοδοι Εξαγωγής Χαρακτηριστικών από το Πρόσωπο

Τα ανθρώπινα χαρακτηριστικά του προσώπου παίζουν ένα σημαντικό ρόλο στην αναγνώριση του προσώπου και στη Νευροφυσιολογική έρευνα. Οι Bhumika et al. [8] αναφέρουν πως στις περισσότερες μελέτες, τα μάτια, το στόμα και η μύτη είναι από τα πιο σημαντικά χαρακτηριστικά για αναγνώριση προσώπου. Η εξαγωγή κάποιων χαρακτηριστικών σημείων από το πρόσωπο είναι αντικείμενο χρήσης σε πολλές εφαρμογές όπως η ανίχνευση και η αναγνώριση προσώπου, η αναγνώριση έκφρασης κ.α. Παράλληλα επειδή τα συστήματα αυτά χρησιμοποιούν γεωμετρία του χώρου για να διαχωρίσουν τα

χαρακτηριστικά του προσώπου, δε λαμβάνουν υπόψη τους χαρακτηριστικά όπως π.χ. τα μαλλιά. Αρκετά από τα προβλήματα της εξαγωγής χαρακτηριστικών από το πρόσωπο είναι όμοια με αυτά της ανίχνευσης προσώπου που αναφέρθηκαν στο προηγούμενο κεφάλαιο. Και εδώ παράγοντες όπως το μέγεθος του προσώπου μέσα στην εικόνα, η κατεύθυνση στην οποία κοιτά, ο φωτισμός καθώς και η ύπαρξη επιπλέον αντικειμένων στο πρόσωπο όπως γυαλιά δυσχεραίνουν το έργο της εξαγωγής χαρακτηριστικών. Επίσης οι περισσότερες μέθοδοι εξαγωγής χαρακτηριστικών από το πρόσωπο είναι ευαίσθητες σε ορισμένες μη ιδεατές συνθήκες όπως ο φωτισμός, ο θόρυβος αλλά και ο χρωματικός χώρος που χρησιμοποιείται.

4.3.1 Εξαγωγή Χαρακτηριστικών από το Πρόσωπο με Gabor κυματίδια

Οι Zhao et al. [53] αναφέρουν πως τα τοπικά, σε αντίθεση με τα γενικά, χαρακτηριστικά στις εικόνες προσώπου είναι πιο εύρωστα στις παραπάνω πιθανές διαταραχές - μεταβολές και επομένως μια χωρικής-συχνότητας ανάλυση (spatial–frequency analysis) είναι συχνά επιθυμητή ώστε να εξάγει τέτοια χαρακτηριστικά. Με τη χρήση καλών χαρακτηριστικών από τον εντοπισμό χωρικής συχνότητας (space–frequency localization), η ανάλυση με κυματίδια και ειδικότερα με τη χρήση των Gabor wavelets φαίνεται να είναι η βέλτιστη βάση [12] για την εξαγωγή τοπικών χαρακτηριστικών από το πρόσωπο.

Η scaling ιδιότητα του μετασχηματισμού Fourier υποδεικνύει πως μειώνοντας τη διάρκεια ενός σήματος στο πεδίο του χρόνου, αυξάνεται το πλάτος της συχνότητας του εύρους ζώνης στο πεδίο της συχνότητας και αντίστροφα. Ένα τυπικό παράδειγμα είναι το Γκαουσσιανό ζεύγος:

$$G_\sigma(t) = \frac{1}{\sigma * \sqrt{2*\pi}} * \exp\left(\frac{-t^2}{2*\sigma^2}\right) \iff \exp\left(\frac{-\sigma^2\omega^2}{2}\right) = \frac{\sqrt{2*\pi}}{\sigma} * G_{\frac{1}{\sigma}}(\omega)$$

όπου τόσο το σήμα όσο και ο μετασχηματισμός Fourier του είναι Γκαουσσιανές συναρτήσεις των οποίων οι τυπικές αποκλίσεις στο πεδίο του χρόνου και της συχνότητας είναι σ και $\frac{1}{\sigma}$ αντίστοιχα. Το παραπάνω παράδειγμα, υποδεικνύει πως αν αυξηθεί η ανάλυση στο χρόνο ενός σήματος μειώνοντας τη διάρκεια της απόκρισης της συχνότητας $h(t)$, τότε το εύρος ζώνης του αυξάνεται και κατ' επέκταση μειώνεται η ανάλυση στη συχνότητα. Ο Gabor [20] ποσοτικοποίησε αυτή την ιδέα αποδεικνύοντας πως το γινόμενο της διάρκειας και του εύρους ζώνης ενός σήματος δε μπορεί ποτέ να γίνει μικρότερο από ένα γενικό ολικό ελάχιστο και ανακάλυψε επίσης, μια κλάση από σήματα τα οποία επιτυγχάνουν αυτό το ολικό ελάχιστο. Αρχικά ορίζουμε ένα τυχαίο πραγματικό ή μιγαδικό σήμα $f(t)$ με πεπερασμένη ενέργεια $\int_{-\infty}^{\infty} |f(t)|^2 \, dt < \infty$. Επίσης υποθέτουμε πως $\lim_{t\to\pm\infty} |f(t)|\sqrt{|t|} = 0$. Τέλος ορίζουμε κάποιες στιγμές του σήματος $\bar{t^n}$, $n = 1., 2, ...$ στο πεδίο του χρόνου :

$$\bar{t} \equiv \frac{\int t * |f(t)|^2 dt}{\int |f(t)|^2 dt} \ , \ \bar{t^2} \equiv \frac{\int t^2 * |f(t)|^2 dt}{\int |f(t)|^2 dt}$$

με τις αντίστοιχες στιγμές $\bar{\omega^n}$, $n = 1., 2, ...$ στο πεδίο της συχνότητας:

$$\bar{\omega} \equiv \frac{\int \omega * |F(\omega)|^2 d\omega}{\int |F(\omega)|^2 d\omega} \ , \ \bar{\omega^2} \equiv \frac{\int \omega^2 * |F(\omega)|^2 d\omega}{\int |F(\omega)|^2 d\omega}$$

Τότε η rms διάρκεια Δt του σήματος και η rms τιμή του εύρους ζώνης του θα είναι:

$$\Delta t \equiv \sqrt{\overline{(t-\bar{t})^2}} \ , \ \Delta\omega \equiv \sqrt{\overline{(\omega-\bar{\omega})^2}}$$

Ο Gabor απέδειξε πως η rms διάρκεια και το εύρος ζώνης του σήματος $f(t)$ με τις παραπάνω ιδιότητες ικανοποιούν την παρακάτω ανισότητα:

$$\Delta t * \Delta\omega \ \geq \frac{1}{2}$$

Η παραπάνω ανισότητα εκφράζει την αρχή της αβεβαιότητας για τα σήματα μίας διάστασης και του μετασχηματισμού Fourier τους. Ο Gabor επίσης ανακάλυψε ότι υπάρχει μια κλάση από σήματα τα οποία πετυχαίνουν την ελάχιστη τιμή $(\frac{1}{2})$ της αρχής της αβεβαιότητας. Αυτά τα σήματα είναι σύνθετα ημίτονα των οποίων τα πλάτη διαμορφώνονται από Γκαουσσιανές και ονομάζονται Gabor σήματα ή Gabor wavelets:

$$f(t) = A \exp\left[\frac{-(t-t_c)^2}{2\sigma^2}\right] \exp\left[j(\omega_c t + \phi)\right]$$

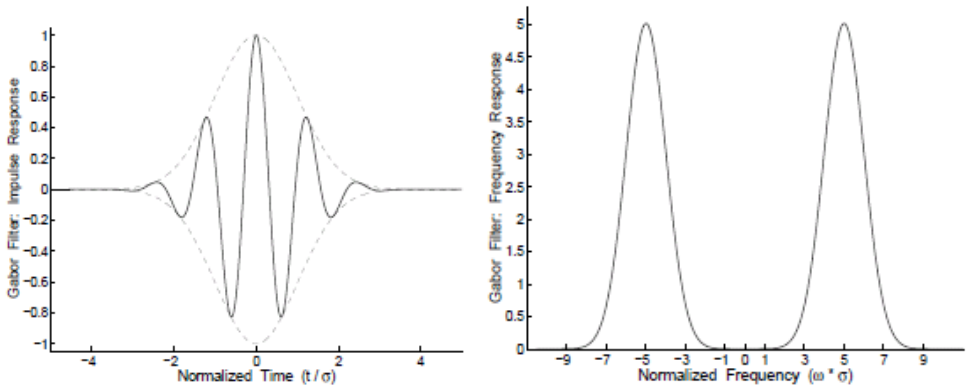
με μετασχηματισμό Fourier:

$$F(\omega) = 2A\pi\sigma\sqrt{2\pi} \exp\left[\frac{-(\omega-\omega_c)^2\sigma^2}{2}\right] \exp\left[j(\omega-\omega_c)t_c + \phi\right]$$

Από αυτές τις πέντε παραμέτρους, οι τρεις σταθερές σ, t_c, A εκφράζουν την τυπική απόκλιση, τη θέση της κορυφής και το ύψος της κορυφής, ενώ οι ω_c, ϕ είναι η συχνότητα και η φάση του διαμορφωμένου ημιτόνου. Για το σύνθετο Gabor κυματίδιο η rms διάρκεια και το εύρος ζώνης είναι:

$$\Delta t = \frac{\sigma}{\sqrt{2}}, \Delta\omega = \frac{1}{\sigma\sqrt{2}}$$

τα οποία, όπως είναι εμφανές, έχουν γινόμενο ίσο με $\frac{1}{2}$. Στο παρακάτω σχήμα παρουσιάζεται η απόκριση πλάτους και η απόκριση συχνότητας ενός πραγματικού Gabor φίλτρου:



Σχήμα 24: Απόκριση πλάτους και απόκρισης συχνότητας ενός πραγματικού Gabor φίλτρου όπου με διακεκομμένη γραμμή παρουσιάζεται ο Gabor φάκελος. (επανεκτύπωση από [34])

Όσον αφορά τις δύο διαστάσεις, θεωρούμε ένα σήμα δύο διαστάσεων $f(x, y)$ και το Fourier μετασχηματισμό του $F(\omega_1, \omega_2)$. Στη συνέχεια θέτουμε x_c, y_c τις χωρικές στιγμές πρώτης τάξης της κατανομής της ενέργειας $|f|^2$ και u, v τις φασματικές στιγμές τις κατανομής της ενέργειας $|F|^2$. Έτσι

τα (x_c, y_c) και (u, v) είναι αντίστοιχα οι μέσες θέσεις της κατανομής της ενέργειας του σήματος στο χώρο και στη συχνότητα αντίστοιχα. Έτσι ορίζονται οι παρακάτω κεντρικές χωρικές και φασματικές αντίστοιχα στιγμές δεύτερης τάξης :

$$(\Delta x)^2 = \frac{\int\int (x - x_c)^2 |f(x,y)|^2 dxdy}{\int\int |f(x,y)|^2 dxdy} \ , \ (\Delta y)^2 = \frac{\int\int (y - y_c)^2 |f(x,y)|^2 dxdy}{\int\int |f(x,y)|^2 dxdy}$$

$$(\Delta \omega_1)^2 = \frac{\int\int (\omega_1 - u)^2 |F(\omega_1,\omega_2)|^2 d\omega_1 d\omega_2}{\int\int |F(\omega_1,\omega_2)|^2 d\omega_1 d\omega_2} \ , \ (\Delta \omega_2)^2 = \frac{\int\int (\omega_2 - v)^2 |F(\omega_1,\omega_2)|^2 d\omega_1 d\omega_2}{\int\int |F(\omega_1,\omega_2)|^2 d\omega_1 d\omega_2}$$

Τα Δx και Δy αναπαριστούν τα πεπερασμένα πλάτη του σήματος γύρω από τη μέση θέση του (x_c, y_c) και είναι ίσα με την τυπικές αποκλίσεις του σήματος σε συνδυασμό με την κατανομή της ενέργειάς του $|f(x,y)|^2$. Ομοίως τα $\Delta\omega_1$ και $\Delta\omega_2$ αναπαριστούν τα πεπερασμένα πλάτη του φάσματος του σήματος γύρω από το μέσο διάνυσμα συχνότητας (u, v).

Ο Daugmann [13] εντόπισε μια εντυπωσιακή ισοδυναμία μεταξύ των δισδιάστατων Gabor συναρτήσεων και των αποκρίσεων των κυττάρων στον οπτικό φλοιό που τον οδήγησε στο να εκφράσει τη δισδιάστατη αβεβαιότητα που παρουσιάζουν τα πεπερασμένα πλάτη στο χώρο και στη συχνότητα:

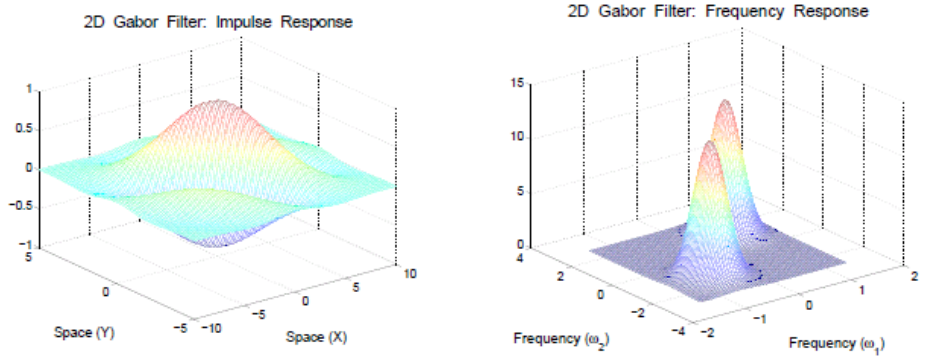
$$(\Delta x)(\Delta y)(\Delta \omega_1)(\Delta \omega_2) \ \geq \ \frac{1}{4}$$

Επιπλέον έδειξε ότι τα δισδιάστατα φίλτρα που επιτυγχάνουν το ελάχιστο γινόμενο χωρικής και φασματικού πλάτους είναι δισδιάστατες επεκτάσεις των σύνθετων μονοδιάστατων Gabor φίλτρων. Μια τυπική μορφή ενός δισδιάστατου Gabor σήματος είναι:

$$f(x,y) = \exp\left[-\frac{(x - x_c)^2}{2\sigma_1^2} - \frac{(y - y_c)^2}{2\sigma_2^2}\right] \exp[ju(x - x_c) + jv(y - y_c)]$$

με τον 2D Fourier μετασχηματισμό να είναι:

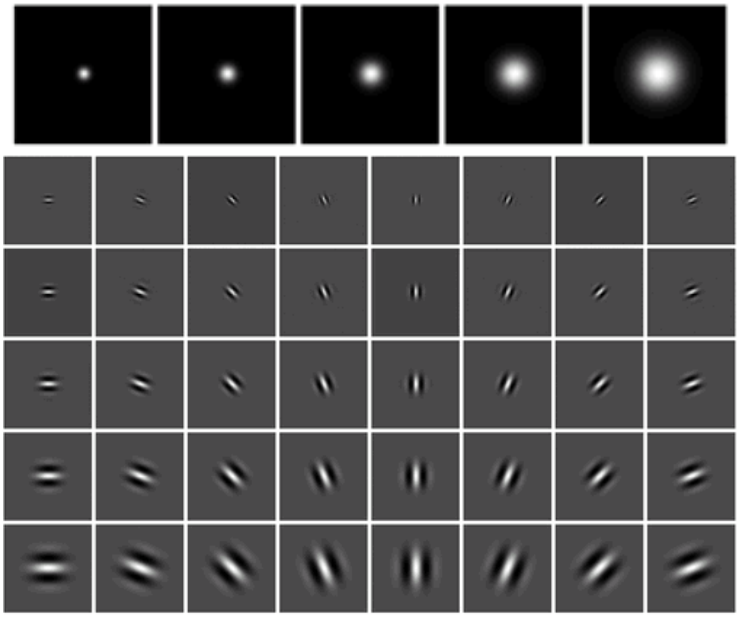
$$F(\omega_1,\omega_2) = 2\pi\sigma_1\sigma_2 \exp\left[-\frac{(\omega_1 - u)^2\sigma_1^2}{2} - \frac{(\omega_2 - v)^2\sigma_2^2}{2}\right] \exp[-j\omega_1 x_c - j\omega_2 y_c]$$



Σχήμα 25: Απόκριση πλάτους και απόκρισης συχνότητας ενός πραγματικού 2D Gabor φιλτρου. (επανεκτύπωση από [34])

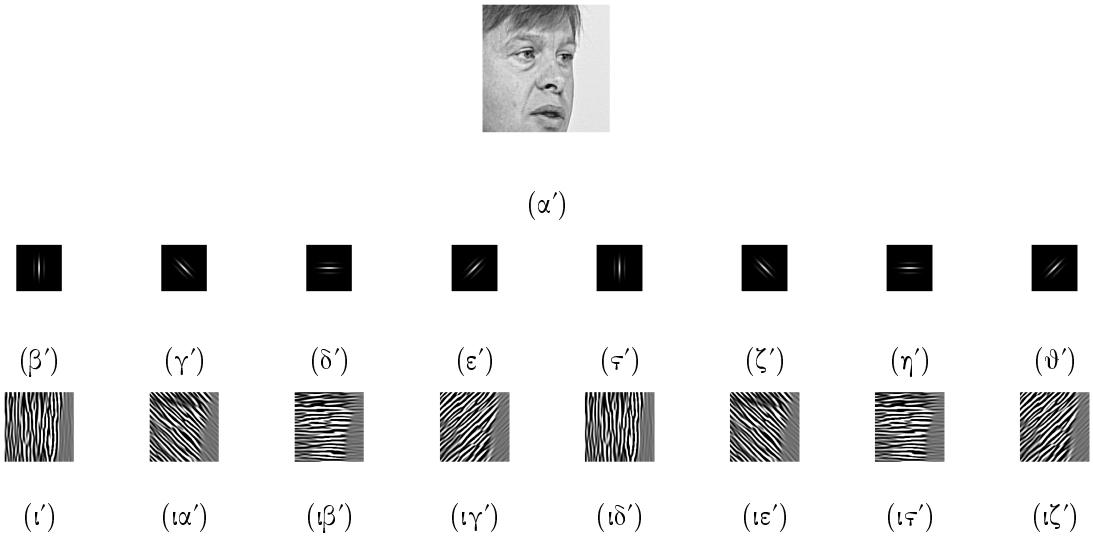
Δεδομένου ότι η τοπική συχνότητα και η κατεύθυνση είναι άγνωστες ποσότητες, απαιτούνται μια οικογένεια από $U \times V$ Gabor wavelets ώστε να εξάγουν τα χαρακτηριστικά από την εικόνα. Τα κυματίδια παρουσιάζουν επιθυμητά χαρακτηριστικά όσον αφορά την χωρική συχνότητα και την επιλογή του προσανατολισμού. Στη βιβλιογραφία [41], χρησιμοποιούνται κατά κόρον πέντε διαφορετικές κλίμακες ($U = 5$) και οκτώ διαφορετικοί προσανατολισμοί ($V = 8$):

$$\theta(v) = \frac{v}{8}\pi \ , \ u = 0, 1, \ldots, U-1 \ , \ v = 0, 1, \ldots, V-1$$



Σχήμα 26: Πλάτη των Gabor wavelets στις 5 διαφορετικές κλίμακες και τα πραγματικά μέρη τους στις ίδιες 5 κλίμακες και σε 8 διαφορετικούς προσανατολισμούς

Η αναπαράσταση των εικόνων με τη χρήση Gabor wavelets πραγματοποιείται με τη συνέλιξη της εικόνας με την οικογένεια των κυματιδίων. Για ένα καρέ από το dataset που διαθέτουμε αφού κάνουμε ανίχνευση προσώπου κρατάμε την παρακάτω εικόνα σε κλίμακα γκρι:



Σχήμα 27: Αρχική εικόνα(α), Gabor κυματίδια(β-ϑ) και αποτελέσματα συνέλιξης(ι-ιζ) με την εικόνα για 1 κλίμακα και 8 κατευθύνσεις

Παρακινούμενοι από την βιολογική ομοιότητα με τον οπτικό φλοιό, η χρήση των Gabor κυματιδίων σχηματίζει μια βέλτιστη βάση για τη μέτρηση των τοπικών χαρακτηριστικών υφής και για την αναπαράσταση των εικόνων. Από τη φύση τους, οι αναπαραστάσεις με Gabor κυματίδια παραμένουν σε ένα βαθμό ανεπηρέαστες σε μεταβολές του φωτισμού και σε τοπικές παραμορφώσεις που οφείλονται στη θέση και την έκφραση του προσώπου καθιστώντας τη χρήση τους μια από τις καλύτερες τεχνικές εξαγωγής χαρακτηριστικών από το πρόσωπο. Παρόλο που η συγκεκριμένη προσέγγιση αναπαράστασης του προσώπου χαρακτηρίζεται για τη διακριτική της ικανότητα και την ευρωστία, πρέπει να επισημανθεί πως ο τρόπος με τον οποίο θα γίνει η επιλογή της βέλτιστης βάσης για τα κυματίδια (δηλαδή η επιλογή του αριθμού των κλιμάκων και των προσανατολισμών) αποτελεί ένα πεδίο που απαιτεί περισσότερη έρευνα.

4.3.2 Αμετάβλητος ως προς την Κλίμακα Μετασχηματισμός Χαρακτηριστικών(SIFT)

Η αντιστοίχιση μεταξύ των εικόνων είναι ένα θεμελιώδες πεδίο έρευνας στην Όραση Υπολογιστών και περιλαμβάνει αναγνώριση αντικειμένων ή σκηνών, καταγραφή της τροχιάς της κίνησης κ.α. Είναι απαραίτητα συνεπώς κάποια χαρακτηριστικά μέσα στην εικόνα που θα έχουν αρκετές ιδιότητες ώστε να τα κάνουν κατάλληλα για αντιστοίχιση διαφορετικών εικόνων ενός αντικειμένου, προσώπου ή σκηνής. Τα χαρακτηριστικά αυτά είναι αμετάβλητα σε αλλαγή της κλίμακας ή περιστροφή της εικόνας αλλά κυρίως αμετάβλητα σε αλλαγές στο φωτισμό και στη λήψη της κάμερας. Ταυτόχρονα έχουν ιδιαίτερη διακριτική ικανότητα κάνοντας εφικτό το σωστό ταίριασμα (με μεγάλη πιθανότητα) ενός χαρακτηριστικού με μια βάση δεδομένων από χαρακτηριστικά. Το κόστος της εξαγωγής τους, ελαχιστοποιείται χρησιμοποιώντας μια προσέγγιση ακολουθιακού φιλτραρίσματος (σαν αυτόν που χρησιμοποιείται για την ανίχνευση προσώπου στον αλγόριθμο των Viola & Jones) κατά την οποία οι πιο πολύπλοκες και υπολογιστικά ακριβές διεργασίες εκτελούνται μόνο στις περιοχές οι οποίες έχουν περάσει τα προηγούμενα τεστ. Ο Lowe [33] χρησιμοποίησε τα ακόλουθα στάδια για την παραγωγή του σετ των χαρακτηριστικών της εικόνας. Αρχικά πραγματοποιείται ανίχνευση ακρότατων στο χώρο κλίμακας ακολουθούμενη από εντοπισμό των θέσεων των keypoints μέσα στην εικόνα. Στη συνέχεια σε κάθε keypoint δίνονται ένας ή περισσότεροι προσανατολισμοί και εν τέλει δημιουργείται ένας περιγραφέας των σημείων αυτών.

Η προσέγγιση αυτή ονομάστηκε από τον Lowe ως Scale-invariant feature transform (SIFT) καθώς μετασχηματίζει τα δεδομένα της εικόνας σε ανεξάρτητες από την κλίμακα συντεταγμένες σχετικές με τα τοπικά χαρακτηριστικά. Μια σημαντική πτυχή αυτής της προσέγγισης είναι ότι παράγει μεγάλο αριθμό από χαρακτηριστικά τα οποία καλύπτουν πυκνά την εικόνα σε ένα πλήρες εύρος από κλίμακες και τοποθεσίες. Μια τυπική εικόνα διαστάσεων 500×500 θα αποδώσει περίπου 2000 σταθερά χαρακτηριστικά (παρόλο που αυτός ο αριθμός εξαρτάται τόσο από το περιεχόμενο της εικόνας όσο και από τις επιλεγόμενες τιμές για διάφορες παραμέτρους).

Για την αναγνώριση προσώπου τα χαρακτηριστικά SIFT εξάγονται πρώτα από το σετ με τις εικόνες που υπάρχουν ως αναφορά και αποθηκεύονται σε μια βάση δεδομένων. Ένα νέο πρόσωπο αντιστοιχίζεται συγκρίνοντας κάθε χαρακτηριστικό της νέας εικόνας με την βάση δεδομένων βρίσκοντας υποψήφια χαρακτηριστικά που ταιριάζουν βασιζόμενοι στην Ευκλείδεια απόσταση των διανυσμάτων χαρακτηριστικών τους. Από το πλήρες σετ ταιριασμάτων, ανιχνεύονται υποσύνολα από keypoints τα οποία συμφωνούν γύρω από το αντικείμενο, τη θέση, την κλίμακα του και τον προσανατολισμό του στη νέα εικόνα ώστε να κρατούν τα καλά ταιριάσματα. Ο προσδιορισμός των clusters πραγματοποιείται γρήγορα μέσω μιας αποδοτικής υλοποίησης με hash πίνακα του γενικευμένου μετασχηματισμού Hough. Κάθε cluster από τρία ή περισσότερα χαρακτηριστικά τα οποία συμφωνούν για ένα αντικείμενο και την πόζα του, υπόκειται σε περαιτέρω λεπτομερή εξακρίβωση. Τελικώς, δοσμένης της ακρίβειας ταιριάσματος και του αριθμού των πιθανών λανθασμένων αντιστοιχίσεων, υπολογίζεται η τελική πιθανότητα ένα συγκεκριμένου σετ από χαρακτηριστικά να υποδεικνύει την πραγματική ύπαρξη ενός αντικειμένου. Αν ένα ταίριασμα αντικειμένων περνάει όλα τα παραπάνω τεστ τότε θεωρείται σαν θετικό αποτέλεσμα με υψηλή πιθανότητα.

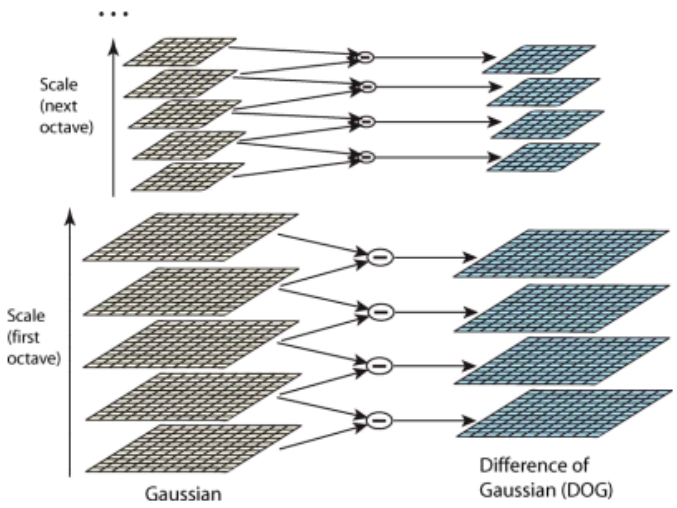
Το πρώτο στάδιο επομένως για την ανίχνευση των SIFT χαρακτηριστικών είναι η ανίχνευση των ακρότατων στο χώρο κλίμακας ανιχνεύοντας τοποθεσίες και κλίμακες οι οποίες μπορούν επαναληλμμένως να αποδίδονται στο ίδιο αντικείμενο υπό διαφορετικές οπτικές γωνίες. Η ανίχνευση από περιοχές που είναι ανεξάρτητες από μεταβολές της κλίμακας μέσα στην εικόνα μπορεί να επιτευχθεί αναζητώντας για σταθερά χαρακτηριστικά μεταξύ όλων των πιθανών κλιμάκων χρησιμοποιώντας μια συνεχή συνάρτηση της κλίμακας που ονομάζεται χώρος κλίμακας. Έχει αποδειχθεί πως η Γκαουσ-

σιανή συνάρτηση είναι το μόνο πιθανό kernel του χώρου κλίμακας με αποτέλεσμα, ο χώρος κλίμακας μιας εικόνας να ορίζεται σαν η συνάρτηση $L(x, y, \sigma)$ που ισούται με τη συνέλιξη μιας μεταβλητής κλίμακας Γκαουσσιανής $G(x, y, \sigma)$ με την εικόνα εισόδου:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y)$$

Για να γίνει αποδοτική η ανίχνευση των σταθερών σημείων των kernels στο χώρο κλίμακας ο Lowe προτείνει αντί για τη Γκαουσσιανή συνάρτηση τη χρήση της διαφοράς δύο Γκαουσσιανών συναρτήσεων μεταξύ δύο γειτονικών κλιμάκων που χωρίζονται από μια πολλαπλασιαστική σταθερά k:

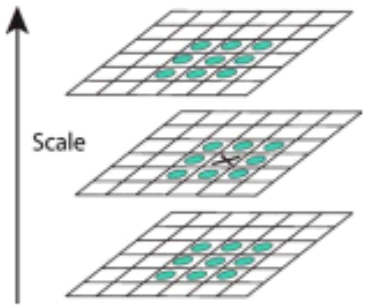
$$D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) = L(x, y, k\sigma) - L(x, y, \sigma)$$



Σχήμα 28: Για κάθε οκτάβα του χώρου κλίμακας, η αρχική εικόνα συνελίσσεται με Γκαουσσιανές ώστε να παραχθεί το σετ από εικόνες στο χώρο κλίμακας στα αριστερά. Αφαιρούνται γειτονικές Γκαουσσιανές εικόνες ώστε να παραχθεί το σετ στα δεξιά. Μετά από κάθε οκτάβα η Γκαουσσιανή εικόνα υποδειγματοληπτείται με συντελεστή δύο και η διαδικασία επαναλαμβάνεται. (επανεκτύπωση από [33])

Για να επιλεγούν τα τοπικά μέγιστα και ελάχιστα, του $D(x, y, \sigma)$, κάθε σημείο συγκρίνεται με τους οκτώ γείτονες του στην τρέχουσα κλίμακα καθώς και με τους εννιά του γείτονες στην παραπάνω και στην παρακάτω κλίμακα. Επιλέγεται το συγκεκριμένο σημείο μόνο αν είναι μεγαλύτερο ή μικρότερο από όλους τους γείτονες. Το κόστος αυτού του ελέγχου είναι αρκετά χαμηλό λόγω του ότι τα περισσότερα σημεία θα απορριφθούν στους πρώτους ελέγχους.

Αφού επιλεγεί το υποψήφιο keypoint συγκρίνοντας ένα pixel με τους γείτονες του, το επόμενο βήμα είναι ένα λεπτομερές ταίριασμα στα γειτονικά δεδομένα που σχετίζονται με τη θέση, την κλίμακα και τον λόγο των κύριων καμπυλοτήτων. Αυτή η πληροφορία καθιστά δυνατή την απόρριψη των σημείων τα οποία έχουν χαμηλή αντίθεση (και συνεπώς είναι ευαίσθητα στο θόρυβο). Αυτό το μεταγενέστερο στάδιο επεξεργασίας είναι ιδιαίτερα σημαντικό ώστε να αυξηθεί η ακρίβεια των εκτιμήσεων της κλίμακας και την κανονικοποίηση της κλίμακας. Όμως, ως προς την ευστάθεια δεν είναι επαρκές να απορρίπτονται keypoints με χαμηλή αντίθεση δεδομένου ότι η συνάρτηση διαφοράς των



Σχήμα 29: Τα μέγιστα και ελάχιστα της διαφοράς των Γκαουσσιανών δύο εικόνων ανιχνεύονται συγκρίνοντας κάθε pixel με τους 26 γείτονές του σε περιοχές διαστάσεων 3x3 στην τρέχουσα και στις γειτονικές κλίμακες.(επανεκτύπωση από [33])

Γκαουσσιανών θα έχει υψηλή απόκριση στις περιοχές των ακμών ακόμα και αν η περιοχή γύρω από μια ακμή είναι ελλιπώς καθορισμένη και συνεπώς ασταθής σε μικρές ποσότητες θορύβου. Μια ελλιπώς ορισμένη κορυφή στη συνάρτηση διαφοράς των Γκαουσσιανών, θα έχει μεγάλη κυρίαρχη καμπυλότητα κατά μήκος της ακμής αλλά μικρή στην κάθετη κατεύθυνση. Οι κυρίαρχες καμπυλότητες μπορούν να υπολογιστούν από έναν 2×2 Hessian πίνακα όπου το H υπολογίζεται στη θέση και την κλίμακα του keypoint.

$$H = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{yx} & D_{yy} \end{bmatrix}$$

με τις παραγώγους να υπολογίζονται λαμβάνοντας τις διαφορές των γειτονικών σημείων. Ενώ οι ιδιοτιμές του H είναι ανάλογες των κυρίαρχων καμπυλοτήτων του D, αποφεύγεται ο υπολογισμός των ιδιοτιμών του H καθώς μας ενδιαφέρει μόνο ο λόγος τους. Αν θέσουμε α την ιδιοτιμή με το μεγαλύτερο πλάτος και β τη μικρότερη θα έχουμε:

$$Tr(H) = D_{xx} + D_{yy} = \alpha + \beta$$
$$Det(H) = D_{xx}D_{yy} - D_{xy}^2 = \alpha\beta$$

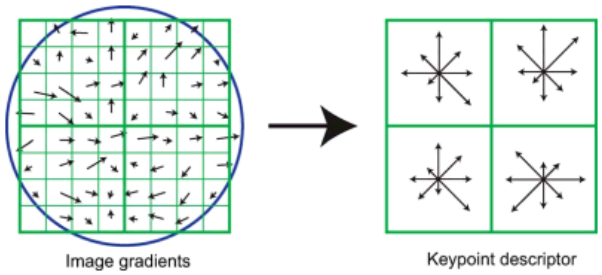
Στη σπάνια περίπτωση, όπου η ορίζουσα θα είναι αρνητική οι καμπυλότητες θα έχουν διαφορετικά πρόσημα έτσι ώστε το σημείο να απορριφθεί και να μην είναι ακρότατο. Αν r είναι ο λόγος μεταξύ της ιδιοτιμής με το μεγαλύτερο πλάτος και της μικρότερης ώστε $\alpha = r\beta$ θα έχουμε:

$$\frac{Tr(H)^2}{Det(H)} = \frac{(\alpha+\beta)^2}{\alpha\beta} = \frac{(r\beta+\beta)^2}{r\beta^2} = \frac{(r+1)^2}{r}$$

Όπως φαίνεται, το τελικό αποτέλεσμα εξαρτάται μόνο από το λόγο των ιδιοτιμών και όχι από τις τιμές που αυτές λαμβάνουν. Η ποσότητα $\frac{(r+1)^2}{r}$ είναι ελάχιστη όταν οι δύο ιδιοτιμές είναι ίσες και αυξάνεται ανάλογα με το r. Επομένως για να ελέγξει κανείς αν ο λόγος των κυρίαρχων καμπυλοτήτων είναι μικρότερος από ένα κατώφλι r, χρειάζεται να ελέγξει μόνο αν:

$$\frac{Tr(H)^2}{Det(H)} < \frac{(r+1)^2}{r}$$

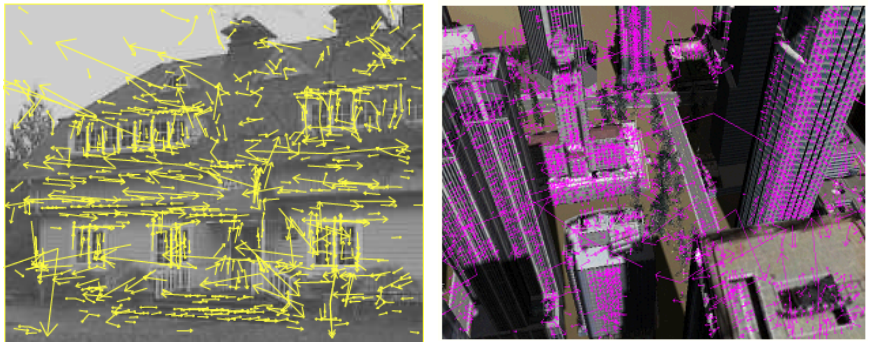
Σε κάθε σημείο ενδιαφέροντος υπολογίζεται ένας περιγραφέας της εικόνας. Ο SIFT περιγραφέας που πρότεινε ο Lowe μπορεί να θεωρηθεί σαν ένα εξαρτώμενο από τη θέση ιστόγραμμα από τοπικές gradient κατευθύνσεις γύρω από το σημείο ενδιαφέροντος. Για να εξαχθεί ένας περιγραφέας ανεξάρτητος της κλίμακας, ο μέγεθος της παραπάνω γειτονικής περιοχής πρέπει να κανονικοποιηθεί με ένα ανεξάρτητο της κλίμακας τρόπο. Αρχικά τα gradient πλάτη και οι κατευθύνσεις δειγματοληπτούνται γύρω από την περιοχή του keypoint χρησιμοποιώντας την κλίμακα του ώστε να επιλεγεί το Γκαουσσιανό θόλωμα της εικόνας. Για να εξασφαλιστεί ανεξαρτησία ως προς την περιστροφή, οι συντεταγμένες του περιγραφέα καθώς και οι κατευθύνσεις του gradient περιστρέφονται ανάλογα με την κατεύθυνση του keypoint. Επιπροσθέτως, χρησιμοποιείται μια Γκαουσσιανή - με βάρη - συνάρτηση με σ ίσο με το μισό του πλάτους του παραθύρου του περιγραφέα ώστε να αναθέσει ένα βάρος στο πλάτος κάθε σημείου. Αυτό φαίνεται στο παρακάτω σχήμα με ένα κυκλικό παράθυρο στο αριστερό μέρος. Ο σκοπός αυτού του Γκαουσσιανού παραθύρου, είναι να αποφευχθούν οι απότομες αλλαγές στον περιγραφέα με μικρές αλλαγές στη θέση του παραθύρου και παράλληλα να δοθεί λιγότερη έμφαση στα gradients τα οποία βρίσκονται μακριά από το κέντρο του περιγραφέα. Ο keypoint περιγραφέας φαίνεται στο δεξί μέρος του σχήματος και επιτρέπει σημαντική μετακίνηση των θέσεων των gradients δημιουργώντας προσανατολισμένα ιστογράμματα πάνω σε 4×4 περιοχές. Το παρακάτω σχήμα παρουσιάζει, οκτώ κατευθύνσεις κάθε κατευθυνόμενο ιστόγραμμα με το μήκος του κάθε τόξου να αντιστοιχεί στο πλάτος του αντίστοιχου ιστογράμματος. Ένα gradient δείγμα στα αριστερά μπορεί να μετακινηθεί μέχρι 4 θέσεις συνεχίζοντας να συνεισφέρει στο ίδιο ιστόγραμμα στα δεξιά, πετυχαίνοντας με αυτόν τον τρόπο το να επιτρέπονται μεγαλύτερες τοπικές μετακινήσεις.



Σχήμα 30: Ένας keypoint περιγραφέας δημιουργείται υπολογίζοντας αρχικά τα gradient πλάτη και τους προσανατολισμούς σε κάθε σημείο της εικόνας σε ένα τμήμα γύρω από την περιοχή του keypoint όπως φαίνεται στα αριστερά. Σε αυτά εφαρμόζεται ένα Γκαουσσιανό παράθυρο όπως υποδηλώνει ο κύκλος και στη συνέχεια προσθέτονται τα δείγματα σε προσανατολισμένα ιστογράμματα συνοψίζοντας τα περιεχόμενα σε 4x4 τμήματα όπως φαίνεται στα δεξιά. Το μήκος κάθε τόξου αντιστοιχεί στο άθροισμα των gradient πλατών κοντά στην κατεύθυνση του τμήματος της εικόνας. Το τελικό αποτέλεσμα είναι ένας 2x2 πίνακας του περιγραφέα που υπολογίζεται σε ένα 8x8 σετ από δείγματα.(επανεκτύπωση από [33])

Συνοψίζοντας, τα SIFT keypoints είναι ιδιαίτερα χρήσιμα εξαιτίας της διακριτικότητας τους, που επιτρέπει το σωστό ταίριασμα ενός σημείου που επιλέγεται με μια μεγάλη βάση δεδομένων από τέτοια σημεία. Η διακριτικότητα αυτή εξασφαλίζεται φτιάχνοντας ένα διάνυσμα πολλών διαστάσεων που αναπαριστά τα gradients της εικόνας μέσα σε μια τοπική περιοχή της. Τα keypoints παρουσιάζουν ανεξαρτησία όσον αφορά την περιστροφή της εικόνας, την κλίμακα καθώς και ευρωστία σε περιπτώσεις όπου αθροίζεται θόρυβος, αλλάζει ο φωτισμός ή υπάρχει αφινική παραμόρφωση. Ταυτόχρονα, ένας μεγάλος αριθμός από keypoints μπορεί να εξαχθεί από μια τυπική εικόνα γεγονός που οδηγεί σε

ευρωστία στην εξαγωγή μικρών αντικειμένων από ένα μη τακτοποιημένο σκηνικό. Η ανίχνευση των keypoints σε ένα πλήρες εύρος από κλίμακες σημαίνει ότι μικρά τοπικά χαρακτηριστικά είναι διαθέσιμα για ταίριασμα μικρών και εν μέρει κρυμμένων αντικειμένων ενώ τα μεγάλα keypoints ανταποκρίνονται καλά για εικόνες με θόρυβο ή θόλωμα. Τέλος ο υπολογισμός τους είναι αποδοτικός έτσι ώστε μερικές χιλιάδες από τέτοια σημεία να μπορούν να εξαχθούν από μια τυπική εικόνα και να επεξεργάζονται σε χρόνο πολύ κοντά στον πραγματικό από έναν Η/Υ.



Σχήμα 31: SIFT χαρακτηριστικά σε εικόνες σπιτιών.

Η χρήση των SIFT χαρακτηριστικών έχει βρει ιδιαίτερο ερευνητικό ενδιαφέρον που σχετίζεται τόσο με αναγνώριση όσο και σε βασιζόμενο στην εικόνα ταίριασμα. Βασιζόμενοι κυρίως σε καλά θεμελιωμένες θεωρητικά πράξεις στο χώρο κλίμακας ή σε προσεγγίσεις αυτών, αυτές οι προσεγγίσεις έχουν επιτρέψει τον εύρωστο υπολογισμό χαρακτηριστικών και περιγραφέων των εικόνων από πραγματικά δεδομένα μέσα στις εικόνας. Μερικά από τα πεδία που βρίσκουν εφαρμογή τα συγκεκριμένα χαρακτηριστικά είναι η αναγνώριση και η κατηγοριοποίηση αντικειμένων, το ταίριασμα σημείων από εικόνες διαφορετικών λήψεων μιας σκηνής καθώς και η Ρομποτική. Ο κλάδος της Ρομποτικής χρησιμοποιεί τα SIFT χαρακτηριστικά ώστε να εντοπίζει τη θέση του ρομπότ σε σχέσεις με ένα σετ γνωστών αναφορών καθώς και για να αναγνωρίζει την ύπαρξη γεωμετρικών σχέσεων στα αντικείμενα του περιβάλλοντος για την καλύτερο χειρισμό του.

4.3.3 Εύρωστος Τοπικός Ανιχνευτής Χαρακτηριστικών(SURF)

Σε αντίθεση με τους παραδοσιακούς αλγορίθμους, με τη χρήση του SIFT γίνεται δυνατή η εξαγωγή εξατομικευμένων χαρακτηριστικών με αποτέλεσμα να χρησιμοποιείται συχνά σαν μέθοδος αναγνώρισης προσώπου. Όμως, εξαιτίας του υψηλού υπολογιστικού του κόστους του ταιριάσματος με SIFT έχουν προταθεί αρκετές μέθοδοι ώστε να επιταχύνουν τη διαδικασία. Για παράδειγμα, χρησιμοποιείται ένα kd δέντρο[4] στο στάδιο της αναζήτησης του k κοντινότερου γείτονα ή γίνεται χρήση της PCA ώστε να μειωθούν οι διαστάσεις των SIFT χαρακτηριστικών. Παρόλα αυτά, αυτές οι μέθοδοι, δε καθιστούν τον SIFT ικανοποιητικά καλό ως προς τις απαιτήσεις της ταχύτητας σε εφαρμογές πραγματικού χρόνου. Ο Bay [5] πρότεινε ένα καινούργιο ανιχνευτή και περιγραφέα που τον ονόμασε SURF (Speeded-Up Robust Features), ο οποίος παραμένει αμετάβλητος σε μεταβολές κλίμακας ή σε in-plane περιστροφές με συγκρίσιμη ή ακόμα και καλύτερη απόδοση από τον SIFT. Όπως και ο SIFT, οι SURF ανιχνευτές χρησιμοποιούνται αρχικά για να βρουν τα σημεία ενδιαφέροντος μέσα σε μια εικόνα και στη συνέχεια οι περιγραφείς εξάγουν τα διανύσματα χαρακτηριστικών από κάθε σημείο ενδιαφέροντος. Ωστόσο, αντί του φίλτρου της διαφοράς των Γκαουσσιανών, ο SURF χρησιμοποιεί μια προσέγγιση του Hessian πίνακα στην ολοκληρωτική εικόνα ώστε να εντοπίσει τα σημεία ενδιαφέροντος μειώνοντας με αυτόν τον τρόπο δραστικά τον υπολογιστικό χρόνο. Όσον αφορά τον περιγραφέα, χρησιμοποιούνται οι αποκρίσεις στον x και τον y άξονα Haar κυματιδίων πρώτης τάξης ώστε να περιγράψουν τη διανομή της έντασης εντός της γειτονιάς ενός σημείου ενδιαφέροντος σε αντίθεση με τον SIFT που κάνει χρήση του gradient. Επιπλέον, είναι σύνηθες να χρησιμοποιούνται μόνο 64 διαστάσεις από τον SURF ώστε να μειώσουν το χρόνο τόσο για τον υπολογισμό των χαρακτηριστικών όσο και για το ταίριασμα. Η ύπαρξη του παραπάνω αριθμού διαστάσεων για κάθε χαρακτηριστικό σε συνδυασμό με το ότι έχει φτιαχτεί ένα indexing σχήμα κάνοντας χρήση του σήματος του ης Λαπλασιανής, καθιστούν τον SURF πολύ γρηγορότερο στο στάδιο του ταιριάσματος από τον SIFT 128 διαστάσεων.

Το πρώτο στάδιο των μεθόδων που χρησιμοποιούν SURF χαρακτηριστικά είναι η ανίχνευση των σημείων ενδιαφέροντος. Σε αντίθεση με τον SIFT που εκμεταλλεύεται τις ιδιότητες της διαφοράς των Γκαουσσιανών, εδώ χρησιμοποιείται η ορίζουσα του προσεγγιστικού Hessian πίνακα σαν βάση για τον ανιχνευτή. Για να εντοπιστεί το σημείο ενδιαφέροντος, ανιχνεύονται δομές που μοιάζουν με μεγάλα δυαδικά αντικείμενα (blobs) σε περιοχές όπου η ορίζουσα λαμβάνει μέγιστες τιμές. Δοσμένου ενός σημείου $x = (x, y)$ σε μια εικόνα I ο Hessian πίνακας στο σημείο x και στην κλίμακα σ δίνεται από τον παρακάτω τύπο:

$$H(x,\sigma) = \begin{bmatrix} L_{xx}(x,\sigma) & L_{xy}(x,\sigma) \\ L_{xy}(x,\sigma) & L_{yy}(x,\sigma) \end{bmatrix}$$

όπου L_{xx}, L_{yy}, L_{xy} οι συνελίξεις των μερικών παραγώγων δεύτερης τάξης της Γκαουσσιανής με την εικόνα I στο σημείο x. Για να μειωθεί ο υπολογιστικός χρόνος, χρησιμοποιείται ένα σετ απο 9×9 τετραγωνικά φίλτρα, καθώς οι προσεγγίσεις μιας Γκαουσσιανής με $\sigma = 1.2$ αποτελεί τη μικρότερη κλίμακα (ή τη μεγαλύτερη χωρική ανάλυση) για τον υπολογισμό των χαρτών αποκρίσεων των blobs και θα συμβολίζονται με $D_{xx}(x,\sigma), D_{yy}(x,\sigma), D_{xy}(x,\sigma)$.

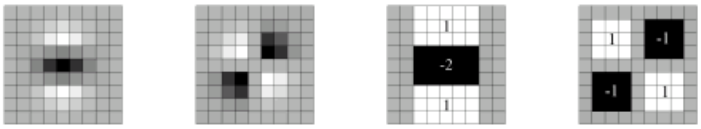
[4]Το kd δέντρο το οποίο αποτελεί συντόμευση για ένα δέντρο k διαστάσεων, είναι μια δομή δεδομένων που χωρίζει το χώρο σε επιμέρους υποσύνολα και χρησιμοποιείται για την οργάνωση των σημείων σε ένα χώρο k διαστάσεων

$$\det(H_{approx}) = D_{xx}D_{yy} - (\omega D_{xy})^2$$

με το ω να συμβολίζει το βάρος της διατήρησης της ενέργειας μεταξύ των Γκαουσσιανών kernels και των προσεγγίσεων τους ενώ ο τύπος με τον οποίο εκφράζεται είναι:

$$\omega = \frac{|L_{xy}(1.2)|_F}{|L_{xy}(1.2)|_F} \frac{|D_{yy}(9)|_F}{|D_{xy}(9)|_F} [5] \approx 0.9$$

Καθώς είναι ανεξάρτητος από την κλίμακα, ο SURF κατασκευάζει ένα χώρο κλίμακας σε σχήμα πυραμίδας όπως και ο SIFT με τη διαφορά ότι αντί να εξομαλύνει επανειλημμένα την εικόνα με μια Γκαουσσιανή και στη συνέχεια να της κάνει υποδειγματοληψία, αλλάζει απευθείας την κλίμακα των τετραγωνικών φίλτρων ώστε να υλοποιήσει το χώρο κλίμακας.



Σχήμα 32: Από αριστερά προς τα δεξιά: οι διακριτοποιημένες και κομμένες μερικές παράγωγοι δεύτερης τάξης της Γκαουσσιανής στην y και xy κατεύθυνση και οι αντίστοιχες προσεγγίσεις με τη χρήση των τετραγωνικών φίλτρων. Οι γκρί περιοχές είναι ίσες με μηδέν.(επανεκτύπωση από [5])

Το δεύτερο στάδιο είναι η περιγραφή των σημείων ενδιαφέροντος, όπου ο SURF χρησιμοποιεί το άθροισμα των αποκρίσεων των Haar κυματιδίων ώστε να περιγράψει ένα χαρακτηριστικό ενός σημείου ενδιαφέροντος. Για την εξαγωγή του περιγραφέα πρωτίστως, κατασκευάζεται μια τετράγωνη περιοχή με κέντρο το σημείο ενδιαφέροντος και προσανατολισμένη κατά τον τρόπο που προτείνεται από τον Bay [5]. Η περιοχή χωρίζεται σε μικρότερες υποπεριοχές σχήματος τετραγώνου και διαστάσεων 4×4 διατηρώντας έτσι τη σημαντική χωρική πληροφορία. Για κάθε υποπεριοχή υπολογίζονται οι αποκρίσεις των Haar κυματιδίων. Ο Bay ονομάζει d_x και d_y τις αποκρίσεις του Haar κυματιδίου στον οριζόντιο και τον κάθετο άξονα αντίστοιχα.

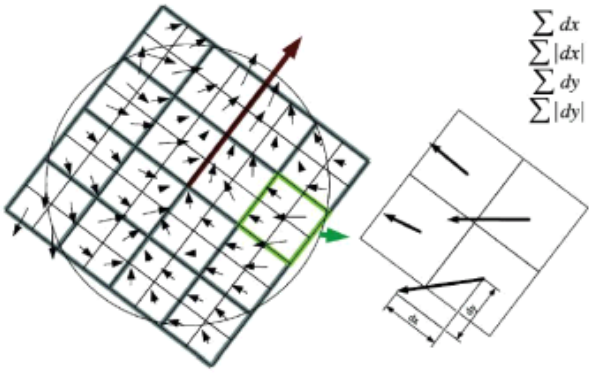
Σχήμα 33: Haar wavelet φίλτρα για την περιγραφή των σημείων ενδιαφέροντος (επανεκτύπωση από [5])

Στη συνέχεια οι κυματιδιακές αποκρίσεις d_x και d_y προστίθενται σε κάθε υποπεριοχή ώστε να σχηματιστεί ένα πρώτο σετ από εισόδους στο διάνυσμα χαρακτηριστικών. Για να εκμεταλλευθεί κανείς πληροφορία για την πολικότητα των αλλαγών της έντασης εξάγονται επίσης και οι απόλυτες τιμές των αθροισμάτων των αποκρίσεων στις δύο διαστάσεις. Κατ' επέκταση, κάθε υποπεριοχή έχει ένα διάνυσμα περιγραφής τεσσάρων διαστάσεων:

[5] $|X|_F$ είναι η Frobenius νόρμα

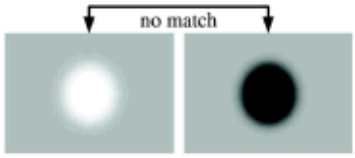
$$v = (\sum dx, \sum dy, \sum |dx|, \sum |dy|)$$

Αν ενωθούν αυτά τα διανύσματα το ένα κάτω από το άλλο για όλες τις 4×4 υποπεριοχές προκύπτει ένα διάνυσμα περιγραφής 64 διαστάσεων.



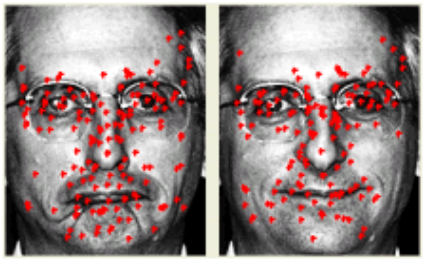
Σχήμα 34: Αναπαράσταση της κατασκευής του περιγραφέα των SURF χαρακτηριστικών (επανεκτύπωση από [5])

Το τελευταίο βήμα που πραγματοποιείται είναι το γρήγορο ταίριασμα των αποτελεσμάτων. Για να επιταχυνθεί η διαδικασία του ταιριάσματος γίνεται χρήση του πρόσημου της Λαπλασιανής (το ίχνος του Hessian πίνακα) για το σημείο ενδιαφέροντος. Μόνο τα ζευγάρια σημείων με το ίδιο πρόσημο αντιστοιχίζονται σε χαρακτηριστικά.



Σχήμα 35: Παραδείγματα από blobs του πρόσημου για γρήγορο ταίριασμα(επανεκτύπωση από [5])

Όμοια με τα SIFT χαρακτηριστικά που χρησιμοποιούνται για την αναγνώριση προσώπου, τα SURF χαρακτηριστικά εξάγονται από εικόνες μέσω των SURF ανιχνευτών και περιγραφέων. Αρχικά εξάγονται τα σημεία ενδιαφέροντος (από τριάντα μέχρι εκατό σε αριθμό) από κάθε εικόνα αφού έχει προηγηθεί κάποια προ-επεξεργασία όπως κανονικοποίηση ή εξίσωση ιστογραμμάτων. Κατόπιν με στόχο την περιγραφή της εικόνας, υπολογίζονται τα αντίστοιχα διανύσματα χαρακτηριστικών, του σετ των σημείων ενδιαφέροντος, τα οποία στη συνέχεια κανονικοποιούνται στη μονάδα. Αυτά τα χαρακτηριστικά εξαρτώνται από το πρόσωπο που βρίσκεται μέσα στην εικόνα καθώς ο αριθμός και οι θέσεις των σημείων που επιλέγονται από τον ανιχνευτή καθώς και τα χαρακτηριστικά γύρω από αυτά τα σημεία που υπολογίζονται από τον περιγραφέα δεν είναι ίδια σε κάθε περίπτωση.



Σχήμα 36: Σημεία ενδιαφέροντος για εξαγωγή χαρακτηριστικών σε δύο διαφορετικές εικόνες του ίδιου προσώπου

Το ταίριασμα των σημείων στην αναγνώριση προσώπου γίνεται συνήθως με τον SIFT. Για να αυξηθεί όμως η ταχύτητα του ταιριάσματος και η ευρωστία προτείνεται συχνά η χρήση των SURF χαρακτηριστικών σε συνδυασμό με κάποιους γεωμετρικούς περιορισμούς. Επειδή όμως η Ημερολογιοποίηση Ομιλητών δεν κάνει αναγνώριση ομιλητή και συνεπώς το ερευνητικό πεδίο της Αναγνώρισης Προσώπου δεν συνδέεται άμεσα με το πρόβλημα που καλούμαστε να αντιμετωπίσουμε παραπέμπουμε τον αναγνώστη στο Παράρτημα Α όπου γίνεται μια αναφορά μεθόδων αναγνώρισης προσώπου.

4.4 Ημερολογιοποίηση Ομιλητών και Εξαγωγή Χαρακτηριστικών

Έχοντας αναλύσει κάποιες τεχνικές εξαγωγής χαρακτηριστικών που χρησιμοποιούνται στη βιβλιογραφία καταλήξαμε στη χρήση των Gabor κυματιδίων για να εξάγουμε χαρακτηριστικά από το πρόσωπο. Οι ιδιότητες και τα πλεονεκτήματα της παραπάνω μεθόδου συνοψίζονται αρχικά στην ομοιότητα που παρουσιάζουν με τον οπτικό φλοιό του ανθρώπου και στη συνέχεια στο ότι παραμένουν ανεπηρέαστα σε μεγάλο βαθμό από αλλαγές φωτισμού, προσανατολισμού και έκφρασης του προσώπου. Επομένως μέσα σε ένα shot, σε κάθε καρέ που έχουμε διαβάσει και έχουμε εφαρμόσει αναγνώριση προσώπου (αν δεν έχουμε βρει πρόσωπο προχωράμε στο επόμενο καρέ), πραγματοποιούμε στη συνέχεια εξαγωγή χαρακτηριστικών από το πρόσωπο τα οποία τα τοποθετούμε το ένα κάτω από το άλλο σε ένα πίνακα του οποίου οι τελικές διαστάσεις θα είναι: $\#FramesRead \times 2704$ όπου 2704 είναι ο αριθμός των χαρακτηριστικών που εξάγεται για κάθε καρέ. Η εξαγωγή των χαρακτηριστικών από το πρόσωπο εκτός του ότι μας δίνει μια πιο περιεκτική πληροφορία λιγότερων διαστάσεων σε σχέση με την επεξεργασία όλης της εικόνας, επικεντρώνεται στα χαρακτηριστικά που μας ενδιαφέρουν για την Ημερολογιοποίηση Ομιλητών καθώς η συμπεριφορά τους θα μας βοηθήσει αρχικά να τα ομαδοποιήσουμε σε κάποια γκρουπ (συστάδες) καθένα από τα οποία θα αντιστοιχούν σε ένα μοναδικό πρόσωπο. Ταυτόχρονα κρατώντας τα χαρακτηριστικά που αντιστοιχούν στο κάτω μέρος του προσώπου και αναφέρονται στην περιοχή του στόματος, θα μπορέσουμε να ανιχνεύσουμε την κίνηση των χειλιών του προσώπου που φαίνεται στο τρέχων καρέ ώστε να διασαφηνιστεί αν αυτό μιλάει ή όχι.

Κεφάλαιο 5

Μείωση των Διαστάσεων και Ομαδοποίηση

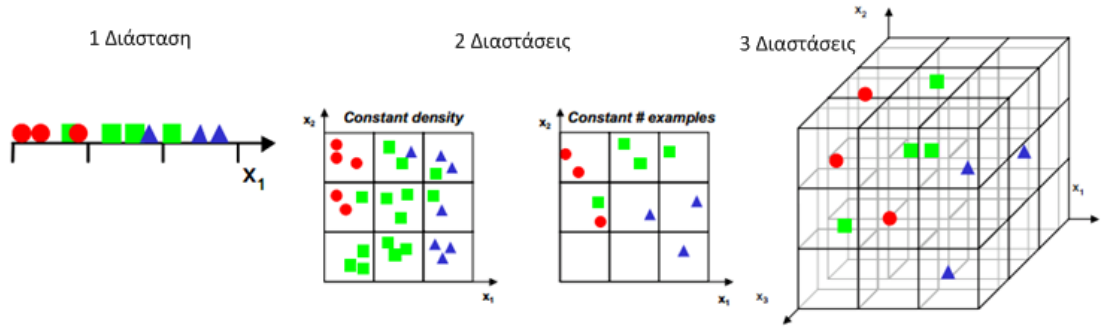
5.1 Μείωση των Διαστάσεων

Η ανάπτυξη που έχει πραγματοποιηθεί τις τελευταίες δεκαετίες στη συλλογή δεδομένων καθώς και στις δυνατότητες που υπάρχουν ώστε αυτά να αποθηκεύονται, έχει οδηγήσει στη συσσώρευση υπερβολικής πληροφορίας στις περισσότερες επιστήμες. Οι ερευνητές που δουλεύουν σε τομείς όπως η αστρονομία, η βιολογία, τα οικονομικά καθώς και οι μηχανικοί, καλούνται να αντιμετωπίσουν όλο και μεγαλύτερο αριθμό παρατηρήσεων και προσομοιώσεων σε καθημερινή βάση. Τέτοιες βάσεις δεδομένων, σε αντίθεση με μικρότερες πιο παραδοσιακές που έχουν μελετηθεί σε βάθος στο παρελθόν, δημιουργούν νέες προκλήσεις στον τομέα της ανάλυσης δεδομένων. Οι παραδοσιακές στατιστικές μέθοδοι αποτυγχάνουν εν μέρει εξαιτίας της αύξησης του αριθμού των παρατηρήσεων, αλλά κυρίως λόγω της αύξησης του αριθμού των μεταβλητών που σχετίζονται με κάθε παρατήρηση. Ο Fodor [17] ορίζει ως διαστάσεις των δεδομένων, τον αριθμό των μεταβλητών οι οποίες μετρώνται για κάθε παρατήρηση. Ένα από τα προβλήματα των βάσεων δεδομένων με πολλές διαστάσεις είναι ότι, σε πολλές περιπτώσεις, δεν είναι απαραίτητες όλες οι μεταβλητές που έχουν μετρηθεί για να γίνουν κατανοητές οι συμπεριφορές που μας ενδιαφέρουν. Ενώ κάποιες υπολογιστικά ακριβές μέθοδοι της βιβλιογραφίας μπορούν να δημιουργήσουν προβλεπτικά μοντέλα μεγάλης ακρίβειας από δεδομένα πολλών διαστάσεων, η μείωση των διαστάσεων των αρχικών δεδομένων πριν τη μοντελοποίηση τους, βρίσκει εφαρμογή σε πληθώρα εφαρμογών.

Συνεπώς το πρόβλημα της μείωσης των διαστάσεων θα μπορούσε να αποτυπωθεί: Δοσμένης μιας τυχαίας μεταβλητής p διαστάσεων $x = (x_1, \ldots, x_p)^T$ προκύπτει μια αναπαράσταση της λιγότερων διαστάσεων, $s = (s_1, \ldots, s_k)^T$ με $k \leq p$ η οποία "αντιλαμβάνεται" το περιεχόμενο των αρχικών δεδομένων σύμφωνα με κάποιο κριτήριο. Πριν γίνει αναφορά στις βασικότερες μεθόδους που εφαρμόζονται για μείωση των διαστάσεων των δεδομένων θα πρέπει να επισημανθεί το πρόβλημα του curse of dimensionality ώστε να γίνει και από μια άλλη σκοπιά κατανοητή στον αναγνώστη η ανάγκη μείωσης των διαστάσεων.

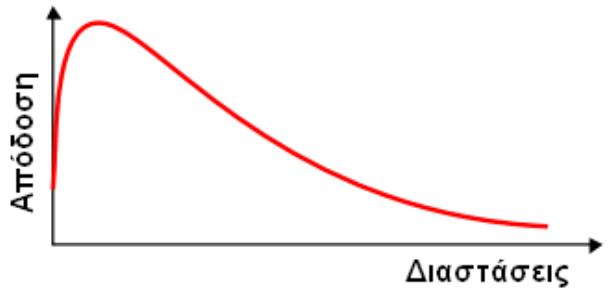
5.1.1 Η κατάρα της διαστατικότητας

Για ένα απλό πρόβλημα Αναγνώρισης Προτύπων τριών κλάσεων, είναι επιθυμητό να κατηγοριοποιηθούν τρία είδη από αντικείμενα ανάλογα με την τιμή ενός χαρακτηριστικού. Μια απλή διαδικασία θα ήταν, ο χωρισμός του χώρου των χαρακτηριστικών σε ομοειδή bins, ο υπολογισμός του λόγου των παραδειγμάτων που αντιστοιχούν σε κάθε κλάση του κάθε bin και συνεπώς για κάθε νέο παράδειγμα να αναζητάται το bin του και να επιλέγεται η κλάση του μέσα σε αυτό. Ξεκινώντας με ένα μόνο χαρακτηριστικό, χωρίζεται η γραμμή σε 3 bins. Επειδή όπως είναι εμφανές και από το παρακάτω σχήμα στη μία διάσταση υπάρχει αρκετή επικάλυψη μεταξύ των κλάσεων εισάγεται ένα ακόμα χαρακτηριστικό. Στις δύο διαστάσεις ο αριθμός των bins αυξάνεται από 3 σε $3^2 = 9$. Αν διατηρηθεί σταθερή η πυκνότητα των παραδειγμάτων μέσα σε κάθε bin θα έχει ως αποτέλεσμα την αύξηση των παραδειγμάτων από 9 σε 27. Από την άλλη, αν μείνει σταθερός ο συνολικός αριθμός των παραδειγμάτων, θα προκύψει μια γραφική παράσταση διασποράς σε δύο διαστάσεις που θα είναι αραιή. Προχωρώντας στις τρεις διαστάσεις τα bins γίνονται 27, για να διατηρηθεί σταθερή η πυκνότητα των παραδειγμάτων θα πρέπει να αυξηθούν τα παραδείγματα σε 81 ενώ για τον ίδιο αριθμό από παραδείγματα η διασπορά στις τρεις διαστάσεις θα είναι σχεδόν άδεια.



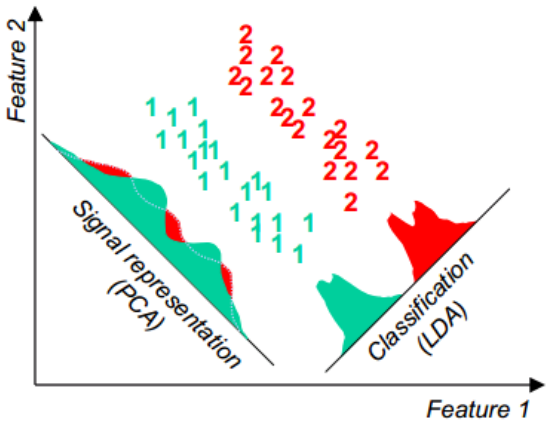
Σχήμα 37: Παράδειγμα κατηγοριοποίησης τριών ειδών δεδομένων σε 1,2 και 3 διαστάσεις

Η Η κατάρα της διαστατικότητας (curse of dimensionality) αναφέρεται σε προβλήματα που εμφανίζονται σε πολυμεταβλητά δεδομένα, καθώς αυξάνεται ο αριθμός των διαστάσεων. Δοσμένου ενός συγκεκριμένου αριθμού δεδομένων, υπάρχει ένας μέγιστος αριθμός από χαρακτηριστικά πάνω από τον οποίο η απόδοση του ταξινομητή φθίνει αντί να αυξάνεται. Στις περισσότερες περιπτώσεις, η πληροφορία που χάνεται πετώντας ορισμένα από τα χαρακτηριστικά αντισταθμίζεται από μια πιο ακριβή αντιστοίχιση στο χώρο των διαστάσεων. Η εξάλειψη του παραπάνω φαινομένου μπορεί να γίνει τόσο μειώνοντας τις διαστάσεις όσο και ενσωματώνοντας πρότερη πληροφορία ή πραγματοποιώντας εξομάλυνση στην τελική συνάρτηση.



Σχήμα 38: Η απόδοση του ταξινομητή αυξάνεται μέχρι ένα συγκεκριμένο αριθμό χαρακτηριστικών και στη συνέχεια μειώνεται καθώς τα χαρακτηριστικά συνεχίζουν να αυξάνονται

Μια αρχική μέθοδος που μειώνει τον αριθμό των διαστάσεων έχει αναφερθεί στο προηγούμενο κεφάλαιο και είναι η εξαγωγή των χαρακτηριστικών από την εικόνα (υποσύνολο της οποίας είναι και η επιλογή των χαρακτηριστικών) που αποσκοπεί στην εύρεση μιας αναπαράστασης των δεδομένων σε λιγότερες διαστάσεις η οποία να διατηρεί όσο το δυνατόν περισσότερη από την πληροφορία της δομής τους. Για την εύρεση της βέλτιστης αντιστοίχισης $y = f(x)$ της εξαγωγής χαρακτηριστικών χρησιμοποιούνται δύο διαφορετικά κριτήρια. Το πρώτο αφορά την αναπαράσταση του σήματος όπου στόχος της εξαγωγής χαρακτηριστικών είναι η ακριβής αναπαράσταση των δειγμάτων σε ένα χώρο μειωμένων διαστάσεων, ενώ το δεύτερο επικεντρώνεται στην ταξινόμηση (classification) και στοχεύει στην ενίσχυση της διακριτικής πληροφορίας μεταξύ των κλάσεων στο χώρο μειωμένων διαστάσεων. Οι δύο βασικότερες τεχνικές που βασίζονται στα προηγούμενα κριτήρια είναι η PCA (Principal Components Analysis) που βασίζεται στο πρώτο και η FLD (Fisher's Linear Discriminant) που βασίζεται στο δεύτερο κριτήριο (λέγεται επίσης και LDA).



Σχήμα 39: Δύο διαφορετικά κριτήρια μείωσης των διαστάσεων των χαρακτηριστικών

Οι δύο αυτές μέθοδοι σε συνδυασμό με την τεχνική της Τυχαίας Προβολής που χρησιμοποιείται σαν στάδιο προ-επεξεργασίας θα αναπτυχθούν στη συνέχεια. Πρέπει να επισημανθεί πως στη βιβλιογραφία υπάρχει μια μεγάλη γκάμα μεθόδων μείωσης των διαστάσεων όπως η ICA, η Παραγοντική Ανάλυση (Factor Analysis) κ.α. από τις γραμμικές καθώς και οι Κύριες Καμπύλες τα Νευρωνικά Δίκτυα κ.α. από μη γραμμικές. Μια καλή επισκόπηση των μεθόδων μείωσης των διαστάσεων γίνεται από τον Fodor [17].

5.1.2 Ανάλυση σε κύριες συνιστώσες (PCA)

Η ανάλυση σε κύριες συνιστώσες PCA είναι ένας τρόπος αναγνώρισης κάποιων προτύπων στα δεδομένα καθώς επίσης και η έκφρασή τους με τέτοιο τρόπο ώστε να γίνουν εμφανείς οι ομοιότητες και οι διαφορές τους. Καθώς η εύρεση προτύπων μπορεί να είναι μια δύσκολη διαδικασία σε δεδομένα πολλών διαστάσεων όπου η πολυτέλεια της γραφικής αναπαράστασης δεν είναι υπαρκτή, η PCA αποτελεί ένα ισχυρό εργαλείο ανάλυσης τους. Το άλλο σημαντικό πλεονέκτημά της κατά τον Smith [43], το οποίο βρίσκει εφαρμογή στη συμπίεση εικόνων, είναι ότι άπαξ και βρεθούν τα πρότυπα στα δεδομένα, στη συνέχεια μπορούν να συμπιεστούν (μειώνοντας τον αριθμό των διαστάσεων) χωρίς μεγάλη απώλεια πληροφορίας.

Η συγκεκριμένη τεχνική βασίζεται στον πίνακα συνδιακύμανσης των μεταβλητών και είναι η καλύτερη γραμμική μέθοδος (δεύτερης τάξης) μείωσης των διαστάσεων όσον αφορά το μέσο τετραγωνικό λάθος [28]. Στόχος της είναι να μειώσει τις διαστάσεις των δεδομένων βρίσκοντας λίγους ορθογώνιους γραμμικούς συνδυασμούς (κύριες συνιστώσες) από τις αρχικές μεταβλητές με τη μεγαλύτερη διασπορά. Θα έχουμε επομένως $s_1 = x^T w_1$ όπου το p διαστάσεων διάνυσμα συντελεστών $w_1 = (w_{1,1}, \ldots, w_{1,p})^T$ επιλύει τη σχέση:

$$w_1 = \underset{\|w=1\|}{\arg\max}\, Var\{x^T w\}$$

Η δεύτερη κύρια συνιστώσα, θα είναι ο γραμμικός συνδυασμός με τη δεύτερη μεγαλύτερη διασπορά ενώ θα είναι ταυτόχρονα ορθογωνικό με την πρώτη κύρια συνιστώσα, με τη διαδικασία αυτή να συνεχίζεται επαναλαμβανόμενα. Υπάρχουν τόσες κύριες συνιστώσες όσες και ο αριθμός των αρχικών μεταβλητών. Για πολλά σετ δεδομένων, οι πρώτες κύριες συνιστώσες εκφράζουν το μεγαλύτερο μέρος της διασποράς έτσι ώστε οι υπόλοιπες να μπορούν να απαλειφθούν με ελάχιστη απώλεια πληροφορίας.

Δεδομένου ότι η διασπορά εξαρτάται από την κλίμακα των μεταβλητών, είναι σύνηθες, κάθε μεταβλητή να τυποποιείται, ώστε να έχει μηδενική μέση τιμή και τυπική απόκλιση ίση με μονάδα. έτσι οι αρχικές μεταβλητές, που μπορεί να έχουν διαφορετικές μονάδες μέτρησης, είναι όλες σε συγκρίσιμες μονάδες. Για τυποποιημένα δεδομένα, με πίνακα συνδιακύμανσης:

$$\Sigma_{p\times p} = \frac{1}{n} X X^T$$

όπου $X = \{x_{i,j} : 1 \leq i \leq p, 1 \leq j \leq n\}$ ο πίνακας των παρατηρήσεων. Χρησιμοποιώντας το θεώρημα της φασματικής αποσύνθεσης το Σ γράφεται:

$$\Sigma = U \Lambda U^T$$

με το $\Lambda = diag(\lambda_1, \ldots, \lambda_p)$ να είναι ο διαγώνιος πίνακας των ζητούμενων ιδιοτιμών $\lambda_1 \leq \cdots \leq \lambda_p$ ενώ το U είναι ένας $p \times p$ ορθογωνικός πίνακας που περιέχει τα ιδιοδιανύσματα. Οι κύριες συνιστώσες δίνονται από τις p γραμμές του $p \times p$ ορθογωνικού πίνακα S όπου $S = U^T X$

Η ερμηνεία των κυρίων συνιστωσών δεν είναι πάντα εύκολη. Παρόλο που πρόκειται για ασυσχέτιστες μεταβλητές που έχουν κατασκευαστεί σαν γραμμικοί συνδυασμοί των αρχικών μεταβλητών με τις επιθυμητές ιδιότητες, δεν αντιστοιχούν απαραίτητα σε ουσιαστικές φυσικές ποσότητες. Σε κάποιες περιπτώσεις η απώλεια της δυνατότητας ερμηνείας των αποτελεσμάτων δεν είναι ικανοποιητική για τους ερευνητές. Ένας εναλλακτικός τρόπος μείωσης των διαστάσεων σε ένα σετ δεδομένων με τη βοήθεια της PCA είναι αντί να χρησιμοποιούνται οι κύριες συνιστώσες σαν νέες μεταβλητές, να χρησιμοποιείται η πληροφορία των κύριων συνιστωσών, με στόχο να βρεθούν σημαντικές και ουσιαστικές μεταβλητές στο αρχικό σετ δεδομένων. Όπως και πριν, υπολογίζονται τα κύρια συστατικά της scree[6] γραφικής παράστασης ώστε να αποφασιστεί ο αριθμός των k πιο σημαντικών μεταβλητών που θα κρατηθούν. Εν συνεχεία, εντοπίζεται το ιδιοδιάνυσμα που αντιστοιχεί στη μικρότερη ιδιοτιμή (η λιγότερο σημαντική κύρια συνιστώσα) και αποβάλλεται η μεταβλητή που αντιστοιχεί στο συντελεστή με τη μεγαλύτερη απόλυτη τιμή με τη διαδικασία αυτή να επαναλαμβάνεται μέχρι να μείνουν k στον αριθμό μεταβλητές.

5.1.3 Γραμμική Διαχωριστική Ανάλυση (LDA)

Η συγκεκριμένη τεχνική της αναγνώρισης προτύπων για μείωση των διαστάσεων πρωτοαναπτύχθηκε από τον Robert Fisher το 1936 και έχει βρει μεγάλη εφαρμογή τόσο στην Όραση Υπολογιστών όσο και στην Αναγνώριση Προτύπων. Στη γενικευμένη περίπτωση που θα έχουμε K κλάσεις, ο αριθμός των διαστάσεων D του χώρου στην είσοδο θα πρέπει να είναι μεγαλύτερος από τον αριθμό K των κλάσεων[7]. Στη συνέχεια θεωρούνται $D' > 1$ γραμμικά χαρακτηριστικά $y_k = w_k^T x$, $k = 1 \ldots D'$. Αυτές οι τιμές των χαρακτηριστικών μπορούν χωρίς δυσκολία να ομαδοποιηθούν μαζί ώστε να σχηματιστεί ένα διάνυσμα y. Ομοίως, τα - με βάρη - διανύσματα w_k μπορούν να θεωρηθούν ως οι στήλες του πίνακα W ώστε να ισχύει:

$$y = W^T x$$

[6] Μια scree γραφική είναι ένα απλό ευθύγραμμο τμήμα που παρουσιάζει το λόγο της συνολικής διασποράς στα δεδομένα όπως την αναπαριστά κάθε κύριο χαρακτηριστικό. Τα κύρια συστατικά είναι ταξινομημένα και αριθμημένα κατά φθίνουσα σειρά ανάλογα με τη συνεισφορά τους στη συνολική διασπορά. Έτσι η γραφική παράσταση όταν διαβάζεται από αριστερά προς τα δεξιά μπορεί να παρουσιάσει ξεκάθαρα το διαχωριστικό εκείνο σημείο, στο οποίο, τα πιο σημαντικά χαρακτηριστικά σταματούν και ξεκινάνε τα λιγότερα σημαντικά.

[7] Το συγκεκριμένο κεφάλαιο βασίστηκε στο αντίστοιχο υποκεφάλαιο του βιβλίου του Bishop [9]

Η γενίκευση του - εντός της κλάσης - πίνακα συνδιακύμανσης στην περίπτωση των K κλάσεων θα είναι:

$$S_w = \sum_{k=1}^{K} S_k$$

όπου

$$S_k = \sum_{n \in C_k} (x_n - m_k)(x_n - m_k)^T$$

με

$$m_k = \frac{1}{N_k} \sum_{n \in C_k} x_n$$

ενώ το N_k είναι ο αριθμός των προτύπων στην κλάση C_k. Για να βρεθεί μια γενίκευση του - μεταξύ των κλάσεων - πίνακα συνδιακύμανσης, οι Duda & Hart [14] ορίζουν το συνολικό πίνακα συνδιακύμανσης:

$$S_T = \sum_{n=1}^{N} (x_n - m)(x_n - m)^T m = \frac{1}{N} \sum_{n=1}^{N} x_n = \frac{1}{N} \sum_{k=1}^{K} N_k m_k$$

όπου m είναι η μέση τιμή του συνολικού σετ δεδομένων

$$m = \frac{1}{N} \sum_{n=1}^{N} x_n = \frac{1}{N} \sum_{k=1}^{K} N_k m_k$$

και $N = \Sigma_k N_k$ είναι ο συνολικός αριθμός σημείων από δεδομένα. Ο συνολικός πίνακας συνδιακύμανσης αποσυντίθεται στο άθροισμα του - εντός της κλάσης - πίνακα συνδιακύμανσης καθώς και σε έναν επιπλέον πίνακα S_B ο οποίος χρησιμοποιείται ως μετρική της συνδιακύμανσης μεταξύ των κλάσεων και δίνεται από τον τύπο:

$$S_B = \sum_{k=1}^{K} N_k (m_k - m)(m_k - m)^T$$

Συνεπώς θα έχουμε:

$$S_T = S_W + S_B$$

Οι παραπάνω πίνακες συνδιακύμανσης έχουν οριστεί στον αρχικό x χώρο. Μπορούν αντίστοιχα να οριστούν όμοιοι πίνακες στον προβαλλόμενο y χώρο D' διαστάσεων. Θα ισχύει συνεπώς:

$$s_W = \sum_{k=1}^{K} \sum_{n \in C_k} (y_n - \mu_k)(y_n - \mu_k)^T$$

καθώς και

$$s_B = \sum_{k=1}^{K} N_k (\mu_k - \mu)(\mu_k - \mu)^T$$

όπου

$$\mu_k = \frac{1}{N_k} \sum_{n \in C_k} y_n \ , \ \mu = \frac{1}{N} \sum_{k=1}^{K} N_k \mu_k$$

Η παραπάνω διαδικασία εκτελείται με σκοπό την κατασκευή ενός βαθμωτού μεγέθους που θα λαμβάνει μεγάλες τιμές όταν η διακύμανση μεταξύ των κλάσεων είναι μεγάλη και μικρές όταν η διακύμανση εντός της κλάσης είναι μικρή. Ο Fukunaga [19] επισημαίνει πως υπάρχει μια πληθώρα επιλογών όσον αφορά το κριτήριο. Ένα παράδειγμα είναι:

$$J(W) = Tr\{s_W^{-1} s_B\}$$

το οποίο μπορεί να γραφτεί με τη μορφή μιας συνάρτησης του πίνακα προβολής W:

$$J(W) = Tr\{(W s_W W^T)^{-1} (W s_B W^T)\}$$

Οι τιμές για τα βάρη επιλέγονται από τα ιδιοδιανύσματα εκείνα του $S_W^{-1} S_B$ που αντιστοιχούν στις D' μεγαλύτερες ιδιοτιμές. Θα πρέπει να τονιστεί πως ανεξάρτητα από το κριτήριο, το S_B αποτελείται από το άθροισμα K πινάκων καθένας από τους οποίους είναι ένα εξωτερικό γινόμενο δύο διανυσμάτων και συνεπώς βαθμό 1. Επίσης, μόνο $K-1$ από αυτούς τους πίνακες είναι ανεξάρτητοι σαν αποτέλεσμα του περιορισμού που δόθηκε παραπάνω για τη μέση τιμή συνολικού σετ δεδομένων. Έτσι, ο S_B θα έχει βαθμό το πολύ ίσο με $K-1$ και κατ' επέκταση, υπάρχουν το πολύ $K-1$ μη μηδενικές ιδιοτιμές. Αυτό δείχνει ότι η προβολή σε ένα υποχώρο $K-1$ διαστάσεων που επεκτείνεται από τα ιδιοδιανύσματα του S_B δεν μεταβάλλει την τιμή του $J(w)$ και συνεπώς δεν είναι δυνατό να βρεθούν περισσότερα από $K-1$ γραμμικά χαρακτηριστικά [19].

5.1.4 Τυχαία Προβολή

Η μέθοδος της τυχαίας προβολής, είναι ιδιαίτερα απλή σε λογική αλλά είναι ταυτόχρονα μια ιδιαίτερα ισχυρή τεχνική μείωσης των διαστάσεων που εκμεταλλεύεται πίνακες τυχαίας προβολής ώστε να προβάλλει τα δεδομένα σε χώρους μειωμένων διαστάσεων [17]. Τα αρχικά δεδομένα $X \in R^p$ μετασχηματίζονται σε $S \in R^k$ με $p \ll k$ μέσω της σχέσης:

$$S = RX$$

όπου οι στήλες του R είναι υλοποιήσεις ανεξάρτητων και όμοια κατανεμημένων μεταβλητών με μηδενική μέση τιμή και με τέτοια κλίμακα ώστε να έχουν μοναδιαίο μήκος. Η μέθοδος προτάθηκε αρχικά για ομαδοποίηση σε έγγραφα κειμένου όπου η αρχική διάσταση p μπορεί να είναι της τάξης του 6000 και η τελική διάσταση k είναι ακόμα σχετικά μεγάλη (της τάξης του 100). Κάτω από τέτοιες συνθήκες ακόμα και η PCA, η πιο απλή εναλλακτική γραμμική μέθοδος μείωσης των διαστάσεων θα είναι υπολογιστικά ακριβή. Οι τυχαίες προβολές εφαρμόζονται σαν ένα στάδιο προ-επεξεργασίας των δεδομένων όταν στα εξαγόμενα δεδομένα μειωμένων διαστάσεων θα πρέπει να γίνει ομαδοποίηση.

5.1.5 Ημιεπιβλεπώμενη Γραμμική Διαχωριστική Ανάλυση

Η εφαρμογή της LDA που αναφέρθηκε παραπάνω απαιτεί δεδομένα στα οποία έχουν δοθεί ετικέτες, και δεδομένου ότι αυτά δεν είναι διαθέσιμα κατά την εκτέλεση των πειραμάτων της Ημερολογιοποίησης Ομιλητών, οι προσεγγίσεις αυτές βασίζονται σε ανεξάρτητα σετ στα οποία έχουν δοθεί ετικέτες

με μη αυτόματο τρόπο και συνεπώς στοχεύουν σε ένα γενικό - ικανό να διαχωρίσει ομιλητές - υποχώρο. Παρόλο που αυτή η προσέγγιση έχει σημαντικά πλεονεκτήματα, ο εξαγώμενος υποχώρος βελτιστοποιείται ούτως ώστε να εντοπίζει τη διαφορά μεταξύ οποιωνδήποτε ομιλητών και όχι μεταξύ των συγκεκριμένων ομιλητών που υπάρχουν στο βίντεο (ή στο audio) που δίνεται στην είσοδο. Επιπλέον, η επιτυχία της μεθόδου βασίζεται στην επιλογή του ανεξάρτητου σετ ενώ παράλληλα ζητήματα όπως η διαφοροποίηση των συνθηκών υπο τις οποίες έχουν ληφθεί τα βίντεο καθώς και η γλώσσα που μιλάνε η ομιλητές μπορεί να μειώσει τη συμμόρφωση του εξαγώμενου υποχώρου σε σχέση με αυτόν που δίνεται ως είσοδος.

Η ημιεπιβλεπώμενη (semi-supervised) παραλλαγή του LDA που προτείνεται από τους Giannakopoulos & Petridis [23] συνδυάζει τα πλεονεκτήματα τόσο της PCA όσο και της LDA καθώς δεν χρειάζονται πλέον δεδομένα στα οποία έχουν δοθεί χειροκίνητα ετικέτες ενώ ταυτόχρονα εξάγει ένα υπόχωρο που εντοπίζει τις διαφορές μεταξύ των ομιλητών για δοθέν σήμα.

Η βασική ιδέα της FLsD (Fisher semi-discriminant analysis) είναι ότι ακόμα και αν λείπει η εκ των προτέρων πληροφορία για τους ομιλητές, υπάρχουν κάποιες σχετικές πληροφορίες που μπορούν να αντληθούν χωρίς πολύ κόπο. Μια χαρακτηριστική περίπτωση είναι η αξιοποίηση της ακολουθιακής δομής του σήματος κάνοντας την παραδοχή ότι γειτονικά δείγματα ομιλίας είναι πολύ πιθανό να αντιστοιχούν στον ίδιο ομιλητη με αποτέλεσμα να επιτυγχάνεται ομαδοποίηση σε γκρούπ από δείγματα που αντιστοιχούν στον ίδιο ομιλητή παρόλο που αυτός δεν είναι γνωστός. Επιπλέον θα πρέπει να τονισθεί πως είναι δυνατή η χρήση τέτοιων περιορισμών ώστε να εξαχθεί ένας FLD προσεγγιστικά βέλτιστος διαχωριστικός υποχώρος κάτω από κάποιες παραδοχές. Επίσης έχει δειχθεί πως η βέλτιστη λύση του FLsD είναι ακριβώς η ίδια βέλτιστη λύση που προκύπτει από την κλασσική FLD ανάλυση.

Ένα σημαντικό αποτέλεσμα όταν εφαρμόζουμε τον FLsD είναι ότι ο εξάγει μια αντιπροσωπευτική αναπαράσταση της ομιλίας με ένα μικρό αριθμό από χαρακτηριστικά.[8] Το συμπέρασμα αυτό, είναι ικανό να μειώσει αρκετά την πολύπλοκότητα μοντέλων που φτιάχνονται σε αυτό το χώρο όπως τα GMMs. Επίσης θέτει τις βάσεις για την δημιουργία λιγότερο πολύπλοκων συστημάτων Ημερολογιοποίησης τα οποία βασίζονται σε απλά μη παραμετρικά μοντέλα.

5.1.6 Ορισμός του FLsD

Όπως αναφέρθηκε και παραπάνω η χρήση του κριτηρίου του FLD απαιτεί την πρότερη γνώση της κλάσης που έχει αντιστοιχιθεί σε κάθε δείγμα. Παρόλα αυτά, μερικές φορές αυτή η πληροφορία μπορεί να μην είναι εξ' ολοκλήρου διαθέσιμη. Η προτεινόμενη μέθοδος, χρησιμοποιεί ένα λιγότερο απαιτητικό στήσιμο καθώς απαιτεί για κάθε δείγμα, μόνο τη γνώση του αριθμού των υπόλοιπων δειγμάτων που αντιστοιχούν στην ίδια κλάση. Για παράδειγμα, όταν ζητούμενο είναι η ομαδοποίηση των ομιλητών, δε γνωρίζουμε εκ των προτέρων όλα τα δείγματα ομιλίας σε ποιον ομιλητή αντιστοιχούν. Από την άλλη θα μπορούσε να πει κανείς ότι για κάθε δείγμα, υπάρχει μεγάλη πιθανότητα όλα τα γειτονικά δείγματα σε ένα μικρό παράθυρο, να ανήκουν στον ίδιο ομιλητή.

Αν θεωρήσουμε πως κάθε κλάση αποτελείται από μια ή περισσότερες μικρές ομάδες(class threads, με τρόπο τέτοιο ώστε όλα τα δείγματα που αντιστοιχούν σε μια μικρή ομάδα v να αντιστοιχούν και στην ίδια κλάση c, ενώ η αντιστοίχιση των μικρών ομάδων σε μια κλάση θα συμβολίζεται με $h(v)$. Δεδομένου όμως ότι το h δεν είναι γνωστό ενώ από την άλλη γνωρίζουμε την αντιστοίχιση των δειγμάτων σε μικρές ομάδες, μπορούμε να εκτιμήσουμε τον μέσο πίνακα διασποράς εντός της μικρής

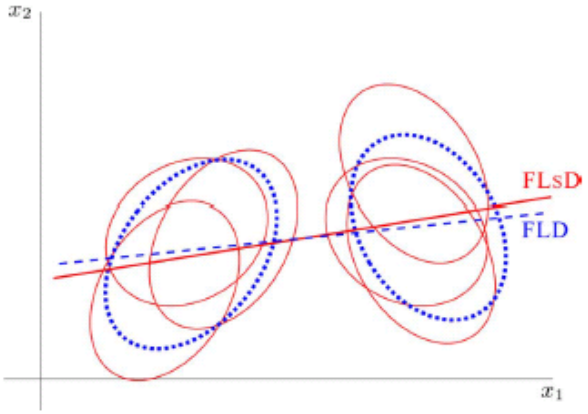
[8] Στα πειράματα μας που παρουσιάζονται στο Κεφάλαιο 7 βλέπουμε πως 4 χαρακτηριστικά είναι αρκετά ώστε να μπορούν να κάνουν διαφοροποίηση μεταξύ 5 ομιλητών

ομάδας S_w^h και μεταξύ των μικρών ομάδων S_b^h και στη συνέχεια να εφαρμόσουμε το κριτήριο του FLD κάνοντας χρήση αυτών των πινάκων.

Έχουμε επομένως ένα σετ από N_x διαστάσεων πραγματικά διανύσματα που τα συμβολίζουμε με $X = \{x^i\}$, το σετ από κλάσεις $C = \{c_k\}$ και την αντιστοίχιση κάθε παρατήρησης στην πραγματική κλάση $\{x^i, c^i\}$. Επίσης έχουμε ένα σετ από μικρές ομάδες V, την αντιστοίχιση των μικρών ομάδων σε κάθε πραγματική κλάση $h : V \to C$ και την αντιστοίχιση του πραγματικού σετ αναφοράς $\{x^i, v^i\}$ ώστε να ισχύει $\forall i : h(v^i) = c^i$. Τέλος θα έχουμε τα S_w^h, S_b^h,S_m που συμβολίζουν τους πίνακες διασποράς εντός, μεταξύ των μικρών ομάδων και την συνδυασμένη κλάση αντίστοιχα. Για κάθε $N_y < N_x$ ο πίνακας που βρίσκουμε από τη βελτιστοποίηση είναι:

$$\hat{A} = \underset{A \in R^{N_x \times N_y}}{\arg\max}\ r(A, S_1, S_2)$$

όπου S_1, S_2 είναι οποιοσδήποτε από τους τρεις συνδυασμούς των παραπάνω πινάκων διασποράς ενώ το r δίνεται από το κριτήριο του lda που περιγράψαμε στο προηγούμενο κεφάλαιο είναι ο βέλτιστος $N_x \times N_y$ FLsD πίνακας.



Σχήμα 40: Παράδειγμα της FLsD σε δύο διαστάσεις με δύο κλάσεις και έξι μικρές ομάδες. Η προβολή που βρίσκουμε με τον FLsD προσεγγίζει την αντίστοιχη που βρίσκουμε με τον FLD (επανεκτύπωση από [23])

5.2 Εκμάθηση χωρίς Επίβλεψη και Ομαδοποίηση

Σε αντίθεση με την Εκμάθηση με Επίβλεψη (Supervised Learing) όπου στα δείγματα που χρησιμοποιούνται από τον ταξινομητή έχει αποδοθεί μια ταμπέλα ανάλογα με την κατηγορία τους, η εκμάθηση χωρίς επίβλεψη (Unsupervised Learing) αναφέρεται στο πρόβλημα εύρεσης κρυφών δομών σε μη μαρκαρισμένα δεδομένα και συνεπώς δεν υπάρχει κάποιος δείκτης λάθους ή επιτυχίας για να αξιολογηθεί η επιτυχία μιας λύσης.

Η επιλογή ενός τέτοιου τρόπου επίλυσης παρόλο που δε φαντάζει ενθαρρυντικός εκ πρώτης όψεως, έχει αρκετά πλεονεκτήματα [14] και για αυτό προτιμάται αρκετά συχνά. Αρχικά, η συλλογή ενός μεγάλου σετ δεδομένων και το να δοθούν ταμπέλες σε αυτά είναι συχνά χρονικά και υπολογιστικά δαπανηρή διαδικασία. Χαρακτηριστικό παράδειγμα είναι ο ηχογραφημένος λόγος στον οποίο για να αποδοθούν ετικέτες με ακρίβεια - δίνοντας έμφαση σε κάθε λέξη ή φώνημα που εκφέρεται κάθε στιγμή

- χρειάζεται αρκετός χρόνος ενώ παράλληλα το ενδεχόμενο λάθος μπορεί να είναι μεγάλο σε κάποιες περιπτώσεις. Αν μπορεί προσεγγιστικά να σχεδιαστεί ένας ταξινομητής βασιζόμενος σε ένα μικρό σετ από μαρκαρισμένα δείγματα και στη συνέχεια να προσαρμοστεί αφού τρέξει χωρίς κάποια επίβλεψη σε ένα μεγάλο σετ από δεδομένα χωρίς ετικέτες μπορεί να εξοικονομηθεί πολύτιμος χρόνος και κόπος. Δευτερευόντως, συχνά είναι επιθυμητή η αντίστροφη διαδικασία δηλαδή η εκπαίδευση ενός ταξινομητή σε ένα μεγάλο αριθμό από (όχι υπολογιστικά ακριβών) αμαρκάριστων δεδομένων και στη συνέχεια να χρησιμοποιηθεί ανθρώπινη επίβλεψη ώστε να δοθούν ετικέτες στα γκρουπ που προέκυψαν. Κάτι τέτοιο μπορεί να είναι κατάλληλο για μεγάλες εφαρμογές του Data Mining όπου τα περιεχόμενα μια μεγάλης βάσης δεδομένων δεν είναι εκ των προτέρων γνωστά. Επίσης, σε πολλές εφαρμογές τα χαρακτηριστικά των προτύπων είναι δυνατόν να αλλάζουν αργά με την πάροδο του χρόνου, όπως για παράδειγμα στην αυτόματη κατηγοριοποίηση του φαγητού καθώς αλλάζουν οι εποχές. Αν αυτές οι αλλαγές αποτυπωθούν από ένα ταξινομητή που τρέχει χωρίς επίβλεψη, τότε θα αυξηθεί η απόδοση του αλγορίθμου. Τέλος, στα πρώτα στάδια της αντιμετώπισης ενός προβλήματος μπορεί να είναι επιθυμητό να αποκτήσει κανείς μια γενική εικόνα για τη φύση και τη δομή των δεδομένων. Η εύρεση ευδιάκριτων υποκλάσεων ή ομοιοτήτων μεταξύ των προτύπων μπορεί να αλλάξει τελείως την προσέγγιση μας για το σχεδιασμό του ταξινομητή. Η απάντηση στο ερώτημα του αν είναι δυνατό να πραγματοποιηθεί εκμάθηση από αμαρκάριστα δεδομένα εξαρτάται από τις παραδοχές που γίνονται. Αυτές όμως οι παραδοχές που γίνονται για τις παραμέτρους καθιστούν το αρχικό πρόβλημα της εκμάθησης χωρίς επίλυση προβληματικό και συνεπώς γίνεται εμφανής η ανάγκη χωρισμού των δεδομένων σε μικρότερα γκρουπ ή ομάδες (clusters).

Οι τεχνικές ομαδοποίησης είναι μέθοδοι χωρίς επίβλεψη που χρησιμοποιούνται για την οργάνωση των δεδομένων σε γκρουπ βασιζόμενες στις ομοιότητες που παρουσιάζουν μεταξύ τους. Η ιδιότητα των περισσοτέρων αλγορίθμων να μη βασίζονται σε συνήθεις παραδοχές που γίνονται σε τυπικές στατιστικές μεθόδους, όπως η στατιστική κατανομή των δεδομένων, τους κάνει χρήσιμους σε περιπτώσεις όπου υπάρχει ελάχιστη πρότερη πληροφορία. Η δυνατότητα των αλγορίθμων ομαδοποίησης να φανερώνουν τις δομές των δεδομένων, τους κάνει ιδιαίτερα χρήσιμους σε μια ποικιλία εφαρμογών όπως η ταξινόμηση, η επεξεργασία εικόνας, η αναγνώριση προτύπων η μοντελοποίηση κ.α.

Όσον αφορά τις ομάδες υπάρχουν πολλοί τρόποι με τους οποίους μπορούν να οριστούν οι οποίοι εξαρτώνται από το αντικείμενο της ομαδοποίησης. Οι Bezdek et al. [7] αντιλαμβάνονται τη ομάδα σαν μια ομάδα από αντικείμενα τα οποία είναι πιο όμοια μεταξύ τους σε σχέση με τα μέλη από άλλες κλάσεις. Ο όρος "ομοιότητα" αντιστοιχεί σε μαθηματική ομοιότητα η οποία μετράται με κάποιο καλά ορισμένο τρόπο και στους μετρικούς χώρους χρησιμοποιείται η νόρμα της απόστασης. Οι τελικές ομάδες δεν είναι γνωστές εκ των προτέρων και δημιουργούνται από τους αλγόριθμους ομαδοποίησης ταυτόχρονα με το χωρισμό των δεδομένων.

Η απόδοση των αλγορίθμων επηρεάζεται όχι μόνο από τα γεωμετρικά σχήματα και τις πυκνότητες των συστάδων, αλλά επίσης και από τις χωρικές σχέσεις και αποστάσεις μεταξύ των συστάδων. Οι ομάδες υπάρχει περίπτωση να είναι καλά διαχωρισμένες μεταξύ τους, να βρίσκονται η μία μετά την άλλη ή να επικαλύπτονται. Στη βιβλιογραφία έχει προταθεί μια πληθώρα από διαφορετικούς αλγορίθμους. Καθώς οι ομάδες μπορούν να θεωρηθούν σαν υποσύνολα του σετ δεδομένων, μια πιθανή κατηγοριοποίηση των μεθόδων ομαδοποίησης θα μπορούσε να είναι ανάλογα με το αν τα υποσύνολα είναι ασαφή (fuzzy) ή αυστηρά ορισμένα (hard). Η δεύτερη κατηγορία θεωρεί πως ένα αντικείμενο είτε ανήκει είτε δεν ανήκει σε μια ομάδα με αποτέλεσμα να χωρίζονται τα δεδομένα σε ένα συγκεκριμένο αριθμό από μοναδικά υποσύνολα. Αντίθετα οι μέθοδοι που χρησιμοποιούν fuzzy ομαδοποίηση, επιτρέπουν στα αντικείμενα να ανήκουν σε αρκετές ομάδες ταυτόχρονα. Σε πολλές περιπτώσεις η ασαφής ομαδοποίηση προτιμάται από την αυστηρά ορισμένη καθώς είναι πιο φυσική. Τα αντικείμενα

που βρίσκονται στα σύνορα μεταξύ συστάδων, δεν ανήκουν υποχρεωτικά σε μια κατηγορία, αλλά έχουν ένα ποσοστό το οποίο υποδηλώνει το κατά πόσο ανήκουν σε όλες τις ομάδες με τις οποίες συνορεύουν. Στο πρόβλημα της Ημερολογιοποίησης Ομιλητών που καλούμαστε να αντιμετωπίσουμε, μετά τη διαδικασία μείωσης των διαστάσεων χρησιμοποιήσαμε 2 διαφορετικές μεθόδους ομαδοποίησης, τον παραδοσιακό αλγόριθμο του k-means που ανήκει στα μοντέλα αλγορίθμων που βασίζονται στα κέντρα των συστάδων (centroid-based clustering) καθώς και μια βελτιωμένη έκδοση του fuzzy αλγορίθμου ομαδοποίησης των Gustafson-Kessel (GK) όπως περιγράφεται από τους Babuka et al. [4].

5.2.1 Ομαδοποίηση με k-means

Η μέθοδος ομαδοποίησης (Clustering) με τον k-means είναι μια από τις πιο ευρέως χρησιμοποιούμενες μεθόδους για γεωμετρική ομαδοποίηση και είναι γνωστή και σαν "Αλγόριθμος του Lloyd's". Πρόκειται για ένα αλγόριθμο τοπικής αναζήτησης που χωρίζει n σημεία δεδομένων σε k ομάδες. Ξεκινώντας με κάποια k στον αριθμό αρχικά κέντρα των συστάδων (μπορούν να επιλεγούν και τυχαία) αποδίδει κάθε σημείο των δεδομένων στο κοντινότερο κέντρο και στη συνέχεια υπολογίζει εκ νέου τα νέα κέντρα κάθε ομάδας σαν τις μέσες τιμές (ή τα κέντρα μάζας) από τα σημεία που έχουν αποδοθεί στην κάθε μία.[9] Η διαδικασία αυτή της απόδοσης σημείων δεδομένων και αναπροσαρμογής των κέντρων επαναλαμβάνεται μέχρι να σταθεροποιηθεί. Παρόλο που ο αλγόριθμος αυτός δεν είναι σύγχρονος, παραμένει πολύ δημοφιλής εξαιτίας της απλότητας και της ταχύτητας του, ενώ βρίσκει εφαρμογές σε τομείς όπως η Τεχνητή Νοημοσύνη, η Αναγνώριση Προτύπων κ.α. Δοθέντος ενός σετ από παρατηρήσεις $x_1, \ldots, x_n$ όπου κάθε παρατήρηση είναι ένα πραγματικό διάνυσμα d διαστάσεων, η ομαδοποίηση με k-means διαχωρίζει τις n παρατηρήσεις σε k σετ με $(k \leq n)$ ώστε $S = \{S_1, \ldots, S_k\}$ έχοντας ως στόχο την ελαχιστοποίηση του αθροίσματος των τετραγώνων εντός της κλάσης:

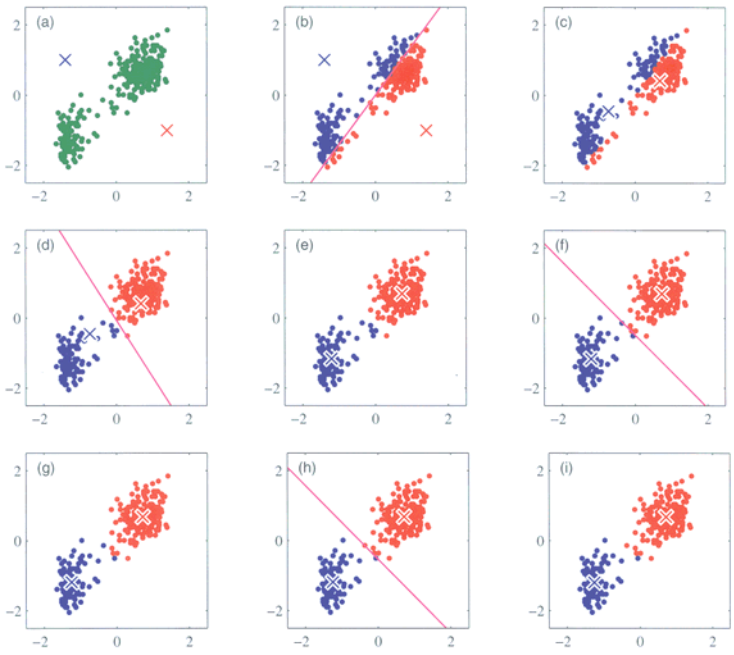
$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x_j \in S_i} \|x_j - \mu_i\|^2$$

όπου μ_i η μέση τιμή των σημείων στη ομάδα S_i.

Αλγόριθμος 3 k-means αλγόριθμος ομαδοποίησης

1: *Επέλεξε k αρχικά κέντρα $\mu_1, \ldots, \mu_k$ των συστάδων*
2: *Στο στάδιο ανάθεσης αντιστοίχισε κάθε παρατήρηση στη ομάδα της οποίας η μέση τιμή είναι πιο κοντά σε αυτή (μπορεί να αποδωθεί μονο σε μία):*
$$S_i^{(t)} = \{x_p : \|x_p - \mu_i^{(t)}\| \leq \|x_p - \mu_j^{(t)}\| \; \forall 1 \leq j \leq k\}$$
3: *Στο στάδιο ανανέωσης υπολόγισε τα νέα των παρατηρήσεων στις νέες ομάδες*
$$\mu_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$
4: *Επανέλαβε τα βήματα 2 και 3 μέχρι τόσο οι ομάδες όσο και τα κέντρα τους να μην αλλάζουν.*

[9] Το όνομα "k-means" οφείλεται σε αυτό τον εκ νέου υπολογισμό των κέντρων του κάθε cluster



Σχήμα 41: Παρουσίαση του k-means αλγορίθμου για δύο ομάδες. Τα πράσινα σημεία υποδηλώνουν το σετ δεδομένων σε Ευκλείδειο χώρο δύο διαστάσεων. Οι αρχικές επιλογές των κέντρων των συστάδων φαίνονται με τον μπλε και τον κόκκινο σταυρό αντίστοιχα. Σε κάθε επανάληψη γίνεται η ανάθεση των σημείων σε κέντρα και η ανανέωση των κέντρων μέχρι να συγκλίνει ο αλγόριθμος. (επανεκτύπωση από [9])

Ο Vattani [48] επισημαίνει πως ο αλγόριθμος υπάρχει περίπτωση να εμφανίσει δύο εκφυλισμένες καταστάσεις. Η πρώτη είναι όταν δεν αποδίδονται καθόλου σημεία σε ένα κέντρο όπου σε αυτή την περίπτωση το κέντρο αφαιρείται και ο τελικός διαχωρισμός οδηγεί σε λιγότερες από k κλάσεις ενώ η δεύτερη είναι όταν ένα σημείο είναι το ίδιο κοντά σε περισσότερα από ένα κέντρα όπου η απόφαση της ανάθεσης λαμβάνεται αυθαίρετα.

5.2.2 Η Fuzzy ομαδοποίηση των Gustafson-Kessel

Ο αλγόριθμος των Gustafson-Kessel (GK) αποτελεί μια ισχυρή τεχνική ομαδοποίησης με μεγάλο αριθμό εφαρμογών σε πολλά πεδία όπως η επεξεργασία εικόνας, η ταξινόμηση και σε συστήματα αναγνώρισης και ταυτοποίησης. Το κυριότερο χαρακτηριστικό της, είναι η τοπική προσαρμογή της μετρικής της απόστασης, στο σχήμα της ομάδα. Αυτό επιτυγχάνεται πραγματοποιώντας μια εκτίμηση του πίνακα συνδιακύμανσης της ομάδας προσαρμόζοντας αντίστοιχα των πίνακα με τις αποστάσεις που δημιουργείται. Ο GK αλγόριθμος βασίζεται στην επαναληπτική βελτιστοποίηση μιας αντικειμενικής συνάρτησης:

$$J(Z, U, V, A\{i\}) = \sum_{i=1}^{K} \sum_{k=1}^{N} (\mu_{ik})^m D_{ikA_i}$$

όπου $U = [\mu_{ik}] \in [0,1]^{K \times N}$ είναι ένας ασαφώς χωρισμένος πίνακας των δεδομένων $Z \in R^{n \times N}$, $V = [v1, \ldots, v_K]$, $v_i \in R^n$ είναι οι K σε αριθμό μοναδικές ομάδες και $m \in [1, \infty)$ είναι μια διανυσμα-

τική παράμετρος που καθορίζει την ασάφεια των εξαγόμενων συστάδων. Τέλος η νόρμα απόστασης εκφράζεται: $D_{ikA_i} = (z_k - v_i)^T A_i (z_k - v_i)$.

Η μετρική της κάθε ομάδας ορίζεται από ένα τοπικό πίνακα των παραγόμενων νορμών A_i ο οποίος χρησιμοποιείται σαν μια από τις μεταβλητές βελτιστοποίησης στην παραπάνω συνάρτηση. Αυτό επιτρέπει στη νόρμα της απόστασης να προσαρμόζεται στην τοπική τοπολογική δομή των δεδομένων. Η ελαχιστοποίηση της GK αντικειμενικής συνάρτησης επιτυγχάνεται αφού πρώτα τροποποιηθεί ο αλγόριθμος από την αρχική του μορφή ώστε να μην αντιμετωπίζει προβλήματα όταν τα δείγματα από δεδομένα είναι λίγα ή όταν είναι γραμμικώς συσχετισμένα μέσα σε μια ομάδα. Η βελτίωση των Babuka et al. [4] διορθώνει το λόγο ανάμεσα στη μέγιστη και την ελάχιστη ιδιοτιμή του πίνακα συνδιακύμανσης και ο αλγόριθμος τους είναι αυτός που ακολουθήσαμε και στην πειραματική διαδικασία.

5.3 Η συμβολή της μείωσης των διαστάσεων και της ομαδοποίησης στην Ημερολογιοποίηση Ομιλητών

Πριν γίνει αναφορά στη διαδικασία που ακολουθήσαμε για να μειώσουμε τις διαστάσεις των χαρακτηριστικών και στη συνέχεια να της ομαδοποιήσουμε σε ομάδες θα πρέπει να επισημανθεί μια μετρική επαλήθευσης των αποτελεσμάτων της ομαδοποίησης που ονομάζεται Silhouette.

5.3.1 Μετρική ομαδοποίησης Silhouette

Η Silhouette αποτελεί μια μέθοδο ερμηνείας και επαλήθευσης των ομάδων από δεδομένα. Η τεχνική αυτή παρέχει μια σαφή και συνάμα σύντομη γραφική αναπαράσταση του πόσο καλά κάθε αντικείμενο βρίσκεται μέσα στη ομάδα του και αναπτύχθηκε από τον Rousseeuw το 1986 [39].

Αφού γίνει ομαδοποίηση των δεδομένων με κάποια τεχνική, όπως για παράδειγμα ο κ-μεανς, θα προκύψει πως για κάθε είσοδο i, υπάρχει η $\alpha(i)$ που είναι η μέση ανομοιότητα της εισόδου i με όλα τα υπόλοιπα δεδομένα εντός της ίδιας ομάδας. Μπορεί να χρησιμοποιηθεί οποιαδήποτε μετρική ανομοιότητας αλλά συνήθως προτιμώνται μετρικές απόστασης. Το $\alpha(i)$ μπορεί να εκφραστεί ως το πόσο καλά έχει ταιριαστεί το i στην ομάδα του. Στη συνέχεια υπολογίζεται η ανομοιότητα του i με τα δεδομένα κάποιας άλλης ομάδας με τη διαδικασία αυτή να επαναλαμβάνεται για όλες τις ομάδες (εκτός αυτής στην οποία ανήκει το i. Η ελάχιστη μέση ανομοιότητα σε κάποια από αυτές τις ομάδες θα συμβολίζεται με $\beta(i)$. Η ομάδα με αυτή την ελάχιστη μέση ανομοιότητα θα ονομάζεται "γειτονική ομάδα" του i καθώς είναι εκείνη η ομάδα - εκτός αυτής στην οποία έχει ήδη αποδοθεί - όπου ταιριάζει καλύτερα. Ορίζεται συνεπώς ως:

$$s(i) = \frac{\beta(i) - \alpha(i)}{\max\{\alpha(i), \beta(i)\}}$$

το οποίο μπορεί να γραφεί και σαν:

$$s(i) = \begin{cases} 1 - \alpha(i)/\beta(i) & \text{αν } \alpha(i) < \beta(i) \\ 0 & \text{αν } \alpha(i) = \beta(i) \\ \beta(i)/\alpha(i) - 1 & \text{αν } \alpha(i) > \beta(i) \end{cases}$$

Από τον παραπάνω ορισμό είναι εμφανές ότι:

$$-1 \leq s(i) \leq 1$$

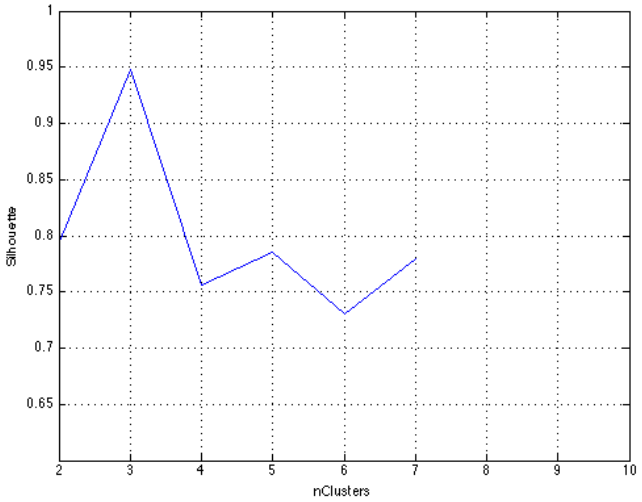
Για $s(i)$ κοντά στη μονάδα θα πρέπει να ισχύει $\alpha(i) \ll \beta(i)$. Καθώς το $\alpha(i)$ είναι μια μετρική του πόσο ανόμοιο είναι το i στη ομάδα στην οποία ανήκει, μια μικρή τιμή του θα σημαίνει ότι έχει γίνει σωστό ταίριασμα. Αντίστοιχα μεγάλη τιμή για το $\beta(i)$ θα σημαίνει πως το i έχει ταιριαστεί λάθος στη γειτονική του ομάδα. Συνεπώς, όταν το $s(i)$ είναι κοντά στο 1 θα σημαίνει πως έχει γίνει ορθή ομαδοποίηση για το συγκεκριμένο χαρακτηριστικό ενώ αν η τιμή του βρίσκεται κοντά στο -1 προκύπτει πως θα ήταν προτιμότερο να έχει αντιστοιχηθεί στη γειτονική του ομάδα. Τέλος τιμή κοντά στο μηδέν σημαίνει ότι το i βρίσκεται στο σύνορο δύο συστάδων.

Συμπερασματικά, το μέσο $s(i)$ μιας ομάδας είναι μια μετρική του πόσο καλά είναι ομαδοποιημένα όλα τα δεδομένα σε μια ομάδα. Έτσι η μέση τιμή όλου του dataset είναι μια μετρική του πόσο κατάλληλα έχουν ομαδοποιηθεί τα δεδομένα. Οι silhouette γραφικές παραστάσεις καθώς και οι μέσοι όροι μπορούν να χρησιμοποιηθούν ώστε να αποφασιστεί ο αριθμός των συστάδων σε σένα σετ δεδομένων.

5.3.2 Περιγραφή της πειραματικής διαδικασίας μείωσης των διαστάσεων και ομαδοποίησης

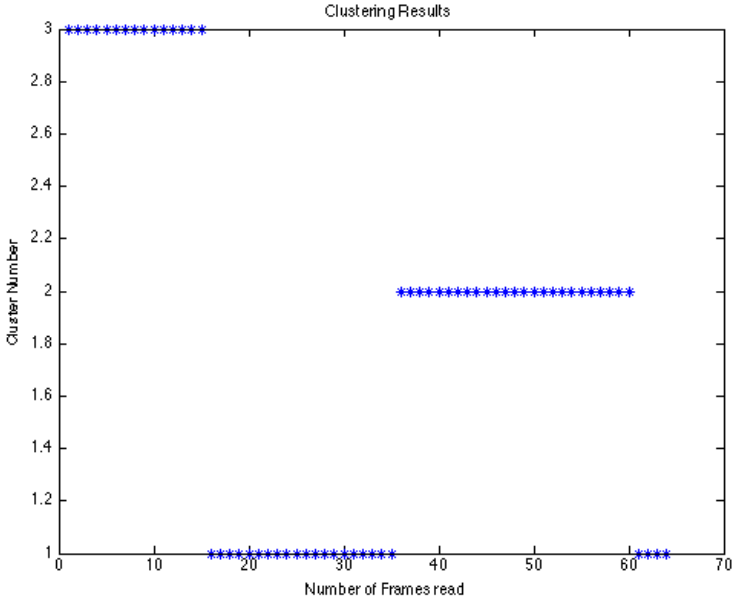
Σε κάθε καρέ του σετ από βίντεο που έχουμε, με την εξαγωγή χαρακτηριστικών χρησιμοποιώντας τα Gabor κυματίδια όπως αναλύθηκε στο προηγούμενο κεφάλαιο, επιστρέφονται 2704 χαρακτηριστικά για κάθε μια από τις κλίμακες συνεπώς όταν χρησιμοποιούνται 5 κλίμακες θα έχουμε 13520 για κάθε καρέ . Ο αριθμός αυτός, στην περίπτωση ενός μεγάλου βίντεο, δεν μας επιτρέπει να δοθούν αυτά τα χαρακτηριστικά στον LDA αλγόριθμο μείωσης των διαστάσεων πόσο μάλλον στον αλγόριθμο ομαδοποίησης. Για αυτό τον λόγο επιλέγεται πρώτα η μέθοδος της τυχαίας προβολής σε ένα χώρο πολύ μικρότερων διαστάσεων και στη συνέχεια τα χαρακτηριστικά που θα εξαχθούν από αυτή δίνονται στον LDA ώστε να μειώσει τις διαστάσεις στο επιθυμητό αποτέλεσμα. Έχουμε δηλαδή μια σταδιακή μείωση αρχικά από τις αρχικές διαστάσεις σε ένα αρκετά μικρότερο αριθμό όπως για παράδειγμα στις 500 με random projection και στη συνέχεια στις τελικές διαστάσεις με τον αλγόριθμο μείωσης των διαστάσεων.

Αφού ολοκληρωθεί η διαδικασία μείωσης των διαστάσεων, στη συνέχεια επιλέγεται ο αλγόριθμος ομαδοποίησης. Οι μέθοδοι που χρησιμοποιήσαμε και συγκρίναμε τα αποτελέσματα τους είναι ο k-means (όπου έχουμε λάβει υπόψη την περίπτωση των άδειων ομάδων τις οποίες και απαλείφουμε) και ο GK fuzzy αλγόριθμος ομαδοποίησης. Αρχικά κάνουμε την παραδοχή πως οι ομιλητές σε ένα βίντεο μπορεί να είναι από 1 έως 5 (στην πραγματικότητα στο dataset του Canal9 που χρησιμοποιήθηκε οι ομιλητές στα βίντεο είναι από 3 έως 5) και στη συνέχεια για κάθε περίπτωση συνολικού αριθμού ομιλητών πραγματοποιούμε πρωτίστως ομαδοποίηση και δευτερευόντως υπολογίζουμε τη silhoutte μετρική παρακάτω ώστε να αξιολογήσουμε τα αποτελέσματα. Αυτή μας επιστρέφει το βέλτιστο αριθμό από ομάδες για κάθε βίντεο καθώς και τα labels που αντιστοιχούν σε αυτή τη βέλτιστη επιλογή από ομάδες. Για παράδειγμα για ένα βίντεο με 3 διαφορετικούς ομιλητές η γραφική παράσταση που επιστρέφει η silhoutte είναι:



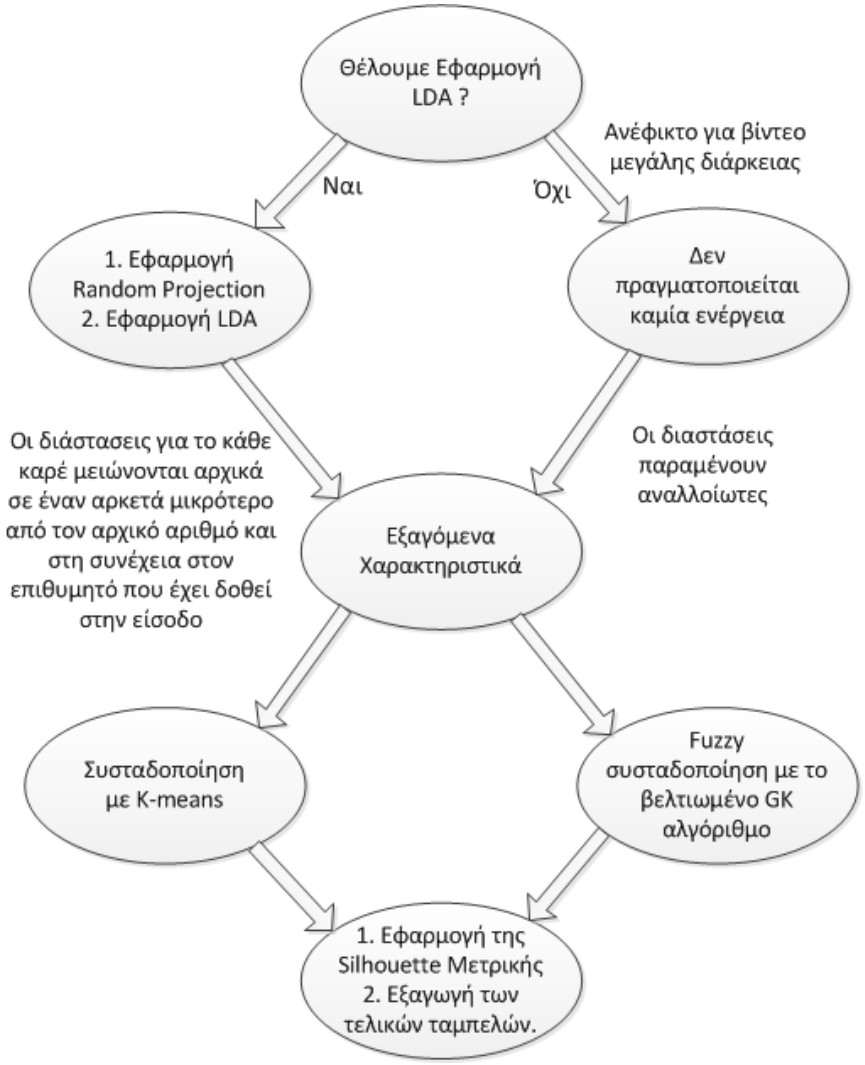
Σχήμα 42: Η γραφική παράσταση που επιστρέφει η silhoutte μετρική για το βέλτιστο αριθμό συστάδων ενός σετ χαρακτηριστικών που έχει εξαχθεί από βίντεο

Η παραπάνω γραφική παρουσιάζει πως για το συγκεκριμένο βίντεο η βέλτιστη επιλογή είναι να θεωρήσουμε πως έχουμε 3 ομάδες, δηλαδή ομιλητές στο βίντεο και επομένως να κρατήσουμε τις ταμπέλες που έχουν αποδοθεί στα χαρακτηριστικά από τη ομαδοποίηση που έγινε για 3 ομάδες. Τα πλεονεκτήματα της χρήσης αυτής της μετρικής είναι ότι αποφεύγεται πρωτίστως η αυθαίρετη επιλογή του τελικού αριθμού των ομάδων και δευτερευόντως ότι έχουμε στα χέρια μας μια μετρική του πόσο καλά είναι τελικά τα αποτελέσματα της ομαδοποίησης. Επομένως η γραφική παράσταση που αντιστοιχεί στην ομαδοποίηση των χαρακτηριστικών σε 3 ομάδες μας οδηγεί στην παρακάτω γραφική παράσταση των τελικών ομάδων ενός βίντεο από το οποίο έχουμε κρατήσει 64 καρέ.



Σχήμα 43: Χρησιμοποιώντας την silhoutte γραφική ομαδοποιούνται τα χαρακτηριστικά σε 3 διαφορετικές ομάδες (κάθετος άξονας) ενώ στον οριζόντιο έχουμε τα καρέ που διαβάστηκαν από το βίντεο

Επομένως το διάγραμμα της πορείας μείωσης των διαστάσεων χαρακτηριστικών και ομαδοποίησης τους που ακολουθήσαμε είναι το ακόλουθο:



Σχήμα 44: Διάγραμμα παρουσίασης των βημάτων που ακολουθήθηκαν για μείωση των διαστάσεων και ομαδοποίηση

5.4 Απόδοση μιας μοναδικής ετικέτας σε κάθε shot

Στο τελευταίο στάδιο επεξεργασίας των χαρακτηριστικών που έχουν εξαχθεί από το πρόσωπο το πρώτο πράγμα που καλούμαστε να κάνουμε είναι να περάσουμε από τα καρέ στα δευτερόλεπτα. Αφού το κάνουμε αυτό και βρούμε τα όρια των shots σε δευτερόλεπτα [10] αποδίδουμε τις τελικές ετικέτες που επεστράφησαν από τη ομαδοποίηση στα πραγματικά καρέ του βίντεο που έχουν διαβαστεί καθώς διαβάζουμε 5 καρέ ανά δευτερόλεπτο. Στη συνέχεια βρίσκουμε σε κάθε shot την ετικέτα που κυριαρχεί πιο πολύ στο συγκεκριμένο διάστημα και την αποδίδουμε σε όλα τα δευτερόλεπτα του συγκεκριμένου shot ανεξάρτητα αν κάποια από αυτά είχαν διαφορετική τιμή πιο πριν. Αυτό το κάνουμε γιατί στο πρόβλημα της Ημερολογιοποίησης Ομιλητών που βασίζεται σε οπτική πληροφορία γίνεται η παραδοχή πως υπάρχει ομοιομορφία του ομιλητή μέσα στο shot. Ταυτόχρονα υπολογίζεται και το ποσοστό ακρίβειας της παραπάνω ετικέτας. Συνεπώς για ένα shot διάρκειας 10 δευτερολέπτων μπορεί οι ετικέτες που επέστρεψε η συσταδοποιηση να είναι [1 1 2 1 3 2 1 1 1 1] και το οποίο μετά από αυτό το στάδιο θα γίνει [1 1 1 1 1 1 1 1 1 1] με την αξιοπιστία του τελικού αποτελέσματος να είναι ίση με 70% αφού σε 3 από τα 10 δευτερόλεπτα του shot υπήρχε κάποια άλλη ετικέτα η οποία υπέστη αλλαγή.

[10] Για παράδειγμα ένα shot από το 27^o καρέ μέχρι το 90^o θα ξεκινά από το 1^o δευτερόλεπτο και θα τελειώνει στο 4^o. (Για Frame Rate ίσο με 25 fps)

Κεφάλαιο 6

Αξιολόγηση των Πειραμάτων της Ημερολογιοποίησης Προσώπου

Αρχικά θα πρέπει να γίνει μια μικρή αναφορά στο σετ δεδομένων που χρησιμοποιήθηκε κατά την πειραματική διαδικασία. Επειδή το αρχικό σετ δεδομένων που είχαμε του Canal9[11] περιείχε annotation με βάση την ακουστική πληροφορία (Ποιος μιλάει και για πόσο) και όχι με βάση την οπτική πληροφορία (Ποιος φαίνεται και για πόσο) δημιουργήσαμε ένα μικρότερο σετ από δεδομένα στα οποία έγινε annotation με βάση την οπτική πληροφορία. Το σετ δεδομένων περιείχε 25 βίντεο διάρκειας από 20 δευτερόλεπτα μέχρι 1.5 λεπτό και τα πρόσωπα που εμφανίζονταν σε καθένα από αυτά ήταν από 2 έως 5 με ίδια συχνότητα (είχαμε δηλαδή από 6 βίντεο για 2,3 και 5 πρόσωπα και 7 για 4 πρόσωπα). Τα δεδομένα εξήχθησαν από 8 διαφορετικά βίντεο του αρχικού σετ δεδομένων με αποτέλεσμα να συγκεντρωθούν 40 διαφορετικά πρόσωπα.

Τα πειράματα που εκτελέσαμε αφορούσαν τη σύγκριση 2 μεθόδων εξαγωγής χαρακτηριστικών (με Gabor) κυματίδια και με σκέτες τις τιμές των pixels) καθώς και την εφαρμογή ή μη της μείωσης των διαστάσεων με την προτεινόμενη FLsD. Σε περίπτωση που γινόταν μείωση των διαστάσεων, πραγματοποιούσαμε πειράματα για τελικές διαστάσεις από 1 μέχρι 6, ενώ ταυτόχρονα εκτελέσαμε πειράματα για τη μέθοδο της ομαδοποίησης (k-means και GK fuzzy). Θα πρέπει να επισημανθεί πως τα αποτελέσματα που παρουσίαζε η ομαδοποίηση με τη μέθοδο k-means εμφάνιζαν κάποιες αποκλίσεις καθώς υπάρχει εξάρτηση από το που επιλέγονται τα αρχικά κέντρα των ομάδων. Για να αντιμετωπισθεί αυτό το φαινόμενο, εφαρμόζαμε τη μέθοδο 50 φορές αντί για μία και κρατούσαμε τα αποτελέσματα εκείνα τα οποία παρουσίαζε την μικρότερη - εντός της ομάδας - απόσταση. Επαναλαμβάνοντας τα πειράματα 5 φορές η απόκλιση που εμφανίζουν τα αποτελέσματα είναι στις περισσότερες περιπτώσεις μικρότερη του 1%. Αντίθετα με τη GK fuzzy μέθοδο ομαδοποίησης τα αποτελέσματα που παρουσιάζονταν για κάθε μια από τις 5 φορές που εκτελέσαμε τα πειράματα, ήταν όμοια. Αυτό συμβαίνει γιατί πριν λάβουν χώρα οι επαναλήψεις του αλγορίθμου έχει προηγηθεί seeding των αριθμών που εξάγει η τυχαία γεννήτρια αριθμών που χρησιμοποιείται κατά την αρχικοποίηση. Η παραπάνω διαδικασία όλων των πειραμάτων πραγματοποιήθηκε τόσο όταν ο αριθμός των προσώπων στο βίντεο ήταν άγνωστος (όπου επιλεγόταν ο βέλτιστος αριθμός ομάδων με τη Silhouette μετρική), όσο και όταν δινόταν ο αριθμός των προσώπων σαν είσοδος στον αλγόριθμο ομαδοποίησης ώστε να συγκριθούν και να αξιολογηθούν καλύτερα τα αποτελέσματα.

Οι μετρικές αξιολόγησης των αποτελεσμάτων των πειραμάτων της Ημερολογιοποίησης Προσώπου ήταν αρχικά το Cluster Accuracy Ratio (CAR), που είναι ο λόγος της διάρκειας των σωστά ομαδοποιημένων τμημάτων του βίντεο προς τη συνολική διάρκειά του.

Στη συνέχεια χρησιμοποιήσαμε τη μετρική cluster purity, που για ένα σετ δεδομένων D στα οποία βρέθηκαν k ομάδες, και το μέγεθος της ομάδας j είναι $|C_j|$ ενώ $|C_j|_{class=i}$ θα είναι ο αριθμός των αντικειμένων της ομάδας i που αποδόθηκαν στην ομάδα j. Συνεπώς το purity μιας ομάδας δίνεται από:

$$purity(C_j) = \frac{1}{|C_j|} max(|C_j|_{class=i})$$

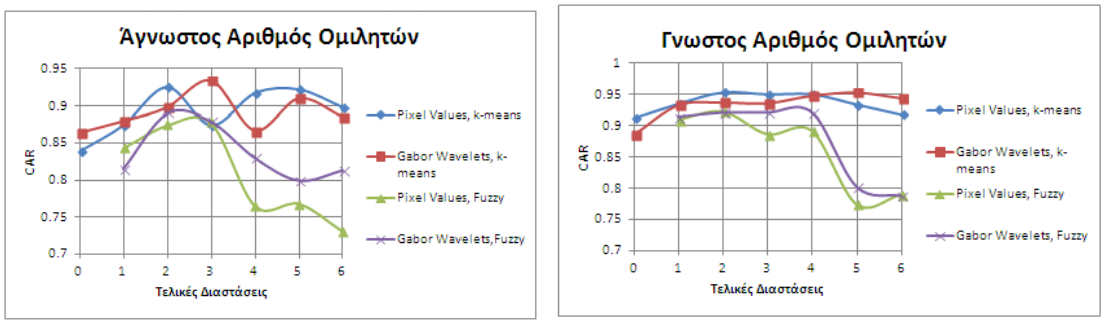
ενώ το συνολικό purity της λύσης της ομαδοποίησης μπορεί να εκφραστεί ως το -με βάρη - άθροισμα των επιμέρους purities ως:

[11] Γίνεται αναφορά σε αυτο στο Παράρτημα Α

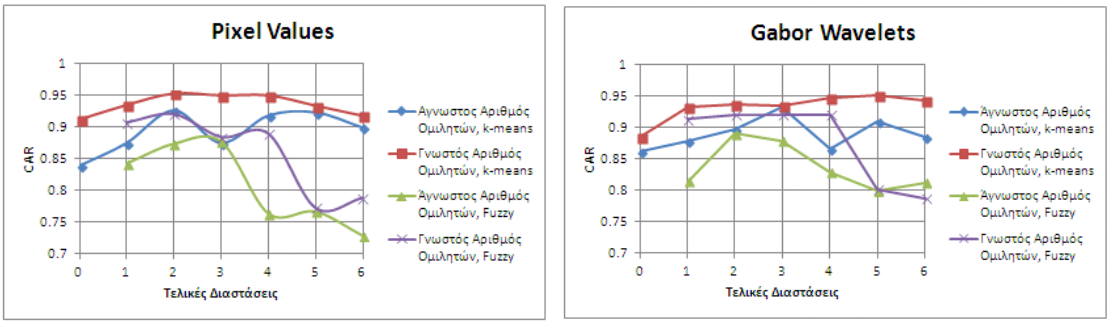
$$Purity = \sum_{j=1}^{k} \frac{|C_j|}{D} purity(C_j)$$

Οι Giannakopoulos & Petridis [23] επισημαίνουν πως η παραπάνω μετρική, επικεντρώνεται στη συχνότητα του πιο συχνού ομιλητή-προσώπου μέσα σε κάθε ομάδα ενώ όσο μεγαλύτερη είναι η τιμή που προκύπτει, τόσο καλύτερη είναι η λύση. Μια ακόμα μετρική που χρησιμοποιήθηκε ήταν το purity του προσώπου-ομιλητή που υποδηλώνει πόσο συχνός είναι ο πιο συχνός ομιλητής που ανιχνεύθηκε σε κάθε μια από τις ομάδες των ομιλητών [23].

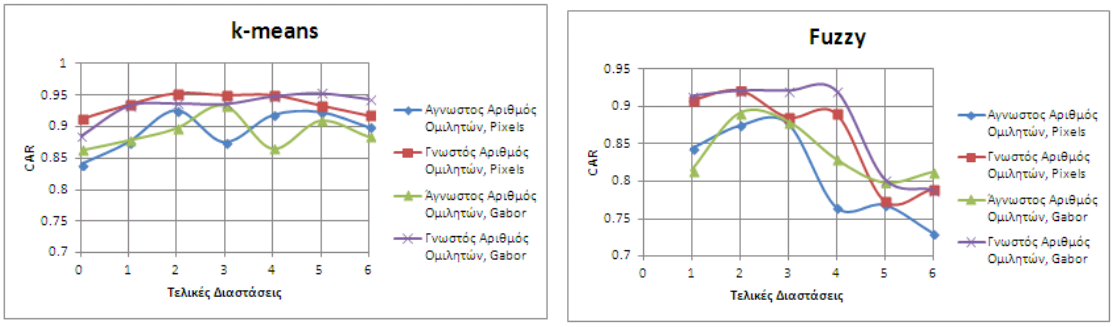
Τα πειραματικά αποτελέσματα της Ημερολογιοποίησης Προσώπου που προέκυψαν είναι τα ακόλουθα:



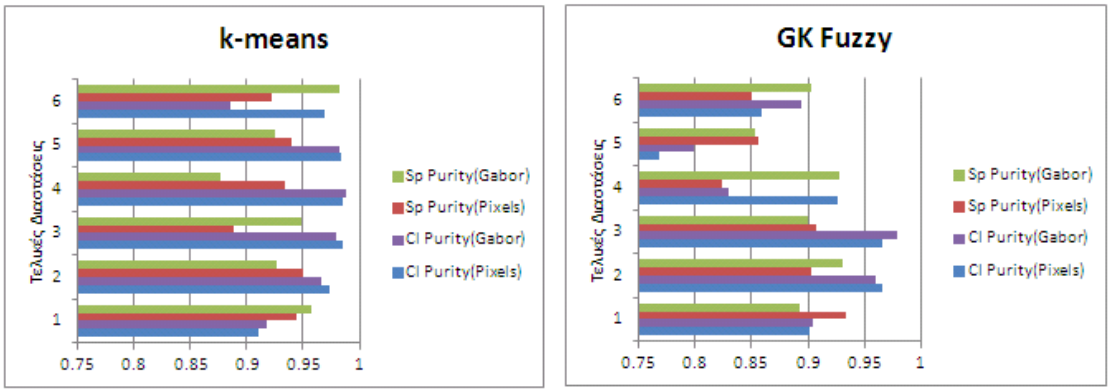
Σχήμα 45: Σύγκριση μεθόδων εξαγωγής χαρακτηριστικών και ομαδοποίησης για Άγνωστο (α) και Γνωστό (β) αριθμό Ομιλητών



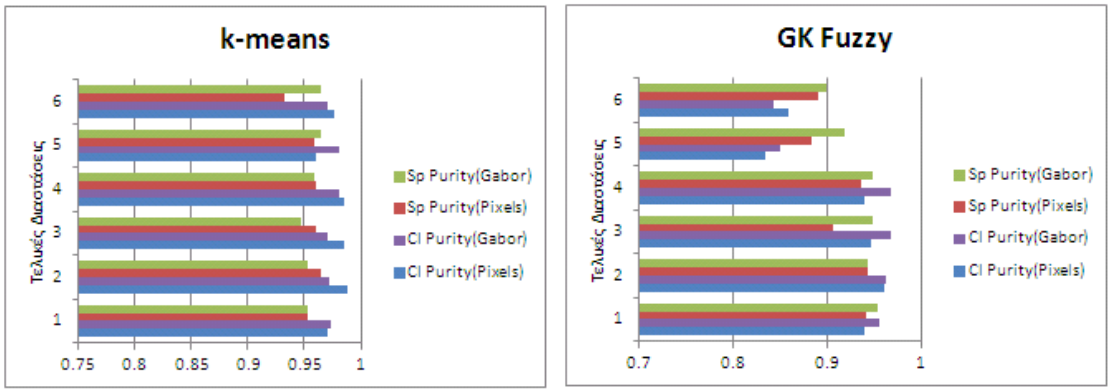
Σχήμα 46: Σύγκριση γνωστού και αγνώστου αριθμού ομιλητών και ομαδοποίησης για Pixels (α) και Gabor κυματιδίων (β)



Σχήμα 47: Σύγκριση γνωστού και αγνώστου αριθμού ομιλητών και μεθόδων εξαγωγής χαρακτηριστικών για k-means (α) και fuzzy (β)



Σχήμα 48: Σύγκριση αποτελεσμάτων cluster & speaker purity για k-means (α) και fuzzy (β) σε άγνωστο αριθμό ομιλητών



Σχήμα 49: Σύγκριση αποτελεσμάτων cluster & speaker purity για k-means (α) και fuzzy (β) σε γνωστό αριθμό ομιλητών

Συγκρίνοντας τις μέγιστες τιμές που λαμβάνει το Cluster Accuracy Ratio για κάθε μια από τις παραπάνω περιπτώσεις εξάγουμε τα ακόλουθα συμπεράσματα. Αρχικά βλέπουμε ότι η χρήση των Gabor κυματιδίων για άγνωστο αριθμό ομιλητών παρουσιάζει καλύτερα αποτελέσματα (5%) στον αρχικό χώρο χαρακτηριστικών (δηλαδή χωρίς μείωση των διαστάσεων). Στη συνέχεια καθώς αυξάνονται οι διαστάσεις τα αποτελέσματα λαμβάνουν κοντινές τιμές καθώς η μείωση των διαστάσεων αρχικά με τυχαία προβολή σε ένα μικρότερο χώρο και στη συνέχεια με τη χρήση FLsD μας οδηγεί σε χαρακτηριστικά που παρουσιάζουν μικρή διακύμανση εντός της ίδιας κλάσης και μεγάλη μεταξύ διαφορετικών κλάσεων. Επιπλέον θα πρέπει να τονίσουμε πως με τη χρήση των Gabor κυματιδίων εξάγουμε καλύτερα αποτελέσματα (μέγιστες τιμές ανά πίνακα) σε όλες τις περιπτώσεις.

Όσον αφορά τη σύγκριση των μεθόδων ομαδοποίησης, συμπεραίνουμε πως ο k-means λειτουργεί καλύτερα από τον Fuzzy με τις τιμές του να είναι μεγαλύτερες κατά $3 \rightarrow 5\%$ Αυτό συμβαίνει γιατί στην περίπτωση του k- means όπου τα κέντρα των ομάδων επιλέγονται κάθε φορά τυχαία, εκτελέσαμε τον αλγόριθμο 10 φορές λαμβάνοντας τα καλύτερα αποτελέσματα.

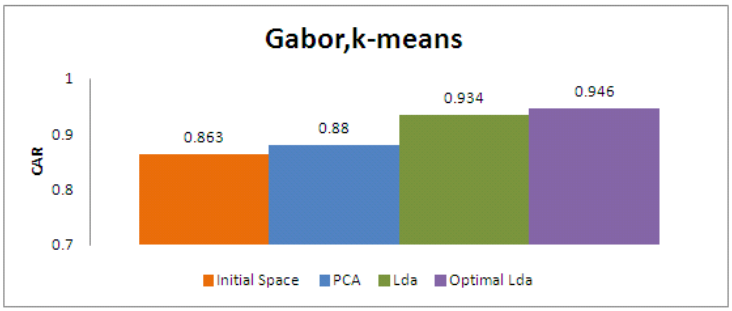
Επιπλέον βλέπουμε πως για το συγκεκριμένο σετ δεδομένων όπου ο αριθμός των προσώπων είναι ομοιόμορφα κατανεμημένος ανάμεσα σε 2 και 5 ομιλητές η επιλογή της μείωσης του χώρου των χαρακτηριστικών στις 2 ή στις 3 διαστάσεις είναι στις περισσότερες περιπτώσεις η καλύτερη ανεξάρτητα από τις μεθόδους που χρησιμοποιούνται. Μόνο όταν χρησιμοποιούμε k-means και Gabor κυματίδια σε εκ των προτέρων γνωστό αριθμό ομιλητών παρουσιάζονται καλύτερα αποτελέσματα στις 5 διαστάσεις. Συνεπώς επειδή τα πειράματα που θα γίνουν για Ημερολογιοποίηση Ομιλητών σε όλο το σετ δεδομένων που περιέχει μεγάλης διάρκειας αρχεία αρκετά από τα οποία έχουν 5 ομιλητές θα κρατήσουμε σαν βέλτιστη διάσταση τις 3 καθώς όσο αυξάνεται ο αριθμός των ομάδων στις οποίες θα καλείται να ομαδοποιήσει τα δεδομένα ο αλγόριθμος, θα απαιτούνται περισσότερες διαστάσεις.

Τέλος όταν δίνεται εκ των προτέρων στον αλγόριθμο ομαδοποίησης ο σωστός αριθμός των προσώπων-ομάδων που υπάρχουν στο βίντεο, τα αποτελέσματα είναι αρκετά βελτιωμένα. Στον αρχικό χώρο διαστάσεων και για σκέτες τιμές των pixels η διαφορά αυτή φτάνει μέχρι και 8.2% ενώ για τις υπόλοιπες διαστάσεις η αύξηση είναι περίπου 3.5%.

Όσον αφορά, το cluster purity, παρατηρείται και εδώ πως μας δίνει τα καλύτερα αποτελέσματα όταν μειώνουμε το χώρο των χαρακτηριστικών στις τρεις διαστάσεις ανεξάρτητα από τη μέθοδο ομαδοποίησης. Μάλιστα όταν είναι η πληροφορία του αριθμού των ομιλητών στο βίντεο δεν είναι εκ των προτέρων γνωστή, τα αποτελέσματα που εξάγουμε για μεγάλο αριθμό διαστάσεων είναι εμφανώς μικρότερα σε σχέση με τα αντίστοιχα για τις 3 διαστάσεις.

Συμπερασματικά, καταλήγουμε πως η βέλτιστη επιλογή είναι η χρήση των Gabor κυματιδίων, η μείωση του χώρου στις τρεις διαστάσεις ενώ για τον αλγόριθμο ομαδοποίησης θα επιλέξουμε τον k-means καθώς μπορεί να εκτελείται αρκετά πιο αργά σε σχέση με τον επαναλμβανόμενο Fuzzy αλλά πετυχαίνει σαφώς μεγαλύτερη ακρίβεια. Από την άλλη επειδή η GK Fuzzy μέθοδος έχει ήδη χρησιμοποιηθεί σε πειράματα Ημερολογιοποίησης Ομιλητών με βάση την ακουστική πληροφορία δε θα αποκλείσουμε τη χρήση της στη συνέχεια καθώς η χρήση του θα μας βοηθήσει στη σύγκριση και στο συνδυασμό των τελικών αποτελεσμάτων της, με τα αντίστοιχα της οπτικής πληροφορίας.

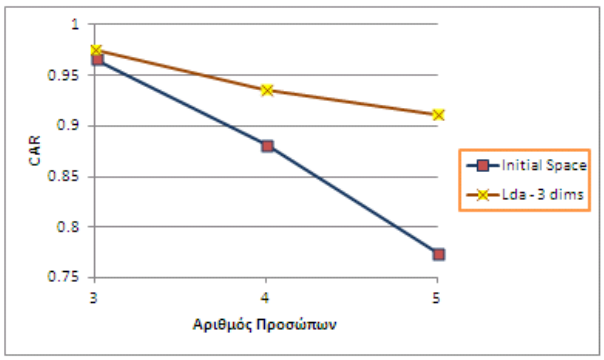
Συγκριτικά αποτελέσματα εξάγαμε επίσης και για τη μέθοδο της μείωσης των διαστάσεων. Εκτελέσαμε τα πειράματα εφαρμόζοντας PCA, FLsD,Τυχαία Προβολή και συγκρίναμε τα αποτελέσματα τους τόσο με αυτά του αρχικού χώρου όσο και με τη βέλτιστη FLsD όπου παίρναμε τις ετικέτες απευθείας από το αρχείο που περιείχε το annotation ώστε να δούμε πόσο απέχουν τα αποτελέσματα μας από το μέγιστο όριο που μπορούν να φτάσουν. Στην περίπτωση που εξάγουμε τα χαρακτηριστικά με τη χρήση Gabor κυματιδίων, κάνουμε ομαδοποίηση με k-means και μειώνουμε τις διαστάσεις σε 3 τα αποτελέσματα που εξάγαμε είναι:



Σχήμα 50: Σύγκριση μεθόδων μείωσης των διαστάσεων σε σχέση με τον αρχικό χώρο

Από το παραπάνω σχήμα βλέπουμε αρχικά πως σε σχέση με τον αρχικό χώρο των χαρακτηριστικών οι μέθοδοι που μας οδηγούνε σε ένα υποχώρο με μικρότερες διαστάσεις για τα χαρακτηριστικά πετυχαίνουν καλύτερα αποτελέσματα. Η PCA μας οδηγεί σε αύξηση 1.7% ενώ η FLsD σε 7.1% καθώς μας ανεβάζει στο 93.4%. Τέλος η βέλτιστη FLsD πετυχαίνει αποτέλεσμα ίσο με 94.6% που είναι μόλις 1.2% καλύτερο από τη μέθοδο που χρησιμοποιούμε.

Τέλος παρουσιάζουμε τα αποτελέσματα μας ανάλογα με το πόσα είναι τα πρόσωπα - ομιλητές που εμφανίζονται στο βίντεο συγκρίνοντας το πόσο καλά αποδίδει η FLsD σε σχέση με τον αρχικό χώρο των χαρακτηριστικών. Τα αποτελέσματα που εξάγαμε είναι τα παρακάτω:



Σχήμα 51: Σύγκριση των αποτελεσμάτων του αρχικού χώρου χαρακτηριστικών και της FLsD τριών διαστάσεων ανάλογα με τον αριθμό των προσώπων ομιλητών που εμφανίζονται στο βίντεο

Από το παραπάνω σχήμα μπορούμε να συμπεράνουμε πως όταν έχουμε 3 ομιλητές η μείωση των διαστάσεων δεν εμφανίζει πολύ καλύτερα αποτελέσματα σε σχέση με τα αντίστοιχα του αρχικού

χώρου. Όμως όσο αυξάνονται οι ομιλητές, τα αποτελέσματα του αρχικού χώρου πέφτουν ραγδαία και για τις δύο μεθόδους εξαγωγής χαρακτηριστικών ενώ αντίθετα με τη χρήση της FLsD η ακρίβεια της ομαδοποίησης εμφανίζει μια αρκετά πιο σταθερή πορεία. Γίνεται συνεπώς εμφανές, πως η προτεινόμενη FLsD αποτελεί ένα αναγκαίο εργαλείο για την αντιμετώπιση του προβλήματος της Ημερολογιοποίησης Ομιλητών καθώς εκτός του ότι εμφανίζει καλύτερα αποτελέσματα από τις υπόλοιπες μεθόδους μείωσης των διαστάσεων, αποτελεί μια πιο αξιόπιστη λύση Ημερολογιοποίησης καθώς παραμένει σε μεγάλο βαθμό ανεπηρέαστη από το πόσοι ομιλητές εμφανίζονται μέσα στο βίντεο ανεξάρτητα από τη μέθοδο ομαδοποίησης που χρησιμοποιείται.

Κεφάλαιο 7

Ανίχνευση Κίνησης Χειλιών και Αξιολόγηση Πειραματικών Αποτελεσμάτων

Το τελευταίο στάδιο επεξεργασίας των δεδομένων που εξάγονται από τα βίντεο της βάσης δεδομένων, είναι η ανίχνευση κίνησης των χειλιών ώστε να αποφανθούμε αν το εικονιζόμενο πρόσωπο μιλάει ή όχι. Μέχρι τώρα η διαδικασία που έχουμε ακολουθήσει ήταν να χωρίσουμε ένα βίντεο σε επιμέρους shot, σε καθένα από αυτά να πραγματοποιήσουμε ανίχνευση προσώπου εξάγοντας στη συνέχεια χαρακτηριστικά από αυτό τα οποία αφού τους μειώσουμε τις διαστάσεις τα ομαδοποιούμε σε γκρουπ με κοινά χαρακτηριστικά καθένα από τα οποία αντιστοιχεί σε ένα πρόσωπο - ομιλητή.

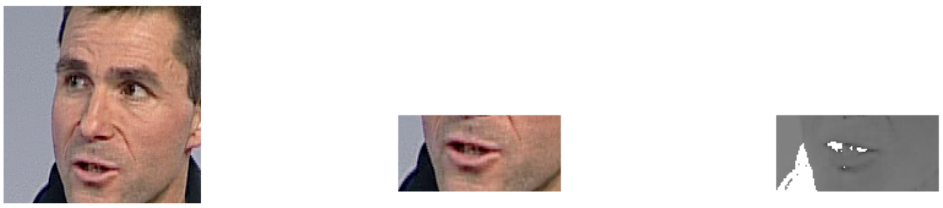
7.1 Ανίχνευση Χειλιών

Κατά τη διάρκεια της διαδικασίας της εξαγωγής χαρακτηριστικών παίρνουμε κάθε πρόσωπο και κρατάμε το κάτω του μέρος κόβοντας παράλληλα τμήματα από τα δεξιά και τα αριστερά του ώστε να προσδιορίσουμε προσεγγιστικά την περιοχή στην οποία βρίσκεται το στόμα. Η μέθοδος αυτή προτιμάται από το να γίνει απευθείας ανίχνευση στόματος με τη μέθοδο των Viola & Jones καθώς αν γινόταν ανίχνευση στόματος θα υπήρχαν αρκετές περιπτώσεις όπου ο αλγόριθμος θα αποτύγχανε να πραγματοποίηση επιτυχή ανίχνευση δεδομένου ότι παρουσιάζει προβλήματα όταν τα πρόσωπα είναι γυρισμένα σε προφίλ. Η μέθοδος που ακολουθήσαμε για την ανίχνευση των χειλιών είναι αυτή που περιγράφεται από τους Soetedjo et al. [3]. Έχοντας επομένως προσδιορίσει την περιοχή μέσα στην οποία βρίσκεται το στόμα, στη συνέχεια παίρνουμε τις δύο παρακάτω κανονικοποιημένες τιμές για το κόκκινο και το πράσινο χρώμα σε κάθε pixel:

$$r = \frac{R}{R+G+B} \ , \ g = \frac{G}{R+G+B}$$

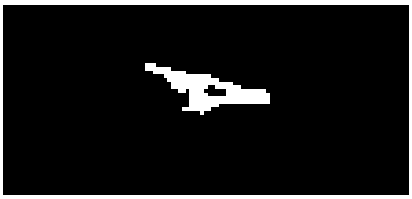
Με τη χρήση των δύο παραπάνω μεταβλητών λαμβάνουμε μια μετασχηματισμένη εικόνα της περιοχής του στόματος σε κλίμακα γκρι όπου η περιοχή των χειλιών εμφανίζεται να έχει πιο σκούρα απόχρωση σε σχέση με τις υπόλοιπες. Επειδή όμως υπάρχει η περίπτωση το στόμα να είναι ανοιχτό και εκεί να εμφανίζονται μαύρα pixel τα οποία δεν ανήκουν στην περιοχή των χειλιών αφαιρούμε τοποθετώντας ένα κατώφλι όλες τις περιοχές στις οποίες η περιοχή του στόματος λαμβάνει πολύ μικρές τιμές οι οποίες αντιστοιχούν στο μαύρο χρώμα. Ο τύπος του μετασχηματισμού αυτού θα είναι:

$$I_{gr} = \frac{1+g-r}{2}$$



Σχήμα 52: Αρχικό πρόσωπο, στόμα και μετασχηματισμένη εικόνα τους στόματος Igr

Στη συνέχεια μετατρέπουμε την παραπάνω μετασχηματισμένη εικόνα κλίμακας γκρι του στόματος σε δυαδική. Αυτό γίνεται αφού πρώτα ταξινομήσουμε σε αύξουσα σειρά την τιμή κάθε pixel της I_{gr} επιλέγοντας ένα κατώφλι ίσο με το 10% των μικρότερων τιμών που λαμβάνει αυτή ώστε να κρατήσουμε τις πιο σκούρες περιοχές της όπου βρίσκονται τα χείλια. Κατόπιν, εφαρμόζεται ένα median φίλτρο διαστάσεων 3×3 ώστε να απαλειφθούν κάποια pixels που βρίσκονται μόνα τους και έχουν τη μορφή θορύβου όσον αφορά το στόμα. Το επόμενο βήμα για είναι να βρεθούν τα connected components στην περιοχή του στόματος ώστε να ομαδοποιηθούν τα δυαδικά αντικείμενα που βρίσκονται μέσα στην εικόνα και να κρατήσουμε το μεγαλύτερο από αυτά που θα αντιστοιχεί στο στόμα. Επομένως το τελικό αποτέλεσμα ανίχνευσης των χειλιών θα έιναι:



Σχήμα 53: Δυαδική εικόνα της περιοχής του στόματος που απεικονίζει με άσπρο τα χείλη

7.2 Κίνηση των Χειλιών

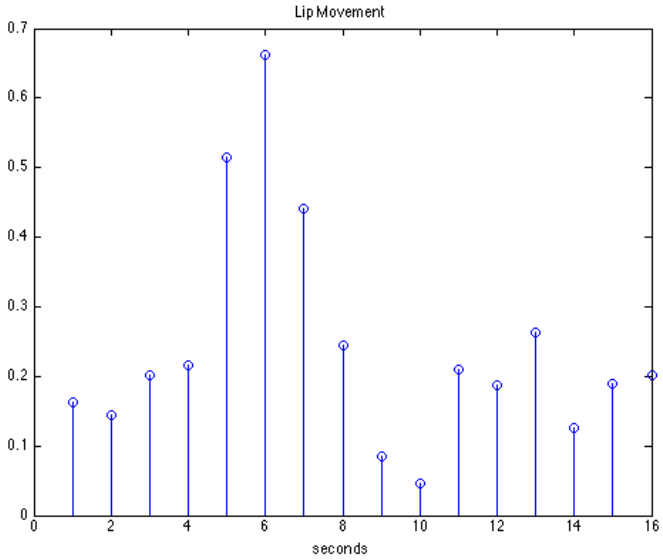
Επειδή η παραπάνω εικόνα είναι μεν ενδεικτική του σχήματος που έχουν τα χείλια στο τρέχων καρέ αλλά δεν είναι από μόνη της αρκετή για να περιγραφεί με ακρίβεια το σχήμα των χειλιών και του στόματος κατά τη διάρκεια της ομιλίας. Για αυτό το λόγο, με τη χρήση των παραπάνω σημείων κατασκευάζεται μια έλλειψη γύρω από αυτά τα σημεία η οποία μοντελοποιεί αισθητά καλύτερα το σχήμα που παίρνει το στόμα όταν ανοίγει και κλείνει. Στο παρακάτω σχήμα έχουμε πάρει τα αρνητικά χρώματα της εικόνας ώστε να είναι εμφανής η έλλειψη που περιβάλλει την περιοχή των χειλιών.

Σχήμα 54: Η χρήση της έλλειψης στη δυαδική εικόνα μοντελοποιεί το σχήμα των χειλιών κατά τη διάρκεια της ομιλίας

Επομένως για κάθε πρόσωπο που ανιχνεύουμε και προσδιορίσουμε την ευρύτερη περιοχή του στόματος, μετά από κάποιους υπολογισμούς, τοποθετούμε μια έλλειψη η οποία περιγράφει με καλή ακρίβεια την περιοχή που υπάρχει το στόμα στην εικόνα. Η μετρική που χρησιμοποιείται για τις ελλείψεις δύο διαδοχικών καρέ θα είναι επομένως:

$$M = mean\left(\frac{|\alpha_{next} - \alpha_{cur}|}{\alpha_{next}}\ ,\ \frac{|\beta_{next} - \beta_{cur}|}{\beta_{next}}\right)$$

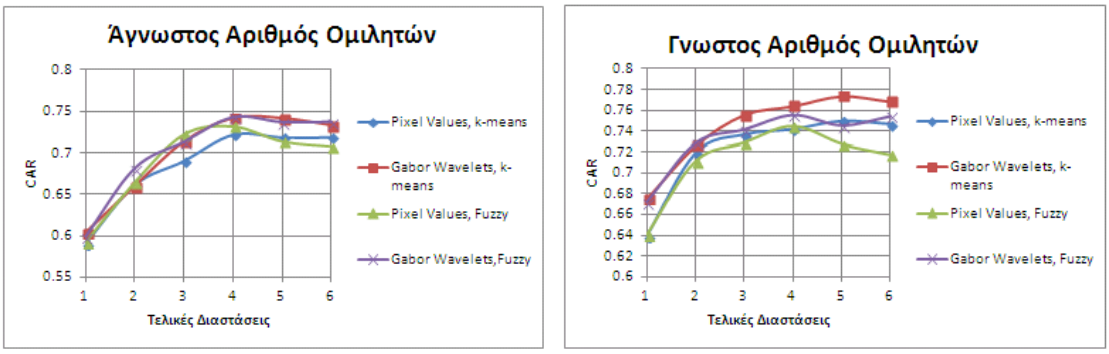
όπου α, β οι δύο ημιάξονες συμμετρίας της. Επειδή όμως για κάθε δευτερόλεπτο του βίντεο διαβάζουμε πέντε καρέ παίρνουμε εν τέλει τη μέση τιμή αυτής της μετρικής για αυτά τα πέντε καρέ ώστε να προκύπτει μια τιμή για κάθε δευτερόλεπτο. Η τελική γραφική παράσταση που προκύπτει για την ανίχνευση της κίνησης των χειλιών θα είναι:



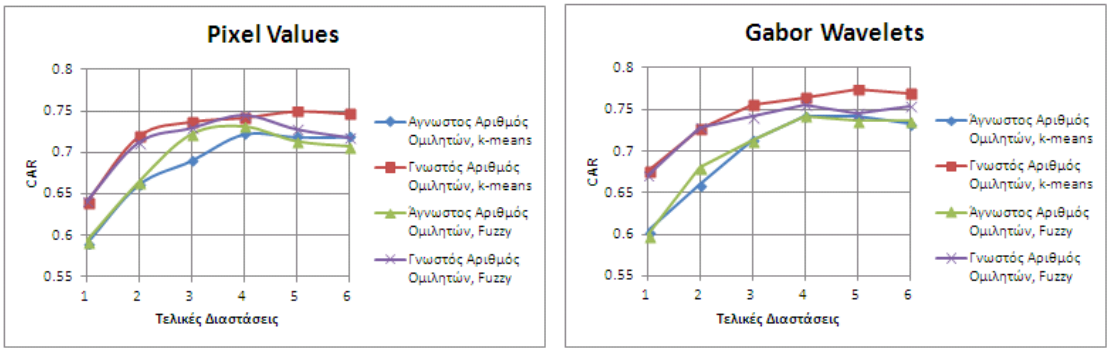
Σχήμα 55: Γραφική Παράσταση πλάτους της μεταβολής της θέσης της έλλειψης που προσδιορίζει την περιοχή των χειλιών

7.3 Αξιολόγηση των Πειραμάτων της Ημερολογιοποίησης Ομιλητών

Η διαδικασία που ακολουθήσαμε για την εκτέλεση των πειραμάτων είναι όμοια με πριν. Η μόνη διαφορά είναι πως εισάγαμε μια ακόμα παράμετρο ανάλογα με το αν τα αποτελέσματα που εξάγουμε είναι φιλτραρισμένα με κάποια μετρική εμπιστοσύνης ή όχι. Έχουμε δηλαδή διαφορετικά αποτελέσματα ανάλογα με το αν επιδιώκουμε να δώσουμε απάντηση για το ποιος είναι ο ομιλητής σε κάθε δευτερόλεπτο του βίντεο ή μόνο σε αυτά για τα οποία είμαστε σίγουροι. Στην πρώτη περίπτωση πηγαίνουμε στον πίνακα που περιέχει τις ετικέτες των ομιλητών για κάθε δευτερόλεπτο και όπου εμφανίζεται η τιμή μηδέν(δηλαδή είτε έχουμε παραπάνω από ένα πρόσωπα στο καρέ είτε επειδή δε βρέθηκε πρόσωπο) αποδίδουμε σε εκείνο το δευτερόλεπτο μια νέα ετικέτα την οποία την λαμβάνουμε έχοντας ως κριτήριο το ποια είναι η πιο κοντινή γειτονικά. Για παράδειγμα αν έχουμε ένα πίνακα $[1100022033]$ αυτός θα μετατραπεί σε $[1111222233]$. Επομένως τα τελικά αποτελέσματα θα είναι:



Σχήμα 56: Σύγκριση μεθόδων εξαγωγής χαρακτηριστικών και ομαδοποίησης για Άγνωστο (α) και Γνωστό (β) αριθμό Ομιλητών



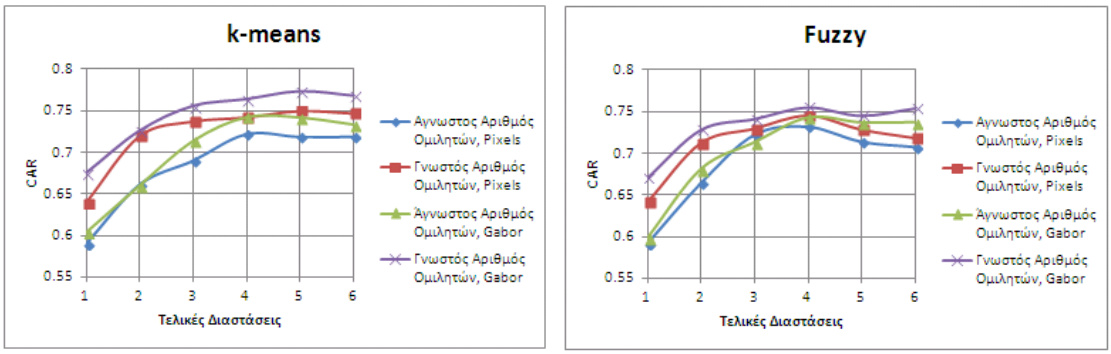
Σχήμα 57: Σύγκριση γνωστού και αγνώστου αριθμού ομιλητών και ομαδοποίησης για Pixels (α) και Gabor κυματιδίων (β)

Από τα παραπάνω διαγράμματα το πρώτο πράγμα που παρατηρεί κανείς είναι ότι η ιδανικότερες διαστάσεις ώστε να μειώσουμε σε αυτές τα δεδομένα μας είναι οι 4. Τα αποτελέσματα για μικρότερες διαστάσεις ξεκινάνε από αρκετά μικρότερες τιμές (περίπου 15% κάτω από τη μέγιστη τιμή) και αυξάνονται σταδιακά, ενώ για μεγαλύτερες διαστάσεις αν και παρουσιάζεται μείωση βλέπουμε πως οι τιμές δεν αποκλίνουν περισσότερο από $3 \rightarrow 4\%$. Θα πρέπει να επισημανθεί πως η βέλτιστη επιλογή των 4 διαστάσεων είναι όμοια με αντίστοιχες έρευνες και πειράματα που έχουν γίνει στο ίδιο σετ δεδομένων με βάση όμως την ακουστική πληροφορία [23].

Όπως και στα πειράματα της Ημερολογιοποίησης Προσώπου έτσι και στα αντίστοιχα της Ημερολογιοποίησης Ομιλητών η εξαγωγή των χαρακτηριστικών με τη βοήθεια των Gabor κυματιδίων αποτελεί την ιδανική επιλογή. Τα αποτελέσματα που παρουσιάζει είναι καλύτερα σε όλες τις περιπτώσεις ανεξάρτητα από τη μέθοδο ομαδοποίησης, τις διαστάσεις στις οποίες μειώνουμε με lda τον αρχικό χώρο καθώς και αν είναι γνωστός εκ των προτέρων ο αριθμός των ομιλητών του βίντεο.

Όσον αφορά τη σύγκριση των μεθόδων ομαδοποίησης όπως και στο προηγούμενο κεφάλαιο έτσι και εδώ τα αποτελέσματα που παρουσιάζουν είναι αρκετά κοντά μεταξύ τους με τη μέθοδο της Fuzzy ομαδοποίησης να είναι καλύτερη από την k-means για περίπου 1%.

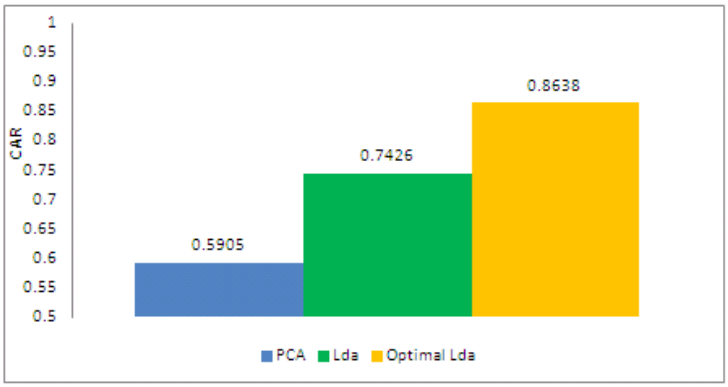
Τέλος όταν στη μέθοδο της ομαδοποίησης δίνουμε εκ των προτέρων τον αριθμό των ομιλητών-ομάδων που πρέπει να βρει αντί να την αφήνουμε να βρει μόνη της με τη βοήθεια της silhouette μετρικής, τα αποτελέσματα που παίρνουμε είναι καλύτερα σε όλες τις περιπτώσεις όπως αναμέναμε.



Σχήμα 58: Σύγκριση γνωστού και αγνώστου αριθμού ομιλητών και μεθόδων εξαγωγής χαρακτηριστικών για k-means (α) και fuzzy (β)

Παρόλα αυτά η διαφορά τους δε ξεπερνά στις περισσότερες περιπτώσεις το 5 → 7% ενώ αν διαλέξουμε τη βέλτιστη επιλογή των παραπάνω παραμέτρων (δηλαδή Gabor κυματίδια, 4 διαστάσεις για τον τελικό χώρο και Fuzzy ομαδοποίηση, η διαφορά που υπάρχει είναι της τάξης του 1%.

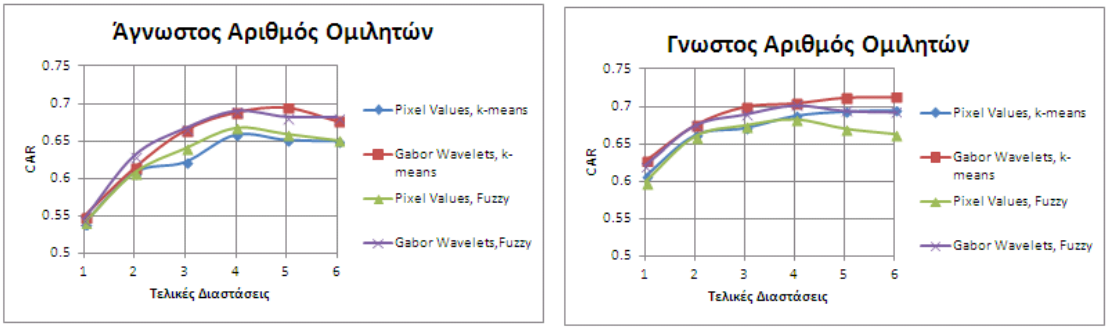
Τέλος για λόγους πληρότητας παρουσιάζονται παρακάτω τα αποτελέσματα που εξάγουμε ανάλογα με τη μέθοδο μείωσης των διαστάσεων και για τη βέλτιστη επιλογή παραμέτρων δηλαδή για Gabor κυματίδια, για 4 τελικές διαστάσεις και για fuzzy ομαδοποίηση.



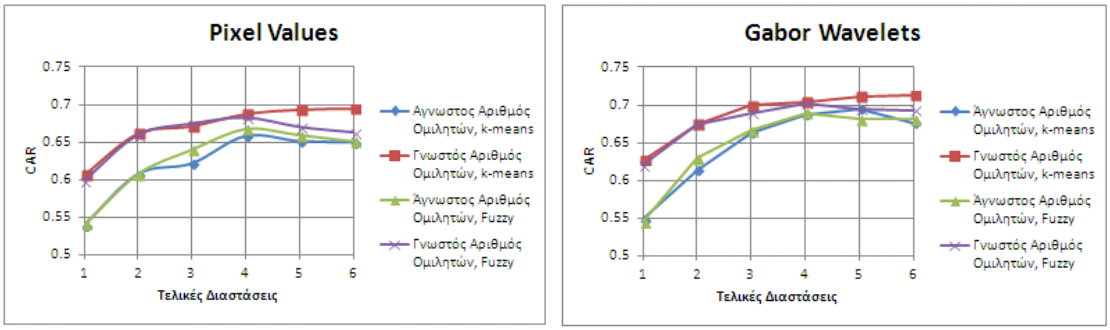
Σχήμα 59: Σύγκριση μεθόδων μείωσης των διαστάσεων όταν απαντάμε για τις χρονικές στιγμές για τις οποίες είμαστε σίγουροι

Βλέπουμε πως συγκριτικά με τη βέλτιστη FLsD όπου πετυχαίνει 86.38% με τη μέθοδο που εφαρμόζουμε πετυχαίνουμε 74.26% ενώ αντίστοιχα η pca πετυχαίνει 59.05%. Σε αυτή την περίπτωση όπου ασχολούμαστε με πειράματα Ημερολογιοποίησης Ομιλητών και όχι Προσώπου είναι αδύνατο να εξαχθούν αποτελέσματα για τον αρχικό χώρο καθώς καμία από τις μεθόδους ομαδοποίησης δε μπορεί να ανταποκριθεί σε ένα πρόβλημα ομαδοποίησης χαρακτηριστικών, διαστάσεων 8000×2500.

Όταν αντίστοιχα είναι επιθυμητό να δοθεί απάντηση για κάθε δευτερόλεπτο μέσα στο βίντεο τα αποτελέσματα που προκύπτουν θα είναι:



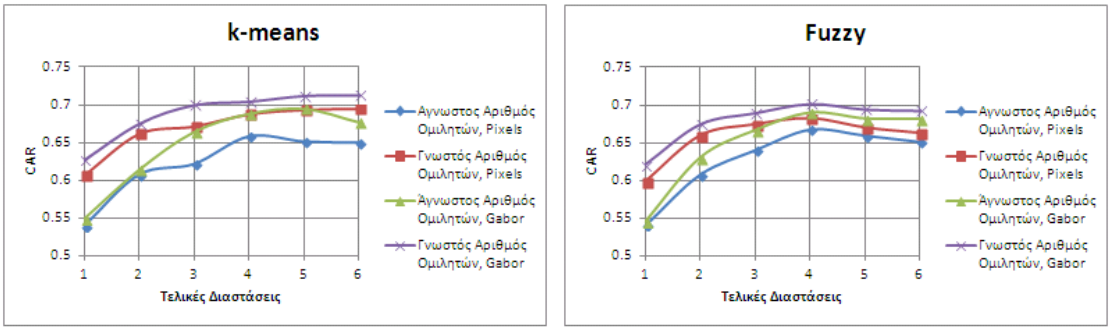
Σχήμα 60: Σύγκριση μεθόδων εξαγωγής χαρακτηριστικών και ομαδοποίησης για Άγνωστο (α) και Γνωστό (β) αριθμό Ομιλητών



Σχήμα 61: Σύγκριση γνωστού και αγνώστου αριθμού ομιλητών και ομαδοποίησης για Pixels (α) και Gabor κυματιδίων (β)

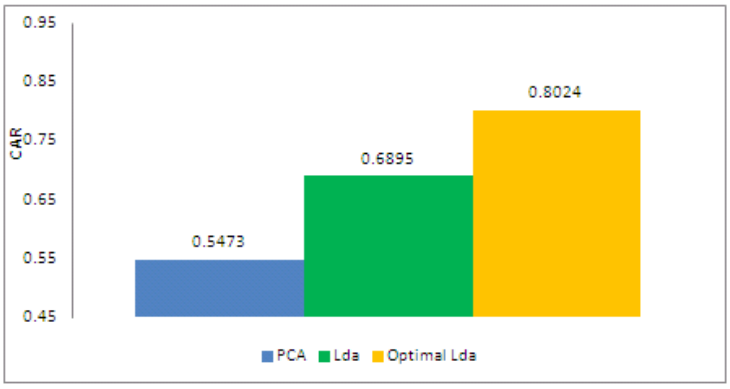
Αρχικά θα πρέπει να αναφέρουμε πως η παραδοχή που γίνεται ότι κάθε στιγμή θα υπάρχει ένας ομιλητής που θα μιλάει είναι ρεαλιστική καθώς στο σετ δεδομένων του Canal9 οι συζητήσεις που γίνονται είναι αρκετά πυκνές σε λόγο με αποτέλεσμα να μην έχουμε κενά διαστήματα όπου δεν υπάρχει ομιλία. Συχνά σε πειράματα Ημερολογιοποίησης Ομιλητών με βάση την ακουστική πληροφορία πραγματοποιείται αφαίρεση των τμημάτων όπου δεν υπάρχει ομιλία από κάθε αρχείο, ώστε το τελικό αποτέλεσμα να εμπεριέχει αποκλειστικά τμήματα λόγου.

Όπως και στην προηγούμενη περίπτωση έτσι και εδώ η βέλτιστη επιλογή μεθόδων είναι η εξαγωγή χαρακτηριστικών με Gabor κυματίδια η μείωση του αρχικού χώρου στις 4 διαστάσεις και η ομαδοποίηση με GK fuzzy. Σε αντίθεση με όταν απαντάμε μόνο για τα τμήματα για τα οποία είμαστε σίγουροι, σε αυτή την περίπτωση τα αποτελέσματα που προκύπτουν είναι μειωμένα κατά $5 \rightarrow 6\%$.



Σχήμα 62: Σύγκριση γνωστού και αγνώστου αριθμού ομιλητών και μεθόδων εξαγωγής χαρακτηριστικών για k-means (α) και fuzzy (β)

Συγκρίνοντας τις μεθόδους μείωσης των διαστάσεων βλέπουμε πως η βέλτιστη FLsD είναι στο 80.24% δηλαδή 6% κάτω σε σχέση με πριν ενώ αντίστοιχα η επιλογή που κάνουμε πετυχαίνει 68.95%. Τέλος θα πρέπει να τονίσουμε πως σε σχέση με την pca όπου δεν υπάρχει επίβλεψη, η μέθοδος μείωσης των διαστάσεων που προτείνουμε όντας ημιεπιβλεπώμενη πετυχαίνει αποτελέσματα $12 \rightarrow 15\%$ καλύτερα.



Σχήμα 63: Σύγκριση μεθόδων μείωσης των διαστάσεων όταν ζητείται να απαντήσουμε για κάθε δευτερόλεπτο

Επιπλέον θα πρέπει να τονισθεί πως το υπόλοιπο 19.76% από τη βέλτιστη FLsD είναι περιπτώσεις όπου στο εικονιζόμενο καρέ εμφανίζεται ένα πρόσωπο ενώ ο ομιλητής είναι διαφορετικός, συμπεριφορά αρκετά συνηθισμένη τόσο στο σετ δεδομένων του Canal9 με το οποίο δουλέψαμε όσο και σε κάποια συνηθισμένη συζήτηση που προβάλλεται σε ένα δελτίο ειδήσεων.

Κεφάλαιο 8

Συμπεράσματα

8.1 Συμβολή της Διπλωματικής Εργασίας

Η παρούσα διπλωματική εργασία ασχολήθηκε με την Ημερολογιοποίηση Ομιλητών με βάση την οπτική πληροφορία που εξάγεται από ένα βίντεο. Πρωταρχικός μας στόχος, ήταν η μελέτη των ήδη υπαρχόντων μεθόδων Ημερολογιοποίησης Ομιλητών που βασίζονται είτε στην ακουστική είτε σε μια πολυμεσική πληροφορία ώστε να επιλεγεί ο συνδυασμός εκείνος των μεθόδων που θα έδινε τα καλύτερα αποτελέσματα ικανοποιώντας παράλληλα και τις αρχικές απαιτήσεις του προβλήματος. Στα πλαίσια της εργασίας, μελετήσαμε σε βάθος

Μια από τις συνεισφορές της διπλωματικής εργασίας είναι ο χωρισμός ενός βίντεο σε επιμέρους shots παίρνοντας τη διαφορά των ιστογραμμάτων μεταξύ δύο διαδοχικών καρέ και σε συνδυασμό με την παραδοχή που κάνουμε πως εντός ενός shot θα υπάρχει αναγκαστικά ένα πρόσωπο ομιλητής (αφού αν υπάρχει κάποιος δεύτερος τότε η κάμερα θα πρέπει να στραφεί σε αυτόν οπότε είτε θα υπάρξει κίνηση της είτε θα αλλάξει τελείως το περιεχόμενο μέσα στο καρέ) βελτιώνουμε αισθητά τα αποτελέσματα σε σχέση με διαφορετικές υλοποιήσεις πετυχαίνοντας - για τα πλαίσια του βίντεο αποκλειστικά - μεγάλη ακρίβεια. Δευτερευόντως, δείξαμε πως η επιλογή του χώρου των χαρακτηριστικών παίζει καθοριστικό ρόλο στην ακρίβεια της Ημερολογιοποίησης. Η χρήση των Gabor κυματιδίων με 5 διαφορετικές κλίμακες και 8 διαφορετικούς προσανατολισμούς εξάγει χαρακτηριστικά από το πρόσωπο ιδιαίτερα περιεκτικά σε πληροφορία. Πραγματοποιώντας πειράματα για διαφορετικές μεθόδους εξαγωγής χαρακτηριστικών διαπιστώσαμε πως η χρήση των Gabor κυματιδίων συμβάλει ευεργετικά στο αρχικό πρόβλημα χωρίς να επιδρά αρνητικά στην ταχύτητα εκτέλεσης των πειραμάτων.

Επιπλέον θα πρέπει να τονιστεί, πως καίρια συμβολή της παρούσας διπλωματικής εργασίας είναι η μείωση των διαστάσεων με τη χρήση της ημιεπιβλεπώμενης FLsD που εκμεταλλεύεται την ύπαρξη σχετικών πληροφοριών. Εκτελώντας πειράματα σύγκρισης των διαφορετικών μεθόδων τόσο με τον αρχικό χώρο των χαρακτηριστικών όσο και με τα βέλτιστα αποτελέσματα που μπορεί να πετύχει η FLsD είδαμε πως για τα πειράματα Ημερολογιοποίησης Προσώπου με τη χρήση της υπάρχει αύξηση από το 86.3% του αρχικού χώρου, σε 93.4% σε ένα χώρο όπου έχουμε αφαιρέσει άσχετη με τον ομιλητή πληροφορία.

Παρόμοια συμπεράσματα, προέκυψαν και από τα πειράματα της Ημερολογιοποίησης Ομιλητών. Με κατάλληλη επιλογή μεθόδων η χρήση της FLsD απαντά σωστά στο αρχικό ερώτημα του "Ποιος μίλησε και πότε" με ακρίβεια 74.26% για τα τμήματα του βίντεο που είμαστε σίγουροι ενώ για όλο το βίντεο επιτυγχάνει ακρίβεια 68.95%. Συγκριτικά με την pca τα αποτελέσματα που παίρνουμε είναι $10 \rightarrow 15\%$ καλύτερα ενώ είναι μόνο 6% χειρότερα σε σχέση με τη βέλτιστη FLsD.

Συμπερασματικά η συμβολή της διπλωματικής εργασίας στο πρόβλημα της Ημερολογιοποίησης Ομιλητών από τη σκοπιά του βίντεο συνοψίζεται στα παρακάτω σημεία:

- Εισαγωγή στο πρόβλημα της Ημερολογιοποίησης Ομιλητών με βάση την οπτική πληροφορία και γίνεται αναφορά στο υπάρχων state-of-the-art καλύπτοντας προσεγγίσεις τόσο από τη σκοπιά του audio όσο και από μια πολύμεσική σκοπιά.

- Παρέχεται μια λεπτομερής επισκόπηση μεθόδων αλλαγής shot σε ένα βίντεο και τονίζεται η καθοριστική συμβολή τους στην Ημερολογιοποίηση Ομιλητών με βάση την οπτική πληροφορία.

- Εξετάζεται σε βάθος η μέθοδος ανίχνευσης προσώπου των Viola & Jones και στη συνέχεια ενσωματώνουμε την ανίχνευση δέρματος ώστε να απορριφθούν περιπτώσεις λανθασμένης ανίχνευσης. Παράλληλα μελετήσαμε τρόπους και μεθόδους εξαγωγής χαρακτηριστικών από το πρόσωπο ώστε να δοθεί μια περιεκτική σε πληροφορία αναπαράσταση του προσώπου.

- Παρουσιάζονται μέθοδοι μείωσης των διαστάσεων του αρχικού χώρου και υλοποιούμε την FLsD μέθοδο μείωσης των διαστάσεων η οποία εμφανίζει πολύ καλύτερα αποτελέσματα από τις κλασσικές υλοποίησεις της pca και lda.

- Ταυτόχρονα μελετάμε τρόπους ομαδοποίησης των δεδομένων σε ομάδες κάθε μία από τις οποίες αντιστοιχεί σε ένα πρόσωπο - ομιλητή.

- Διεξάγονται πειράματα Ημερολογιοποίησης Προσώπου συγκρίνοντας τα αποτελέσματα για πληθώρα διαφορετικών συνδυασμών των παραμέτρων που χρησιμοποιούνται. Αρχικά επιλέγουμε τη βέλτιστη επιλογή παραμέτρων που θα μπορούσε να χρησιμοποιηθεί για την κατασκευή ενός συστήματος που επιθυμεί να δώσει απάντηση στο ερώτημα "Ποιο πρόσωπο εμφανίζεται και πότε".

- Για τις παραμέτρους αυτές βρίσκουμε τόσο ποια μέθοδος μείωσης των διαστάσεων βελτιώνει καλύτερα τον αρχικό χώρο ενώ παράλληλα παρουσιάζουμε διαγράμματα που συγκρίνουν τη συμπεριφορά του αρχικού χώρου σε σχέση με τον μειωμένο όταν αυξάνεται ο αριθμός ομιλητών στο βίντεο.

- Σχεδιάζεται και υλοποιείται μια μέθοδος ανίχνευσης κίνησης των χειλιών ώστε να γίνει η μετάβαση στην Ημερολογιοποίηση Ομιλητών. Επιπλέον πραγματοποιήσαμε μια πληθώρα πειραμάτων Ημερολογιοποίησης Ομιλητών εξάγουμε τις βέλτιστες επιλογές για την κατασκευή ενός μελλοντικού τελικού συστήματος και σχολιάζουμε τα αποτελέσματα.

8.2 Μελλοντικές Κατευθύνσεις

Κρίνοντας από τα αποτελέσματα είμαστε πεπεισμένοι πως υπάρχει μεγάλο περιθώριο βελτίωσης. Ένας τρόπος να γίνει αυτό θα ήταν να πραγματοποιηθεί μια στοιχειώδης αποδόμηση του προσώπου σε επιμέρους μέρη και στη συνέχεια να εξαχθούν τα χαρακτηριστικά. Ταυτόχρονα υπάρχει πολύς χώρος για έρευνα ως προς τη μέθοδο εξαγωγής χαρακτηριστικών που θα χρησιμοποιηθεί παρόλο που για την περιοχή του προσώπου η χρήση των Gabor κυματιδίων αποτελεί την πιο διαδεδομένη μέθοδο στη βιβλιογραφία.

Επιπλέον, η σύγκριση της μεθόδου με τη θεωρητική FLsD όπου παίρνουμε τα βέλτιστα αποτελέσματα μας οδηγεί στο συμπέρασμα ότι υπάρχει και σε αυτό το τμήμα του συστήματος περιθώριο για εξέλιξη. Ένα από τα βασικότερα προβλήματα που καλείται να λύσει η Ημερολογιοποίηση Ομιλητών και δεν έχει αναπτυχθεί σε βάθος είναι οι περιπτώσεις όπου εμφανίζεται επικαλυπτόμενη ομιλία από τους ομιλητές. Ένας τρόπος αντιμετώπισης αυτού του προβλήματος είναι να χρησιμοποιηθεί πρωτίστως κάποιου είδους ανίχνευση των σημείων όπου υπάρχει επικαλυπτόμενη ομιλία, ενώ δευτερευόντως θα μπορούσε να χρησιμοποιηθεί η FLsD ώστε να εξάγει το βέλτιστο εκείνο υποχώρο που θα διαχωρίζει τον επικαλυπτόμενο από την μη επικαλυπτόμενο λόγο.

Εν συνεχεία των παραπάνω, η απόδοση της Ημερολογιοποίησης μπορεί να βελτιωθεί ενσωματώνοντας πιο λεπτομερείς και περίπλοκους τρόπους μοντελοποίησης τόσο της συμπεριφοράς των ομιλητών όταν παίρνουν τον λόγο όσο και του ρόλου που μπορεί να παίζουν στη συζήτηση. Παράλληλα,

θα μπορούσε να γίνει αντικείμενο για περεταίρω έρευνα η επέκταση της μεθόδου Ημερολογιοποίησης Ομιλητών με βάση την οπτική πληροφορία σε πραγματικό χρόνο. Κάτι τέτοιο συνεπάγεται την παρακολούθηση του βέλτιστου υποχώρου που είναι σχετικός με τον ομιλητή κάθε στιγμή σε συνδυασμό με την ταυτόχρονη ομαδοποίηση του σε ομάδες.

Τέλος η μέθοδος που υλοποιήσαμε μπορεί να χρησιμοποιηθεί σαν τμήμα μιας διαδικασίας Ημερολογιοποίησης Ομιλητών που χρησιμοποιεί τόσο την πληροφορία που εξάγεται από το βίντεο όσο και αυτή από το audio. Είναι λογικό η Ημερολογιοποίηση Ομιλητών να δουλεύει καλύτερα όταν πραγματοποιείται με βάση την ακουστική πληροφορία σε σχέση με την οπτική. Με τη χρήση όμως μιας συνδυαστικής μεθόδου είναι δυνατόν να επιτευχθούν ακόμα καλύτερα αποτελέσματα καθώς στις περιοχές όπου θα αποτυγχάνει η Ημερολογιοποίηση Ομιλητών με βάση την ακουστική πληροφορία θα χρησιμοποιούμε τα αποτελέσματα που εξάγουμε από την οπτική ώστε να βελτιώσουμε τη συνολική απόδοση.

Παράρτημα Α: Μέθοδοι Αναγνώρισης Προσώπου από ένα Σετ Δεδομένων

Δοσμένου ενός σετ εικόνων προσώπου στις οποίες έχουν δοθεί ταμπέλες με την ταυτότητα του κάθε ανθρώπου (σετ εκμάθησης) και ενός σετ εικόνων προσώπου χωρίς ταμπέλες από το ίδιο γκρουπ από ανθρώπους (test σετ) στόχος μας είναι η αναγνώριση του κάθε προσώπου στις εικόνες που τεστάρουμε.

SIFT και SURF

Επειδή στην αναγνώριση προσώπου οι εικόνες προσώπου είναι συνήθως σε ευθύ προσανατολισμό και κανονικοποιημένες, προκύπτει πως τα σημεία που χρησιμοποιούνται για ταίριασμα στις δύο εικόνες θα βρίσκονται σε όμοιες θέσεις μέσα στις δύο εικόνες προσώπου. Έτσι, για ένα σημείο ενδιαφέροντος (x, y) της εξεταζόμενης εικόνας, η περιοχή αναζήτησης του αντίστοιχου σημείου περιορίζεται εντός ενός ορθογώνιου παραθύρου με κέντρο το (x, y) της δεύτερης εικόνας. Το ταίριασμα σημείων με τη μικρότερη απόσταση μεταξύ των περιγραφέων θα θεωρείται σαν ένα υποψήφιο ζευγάρι ταιριάσματος. Για να επιβεβαιωθεί η εγκυρότητα του υποψήφιου ταιριάσματος σημείων, η επόμενη ελάχιστη απόσταση που περιέχει το ίδιο σημείο στην εξεταζόμενη εικόνα (και κάποιο άλλο κατ΄ επέκταση στη δεύτερη) αναζητείται σε ολόκληρη την εικόνα και όχι σε μια υποπεριοχή της όπως πριν. Αν ο λόγος των δύο αποστάσεων είναι μικρότερος από ένα προκαθορισμένο κατώφλι τότε το ζευγάρι των σημείων με τη μικρότερη απόσταση επιβεβαιώνεται σαν ένα ζευγάρι ταιριάσματος. Καθώς η πληροφορία της τοποθεσίας εμφανίζεται στην αναζήτηση του ζευγαριού σημείων με την ελάχιστη απόσταση, ενώ ο λόγος της ελάχιστης απόστασης και της επόμενης σε σειρά ελάχιστης απόστασης μετρά την αξιοπιστία της αντιστοιχίας των δύο σημείων ενδιαφέροντος, η παραπάνω μέθοδος μπορεί να αποφύγει αποδοτικά λάθη αναντιστοιχιών. Τέλος, ορίζεται μια μετρική ομοιότητας για το λόγο των αποστάσεων όλων των σημείων ταιριάσματος για την αναγνώριση προσώπου:

$$\text{Μετρική Ομοιότητας} = \begin{cases} (\text{Μέση Απόσταση} + \text{Μέσος Λόγος})/2 & \text{αν } N \geq T \\ (\text{Μέση Απόσταση} + \text{Μέσος Λόγος})/2 + 1 & \text{αν } N < T \end{cases}$$

$$\text{Μέση Απόσταση} = \frac{1}{N} \sum_{n}^{N} \text{Ελάχιστη Απόσταση}$$

$$\text{Μέσος Λόγος} = \frac{1}{N} \sum_{n}^{N} \text{Λόγος Αποστάσεων}$$

όπου N είναι ο αριθμός των σημείων που ταιριάζονται στις δύο εικόνες προσώπου,T το προκαθορισμένο κατώφλι, *Ελάχιστη Απόσταση* είναι η Ευκλείδεια απόσταση μεταξύ δύο σημείων στα οποία υπάρχει αντιστοιχία και *Λόγος Αποστάσεων* είναι ο λόγος των αποστάσεων των σημείων ταιριάσματος. Όταν ο αριθμός των σημείων ταιριάσματος δύο εικόνων είναι μικρότερος από το κατώφλι που έχει τεθεί, θεωρείται πως το αποτέλεσμα δεν μπορεί να θεωρηθεί αξιόπιστο. Στο παρακάτω σχήμα παρουσιάζεται το αποτέλεσμα του ταιριάσματος μεταξύ σημείων σε δύο εικόνες προσώπου με τις κόκκινες γραμμές να υποδηλώνουν την αντιστοιχία μεταξύ των δύο σημείων που ενώνουν.

(α′) (β′)

Σχήμα 64: Ταίριασμα SIFT (α) και SURF (β) χαρακτηριστικών μεταξύ 2 διαφορετικών φωτογραφιών του ίδιου προσώπου

Eigenfaces

Μια σύγχρονη μέθοδος που χρησιμοποιείται στη βιβλιογραφία, αντί να βρίσκει τις περιοχές των διακριτικών χαρακτηριστικών της εικόνας όπως τα μάτια, η μύτη και το στόμα και να μετράει τις αποστάσεις μεταξύ τους, βασίζεται στη σύγκριση μεταξύ εικόνων κλίμακας γκρι που προβάλλονται σε υποχώρους μικρότερων διαστάσεων που ονομάζονται eigenfaces. Τα Eigenfaces ανήκουν στην κατηγορία των προσεγγίσεων στην αναγνώριση προσώπου που βασίζονται στην εμφάνιση (Appearance-based) και στοχεύουν στο να εντοπίσουν τη διασπορά που εμφανίζεται σε μια συλλογή από εικόνες προσώπων ώστε να χρησιμοποιήσουν αυτή την πληροφορία για να κωδικοποιήσουν και να συγκρίνουν εικόνες από πρόσωπα με ένα ολιστικό (holistic)τρόπο (και όχι με έναν που βασίζεται σε συγκεκριμένα τμήματα ή σε κάποια χαρακτηριστικά). Ειδικότερα, τα eigenfaces είναι τα κύρια συστατικά μιας συλλογής από πρόσωπα, ή ισοδύναμα, τα ιδιοδιανύσματα του πίνακα συνδιακύμανσης ενός σετ από εικόνων προσώπων, όπου μια εικόνα με N pixels θεωρείται ως ένα σημείο (ή διάνυσμα) σε ένα χώρο N διαστάσεων.

Η ιδέα του να χρησιμοποιήσει κανείς βασικά συστατικά για να αναπαραστήσει ένα ανθρώπινο πρόσωπο αναπτύχθηκε από τους Sirovich & Kirby [42] και χρησιμοποιήθηκε από τους Turk and Pentland [46] για ανίχνευση και αναγνώριση προσώπου. Τα Eigenfaces συχνά θεωρούνται σαν η baseline μέθοδος σύγκρισης για να αποδείξει κανείς την ελάχιστη αναμενόμενη απόδοση ενός τέτοιου συστήματος. Ο σκοπός των eigenfaces είναι διττός και έχει να κάνει αρχικά με την εξαγωγή της σχετικής πληροφορίας από το πρόσωπο, η οποία μπορεί(ή και όχι) να είναι απευθείας συσχετισμένη με την ανθρώπινη διαίσθηση των χαρακτηριστικών του προσώπου(μάτια, μύτη, στόμα, χείλια κ.α.). Ένας τρόπος για να το κάνει κανείς αυτό είναι να εντοπίσει τη στατιστική μεταβολή μεταξύ των εικόνων του προσώπου. Δευτερευόντως, στόχος μιας τέτοιας προσέγγισης, είναι η αποδοτική αναπαράσταση των εικόνων του προσώπου. Για να μειωθεί η υπολογιστική πολυπλοκότητα, κάθε πρόσωπο μπορεί να αναπαρασταθεί χρησιμοποιώντας ένα μικρό αριθμό από παραμέτρους. Τα eigenfaces μπορούν να θεωρηθούν επομένως, σαν ένα σετ χαρακτηριστικών τα οποία χαρακτηρίζουν την συνολική διασπορά μεταξύ εικόνων από πρόσωπα. Στη συνέχεια, κάθε εικόνα ενός προσώπου μπορεί να προσεγγισθεί χρησιμοποιώντας ένα υποσύνολο από τα eigenfaces, αυτά που αντιστοιχούν στα μεγαλύτερα ιδιοδιανύσματα. Τα χαρακτηριστικά αυτά αντιπροσωπεύουν τη μεγαλύτερη διασπορά στο σετ που έγινε η εκπαίδευση.

Πριν παραχθούν τα eigenfaces, οι εικόνες προσώπου κανονικοποιούνται ώστε να παραταχθούν σε αντιστοιχία τα μάτια και το στόμα και στη συνέχεια κάνουμε σε όλες τις εικόνες δειγματοληψία ώστε να έχουν την ίδια ανάλυση. Τα background των εικόνων(ή πιθανές περιοχές που δεν περιέχουν πρόσωπο όπως τα μαλλιά ή ο λαιμός) είτε αφαιρούνται είτε πρέπει να είναι σταθερά.

Τα eigenfaces στη συνέχεια εξάγονται από τα δεδομένα των εικόνων με τη βοήθεια της PCA με τον παρακάτω τρόπο:

Αλγόριθμος 4 Αλγόριθμος εξαγωγής των eigenfaces από ένα σετ εικόνων προσώπου

1: *Δοσμένου ενός σετ* S *με* M *εικόνες προσώπων διαστάσεων* $h \times w$ *κάθε εικόνα μετασχηματίζεται σε ένα διάνυσμα* $D(= hw)$ *διαστάσεων και τοποθετείται στο σετ:*

$$S = \{\Gamma_1, \Gamma_2, \ldots, \Gamma_M\}$$

2: *Εξάγεται η μέση εικόνα:*

$$\Psi = \frac{1}{M}\sum_{i=1}^{M}\Gamma_i$$

3: *Υπολογίζεται η διαφορά* Φ *της εικόνας που δίνεται σαν είσοδος και της μέσης εικόνας:*

$$\Phi_i = \Gamma_i - \Psi$$

4: *Ορίζεται ένα σετ από ορθοκανονικά διανύσματα* u_n*, το οποίο περιγράφει καλύτερα την κατανομή των δεδομένων. Το* k *-οστό διάνυσμα* u_k *επιλέγεται έτσι ώστε το*

$$\lambda_k = \frac{1}{M}\sum_{i=1}^{M}(u_k^T\Phi_i)^2$$

να είναι μέγιστο υπό τον περιορισμό:

$$u_l^T u_k = \delta_{lk} = \begin{cases} 1 & \text{αν } l = k \\ 0 & \text{Αλλιώς} \end{cases}$$

όπου τα u_k *,* λ_k *είναι ιδιοδιανύσματα και ιδιοτιμές αντίστοιχα του πίνακα συνδιακύμανσης* C*. (Η συνδιακύμανση δύο μεταβλητών είναι η εκτιμώμενη τιμή του γινομένου των αποκλίσεων από τις μέσες τιμές τους και δίνεται από τον ακόλουθο τύπο:* $E[(x-E[x])(y-E[y])]$*)*

5: *Υπολογίζεται ο πίνακας συνδιακύμανσης* C*:*

$$C = \frac{1}{M}\sum_{i=1}^{M}(\Phi_i\Phi_i^T) = AA^T \ , \ A = \{\Phi_1, \Phi_2, \ldots, \Phi_M\} \in R^{D\times M}$$

6: *Επειδή ο υπολογισμός των ιδιοδιανυσμάτων του* C *είναι δύσκολος και μη αποδοτικός για τυπικές εικόνες όταν το* $D \gg M$ *προτιμάται να υπολογίζονται τα ιδιοδιανύσματα του αρκετά μικρότερου σε διαστάσεις* $(M \times M)$ *πίνακα* AA^T*. Τα ιδιοδιανύσματα και οι ιδιοτιμές του* AA^T *είναι:*

$$V = \{v_1, v_2, \ldots, v_r\} \ , \ \Lambda = diag\{\lambda_1, \lambda_2, \ldots, \lambda_r\} \ , \ \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r$$

με r *να είναι το rank του πίνακα* A*. Επίσης τα ιδιοδιανύσματα που αντιστοιχούν σε μηδενικές ιδιοτιμές έχουν απαλειφθεί.*

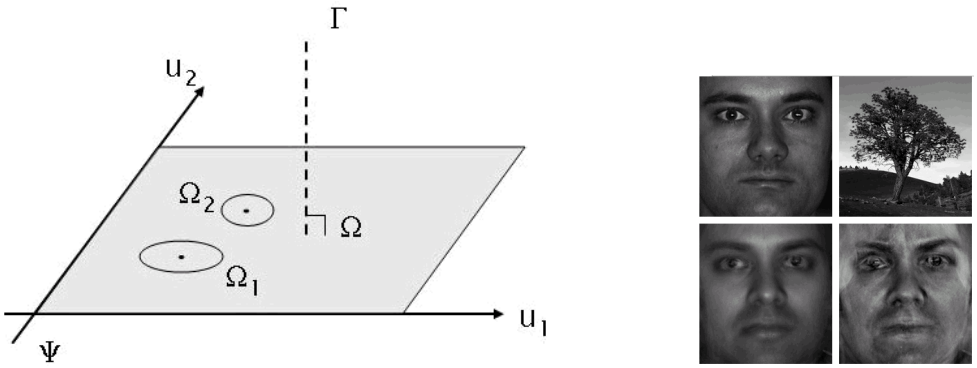
7: *Επομένως ο πίνακας ιδιοδιανυσμάτων του* C *ισουταί με* $U = AV\Lambda^{-1/2}$ *όπου* $U = \{u_i\}$ *είναι η συλλογή από τα eigenfaces*

Σχήμα 65: Αρχικό σετ δεδομένων, μέση εικόνα και εξαγόμενα Eigenfaces

Τα eigenfaces συνεπώς, span ένα υποχώρο m διαστάσεων του χώρου της αρχικής εικόνας επιλέγοντας εκείνο το υποσύνολο από ιδιοδιανύσματα $\hat{U} = \{u_1, \ldots, u_m\}$ το οποίο αντιστοιχεί στις m μεγαλύτερες ιδιοτιμές. Οι Turk & Pentland [47], ονομάζουν ως "χώρο προσώπου" (face space) ένα χώρο που δημιουργείται από το μέσο πρόσωπο και του οποίου οι άξονες είναι τα eigenfaces. Επειδή ο "χώρος του προσώπου" καθορίζει το χώρο των εικόνων προσώπου, η ανίχνευση προσώπου μπορεί να θεωρηθεί σαν η διαδικασία ανίχνευσης τμημάτων της εικόνας τα οποία βρίσκονται κοντά στο "χώρο του προσώπου". Με άλλα λόγια, η προβαλλόμενη απόσταση δ θα πρέπει να βρίσκεται μεταξύ κάποιου κατωφλίου θ_δ. Η απόσταση δ μεταξύ του σημείου και του χώρου, είναι η απόσταση μεταξύ της εικόνας του προσώπου και της προβολής της στο face space και υπολογίζεται:

$$\delta = \|(I - \hat{U}\hat{U}^T)(\Gamma - \Psi)\|$$

όπου I είναι ο μοναδιαίος πίνακας. Όπως φαίνεται και στην παρακάτω εικόνα η απόσταση μεταξύ μιας εικόνας και της προβολής της στο χώρο του προσώπου είναι πολύ μικρότερη για ένα πρόσωπο σε σχέση με μια εικόνα που δεν περιέχει πρόσωπο αλλά κάτι άλλο (π.χ. ένα δέντρο).



(α′) (β′)

Σχήμα 66: Face space δύο διαστάσεων με άξονες που αναπαριστούν δύο eigenfaces(α) και προβολές εικόνων προσώπου και άσχετων εικόνων σε αυτόν(β)

Όταν έρχεται μια νέα φωτογραφία προσώπου για αναγνώριση τότε η διαδικασία που ακολουθείται είναι η ακόλουθη:

Αλγόριθμος 5 Αλγόριθμος ανίχνευσης νέου προσώπου

1: *Το νέο πρόσωπο προβάλλεται στο face space:*

$$\Omega = \hat{U}^T(\Gamma - \Psi)$$

όπου το $\hat{U}$ *αποτελεί το σετ των πιο σημαντικών ιδιοδιανυσμάτων και το διάνυσμα με βάρη* Ω *είναι η αναπαράσταση του καινούργιου προσώπου στο face space.*

2: *Καθορίζεται ποια κλάση προσώπων* Γ *παρέχει την καλύτερη περιγραφή για την εικόνα. Αυτό πραγματοποιείται ελαχιστοποιώντας την Ευκλείδια απόσταση:*

$$\varepsilon_k = \|\Omega - \Omega_k\|$$

όπου Ω_k *είναι το διάνυσμα με βάρη που αντιστοιχεί στην* k *κλάση προσώπων.*

3: *Το πρόσωπο θεωρείται ότι ανήκει στην κλάση αν το* $\varepsilon_k < \theta_\varepsilon$ *και θεωρείται ότι η συγκεκριμένη εικόνα προσώπου συνιστά ένα γνωστό πρόσωπο. Σε αντίθετη περίπτωση κατηγοριοποιείται σαν άγνωστο πρόσωπο, και αν είναι επιθυμητό τοποθετείται στο αρχικό σετ εκπαίδευσης ώστε να χρησιμοποιηθεί σε μελλοντικές εφαρμογές αναγνώρισης προσώπου.*

Ο Szeliski [45] επισημαίνει πως ένα από τα μεγαλύτερα πλεονεκτήματα της χρήσης των eigenfaces είναι ότι μειώνουν τον αριθμό των συγκρίσεων που γίνονται καθώς γίνεται σύγκριση μόνο με τη μέση εικόνα του dataset σε ένα χώρο με λιγότερες διαστάσεις σε σχέση με τον αρχικό. Παρόλα αυτά, η μέθοδος αυτή είναι αρκετά ευαίσθητη σε μεταβολές του φωτισμού, της κλίμακας, της πόζας και της έκφρασης του προσώπου. Συνεπώς για να λειτουργήσει καλά θα πρέπει το πρόσωπο να παρουσιάζεται σε λήψη από μπροστά, σε κατάλληλη κλίμακα, με συγκεκριμένο φωτισμό και με προκαθορισμένη (συνήθως ουδέτερη) έκφραση. Επιπροσθέτως, η χρήση της Ευκλείδιας απόστασης, έχει το μειονέκτημα ότι στους άξονες (δηλαδή στα eigenfaces) - που όπως έχει αναφερθεί είναι ιδιοδιανύσματα κάποιων τυχαίων μεταβλητών - εμφανίζονται διαφορετικές κλίμακες. Η Ευκλείδεια απόσταση αντιμετωπίζει με τον ίδιο τρόπο κάθε μονάδα ανεξάρτητα από το μέγεθος και την κλίμακα του άξονα, παρόλο που οι αποστάσεις σε ένα κοντύτερο άξονα είναι αρκετά πιο ευαίσθητες σε αλλαγές. Για αυτό το λόγο, προτιμάται συχνά η Mahalanobis [12] απόσταση, η οποία έχει σχεδιασθεί ώστε να χρησιμοποιεί τον πίνακα συνδιακύμανσης ώστε να μετράει τις αποστάσεις με διαφορετικά βάρη.

Ένας άλλος τρόπος που προτείνεται στη βιβλιογραφία [36], για να βελτιωθεί η απόδοση των προσεγγίσεων που βασίζονται στα eigenfaces είναι να χωριστεί η εικόνα σε επιμέρους τμήματα όπως τα μάτια η μύτη και το στόμα και στη συνέχεια να γίνει αντιστοίχιση σε καθένα από αυτούς τους μικρότερους χώρους ανεξάρτητα. Το πλεονέκτημα μιας τέτοιας προσέγγισης είναι η ανεκτικότητα της σε ένα ευρύτερο προσανατολισμό του προσώπου μέσα στην εικόνα, επειδή κάθε τμήμα μπορεί να μετακινηθεί σε σχέση με τα υπόλοιπα που περιέχουν το αντίστοιχο περιεχόμενο. Ταυτόχρονα υποστηρίζει μεγαλύτερη ποικιλία συνδυασμών καθώς προσφέρει τη δυνατότητα να μοντελοποιηθεί ένα πρόσωπο με συγκεκριμένα χαρακτηριστικά, χωρίς να απαιτηθεί από τα eigenfaces να κάνουν span όλους τους πιθανούς συνδυασμούς από χαρακτηριστικά του προσώπου.

[12] Δοσμένου του πίνακα συνδιακύμανσης S η Mahalanobis απόσταση μεταξύ δύο τυχαίων διανυσμάτων x και y εκφράζεται: $d(x,y) = \sqrt{(x-y)^T S^{-1}(x-y)}$

Fisherfaces

Εν συνεχεία των παραπάνω, ένας τρόπος για να αναπαρασταθούν τα δεδομένα εισόδου, είναι να βρεθεί ένας υποχώρος ο οποίος αναπαριστά όσο το δυνατόν περισσότερη από τη διασπορά των δεδομένων. Αυτό μπορεί να γίνει με τη χρήση της PCA, η οποία όταν εφαρμόζεται σε εικόνες προσώπου επιστρέφει ένα σετ από eigenfaces. Τα eigenfaces είναι τα ιδιοδιανύσματα εκείνα τα οποία σχετίζονται με τις μεγαλύτερες ιδιοτιμές του πίνακα συνδιακύμανσης των δεδομένων εκπαίδευσης ενώ επίσης αντιστοιχούν και στις λύση των ελαχίστων τετραγώνων. Αυτός είναι ένας πολύ ισχυρός τρόπος αναπαράστασης των δεδομένων, καθώς εξασφαλίζει ότι η διασπορά των δεδομένων διατηρείται ενώ παράλληλα αφαιρεί τις αχρείαστες συσχετίσεις που εμφανίζονται μεταξύ των αρχικών χαρακτηριστικών (διαστάσεων) στα τυχαία διανύσματα. Από την άλλη, η προσέγγιση αυτή βρίσκει τις κατευθύνσεις των προβολών οι οποίες μεγιστοποιούν τη συνολική διασπορά μεταξύ όλων των κλάσεων (δηλαδή μεταξύ όλων των εικόνων από πρόσωπα). Κατά την επιλογή της προβολής που μεγιστοποιεί τη συνολική διασπορά, η PCA διατηρεί ανεπιθύμητες μεταβολές εξαιτίας του φωτισμού ή των εκφράσεων του προσώπου. Οι Moses et al. [1] αναφέρουν χαρακτηριστικά πως οι μεταβολές μεταξύ εικόνων του ίδιου πρόσωπου που οφείλονται στο φωτισμό και στον προσανατολισμό του προσώπου σε σχέση με την κάμερα είναι τις περισσότερες φορές μεγαλύτερες από μεταβολές που οφείλονται σε διαφορετική ταυτότητα του προσώπου. Έτσι, ενώ η προβολές με την PCA είναι οι βέλτιστες σε περιπτώσεις ανακατασκευής από μια βάση λίγων διαστάσεων, δεν παρουσιάζουν αντίστοιχη συμπεριφορά όσον αφορά τη διακριτική τους ικανότητα.

Οι Belhumeur et al. [6] επέλεξαν να χρησιμοποιήσουν τις κατευθύνσεις εκείνες των προβολών οι οποίες είναι σχεδόν ορθογωνικές ως προς την διασπορά εντός της κλάσης, αφαιρώντας μεταβολές που σχετίζονται με το φωτισμό και την έκφραση του προσώπου διατηρώντας ταυτόχρονα την διακριτική ικανότητα. Η μέθοδος τους, την οποία ονόμασαν FisherFaces, και είναι παράγωγο της FLD (Fisher's Linear Discriminant) μεγιστοποιεί το λόγο της διασποράς εντός της κλάσης προς τη διασπορά μεταξύ των κλάσεων.

Οι διαφορές εντός της κλάσης μπορούν να εκτιμηθούν χρησιμοποιώντας έναν πίνακα διασποράς εντός της κλάσης:

$$S_w = \sum_{j=1}^{C} \sum_{i=1}^{n_j} (x_{ij} - \mu_j)(x_{ij} - \mu_j)^T$$

όπου x_{ij} είναι το ι-οστό δείγμα της κλάσης j, μ_j η μέση τιμή της κλάσης j και n_j ο αριθμός των δειγμάτων στην κλάση j. Ομοίως η διαφορές μεταξύ των κλάσεων υπολογίζονται χρησιμοποιώντας τον πίνακα διασποράς μεταξύ των κλάσεων:

$$S_b = \sum_{j=1}^{C} (\mu_j - \mu)(\mu_j - \mu)^T$$

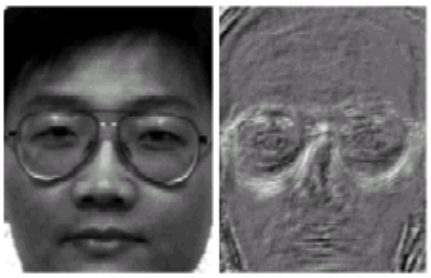
όπου μ είναι η μέση τιμή όλων των κλάσεων. Στη συνέχεια βρίσκονται εκείνα τα διανύσματα βάσης V όπου το S_w ελαχιστοποιείται και το S_b μεγιστοποιείται, όπου V είναι ένας πίνακας του οποίου οι στήλες v_i είναι τα διανύσματα βάσης που ορίζουν τον υπόχωρο. Αυτά δίνονται από:

$$\frac{|V^T S_b V|}{|V^T S_w V|}$$

Η λύση στο παραπάνω πρόβλημα δίνεται από τη γενικευμένη αποσύνθεση ιδιοτιμών:

$$S_b V = S_w V \Lambda$$

με Λ να είναι ο διαγώνιος πίνακας των αντίστοιχων ιδιοτιμών. Τα ιδιοδιανύσματα του V που αντιστοιχούν σε μη μηδενικές ιδιοτιμές είναι τα Fisherfaces. Από τον ορισμό του S_b γίνεται αντιληπτό πως υπάρχει ένας μέγιστος αριθμός από $C-1$ Fisherfaces αφού C διανύσματα ορίζουν ένα υποχώρο με $C-1$ ή λιγότερες διαστάσεις με την ισότητα να ισχύει όταν όλα τα ιδιοδιανύσματα είναι γραμμικώς ανεξάρτητα μεταξύ τους. Μια από τις κυριότερες εφαρμογές των Fisherfaces είναι το να αναγνωρίσουν αν ένα πρόσωπο φοράει γυαλιά όπως φαίνεται και στο παρακάτω σχήμα:



Σχήμα 67: Πρωτότυπη εικόνα και αντίστοιχο Fisherface για να διευκρινιστεί αν το πρόσωπο φοράει γυαλιά (επανεκτύπωση από [6])

Παράρτημα Β: Περιγραφή του Σετ Δεδομένων

Καθώς η αυτόματη ανάλυση των κοινωνικών αλληλεπιδράσεων αποσπά όλο και μεγαλύτερη προσοχή τα τελευταία χρόνια [22], η ανάγκη ύπαρξης κάποιων dataset προσβάσιμα από όλους σαν κοινό σημείο αναφοράς είναι ιδιαίτερα σημαντική για την εξέλιξη του κλάδου της επιστήμης των υπολογιστών. Τα benchmarks επιτρέπουν σε διαφορετικούς ερευτνητές να εφαρμόσουν τα ίδια πειραματικά πρωτόκολλα στα ίδια δεδομένα, καθώς αυτός είναι ο μόνος τρόπος για να πραγματοποιηθούν λεπτομερείς συγκρίσεις χρησιμοποιώντας διαφορετικές τεχνικές.

Το dataset που χρησιμοποιήθηκε για τα πειράματα είναι μια συλλογή από πολιτικά debates του Canal 9 με στόχο την ανάλυση των κοινωνικών αλληλεπιδράσεων και περιγράφεται λεπτομερώς από τους Vinciarelli et al. [49]. Από την σκοπιά της ανάλυσης κοινωνικών αλληλεπιδράσεων τα πολιτικά debates έχουν το θετικό χαρακτηριστικό ότι αποτελούν ρεαλιστικά παραδείγματα σε αντίθεση με τα περισσότερα benchmarks που χρησιμοποιούνται από την ευρευνητική κοινότητα. Οι συμμετέχοντες σε ένα debate συμπεριφέρονται με ιδιαίτερα ρεαλιστικό τρόπο καθώς τα γεγονότα αυτά έχουν ιδιαίτερη επίδραση στην πραγματική τους ζωή (για παράδειγμα μπορεί να επηρεάσουν το αποτέλεσμα των εκλογών). Έτσι, ακόμα και αν η δομή ενός debate μπορεί να εμπεριέχει κάποιους περιορισμούς, οι συμμετέχοντες παρουσιάζουν μια αρκετά αυθόρμητη κοινωνική συμπεριφορά.

Η συλλογή αποτελείται από 70 βίντεο συνολικής διάρκειας 43 ωρών και 10 λεπτών. Κάθε debate στρέφεται γύρω από την θετική ή αρνητική απάντηση σε ένα γενικότερο ερώτημα όπως "Είστε θετικοί απέναντι στους νέους νόμους για την επιστημονική έρευνα;". Τα βίντεο είναι τραβηγμένα σε υψηλή ποιότητα με ανάλυση 720×576, συμπιεσμένα με DV (lossy μέθοδος συμπίεσης), με 25 καρέ ανα δευτερόλεπτο ενώ η διάρκεια του κάθε βίντεο είναι περίπου 40 λεπτά. Σε αντίθεση με άλλα dataset που έχουν δημιουργηθεί υπο εργαστηριακές συνθήκες, σε αυτό του Canal 9 δεν είναι ορατοί όλοι οι συμμετέχοντες καθ' όλη τη διάρκεια του βίντεο.

Συνολικά στο σύνολο των δεδομένων υπάρχουν 190 μοναδικοί συμμετέχοντες εκ των οποίων οι 165 είναι άντρες και οι υπόλοιποι 25 γυναίκες. Το κάθε βίντεο περιέχει από 3 μέχρι 5 συμμετέχοντες (συμπεριλαμβανομένου και του παρουσιαστή) με αποτέλεσμα οι λήψεις ακόμα και της ίδιας κατηγορίας (π.χ. των μοναδικών ομιλητών μέσα στο καρέ) να μην είναι τραβηγμένες με τον ίδιο τρόπο. Έχουμε δηλαδή σε αρκετά καρέ περιπτώσεις στις οποίες οι ομιλητές είναι γυρισμένοι σε προφίλ αντί να μιλάνε κοιτώντας την κάμερα με συνέπεια ο αλγόριθμος ανίχνευσης πρόσωπων να συναντά αρκετές δυσκολίες.

Τα debate πραγματοποιήθηκαν στο ίδιο studio και συνεπώς το background σε όλο το dataset είναι σταθερό. Το 19.7% των δεδομένων αποτελείται από καρέ που παρουσιάζουν όλο το γκρουπ των ομιλητών, το 66.1% εμπεριέχει ένα μόνο ομιλητή ενώ το 11% περιλαμβάνει αρκετούς συμμετέχοντες αλλά όχι όλους. Το υπόλοιπο 3.2% αποτελείται είτε από μικρά τμήματα βίντεο που αναφέρουν το θέμα του debate (συνήθως στην αρχή του βίντεο) είτε από τα credits στην αρχή και στο τέλος του βίντεο. Χαρακτηριστικά παραδείγματα των πιο συνηθισμένων καρέ μέσα στα βίντεο παρουσιάζονται στο παρακάτω σχήμα:

(α′) Όλο το γκρουπ (β′) Ατομική λήψη (γ′) Πολλαπλοί συμμετέχοντες

Σχήμα 68: Οι πιο συνηθισμένες λήψεις της κάμερας στο σετ δεδομένων

Αναφορές

[1] Yael Adini, Yael Moses, and Shimon Ullman. Face recognition: The problem of compensating for changes in illumination direction. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):721–732, 1997. 102

[2] Xavier Anguera Miro, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals. Speaker diarization: A review of recent research. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(2):356–370, 2012. 20

[3] F. Yudi Limpraptono Aryuanto Soetedjo, Koichi Yamada. Lip detection based-on normalized rgb chromaticity diagram. *ITN Malang, Indonesia*, 2010. 86

[4] R Babuka, PJ Van der Veen, and U Kaymak. Improved covariance estimation for gustafson-kessel clustering. In *Fuzzy Systems, 2002. FUZZ-IEEE'02. Proceedings of the 2002 IEEE International Conference on*, volume 2, pages 1081–1085. IEEE, 2002. 72, 74

[5] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *Computer Vision–ECCV 2006*, pages 404–417. Springer, 2006. 59, 60, 61

[6] Peter N. Belhumeur, Joao P. Hespanha, and David J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):711–720, 1997. 102, 103

[7] James C Bezdek. *Pattern recognition with fuzzy objective function algorithms*. Kluwer Academic Publishers, 1981. 71

[8] Bhumika G Bhatt and Zankhana H Shah. Face feature extraction techniques: a survey. In *National conference on recent trends in engineering & technology*, pages 13–14, 2011. 48

[9] Christopher M Bishop et al. *Pattern recognition and machine learning*, volume 1. springer New York, 2006. 66, 73

[10] Patrick Bouthemy, Marc Gelgon, and Fabrice Ganansia. A unified approach to shot change detection and camera motion characterization. *Circuits and Systems for Video Technology, IEEE Transactions on*, 9(7):1030–1044, 1999.

[11] Gary R Bradski. Real time face and object tracking as a component of a perceptual user interface. In *Applications of Computer Vision, 1998. WACV'98. Proceedings., Fourth IEEE Workshop on*, pages 214–219. IEEE, 1998.

[12] Ingrid Daubechies. The wavelet transform, time-frequency localization and signal analysis. *Information Theory, IEEE Transactions on*, 36(5):961–1005, 1990. 49

[13] John G Daugman et al. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Optical Society of America, Journal, A: Optics and Image Science*, 2(7):1160–1169, 1985. 51

[14] Richard O Duda, Peter E Hart, et al. *Pattern classification and scene analysis*, volume 3. Wiley New York, 1973. 67, 70

[15] Donald E Farrar and Robert R Glauber. Multicollinearity in regression analysis: The problem revisited. *The Review of Economics and Statistics*, 49(1):92–107, 1967. 44

[16] Margaret M Fleck, David A Forsyth, and Chris Bregler. Finding naked people. In *Computer Vision - ECCV'96*, pages 593–602. Springer, 1996. 39

[17] Imola K Fodor. A survey of dimension reduction techniques, 2002. 63, 65, 68

[18] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997. 36

[19] Keinosuke Fukunaga. *Introduction to statistical pattern recognition*. Academic press, 1990. 68

[20] Dennis Gabor. Theory of communication. part 1: The analysis of information. *Electrical Engineers-Part III: Radio and Communication Engineering, Journal of the Institution of*, 93(26):429–441, 1946. 49

[21] Ullas Gargi, Rangachar Kasturi, and Susan H. Strayer. Performance characterization of video-shot-change detection methods. *Circuits and Systems for Video Technology, IEEE Transactions on*, 10(1):1–13, 2000. 24

[22] Daniel Gatica-Perez. Automatic nonverbal analysis of social interaction in small groups: A review. *Image and Vision Computing*, 27(12):1775–1787, 2009. 104

[23] Theodoros Giannakopoulos and Sergios Petridis. Fisher linear semi-discriminant analysis for speaker diarization. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(7):1913–1922, 2012. 20, 69, 70, 81, 90

[24] Theodoros D. Giannakopoulos. *Study and application of acoustic information for the detection of harmful content, and fusion with visual information*. PhD thesis, National and Kapodistrian University of Athens, 2009. 16

[25] Yihong Gong and Wei Xu. *Machine learning for multimedia content analysis*, volume 30. Springer, 2007.

[26] Arun Hampapur, Ramesh Jain, and Terry E Weymouth. Production model based digital video segmentation. *Multimedia Tools and Applications*, 1(1):9–46, 1995. 25

[27] Erik Hjelmås and Boon Kee Low. Face detection: A survey. *Computer vision and image understanding*, 83(3):236–274, 2001. 34

[28] J Edward Jackson. *A user's guide to principal components*, volume 244. Wiley-Interscience, 2005. 65

[29] Rangachar Kasturi and Ramesh Jain. Computer vision: principles. *IEEE Computer Society Press*, 1991. 29

[30] Mary Tai Knox and Gerald Friedland. Multimodal speaker diarization using oriented optical flow histograms. In *International Conference of the International Speech Communication Association (Interspeech)*, pages 290–293, 2010. 21

[31] Mee-Sook Lee, Yun-Mo Yang, and Seong-Whan Lee. Automatic video parsing using shot boundary detection and camera operation analysis. *Pattern Recognition*, 34(3):711–719, 2001. 24

[32] Rainer Lienhart, Alexander Kuranov, and Vadim Pisarevsky. Empirical analysis of detection cascades of boosted classifiers for rapid object detection. In *Pattern Recognition*, pages 297–304. Springer, 2003.

[33] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 54, 55, 56, 57

[34] Petros Maragos. *Image Analysis and Computer Vision*. N.T.U.A., 2005. 40, 50, 51

[35] A. Miller. *Subset Selection in Regression*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 2002. 46

[36] Baback Moghaddam and Alex Pentland. Probabilistic visual learning for object representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):696–710, 1997. 101

[37] Akio Nagasaka and Yuzuru Tanaka. Automatic video indexing and full-video search for object appearances. *North-Holland Publishing Co*, 1992. 25

[38] Kazuhiro Otsuka, Shoko Araki, Kentaro Ishizuka, Masakiyo Fujimoto, Martin Heinrich, and Junji Yamato. A realtime multimodal system for analyzing group meetings by combining face pose tracking and speaker diarization. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 257–264. ACM, 2008. 41

[39] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987. 74

[40] Paul R Schrater. Feature selection - extraction, 2009. Patern Recognition Lecture Slides. 45

[41] Linlin Shen and Li Bai. A review on gabor wavelets for face recognition. *Pattern analysis and applications*, 9(2-3):273–292, 2006. 52

[42] Lawrence Sirovich and Michael Kirby. Low-dimensional procedure for the characterization of human faces. *JOSA A*, 4(3):519–524, 1987. 98

[43] Lindsay I Smith. A tutorial on principal components analysis. *Cornell University, USA*, 51:52, 2002. 65

[44] Michael J Swain and Dana H Ballard. Color indexing. *International journal of computer vision*, 7(1):11–32, 1991. 25

[45] Richard Szeliski. *Computer vision: algorithms and applications*. Springer, 2010. 101

[46] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991. 98

[47] Matthew A Turk and Alex P Pentland. Face recognition using eigenfaces. In *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on*, pages 586–591. IEEE, 1991. 100

[48] Andrea Vattani. K-means requires exponentially many iterations even in the plane. *Discrete & Computational Geometry*, 45(4):596–616, 2011. 73

[49] Alessandro Vinciarelli, Alfred Dielmann, Sarah Favre, and Hugues Salamin. Canal9: A database of political debates for analysis of social interactions. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–4. IEEE, 2009. 104

[50] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511. IEEE, 2001. 34

[51] Paul Viola and Michael J Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004. 34

[52] HongJiang Zhang, Atreyi Kankanhalli, and Stephen W Smoliar. Automatic partitioning of full-motion video. *Multimedia systems*, 1(1):10–28, 1993. 26

[53] Wenyi Zhao, Rama Chellappa, P Jonathon Phillips, and Azriel Rosenfeld. Face recognition: A literature survey. *Acm Computing Surveys (CSUR)*, 35(4):399–458, 2003. 49

[54] Mu Zhu. *Feature Extraction and Dimension Reduction with Applications to Classification and the Analysis of Co-occurrence Data*. PhD thesis, Stanford University, June 2001. 43