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

Ανίχνευση ενεργειών σε βίντεο με χρήση σάκου λέξεων και συγχώνευσης χαρακτηριστικών

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΜΑΝΩΛΗΣ ΗΛΙΑΚΗΣ

Επιβλέπων : Ανδρέας – Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιανουάριος 2018

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

Ανίχνευση ενεργειών σε βίντεο με χρήση σάκου λέξεων και συγχώνευσης χαρακτηριστικών

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΜΑΝΩΛΗΣ ΗΛΙΑΚΗΣ

Επιβλέπων : Ανδρέας – Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 25η Ιανουαρίου 2018.

.....	.....	.....
Ανδρέας – Γεώργιος Σταφυλοπάτης	Γεώργιος Στάμου	Παναγιώτης Τσανάκας
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Ιανουάριος 2018

.....

ΜΑΝΩΛΗΣ ΗΛΙΑΚΗΣ

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

Περίληψη

Στη σημερινή εποχή που οι νέες τεχνολογίες της Τεχνητής Νοημοσύνης εισέρχονται με ραγδαίους ρυθμούς στην καθημερινότητα, ο κλάδος της Όρασης Υπολογιστών έχει γνωρίσει άνθιση, με την έρευνα να βελτιώνει συνεχώς τις μεθόδους που οι υπολογιστές αντιλαμβάνονται και αναλύουν τα οπτικά ερεθίσματα που δέχονται. Η ανίχνευση ενεργειών σε πολυμέσα απασχολεί ένα μεγάλο κομμάτι της έρευνας αυτής, και στοχεύει στην αναγνώριση από ένα σύστημα των ανθρωπίνων ενεργειών που εμπεριέχονται σε ένα αρχείο βίντεο, εικόνας κ.λπ. Με τον όρο ενέργεια, εννοούμε μια στοιχειώδη ανθρωποκεντρική αλληλεπίδραση με νόημα και μπορεί να αφορά από απλούστερες ενέργειες, όπως «Περπατάω», μέχρι πιο σύνθετες, όπως «Παίζω Ποδόσφαιρο».

Στην εργασία μας υλοποιούμε ένα σύστημα ανίχνευσης ενεργειών, το οποίο εξάγει χαρακτηριστικά εικόνας, ήχου και κίνησης για την αναπαράσταση των βίντεο και τα κωδικοποιεί χρησιμοποιώντας τη διαδεδομένη τεχνική σάκου λέξεων (Bag of Words), που δημιουργεί ένα λεξικό από κομμάτια των δεδομένων εκπαίδευσης και εκφράζει το σύνολο των δεδομένων με βάση αυτά, δημιουργώντας μια εύρωστη αναπαράσταση με ένα διάνυσμα για κάθε βίντεο. Η τεχνική αυτή ευνοεί την εκπαίδευση ενός ταξινομητή, που στην περίπτωση μας είναι μια Μηχανή Διανυσμάτων Υποστήριξης (SVM) ο οποίος καλείται να κατηγοριοποιήσει τα βίντεο με βάση την κατηγορία ενέργειας που περιέχουν.

Στη συνέχεια, πειραματιστήκαμε με διάφορες μεθόδους συγχώνευσης των εξαγμένων χαρακτηριστικών από τα δεδομένα μας, ώστε να επιτύχουμε πιο αντιπροσωπευτικές αναπαραστάσεις και να βελτιώσουμε την συνολική απόδοση του συστήματός μας. Συγκεκριμένα, υλοποιήσαμε μεθόδους πρώιμης συγχώνευσης, καθώς και μεθόδους όψιμης συγχώνευσης, με ή χωρίς επιπλέον εκπαίδευση. Ακόμη, μελετήσαμε και τις δυνατότητες συνδυασμού των δύο παραπάνω κατηγοριών μεθόδων συγχώνευσης.

Τα αποτελέσματα που εξάγαμε, αναδεικνύουν τη σημασία της σωστής προεπεξεργασίας των δεδομένων μας πριν την εκπαίδευση των ταξινομητών ώστε να επιτύχουμε ένα αποδεκτό επίπεδο γενίκευσης. Ακόμη, συμπεραίνουμε ότι η συγχώνευση διαφορετικών χαρακτηριστικών, συμπληρωματικών μεταξύ τους, ακόμα και με απλές στην υλοποίησή τους μεθόδους, μπορεί να επιφέρει σημαντική βελτίωση στη συνολική απόδοση ενός τέτοιου συστήματος και μάλιστα τα πειραματικά αποτελέσματα ενθαρρύνουν περαιτέρω έρευνα σε αυτή την κατεύθυνση.

<u>Λέξεις Κλειδιά</u>: αναγνώριση ανθρωπίνων ενεργειών, βίντεο, SIFT, MFCC, STIP, Μηχανές Διανυσμάτων Υποστήριξης, Σάκος λέξεων, Ανάλυση Κύριων Συνιστωσών, K-means, πρώιμη συγχώνευση, όψιμη συγχώνευση, UCF101

Abstract

Nowadays, Artificial Intelligence enters our everyday lives in a rapid pace and the field of Computer Vision has experienced great growth, while research constantly improves the way that computers understand and analyze the visual clues which they receive. Multimedia Action Recognition has received attention of the research community. Its aim is to develop a system that detects human actions that appear in a video, picture etc. The term “action” means a basic person-related interaction with meaning and it might include the simplest actions, like “Walking”, or maybe more complex, like “Playing Soccer”.

In this thesis, we develop an action recognition system, which extracts visual, sound and motion features for video representation and uses the well-known Bag of Words framework to represent these features using a codebook consisting of fragments of train data. This codebook is used to encode the train data, creating a robust representation with a single vector for each video. This technique benefits the training process of a classifier, which in our case is a Support Vector Machine. The classifier predicts the action classes in which each video belongs.

Moving on, we have experimented different feature fusion methods in order to achieve a more representative representation and finally to improve the average accuracy of our system. Specifically, we have implemented early fusion methods as well as late fusion methods, with or without a meta-classifier. Furthermore, we checked the combination of different fusion categories.

Our results highlight the significance of a proper preprocessing phase of our data before training the classifiers in order to achieve an acceptable level of generalization. Moreover, we conclude that even the simplest implementation of fusion of complementary features can result an important improvement in the average accuracy of our system. Our experimental results encourage further research towards this direction.

Key Words: human action recognition, video, SIFT, MFCC, STIP, Support Vector Machines, Bag of Words, Principal Components Analysis, K-means, early fusion, late fusion, UCF101

Ευχαριστίες

Η διπλωματική εργασία αυτή σηματοδοτεί και την ολοκλήρωση των σπουδών μου στο Εθνικό Μετσόβιο Πολυτεχνείο και θα ήθελα σε αυτό το σημείο να ευχαριστήσω τους ανθρώπους που με βοήθησαν με τη στήριξή τους στην εκπόνηση της και μου έδωσαν τα ερεθίσματα και τη δύναμη να συνεχίσω τόσο σε επιστημονικό, όσο και σε ψυχολογικό επίπεδο.

Αρχικά, οφείλω να ευχαριστήσω τον κ. Ανδρέα – Γεώργιο Σταφυλοπάτη, καθηγητή Ε.Μ.Π., για την ευκαιρία που μου έδωσε αναθέτοντας μου την εκπόνηση αυτής της διπλωματικής εργασίας. Ευγνωμοσύνη οφείλω και στον κ. Γεώργιο Σιόλα, μέλος Ε.Δ.Ι.Π. του Ε.Μ.Π. για την καθημερινή του βοήθεια και στήριξη σε όλη την πορεία της εργασίας και για τις λιτές και πάντα χρήσιμες συμβουλές του στις δυσκολίες που αντιμετώπισα. Επίσης, θα ήθελα να ευχαριστήσω και όλους τους διδάκτορες και υποψήφιους διδάκτορες του Εργαστηρίου Ευφυών Υπολογιστικών Συστημάτων για τη θετική τους διάθεση και το καλό κλίμα που επικρατεί στο Εργαστήριο και ευνοεί τη μελέτη και την έρευνα.

Τέλος, θα ήθελα να ευχαριστήσω και τους δικούς μου ανθρώπους που μου πρόσφεραν δύναμη όλα τα χρόνια των σπουδών μου, ξεκινώντας από την οικογένειά μου που μου εξασφάλισε τις καλύτερες δυνατές συνθήκες για τη φοιτητική μου ζωή. Επίσης θα ήθελα να ευχαριστήσω τους φίλους μου, Ηλεκτρολόγους και μη, για τη διαρκή στήριξη τους σε μένα σε εύκολες και δύσκολες στιγμές που περάσαμε. Ιδιαίτερα θα ήθελα να ευχαριστήσω την Ελίζα για την ώθηση της, με το μοναδικό της τρόπο, να επιτύχω τους στόχους μου και την Αλεξάνδρα για την ηρεμία που μου προσέφερε στο κρισιμότερο σημείο των σπουδών μου και την πίστη της σε μένα, όποτε συναντούσα εμπόδια.

Μανώλης Ηλιάκης,

Αθήνα, 25η Ιανουαρίου 2018

Περιεχόμενα

Κατάλογος Εικόνων

Κατάλογος Πινάκων

Κεφάλαιο 1 Εισαγωγή

1.1 Μηχανική Μάθηση

Η Μηχανική Μάθηση (Machine Learning) αποτελεί πεδίο της επιστήμης των υπολογιστών που αναπτύχθηκε από τη μελέτη της αναγνώρισης προτύπων και της υπολογιστικής θεωρίας μάθησης στην Τεχνητή Νοημοσύνη. Ορίζεται ως το πεδίο της επιστήμης των υπολογιστών το οποίο δίνει τη δυνατότητα στους υπολογιστές να «μαθαίνουν» και να κάνουν προβλέψεις βάσει δεδομένων χωρίς να προγραμματίζεται ρητά η λειτουργία τους.

Συνήθως, οι εργασίες μηχανικής μάθησης ταξινομούνται σε τρεις κατηγορίες: Επιβλεπόμενη Μάθηση(Supervised Learning), Μη Επιβλεπόμενη Μάθηση(Unsupervised Learning) και Ενισχυτική Μάθηση(Reinforcement Learning). Επίσης, ανάλογα με το επιθυμητό αποτέλεσμα, τα προβλήματα που επιλύει η μηχανική μάθηση χωρίζονται σε: προβλήματα Ταξινόμησης, Παλινδρόμησης, Συσταδοποίησης, Εκτίμησης Πιθανότητας και Μείωσης Διαστατικότητας.

Στην παρούσα εργασία, αντιμετωπίζουμε ένα πρόβλημα Ταξινόμησης χρησιμοποιώντας Επιβλεπόμενη Μάθηση. Όπως θα δούμε παρακάτω, όμως, κατά τη διαδικασία επίλυσης θα χρειαστεί να χρησιμοποιηθούν κι άλλα είδη μάθησης για να αντιμετωπίσουν διαφορετικά κομμάτια του προβλήματος.

1.1.1 Όραση Υπολογιστών

Η Όραση Υπολογιστών αποτελεί πεδίο της Τεχνητής Νοημοσύνης και είναι ένα από τα χαρακτηριστικά παραδείγματα όπου αξιοποιούνται τεχνολογίες Μηχανικής Μάθησης. Αφορά την τεχνολογία και τη θεωρία για τη σχεδίαση και κατασκευή συστημάτων που λαμβάνουν και αναλύουν δεδομένα, αποκτώντας ουσιαστικά «υψηλού επιπέδου» κατανόηση από τις εικόνες ή τα βίντεο που επεξεργάζονται. Αυτή η κατανόηση εικόνας μπορεί να περιγραφεί ως η εξαγωγή συμβολικών πληροφοριών από δεδομένα εικόνας χρησιμοποιώντας μοντέλα που αξιοποιούν έννοιες από τους κλάδους της γεωμετρίας, της φυσικής, των στατιστικών και της θεωρίας γνώσης.

Επί μέρους πεδία της Όρασης Υπολογιστών αποτελούν η Κατανόηση Σκηνής, η Ανίχνευση Ενεργειών, η Ανίχνευση Κίνησης, η Αναγνώριση Αντικειμένων, η Ευρετηριοποίηση, η Αναγνώριση Κίνησης, η Τρισδιάστατη Ανακατασκευή, κ.ά. Στην παρούσα εργασία προσπαθούμε να επιλύσουμε ένα πρόβλημα Ανίχνευσης Ενεργειών.

1.2 Ανίχνευση Ενεργειών σε Πολυμέσα

Στην εποχή που ζούμε με τη διαρκή εισροή νέων τεχνολογιών στην καθημερινή ζωή και την παντοδυναμία του διαδικτύου που πλέον φιλοξενεί σχεδόν όλες τις πτυχές της κοινωνικής ζωής, η δυνατότητα των υπολογιστών να μπορούν να αναγνωρίσουν έννοιες από οπτικά ερεθίσματα μπορεί να επηρεάσει άμεσα πολλές από τις δραστηριότητές μας. Για παράδειγμα σε μια πραγματικότητα με ένα χάος πληροφοριών από βίντεο χρηστών του διαδικτύου, η δυνατότητα αυτοματοποιημένης αναγνώρισης του περιεχομένου τους μπορεί να διευκολύνει την οργάνωση και αναζήτηση τους. Σε πιο σύνθετες εφαρμογές, θα μπορούσε η δυνατότητα αυτή να αξιοποιηθεί για τη δημιουργία αυτόματων περιγραφών των περιεχομένων των βίντεο για άτομα με προβλήματα όρασης. Επίσης, με την ανάπτυξη της τεχνητής νοημοσύνης σε όλα τα επίπεδα και την πορεία προς μια «γενική τεχνητή νοημοσύνη», η δυνατότητα των υπολογιστών να μπορούν να αναγνωρίζουν έννοιες και δραστηριότητες από τα ερεθίσματα που παίρνουν σε πραγματικό χρόνο με τη χρήση κάμερας θα μπορούσε να αξιοποιηθεί σε εφαρμογές όπως αυτοκινούμενα αυτοκίνητα, ρομποτ – οικιακός βοηθός κ.ά., όπου η άμεση αλληλεπίδραση τους με της ανθρώπινες δραστηριότητες θα αποτελεί βασικό παράγοντα για τη λειτουργικότητά τους.

Η ανίχνευση ενεργειών (actions) σε βίντεο, αποτελεί ένα από τα σημαντικότερα προβλήματα της Όρασης Υπολογιστών και της Τεχνητής Νοημοσύνης στις μέρες μας και παρά τα πολλά και ελπιδοφόρα αποτελέσματα της σύγχρονης έρευνας παραμένει ένα ανοιχτό πρόβλημα με μεγάλα περιθώρια βελτίωσης των υπαρχουσών τεχνικών. Στόχος του προβλήματος αυτού, είναι η αυτόματη κατηγοριοποίηση ενός βίντεο εισόδου ανάλογα με μια ενέργεια που εμφανίζεται στο βίντεο. Τέτοιες ενέργειες μπορεί να είναι από πιο απλές (όπως Περπάτημα, Πήδημα κ.λπ.) μέχρι και πιο σύνθετες (όπως Παίζω πιάνο, Ποδόσφαιρο, Πληκτρολογώ κ.λπ.).

Η δυσκολία αυτού του προβλήματος οφείλεται σε διάφορους παράγοντες [1]. Αρχικά, υπάρχουν μεγάλες παραλλαγές στο εσωτερικό της ίδιας κατηγορίας (κλάσης) που προκύπτουν από τις διαφορετικές ταχύτητες της κίνησης, τις αλλαγές οπτικής γωνίας ή τη σύγχυση από το φόντο. Επίσης, η αναγνώριση μιας ενέργειας σχετίζεται με πολλά υψηλού επιπέδου οπτικά στοιχεία, όπως ανθρώπινη στάση, αλληλεπιδρώντα αντικείμενα ή το σκηνικό. Τέτοια υποπροβλήματα είναι δύσκολα στην λύση τους ακόμα και μόνα τους.

Ένας ακόμα παράγοντας της δυσκολίας του προβλήματος έγκειται στην υποκειμενικότητα του καθορισμού της χρονικής διάρκειας μιας ενέργειας, αφού δεν υπάρχει σαφής ορισμός του πότε αρχίζει και πότε τελειώνει η ενέργεια, κάτι το οποίο δεν παρουσιάζεται για παράδειγμα στην ανίχνευση αντικειμένων σε βίντεο, όπου εκεί ο στόχος είναι σαφής. Τέλος, η υψηλή διαστατικότητα και η χαμηλή ποιότητα των δεδομένων βίντεο συνήθως προσθέτει ακόμα μεγαλύτερη δυσκολία στην ανάπτυξη καλών και αποδοτικών αλγορίθμων αναγνώρισης.

Οι αρχικές προσεγγίσεις του προβλήματος ερμήνευαν μια ενέργεια ως σύνολο από 2Δ ή 3Δ χωρο-χρονικές τροχιές ανθρώπινων αρθρώσεων [2], [3], [4], [5]. Τέτοιες μέθοδοι, όμως, χρειάζονταν την αναγνώριση ανθρώπινων μελών του σώματος και τον εντοπισμό τους σε κάθε frame του βίντεο, πράγμα δύσκολο ακόμα και στις μέρες μας για βίντεο του «πραγματικού» κόσμου. Έτσι, πιο σύγχρονες μέθοδοι αναγνώρισης χρησιμοποιούν τοπικά χωρο-χρονικά χαρακτηριστικά [6], [7] και έχουν καταφέρει πολύ καλές αποδόσεις σε δύσκολα datasets ανίχνευσης ενεργειών. Αυτές οι μέθοδοι αντιμετωπίζουν τον «όγκο» που ενεργεί στο βίντεο ως ένα άκαμπτο 3Δ αντικείμενο και εξάγουν κατάλληλα χαρακτηριστικά για να περιγράψουν τα μοτίβα κάθε 3Δ όγκου. Αυτή η προσέγγιση είναι αποδοτική και ξεπερνάει προβλήματα όπως τη σύγχυση από το φόντο, τις αλλαγές στο φωτισμό και το θόρυβο.

Τα τελευταία χρόνια, η έρευνα έχει κατευθυνθεί σε μεγάλο βαθμό σε μεθόδους που χρησιμοποιούν βαθιά μάθηση όπως και συνελικτικά νευρωνικά δίκτυα [8], [9]. Αυτές οι προσπάθειες έχουν επιτύχει τα καλύτερα αποτελέσματα μέχρι τώρα σε απαιτητικά datasets για αναγνώριση ενεργειών. Παρόλα αυτά, τα αποτελέσματα αυτά είναι συγκρίσιμα [1] και με αυτά που έχουν επιτύχει τεχνικές σάκου λέξεων (Bag of Words) με χρήση χειροποίητων χαρακτηριστικών [10]. Τέτοιες τεχνικές, αποτελούσαν το μεγαλύτερο κομμάτι της έρευνας στην αναγνώριση ενεργειών μέχρι πριν 4-5 χρόνια, όπου και έγινε στροφή στην κατεύθυνση των συνελικτικών δικτύων.

Στο επόμενο κεφάλαιο αναφερόμαστε εκτενέστερα στη σχετική δουλειά που έχει γίνει πάνω στο ζήτημα τα τελευταία χρόνια, καθώς και στην πορεία εξέλιξης των διαφορετικών μεθόδων.

1.3 Συνεισφορά της εργασίας

Στην παρούσα εργασία θέτουμε δύο κύριους στόχους στην ενασχόλησή μας με το πρόβλημα της ανίχνευσης ενεργειών σε βίντεο:

- Μια δομημένη μελέτη της διαδικασίας σάκου λέξεων (Bag of Words) που είναι ευρέως διαδιδομένη σε εφαρμογές με δεδομένα εικόνας ή βίντεο ως δεδομένα εκπαίδευσης των μοντέλων
- Πειραματισμό και αξιολόγηση γύρω από διάφορες μεθόδους συγχώνευσης διαφορετικών χαρακτηριστικών για την αναπαράσταση των δεδομένων μας και διαφορετικών συνδυασμών ταξινομητών, ώστε να αποφανθούμε για την αποδοτικότερη αρχιτεκτονική ενός συστήματος ανίχνευσης ενεργειών με χρήση πολλαπλών τρόπων αναπαράστασης.

1.4 Οργάνωση της εργασίας

Στην παρούσα εργασία αναπτύξαμε ένα σύστημα με τη μεθοδολογία Bag of Words και χρήση χειροποίητων χαρακτηριστικών που εξάγονται από τα βίντεο για την εκπαίδευση του ταξινομητή. Για την ταξινόμηση στις επιμέρους κατηγορίες ενεργειών χρησιμοποιούμε Μηχανές Διανυσμάτων Υποστήριξης (SVMs). Εξετάσαμε διαφορετικές επιλογές που μπορούν να γίνουν σε διάφορα στάδια κατά την ανάπτυξη ενός τέτοιου συστήματος και παρουσιάζουμε πειραματικά αποτελέσματα που αναδεικνύουν την επιρροή τους στην απόδοση του συνολικού συστήματος.

Στη συνέχεια, το κείμενο οργανώνεται ως εξής:

Στο 2^{ο} κεφάλαιο παρουσιάζουμε σχετικές εργασίες που έχουν γίνει τα τελευταία χρόνια στον τομέα της αναγνώρισης ενεργειών και αναλύουμε θετικά και αρνητικά χαρακτηριστικά διαφορετικών προσεγγίσεων.

Στο 3^{ο} κεφάλαιο κάνουμε μια εκτενή παρουσίαση της διαδικασίας εξαγωγής χειροποίητων χαρακτηριστικών από τα βίντεο, καθώς και των επιμέρους σταδίων της μεθοδολογίας Bag of Words που ουσιαστικά αποτελεί την προεπεξεργασία των δεδομένων μας πριν την ταξινόμηση.

Στο 4^{ο} κεφάλαιο αναλύουμε εις βάθος τις διάφορες τεχνικές μηχανικής μάθησης που χρησιμοποιούμε στην παρούσα εργασία καθώς και τις διαφορετικές μεθόδους συγχώνευσης των διαφορετικών χαρακτηριστικών που χρησιμοποιούμε.

Στο 5^{ο} κεφάλαιο παρουσιάζουμε όλη την πειραματική διαδικασία που εκτελέσαμε, την ανάπτυξη του συστήματός μας καθώς και αποτελέσματα τόσο για το τελικό σύστημα, όσο και για σύγκριση επί μέρους επιλογών που κάναμε κατά την υλοποίηση.

Τέλος, στο 6^{ο} κεφάλαιο συνοψίζουμε την εργασία εξάγοντας κάποια χρήσιμα συμπεράσματα και προτείνουμε κατευθύνσεις για μελλοντική έρευνα.

Κεφάλαιο 2 Σχετικές Εργασίες

Αρχικά, αξίζει να γίνει μια προσπάθεια ορισμού του τι σημαίνει η έννοια της «ενέργειας», μιας και περιέχει μια ασάφεια. Ένας χρήσιμος ορισμός που προτείνεται στο [11] αναφέρει ότι «Ενέργεια είναι η πιο στοιχειώδης ανθρωποκεντρική αλληλεπίδραση με νόημα». Προφανώς, στο παρόν πρόβλημα, αναφερόμαστε μόνο σε ενέργειες σχετικές τον άνθρωπο, ενώ με τη λέξη «νόημα» υπονοείται η κατηγορία που ταξινομούμε την ανάλογη ενέργεια. Οι προσπάθειες που έχουν γίνει για την αναγνώριση και κατηγοριοποίηση τέτοιων ενεργειών μπορούν να χωριστούν σε δύο υποκατηγορίες, αυτές που χρησιμοποιούν χειροποίητα χαρακτηριστικά και τοπικούς περιγραφείς για να αναπαραστήσουν τα βίντεο και αυτές που χρησιμοποιούν αρχιτεκτονικές βαθιάς μάθησης και συνελικτικά νευρωνικά δίκτυα. Παρακάτω παραθέτουμε χαρακτηριστικά παραδείγματα των κατηγοριών αυτών καθώς και την πορεία εξέλιξής τους.

Στις αναπαραστάσεις με χειροποίητα χαρακτηριστικά ακολουθείται συνήθως μια μεθοδολογία, όπου αρχικά εξάγονται τοπικά στατιστικά από χωρικές προεξοχές (π.χ. γωνίες) και κινήσεις, έπειτα συνδυάζονται σε μια αναπαράσταση σε επίπεδο βίντεο, δημιουργώντας ένα διάνυσμα για κάθε βίντεο και τέλος αυτά τροφοδοτούν ταξινομητές (συνήθως SVM [12]). Τέτοια χαμηλού επιπέδου χαρακτηριστικά εντοπίζονται σε επίπεδο pixel στο βίντεο και έχουν σχεδιαστεί ώστε να αντιμετωπίζουν προκλήσεις όπως αλλαγές κλίμακας, περιστροφής, φωτεινότητας και θορύβου. Παράδειγμα τέτοιων χαρακτηριστικών αποτελούν οι ανεξάρτητοι κλίμακας μετασχηματισμοί (SIFT [13]) που έχουν σχεδιαστεί να ταιριάζουν σε διαφορετικές εικόνες ή αντικείμενα και χρησιμοποιούνταν ευρέως για αναγνώριση αντικειμένων σε εικόνες, ενώ έχουν αναπτυχθεί διάφορες επεκτάσεις των SIFT, όπως χρωματικοί-SIFT (color SIFT [14]) ή 3Δ-SIFT [15]. Σε αντίθεση με τους SIFT που αφορούν κυρίως την αναγνώριση εικόνων, τα χωρό-χρονικά σημεία ενδιαφέροντος (STIP [6]) εντοπίζουν σημεία που έχουν μεταβολές σε χώρο και χρόνο, χρησιμοποιώντας 3Δ-Harris σημεία ενδιαφέροντος για τον εντοπισμό των μεταβολών και στη συνέχεια Ιστογράμματα Προσανατολισμένων Κλίσεων (Histograms of Oriented Gradients-HOG[16]) και Ιστογράμματα Οπτικής Ροής (Histograms of Optical Flow-HOF[17]) για την περιγραφή των σημείων αυτών σε διανύσματα. Ένα πιο πρόσφατο παράδειγμα χαρακτηριστικών αποτελούν οι Πυκνές Τροχιές (Dense Trajectories [18]) και οι βελτιωμένες Πυκνές Τροχιές (improved Dense Trajectories [19]) τα οποία αντί για σημεία, εντοπίζουν τροχιές των σημείων που αναδεικνύουν καλύτερα τις αλλαγές στο χρόνο, ενώ για περιγραφείς εκτός από τα HOG και HOF που προαναφέραμε, χρησιμοποιούν και Ιστογράμματα Οριακής Κίνησης (Motion Boundary Histograms-MBH [20]). Τα STIP και οι iDTs αποτελούν ίσως τα σημαντικότερα είδη χειροποίητων χαρακτηριστικών στις μέρες μας και αντιπροσωπεύουν δύο διαφορετικές λογικές εξαγωγής χαρακτηριστικών, καθώς τα πρώτα εντοπίζουν αραιά σημεία ενδιαφέροντος, ενώ τα δεύτερα πυκνά, με τα δεύτερα να παρουσιάζουν τα καλύτερα αποτελέσματα, αλλά και να έχουν μεγαλύτερο υπολογιστικό κόστος.

Τα χειροποίητα χαρακτηριστικά αξιοποιούνται κυρίως από συστήματα μεθοδολογίας Bag of Words, στις οποίες δημιουργείται ένα «λεξικό» από σημεία ενδιαφέροντος των βίντεο με σκοπό την κωδικοποίηση των χαρακτηριστικών που έχουν εξαχθεί και τελικά την αναπαράσταση των βίντεο με ένα μοναδικό για κάθε βίντεο διάνυσμα κοινού μήκους για όλα, τα οποία θα αποτελούν και τα δεδομένα εκπαίδευσης για τον ταξινομητή. Η μεθοδολογία Bag of Words μπορεί να περιγραφεί, σύμφωνα με το [1], ως μια ακολουθία 5 βημάτων: i) Εξαγωγή Χαρακτηριστικών, ii) Προεπεξεργασία Χαρακτηριστικών, iii) Δημιουργία «λεξικού», iv) Κωδικοποίηση Χαρακτηριστικών, v) Συγκέντρωση και Κανονικοποίηση. Με την πάροδο των χρόνων έχουν υπάρξει νέες ιδέες, διαφορετικές προσεγγίσεις και επιλογές για κάθε ένα από αυτά τα επί μέρους βήματα, πολλές από τις οποίες έφεραν και σημαντικές βελτιώσεις στην απόδοση τέτοιων μεθόδων. Ενδεικτικά να αναφέρουμε τη χρήση Ανάλυσης Κύριων Συνιστωσών (Principal Components Analysis-PCA [21]) για μείωση των διαστάσεων των δεδομένων πριν τη δημιουργία του «λεξικού», τη χρήση του αλγορίθμου Κ-μέσων (k-means [21]) ή Μοντέλου Μείγματος Γκαουσιανών Κατανομών (Gaussian Mixture Model-GMM [21]) για τη δημιουργία του λεξικού, τη χρήση διανυσμάτων Fisher [22] για την κωδικοποίηση των χαρακτηριστικών κ.ά..

Σημαντικό μέρος της έρευνας σε τέτοιες μεθοδολογίες έχει κατευθυνθεί και στη χρήση διαφορετικών χαρακτηριστικών στο ίδιο σύστημα για την αναπαράσταση των βίντεο με διαφορετικούς, συμπληρωματικούς μεταξύ τους τρόπους, ώστε να βελτιωθεί η απόδοση του συνολικού συστήματος. Σε τέτοιες περιπτώσεις, μπορούν να συνδυαστούν όχι μόνο οπτικοί περιγραφείς (όπως τα SIFT, STIP, iDTs), αλλά και χαρακτηριστικά για τον ήχο του βίντεο (π.χ. Mel Frequency Cepstral Coefficients-MFCC [23], Automatic Speech Recognition-ASR [24]) ή και χαρακτηριστικά κειμένου στο βίντεο (π.χ. Histograms of Textual Concepts-HTC [25]). Τέτοια χαρακτηριστικά χρησιμοποιούνται συχνότερα σε συστήματα αναγνώρισης συμβάντων, που αποτελούν πιο υψηλού επιπέδου έννοιες και πιο ασαφείς και απαιτούν συνδυασμό περισσότερων στοιχείων και βαθύτερη κατανόηση (παραδείγματα τέτοιων συμβάντων μπορεί να είναι Φτιάχνω ένα σάντουιτς ή Πάρτι γενεθλίων). Ο τρόπος που συνδυάζονται αυτά τα διαφορετικά χαρακτηριστικά παίζει κομβικό ρόλο στην τελική απόδοση που θα επιτευχθεί. Σε μια απλή κατηγοριοποίηση των μεθόδων [1] η συγχώνευση των χαρακτηριστικών μπορεί να γίνει: α) σε επίπεδο περιγραφέων, β) σε επίπεδο αναπαράστασης και γ) σε επίπεδο σκορ (ποσοστού επιτυχίας ταξινομητή). Στην πρώτη περίπτωση οι διαφορετικοί περιγραφείς χαρακτηριστικών συγχωνεύονται στο ίδιο διάνυσμα και έπειτα δημιουργείται το λεξικό για την κωδικοποίηση τους, στην επόμενη κωδικοποιούνται με ξεχωριστά λεξικά και στη συνέχεια συγχωνεύονται τα κωδικοποιημένα διανύσματα και τροφοδοτούν τον ταξινομητή, ενώ στην τελευταία περίπτωση εκπαιδεύεται ένας ξεχωριστός ταξινομητής για κάθε είδος χαρακτηριστικών και στη συνέχεια τα αποτελέσματα τους συγχωνεύονται με κάποιο σχήμα ώστε να αποφασιστεί η τελική ταξινόμηση του βίντεο. Στα [7], [26], [27], [28] φαίνονται κάποια ενδεικτικά παραδείγματα τέτοιων μεθόδων.

Παρά τα πολύ αξιόλογα αποτελέσματα των παραπάνω μεθόδων, τα τελευταία χρόνια το μεγαλύτερο ποσοστό της έρευνας έχει στραφεί σε τεχνολογίες βαθιάς μάθησης (deep

learning), που αποτελούν στατικά μοντέλα νευρωνικών δικτύων πολλών επιπέδων. Ένα τέτοιο παράδειγμα είναι και τα Συνελικτικά Νευρωνικά Δίκτυα (Convolutional Neural Networks-CNN [29]) που αρχικά σχεδιάστηκαν για την αναγνώριση χειρόγραφων χαρακτήρων. Μια εργασία που έδειξε τη δύναμη τέτοιων μοντέλων ήταν η [30] που πέτυχε κορυφαία για την εποχή απόδοση στην κατηγοριοποίηση εικόνων στο απαιτητικό dataset Imagenet. Τα CNN ήταν πολύ αποδοτικά στην αναγνώριση εικόνων, αλλά δεν μπορούσαν να συμπεριλάβουν την παράμετρο του χρόνου περιορίζοντας τα στην επεξεργασία βίντεο. Για να ξεπεραστεί αυτό το εμπόδιο αναπτύχθηκαν 3Δ Συνελικτικά Νευρωνικά Δίκτυα [31] που χρησιμοποιούν 3Δ πυρήνες για να εξάγουν χαρακτηριστικά σε χώρο και χρόνο, τα οποία σε άλλες εργασίες [32], [33] εξελίχθηκαν ως προς το πώς διαχειρίζονται και εισάγουν τις χρονικές πληροφορίες στο συνελικτικό δίκτυο. Αξιοσημείωτες είναι και οι εργασίες [34], [35] που αξιοποίησαν ένα είδος Αναδρομικών Νευρωνικών Δικτύων (Recurrent Neural Networks-RNN [36]) τα Νευρωνικά Μακρο-βραχυπρόθεσμης Μνήμης (Long-Short Term Memory-LSTM [37]) για να εκμεταλλευτούν τη χρονική πληροφορία. Μια διαφορετική προσέγγιση στην εισαγωγή του χρόνου στην ανάλυση ήταν αυτή του [38] που ανέπτυξε ένα Συνελικτικό Νευρωνικό Δίκτυο δύο ροών που δέχονται σαν είσοδο καρέ του βίντεο στη μια ροή και χαρακτηριστικά οπτικής ροής στην άλλη. Άλλη μια ενδιαφέρουσα προσέγγιση είναι η [39] που εξάγει χωροχρονικά χαρακτηριστικά (τα ονομάζει C3D) από ένα 3D-CNN και τα εισάγει σε ένα SVM για ταξινόμηση.

Αξίζει τέλος να σημειωθεί ότι στο [40] που αφορά μια βελτιωμένη εκδοχή Συνελικτικού Δικτύου 2 ροών παρατηρείται από τα πειράματα ότι μια απλή συγχώνευση σε επίπεδο σκορ ενός τέτοιου δικτύου με ένα σύστημα χειροποίητων χαρακτηριστικών (iDTs) πετυχαίνει βελτιωμένα αποτελέσματα σε σχέση με αυτά του Συνελικτικού Δικτύου από μόνο του και μάλιστα, ίσως τα μεγαλύτερα που έχουν επιτευχθεί ως τώρα σε απαιτητικά datasets. Αυτό μας δείχνει τη συμπληρωματικότητα που παρατηρείται σε αυτές τις δύο διαφορετικές προσεγγίσεις, ενώ μπορούμε να συμπεράνουμε ότι υπάρχει περιθώριο βελτίωσης από τη συνέχιση της έρευνας και προς τις δύο κατευθύνσεις.

Κεφάλαιο 3 Εξαγωγή χαρακτηριστικών

Στον παρόν κεφάλαιο περιγράφουμε κάποια διαδεδομένα είδη χαρακτηριστικών για την αναπαράσταση των βίντεο σε διανύσματα και παρουσιάζουμε τη λειτουργία τους. Στη συνέχεια παρουσιάζουμε συνοπτικά τη μέθοδο του «σάκου λέξεων» που ακολουθεί την εξαγωγή των χαρακτηριστικών και κατασκευάζει τις τελικές αναπαραστάσεις των βίντεο που αποτελούν και τα δεδομένα εκπαίδευσης που τροφοδοτούν τον ταξινομητή μας.

3.1 Είδη χαρακτηριστικών

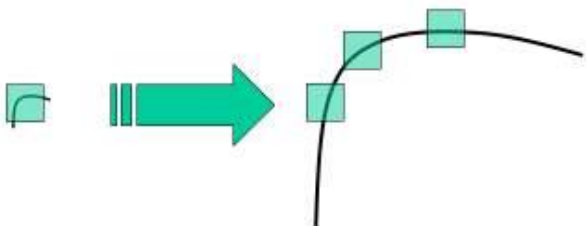
Τα χαρακτηριστικά που χρησιμοποιήσαμε είναι οι Ανεξάρτητοι Κλίμακας Μετασχηματισμοί (SIFT) για δεδομένα εικόνας, οι Συντελεστές Cepstral Mel Συχνοτήτων (MFCC) για δεδομένα ήχου και τα Χώρο-χρονικά Σημεία Ενδιαφέροντος (STIP) για δεδομένα κίνησης.

3.1.1 Scale-Invariant Feature Transform (SIFT)

Ο αλγόριθμος SIFT δημιουργήθηκε το 2004 από τον D.Lowe [13] εξελίσσοντας μια προηγούμενη δουλειά του ίδιου ήδη από το 1999 [41]. Ένα σημαντικό χαρακτηριστικό των ανιχνευτών γωνιών, που ήδη χρησιμοποιούνταν για την αναπαράσταση εικόνας, είναι ότι είναι ανεξάρτητοι περιστροφής, αφού όπως και να περιστραφεί μια εικόνα, η γωνία ενός σχήματος παραμένει ίδια. Το ίδιο όμως δεν ισχύει για αλλαγές στην κλίμακα, όπου εκεί οι γωνίες θα αλλάξουν σχήμα ανάλογα το ζουμ, όπως φαίνεται στην Εικ.2. Ο αλγόριθμος SIFT επιλύει αυτό ακριβώς το πρόβλημα.

1 Παράδειγμα εντοπισμένων σημείων με SIFT

Αρχικά, από την Εικ.2 φαίνεται ότι το ίδιο παράθυρο δεν αρκεί για γωνίες διαφορετικών μεγεθών. Έτσι, ο αλγόριθμος, πέρα από το να βρει σημεία στο χώρο που μένουν αμετάβλητα στις αλλαγές οπτικής, βρίσκει και την κλίμακα στην οποία αυτά τα σημεία παρουσιάζονται κατάλληλα. Για το λόγο αυτό χρησιμοποιείται το Λαπλασιανό του Γκαουσιανού (Laplacian of Gaussian-LoG) της εικόνας σαν φίλτρο κλίμακας-χώρου που συμπεριλαμβάνει τον παράγοντα $\boldsymbol{\sigma}$ για διαφορετικές κλίμακες. Σύμφωνα με το [13] τα

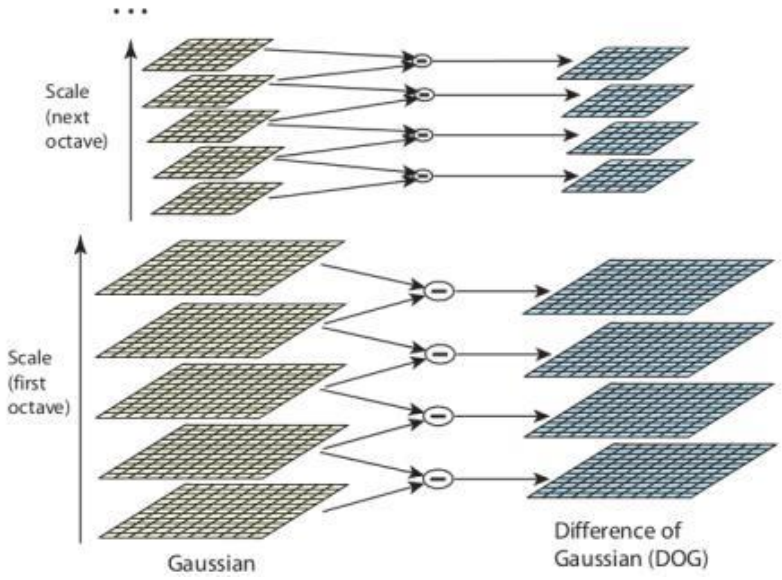


Εικόνα 2

τοπικά μέγιστα της συνάρτησης αυτής είναι τα πιθανά σημεία κλειδιά που ψάχνουμε. Λόγω όμως του κόστους υπολογισμού του LoG της εικόνας για όλες τις διαφορετικές τιμές χρησιμοποιείται η Διαφορά του Γκαουσιανού (Difference of Gaussian-DoG) της εικόνας για διαφορετικά $\boldsymbol{\sigma}$ που αποτελεί μια καλή προσέγγιση του LoG. Έτσι έχουμε:

$$\boldsymbol{D(x, y, \sigma) = L(x, y, k_i\sigma) - L(x, y, k_j\sigma)}$$

με $\boldsymbol{L(x, y, k_i\sigma) = G(x, y, k\sigma) * I(x, y)}$, όπου $\boldsymbol{I(x, y)}$ είναι η αρχική εικόνα και $\boldsymbol{G(x, y, k\sigma)}$ το Γκαουσιανό φίλτρο στην κλίμακα $\boldsymbol{k\sigma}$. Υπολογίζονται οι DoG για



Εικόνα 3

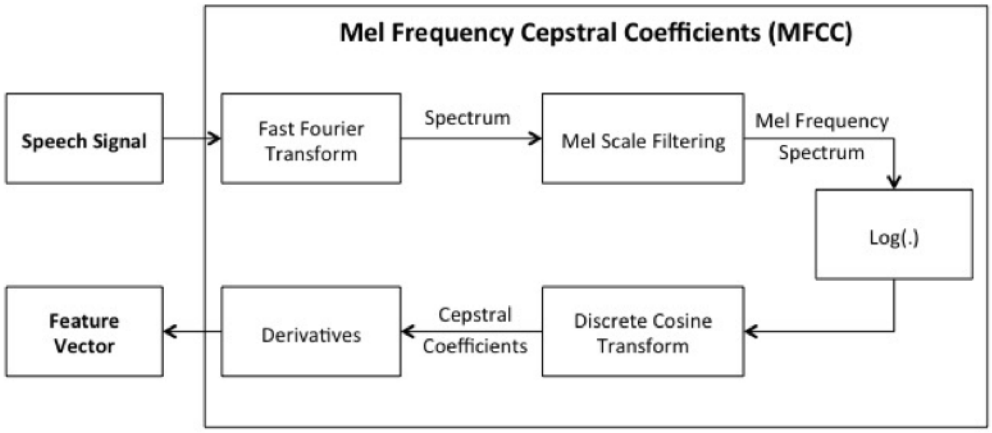
διαφορετικές οκτάβες της Γκαουσιανής Πυραμίδας, όπως φαίνεται και στην Εικ.3. Σε αυτές αναζητούμε τοπικά μέγιστα κλίμακα και χώρο. Έτσι, κάθε πίξελ συγκρίνεται με τα 8 γειτονικά του και με τα 9 της προηγούμενης κλίμακας και με τα άλλα 9 της επόμενης. Φτιάχνεται λοιπόν μια λίστα από $\boldsymbol{(x, y, \sigma)}$ που αποτελούν τα πιθανά σημεία κλειδιά. Από αυτά αποκλείονται όσα έχουν αντίθεση κάτω από ένα σταθερό κατώφλι (0.03 στο [13]) καθώς και όσα αφορούν ακμές, που εντοπίζονται με μια απλή συνάρτηση, ενώ μέσω του αναπτύγματος Taylor των DoG βρίσκονται με μεγαλύτερη ακρίβεια οι θέσεις των μεγίστων.

Στη συνέχεια δίνεται σε κάθε σημείο κλειδί ένας ή παραπάνω προσανατολισμός με βάση την τοπική κλίση της εικόνας. Με αυτό τον τρόπο εξασφαλίζεται και η ανεξαρτησία περιστροφής των χαρακτηριστικών, καθώς συμπεριλαμβάνουν την πληροφορία αυτή στην αναπαράστασή τους. Για να γίνει αυτό υπολογίζονται το μέτρο και η κατεύθυνση της κλίσης για κάθε πίξελ στη γειτονιά του σημείου κλειδιού. Δημιουργείται ένα ιστόγραμμα με 36 θέσεις που η καθεμία αντιπροσωπεύει 10 μοίρες και καθεμιά από τις παραπάνω κατευθύνσεις προστίθεται, λαμβάνοντας υπόψη και το μέτρο, αλλά και τη θέση του στη γειτονιά. Η γωνία με το μέγιστο στο ιστόγραμμα αποτελεί και τον προσανατολισμό του σημείου κλειδιού, ενώ αν υπάρχουν τοπικά μέγιστα στο 80% του ολικού μέγιστου, τότε του αποδίδονται και αυτές.

Τέλος, δημιουργείται το διάνυσμα/περιγραφέας του κάθε σημείου κλειδιού παίρνοντας μια γειτονιά 16x16 γύρω από το σημείο η οποία χωρίζεται σε 16 υπό-μπλοκ μεγέθους 4x4 το καθένα από τα οποία έχει ένα ιστόγραμμα 8 θέσεων. Σε αυτές προστίθενται οι κατευθύνσεις των σημείων του υπό-μπλοκ (με παρόμοιο τρόπο με αυτόν που αναφέραμε παραπάνω) και λαμβάνονται κάποια πρόσθετα μέτρα για να εξασφαλιστεί η ανεξαρτησία από αλλαγές στο φωτισμό. Τελικά δημιουργείται ένα διάνυσμα 128 στηλών για κάθε σημείο κλειδί.

3.1.2 Mel Frequency Cepstral Coefficients (MFCC)

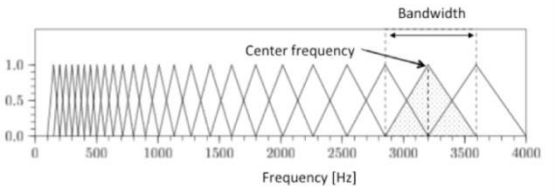
Αποτελεί την πιο διαδεδομένη μέθοδο εξαγωγής δεδομένων στην Αυτόματη Αναγνώριση Φωνής. Η μέθοδος αναπτύχθηκε από τον Mermelstein το 1976 [42] και βασίζεται σε πειράματα πάνω στην ανθρώπινη παρερμηνεία λέξεων. Ο αλγόριθμος MFCC προσπαθεί να μιμηθεί κομμάτια του τρόπου που ο άνθρωπος παράγει και αντιλαμβάνεται την ομιλία ενώ αποκλείει χαρακτηριστικά σχετικά με το μεγάφωνο, εξαιρώντας την κύρια συχνότητα και τις αρμονικές τους. Το τελικό διάνυσμα περιλαμβάνει και την αλλαγή του διανύσματος στο χρόνο για να αναπαραστήσει την δυναμική φύση της ομιλίας.



Εικόνα 4

Στο μπλοκ διάγραμμα της Εικ.4 φαίνονται τα βήματα της διαδικασίας υπολογισμού του τελικού διανύσματος. Αρχικά υπολογίζεται ο Διακριτός Μετασχηματισμός Fourier του σήματος εισόδου ώστε να πάρουμε το φάσμα συχνότητας, ώστε να υπολογίσουμε και το περιοδόγραμμα φάσματος ισχύος. Στη συνέχεια υπολογίζεται το φάσμα των συχνοτήτων Mel φιλτράροντας το περιοδόγραμμα με με μια τράπεζα συχνοτήτων Mel που αποτελείται από πολλαπλά ζωνοπερατά φίλτρα (παράδειγμα τέτοιας τράπεζας στην Εικ.5) και υπολογίζοντας τις εντάσεις των συχνοτήτων αυτών. Ο αριθμός, το σχήμα και οι κεντρικές συχνότητες των ζωνοπερατών φίλτρων ποικίλουν, ανάλογα την εφαρμογή. Η κλίμακα Mel είναι μια μη γραμμική κλίμακα και επιλέγεται γιατί μπορεί να προσαρμοστεί στη μη γραμμική τονική αντίληψη του ανθρώπινου ακουστικού

συστήματος. Έπειτα υπολογίζουμε το λογάριθμο των παραπάνω εντάσεων. Αυτή η επιλογή βασίζεται σε πειράματα που δείχνουν ότι η ανθρώπινη αντίληψη της έντασης του ήχου είναι σε λογαριθμική κλίμακα. Ακολούθως, υπολογίζουμε τους Cepstral συντελεστές του σήματος υπολογίζοντας τον Διακριτό Μετασχηματισμό Συνημιτόνου (Δ.Μ.Σ.) του και κρατώντας τις χαμηλής έντασης συχνότητες του cepstrum (προκύπτει από τον Δ.Μ.Σ. και μπορεί να ερμηνευτεί σαν φάσμα του φάσματος). Η διαδικασία αυτή γίνεται ώστε να απομονοθεί η συχνότητα και οι αρμονικές της που σχετίζονται με το μεγάφωνο που παράγει το σήμα ήχου, καθώς οι συχνότητες αυτές μετασχηματίζονται σε συντελεστές Cepstral μεγαλύτερης έντασης και έτσι ανιχνεύονται και αγνοούνται στη συνέχεια. Τέλος, για να συμπεριληφθεί η χρονική μεταβολή των συντελεστών, το τελικό διάνυσμα πέραν των συντελεστών για τα διάφορα χρονικά παράθυρα και συχνότητες που υπολογίστηκαν επεκτείνεται και με της πρώτης και δεύτερης τάξης παραγώγους των συντελεστών αυτών. Έτσι:



Εικόνα 5

Αν $\boldsymbol{c_{\tau,j}}$ ο Cepstral συντελεστής του χρονικού παραθύρου $\boldsymbol{\tau}$ και της συχνότητας $\boldsymbol{j}$, τότε $\boldsymbol{\Delta c_{\tau,j} = c_{\tau+1,j} - c_{\tau-1,j}}$ και $\boldsymbol{\Delta\Delta c_{\tau,j} = \Delta c_{\tau+1,j} - \Delta c_{\tau-1,j}}$ θα είναι οι παράγωγοι πρώτης και δεύτερης τάξης, ενώ το τελικό διάνυσμα θα είναι της μορφής $\boldsymbol{[c_{\tau,j}, \Delta c_{\tau,j}, \Delta\Delta c_{\tau,j}]}$.

3.1.3 Spatio Temporal Interest Points (STIP)

Τα χώρο-χρονικά σημεία ενδιαφέροντος δημιουργήθηκαν το 2005 από τον I.Laptev [6] με σκοπό την επέκταση υπαρχόντων ανιχνευτών γωνιών Harris για δεδομένα εικόνας και στη διάσταση του χρόνου που εμφανίζεται στα δεδομένα βίντεο. Τα βίντεο σε αυτή την προσέγγιση θεωρούνται σαν όγκοι από pixel, ενώ τα pixel αποτελούν ουσιαστικά κύβους με το χρόνο να αντιστοιχεί στη διάσταση του βάθους τους. Όπως και στη 2Δ περίπτωση των SIFT χαρακτηριστικών, τα χαρακτηριστικά αυτά παραμένουν σταθερά σε αλλαγές περιστροφής, οπτικής γωνίας, κλίμακας και φωτισμού. Υπάρχουν διάφορες παραλλαγές χώρο-χρονικών χαρακτηριστικών, αλλά στην υλοποίησή του I.Laptev που παρουσιάζουμε παρακάτω χρησιμοποιούνται οι Harris3D για ανιχνευτές των σημείων ενδιαφέροντος και ο συνδυασμός Ιστογραμμάτων Προσανατολισμένων Κλίσεων (HOG) και Οπτικής Ροής (HOF) για περιγραφείς των σημείων που ανιχνεύτηκαν.

6 Παράδειγμα STIP χαρακτηριστικών

Αρχικά πρέπει να δώσουμε τους παρακάτω ορισμούς:

Έστω ότι το βίντεο αναπαρίσταται με μια συνάρτηση $\boldsymbol{f(x, y, t)}$.

$\boldsymbol{g(x, y, t; \sigma^2, \tau^2) = \frac{1}{\sqrt{(2\pi)^3 \sigma^4 \tau^2}} \exp(-\frac{x^2+y^2}{2\sigma^2} - \frac{t^2}{2\tau^2})}$ είναι ένας χώρο-χρονικός διαχωρίσιμος Γκαουσιανός πυρήνας, όπου $\boldsymbol{\sigma^2, \tau^2}$ αποτελούν τις παραμέτρους κλίμακας για το χώρο και το χρόνο αντίστοιχα.

$\boldsymbol{L(x, y, t; \sigma^2, \tau^2) = f(x, y, t) * g(x, y, t; \sigma^2, \tau^2)}$ είναι το Γκαουσιανό του βίντεο για αναπαράσταση σε χώρο-κλίμακα, ενώ η $\boldsymbol{\nabla L = (L_x, L_y, L_t)^T}$ αποτελεί την χώρο-χρονική κλίση για κάθε διαφορετικό $\boldsymbol{(x, y, t)}$ του βίντεο.

$$\boldsymbol{\mu = g(\cdot; \sigma^2, \tau^2) * \left(\left(\nabla L(\cdot; \sigma^2, \tau^2) \right) ((\nabla L(\cdot; \sigma^2, \tau^2))^T \right) = g(\cdot; \sigma^2, \tau^2) * \begin{pmatrix} L_x^2 & L_y L_x & L_t L_x \\ L_x L_y & L_y^2 & L_t L_y \\ L_x L_t & L_y L_t & L_t^2 \end{pmatrix}}$$

αποτελεί τον πίνακα ροπών 2ης τάξης της χώρο-χρονικής αναπαράστασης του βίντεο.

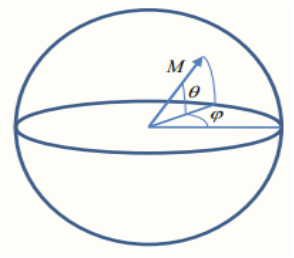
Σκοπός του ανιχνευτή γωνιών είναι να βρει σημεία της $\boldsymbol{f}$ με σημαντικές αλλαγές σε καθεμιά από τις τρεις διαστάσεις της. Ο πίνακας $\boldsymbol{\mu}$ αποτελεί ουσιαστικά τη διανομή των κλίσεων σε μια γειτονιά ενός σημείου του βίντεο. Έτσι, οι ιδιοτιμές του πίνακα αυτού δείχνουν τις μεταβολές της $\boldsymbol{f}$ στην κάθε διάσταση. Συγκεκριμένα, αν παρατηρήσουμε μεγάλες τις ιδιοτιμές $\boldsymbol{\lambda_1, \lambda_2, \lambda_3}$ σε ένα σημείο, το σημείο αυτό θεωρείται από τα ζητούμενα σημεία ενδιαφέροντος. Αντίστοιχα με το [43] για να ανιχνεύσουμε τέτοια σημεία πρέπει η συνάρτηση $\boldsymbol{H = \det(\mu) - k\, trace^3(\mu) = \lambda_1 \lambda_2 \lambda_3 - k(\lambda_1 + \lambda_2 + \lambda_3)^3}$ να παρουσιάζει τοπικό μέγιστο. Το $\boldsymbol{k}$ αποτελεί μια σταθερά. Στη συνέχεια, για να βρεθούν οι χαρακτηριστικές κλίμακες χώρου και χρόνου για το σημείο που ανιχνεύτηκε υπολογίζονται τα κανονικοποιημένα Λαπλασιανά ($\boldsymbol{\nabla^2_{norm} L}$) στην τοποθεσία του και σε γειτονικές κλίμακες και επιλέγονται αυτές που μεγιστοποιούν το Λαπλασιανό.

Για τη δημιουργία του περιγραφέα των παραπάνω σημείων ενδιαφέροντος, ακολουθείται μια διαδικασία αρκετά παρόμοια με αυτή που είδαμε προηγουμένως με τα χαρακτηριστικά SIFT. Αρχικά, ορίζεται μια κυβική γειτονιά του σημείου της μορφής $\boldsymbol{(k\sigma)(k\sigma)(k'\tau)}$, η οποία χωρίζεται εκ νέου σε μικρότερες κυβικές υπό-γειτονιές. Για κάθε pixel της γειτονιάς υπολογίζεται μια συνάρτηση και τέλος δημιουργείται το ιστόγραμμα. Στην περίπτωση των STIP, όπως προαναφέραμε, υπολογίζονται δύο ειδών ιστογράμματα.

Για τα Ιστογράμματα 3Δ Προσανατολισμένων Κλίσεων (HOG3D) οι κλίσεις υπολογίζονται για καθεμιά από τις μεταβλητές από τις διαφορές της συνάρτησης εικόνας:

$\boldsymbol{G_x(x, y, t) = I(x + 1, y, t) - I(x - 1, y, t)}$
$\boldsymbol{G_y(x, y, t) = I(x, y + 1, t) - I(x, y - 1, t)}$
$\boldsymbol{G_t(x, y, t) = I(x, y, t + 1) - I(x, y, t - 1)}$

Από αυτές υπολογίζεται το μέτρο $\boldsymbol{M = \sqrt{G_x^2 + G_y^2 + G_t^2}}$, και οι κατευθύνσεις $\boldsymbol{\theta = \tan^{-1}(G_t / \sqrt{G_x^2 + G_y^2})}$ και $\boldsymbol{\varphi = \tan^{-1}(G_y / G_x)}$ (Εικ.7). Το ιστόγραμμα δημιουργείται με κάθε σημείο της γειτονιάς να αντιστοιχίζεται σε κάποια θέση του ιστογράμματος με βάση τις κατευθύνσεις του, αλλά και με διαφορετικό βάρος ανάλογα το μέτρο, τη θέση του στη γειτονιά (αν είναι στις άκρες ή στο κέντρο) κ.ά.



Εικόνα 7

Τα Ιστογράμματα Οπτικής Ροής (HOF) δημιουργούνται με παρόμοιο τρόπο, με τη διαφορά ότι εκεί χρειάζεται να απεικονιστεί η εμφανής κίνηση ενός pixel μεταξύ δύο frames. Έτσι υπολογίζονται διανύσματα ταχυτήτων για κάθε pixel $(\boldsymbol{V_x, V_y})$ με βάση τις μεταβολές της έντασης της εικόνας. Από αυτά υπολογίζονται το μέτρο $\boldsymbol{M = \sqrt{{V_x}^2 + {V_y}^2}}$ και η κατεύθυνση $\boldsymbol{\theta = \tan^{-1}(V_x / V_y)}$ του διανύσματος της ταχύτητας, ενώ το ιστόγραμμα δημιουργείται με θέσεις για τις διαφορετικές τιμές της κατεύθυνσης των pixel της γειτονιάς, συνυπολογίζοντας και το μέτρο, ενώ υπολογίζεται και θέση για έλλειψη κίνησης.

3.2 Σάκος λέξεων

Αφού ολοκληρωθεί η διαδικασία εξαγωγής των παραπάνω χαρακτηριστικών βρισκόμαστε αντιμέτωποι με μια πρόκληση. Κάθε βίντεο του συνόλου εκπαίδευσης μας έχει πλέον αναπαρασταθεί από ένα σύνολο διανυσμάτων, καθένα από τα οποία περιγράφει ένα σημείο ενδιαφέροντος του βίντεο. Προκειμένου, όμως, τα δεδομένα αυτά να τροφοδοτήσουν τον ταξινομητή μας για τη διαδικασία της εκπαίδευσης θα πρέπει να έχουμε καταλήξει σε ένα αντιπροσωπευτικό διάνυσμα για κάθε βίντεο. Για να το επιτύχουμε αυτό, χρησιμοποιούμε την τεχνική του «σάκου λέξεων» (Bag of Words) που θα αναλύσουμε παρακάτω. Η μέθοδος αυτή είναι εμπνευσμένη από τον κλάδο της επεξεργασίας και ανάκτησης κειμένου [44] και κάποιες από τις πρώτες προσεγγίσεις της μεθόδου στην αναγνώριση βίντεο είναι των J.Sivic και A.Zisserman [45] και των G.Csurka et al. [46] λίγο αργότερα.

Ο σάκος λέξεων, όταν πρόκειται για αναπαράσταση κειμένου, εντοπίζει όλες ή τις περισσότερες διαφορετικές λέξεις που εμφανίζονται στο κείμενο (αγνοώντας άρθρα, συνδέσμους κ.λπ.) και στη συνέχεια δημιουργεί ένα ιστόγραμμα με μια θέση για καθεμιά από τις διαφορετικές λέξεις που αντιπροσωπεύει τη συχνότητα εμφάνισής της. Έτσι συμπυκνώνεται ένα ολόκληρο κείμενο σε ένα διάνυσμα, πράγμα ιδιαίτερα αποδοτικό σε εργασίες κατηγοριοποίησης, αναγνώρισης κ.ά. Παρακάτω παρουσιάζουμε πως η μέθοδος αυτή εφαρμόζεται σε δεδομένα βίντεο.

Σύμφωνα με το [1] η διαδικασία μπορεί να περιγραφεί απλά σε 5 βήματα:

- **Εξαγωγή Χαρακτηριστικών**: Είναι η διαδικασία που προαναφέραμε όπου εντοπίζονται τα σημεία ενδιαφέροντος στο βίντεο (ανάλογα με τον αλγόριθμο που έχουμε επιλέξει να τα εξάγουμε). Συνήθως αναφερόμαστε σε τοπικά χαρακτηριστικά χαμηλού επιπέδου, τα οποία στις περισσότερες περιπτώσεις αποτελούνται από έναν ανιχνευτή των χαρακτηριστικών και έναν ή περισσότερους περιγραφείς.
- **Προεπεξεργασία Χαρακτηριστικών**: Αποτελεί ένα στάδιο που άρχισε να συμπεριλαμβάνεται στη διαδικασία τα τελευταία χρόνια και παρά το ότι μπορεί να μοιάζει προαιρετικό, έχει παρουσιάσει σημαντική βελτίωση στην απόδοση των συστημάτων. Οι χαμηλού επιπέδου περιγραφείς των χαρακτηριστικών που έχουν εξαχθεί συνήθως έχουν μεγάλη διαστατικότητα και ισχυρή συσχέτιση πράγμα που δυσκολεύει πολύ τους αλγόριθμους μη επιβλεπόμενης μάθησης για τη δημιουργία του λεξικού. Έτσι, προκρίνεται η Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis-PCA) ώστε να καταστήσει γραμμικά ανεξάρτητες τις μεταβλητές των διανυσμάτων, ενώ συνήθως επιλέγεται αριθμός κύριων συνιστωσών μικρότερος από τις αρχικές μεταβλητές, άρα επιτυγχάνεται και μείωση διαστατικότητας. Στη συνέχεια, εφαρμόζεται ένας γραμμικός μετασχηματισμός στο σύνολο των χαρακτηριστικών ώστε ο πίνακας συνδιακύμανσης τους να είναι μοναδιαίος, να έχουν δηλαδή διακύμανση ίση με 1 σε κάθε διάστασή. Η διαδικασία αυτή ονομάζεται «άσπρισμα» δεδομένων (Whitening data).
- **Δημιουργία Λεξικού**: Υπάρχουν διάφορες προσεγγίσεις, ανάλογα την εφαρμογή, για τη δημιουργία του λεξικού. Αυτή που χρησιμοποιούμε στην παρούσα εργασία εφαρμόζει τον αλγόριθμο K-μέσων (K-means) σε ένα δείγμα των δεδομένων εκπαίδευσης από όλες τις κλάσεις για την συσταδοποίηση των σημείων ενδιαφέροντος που έχουμε εξάγει. Κάθε μία από τις K συστάδες που δημιουργούνται αντιπροσωπεύεται από ένα μέλος-πρότυπο της και οι αντιπρόσωποι αυτοί αποτελούν τελικά το λεξικό με το οποίο θα κωδικοποιηθούν στη συνέχεια τα δεδομένα. Αναλυτικά η λειτουργία του αλγορίθμου παρουσιάζεται στο επόμενο κεφάλαιο.
- **Κωδικοποίηση Χαρακτηριστικών**: Σε αυτό το βήμα επιλέγουμε τον αλγόριθμο με τον οποίο θα κωδικοποιηθούν τα δεδομένα εκπαίδευσης. Με βάση την αντιστοίχιση του κάθε σημείου ενδιαφέροντος με μια λέξη του λεξικού δημιουργείται ένα ιστόγραμμα με θέσεις για κάθε λέξη. Στην απλή περίπτωση που εφαρμόζουμε στην παρούσα εργασία χρησιμοποιούμε τον Κβαντισμό Διανύσματος (Vector Quantization-VQ) και όταν μια λέξη (σημείο ενδιαφέροντος) μιας συστάδας εμφανίζεται στο βίντεο θέτουμε **1** στην αντίστοιχη θέση του ιστογράμματος και **0** στις υπόλοιπες θέσεις. Έτσι πλέον κάθε σημείο ενδιαφέροντος έχει εκφραστεί από ένα διάνυσμα, με κοινό μήκος (όσες και οι λέξεις του λεξικού) για όλα. Στο επόμενο βήμα βλέπουμε το πώς συνδυάζουμε τα διανύσματα αυτά για να δημιουργήσουμε την αναπαράσταση του βίντεο με ένα και μοναδικό διάνυσμα. Αξίζει να σημειωθεί ότι η επιλογή αλγορίθμου για την

δημιουργία του λεξικού, σε κάποιες περιπτώσεις σχετίζεται άμεσα και με την επιλογή κωδικοποίησης σε αυτό το βήμα και η μία επιλογή υπαγορεύει και την άλλη.

- **Συγκέντρωση και Κανονικοποίηση**: Στο τελευταίο βήμα συγκεντρώνουμε τα κωδικοποιημένα πλέον σημεία ενδιαφέροντος του βίντεο. Στην παρούσα εργασία επιλέγουμε να αναπαραστήσουμε το βίντεο με το άθροισμα των κωδικοποιημένων περιγραφέων των σημείων ενδιαφέροντός του. Στο τέλος της διαδικασίας εφαρμόζουμε μια κανονικοποίηση στις αναπαραστάσεις αυτές, ώστε να περιορίσουμε την αναντιστοιχία σε ποσότητα εξαγμένων χαρακτηριστικών που μπορεί να έχουν τα βίντεο μεταξύ τους. Στην υλοποίησή μας προκρίνεται η L2-κανονικοποίηση που διαιρεί το κάθε διάνυσμα με την Ευκλείδια νόρμα του.

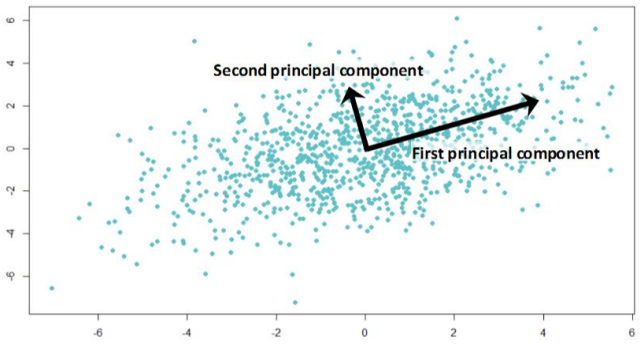
Στο τέλος της διαδικασίας αυτής έχουμε κατασκευάσει ένα διάνυσμα αναπαράστασης για κάθε βίντεο των δεδομένων μας που εμπεριέχει την πληροφορία όλων των χαρακτηριστικών του που εξάγαμε στην αρχή της διαδικασίας. Τα διανύσματα αυτά αποτελούν και το σύνολο δεδομένων εκπαίδευσης που θα τροφοδοτήσουν τον ταξινομητή που μέσω εκπαίδευσης θα επιλύσει το πρόβλημα κατηγοριοποίησης που αντιμετωπίζουμε.

Κεφάλαιο 4 Αλγόριθμοι και Ευφυείς Τεχνικές

Στο παρόν κεφάλαιο παρουσιάζουμε την Ανάλυση Κύριων Συνιστωσών (PCA) που χρησιμοποιούμε κατά την προεπεξεργασία των δεδομένων μας στη φάση της εφαρμογής του σάκου λέξεων, τους αλγορίθμους μηχανικής μάθησης που χρησιμοποιούνται στην υλοποίησή μας και στη συνέχεια αναλύουμε τεχνικές που χρησιμοποιήσαμε για τη συγχώνευση των διαφορετικών εξαγμένων χαρακτηριστικών, καθώς όπως προαναφέρθηκε υποστηρίζουμε ότι λειτουργούν συμπληρωματικά μεταξύ τους. Θα δούμε παρακάτω, ότι πέραν των αλγορίθμων μάθησης που μοντελοποιούν και επιλύουν το πρόβλημα που αντιμετωπίζουμε, σημαντικό ρόλο στην απόδοση του συστήματος παίζουν και οι μέθοδοι με τις οποίες θα συνδυάσουμε τα αποτελέσματα τους.

4.1 *Principal Components Analysis (PCA)*

Η Ανάλυση Κύριων Συνιστωσών [21] αποτελεί μια στατιστική διαδικασία που χρησιμοποιεί έναν γραμμικό μετασχηματισμό για να μετατρέψει ένα σύνολο δεδομένων συσχετισμένων μεταξύ τους (όπως συμβαίνει στα δεδομένα εικόνας, όπου γειτονικά pixel σχετίζονται μεταξύ τους) σε ένα σύνολο γραμμικά ανεξάρτητων μεταβλητών που ονομάζονται Κύριες Συνιστώσες. Ο αριθμός των συνιστωσών αυτών είναι ίσος με τον αριθμό των ιδιοδιανυσμάτων του πίνακα δεδομένων, δηλαδή ίσος με τη μικρότερη διάσταση του πίνακα. Η πρώτη κύρια συνιστώσα είναι αυτή που έχει και τη μεγαλύτερη δυνατή διακύμανση και οι υπόλοιπες κατατάσσονται με βάση τη διακύμανση τους σε φθίνουσα σειρά και κάθε επόμενη οφείλει να είναι κάθετη στις προηγούμενες. Ο αλγόριθμος χρησιμοποιείται συχνά για μείωση διαστατικότητας σε δεδομένα, είτε με σκοπό την οπτικοποίηση των δεδομένων σε λιγότερες διαστάσεις, είτε ως ένα χρήσιμο βήμα προεπεξεργασίας για μη επιβλεπόμενη μάθηση, καθώς οι κύριες συνιστώσες περιέχουν την ίδια στατιστική πληροφορία με τα αρχικά δεδομένα, συμπυκνωμένη όμως στις πρώτες συνιστώσες με αποτέλεσμα να μπορούμε να παραλείψουμε πολλές από αυτές.



8 Παράδειγμα κύριων συνιστωσών σε ένα σύνολο δεδομένων

Έστω ότι έχουμε ένα σύνολο δεδομένων n διαστάσεων. Αρχικά, αφαιρούμε το μέσο όρο κάθε διάστασης από την αντίστοιχη στήλη κάθε δείγματος με σκοπό τα νέα δεδομένα να έχουν όλα κοινό μέσο όρο ίσο με 0. Έπειτα, υπολογίζουμε τον πίνακα συνδιακύμανσης

των δεδομένων που προκύπτει από τη σχέση: $\boldsymbol{\Sigma} = \frac{\mathbf{1}}{\boldsymbol{m}} \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{m}} (\boldsymbol{x_i})(\boldsymbol{x_i})^{\boldsymbol{T}}$, με $\boldsymbol{m}$ το πλήθος των δειγμάτων. Από αυτόν τον πίνακα βρίσκουμε τα ιδιοδιανύσματά του, όπως και τις αντίστοιχες ιδιοτιμές τους και δημιουργούμε τον παρακάτω πίνακα, όπου κάθε στήλη είναι ένα ιδιοδιάνυσμα, ταξινομημένα από τη μεγαλύτερη προς τη μικρότερη αντίστοιχη ιδιοτιμή: $\boldsymbol{U} = \begin{bmatrix} | & | & & | \\ \boldsymbol{u_1} & \boldsymbol{u_2} & \dots & \boldsymbol{u_n} \\ | & | & & | \end{bmatrix}$. Τέλος, σκοπός μας ουσιαστικά είναι να εκφράσουμε τα αρχικά δεδομένα μας σε ένα νέο σύστημα με άξονες τα ιδιοδιανύσματα αυτά, που αντιπροσωπεύουν τις διευθύνσεις μεταβολής των δεδομένων μας. Εφαρμόζουμε τον παρακάτω μετασχηματισμό που πρακτικά «περιστρέφει» τα δεδομένα μας με βάση τους άξονες των ιδιοδιανυσμάτων: $\boldsymbol{x_{PCA}} = \boldsymbol{U^T x} = \begin{bmatrix} \boldsymbol{u_1^T x} \\ \boldsymbol{u_2^T x} \\ \vdots \\ \boldsymbol{u_n^T x} \end{bmatrix}$. Αν θέλουμε να επιστρέψουμε στα αρχικά μας δεδομένα, εφαρμόζουμε τον αντίστροφο μετασχηματισμό, δηλαδή πολλαπλασιάζουμε το $\boldsymbol{x_{PCA}}$ με τον πίνακα $\boldsymbol{U}$. Αυτό συμβαίνει επειδή ο πίνακας $\boldsymbol{U}$ είναι ορθογώνιος.

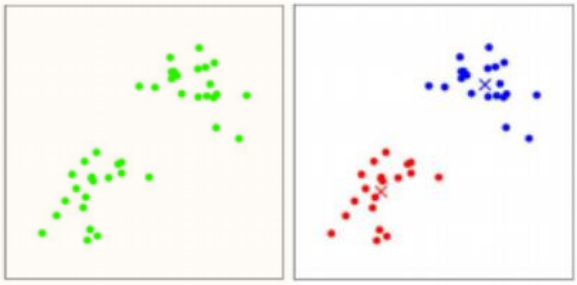
Αφού ολοκληρωθεί ο μετασχηματισμός αυτός, τα δεδομένα μας έχουν εκφραστεί με βάση όλα τα ιδιοδιανύσματα του πίνακα συνδιακύμανσης και έτσι ο καινούριος πίνακας έχει τις ίδιες διαστάσεις με τον αρχικό. Για να μειώσουμε τις διαστάσεις, που είναι ο στόχος της διαδικασίας, απλά επιλέγουμε τα $\boldsymbol{k}$ πρώτα από τα ιδιοδιανύσματα του πίνακα $\boldsymbol{U}$ πριν κάνουμε το μετασχηματισμό και μηδενίζουμε τα υπόλοιπα. Αυτό αποτελεί μια προσέγγιση του πίνακα δεδομένων, καθώς όπως προαναφέρθηκε τα ιδιοδιανύσματα είναι ταξινομημένα από την μεγαλύτερη προς τη μικρότερη ιδιοτιμή και άρα από τη μεγαλύτερη διακύμανση προς τη μικρότερη. Για να επιλέξουμε κατάλληλο $\boldsymbol{k}$ που θα επιφέρει σημαντική μείωση των διαστάσεων, χωρίς να χάσουμε σημαντική πληροφορία από τον πίνακα, υπολογίζουμε το άθροισμα των ιδιοτιμών των $\boldsymbol{k}$ επιλεγμένων ιδιοδιανυσμάτων και το διαιρούμε με το άθροισμα όλων των ιδιοτιμών. Αυτό αποτελεί και κατά προσέγγιση το ποσοστό διακύμανσης που εμπεριέχουν οι επιλεγμένες συνιστώσες. Σε εφαρμογές εικόνας, ένα αποδεκτό ποσοστό της διακύμανσης είναι κοντά στο 99%. Αξίζει εδώ να σημειωθεί, ότι το μεγαλύτερο μέρος του ποσοστού που παίρνουμε από την επιλογή αυτή προκύπτει από τις λίγες πρώτες συνιστώσες και κάθε επόμενη προσφέρει όλο και λιγότερο, με αποτέλεσμα να μπορούμε να μειώσουμε τις διαστάσεις των δεδομένων ακόμα και στο 25% των αρχικών, ανάλογα την εφαρμογή.

Συνήθως, μετά το τέλος της παραπάνω διαδικασίας, εφαρμόζεται άλλος ένας μετασχηματισμός που ονομάζεται «άσπρισμα» δεδομένων (whitening) και στοχεύει στο να έχουν όλα τα δεδομένα κοινή διακύμανση και ίση με 1. Αυτό χρησιμεύει στην βελτίωση της ομοιομορφίας των δεδομένων μας πράγμα που βοηθάει στην απόδοση και ταχύτητα αλγορίθμων συσταδοποίησης. Για να το επιτύχουμε αυτό, διαιρούμε κάθε στήλη του πίνακα δεδομένων που προέκυψε από την Ανάλυση Κύριων Συνιστωσών με τη ρίζα της αντίστοιχης ιδιοτιμής του πίνακα συνδιακύμανσης, που έχουν υπολογιστεί από την προηγούμενη διαδικασία.

4.2 Αλγόριθμοι μηχανικής μάθησης

Οι αλγόριθμοι μηχανικής μάθησης που χρησιμοποιούνται στην υλοποίησή μας είναι ο αλγόριθμος Κ-μέσων (K-means) και οι Μηχανές Διανυσμάτων Υποστήριξης (SVM).

4.2.1 K-means clustering



9 Παράδειγμα Κ-μέσων

Ο αλγόριθμος Κ-μέσων (MacQueen, 1967) [21] είναι ένας από τους απλούστερους αλγορίθμους μη επιβλεπόμενης μάθησης για το πρόβλημα συσταδοποίησης. Είναι ένας διαχωριστικός αλγόριθμος που στοχεύει στην κατηγοριοποίηση ενός συνόλου δεδομένων σε Κ συστάδες, με τον αριθμό των συστάδων να επιλέγεται εξ αρχής. Τα δεδομένα ομαδοποιούνται με βάση την ομοιότητα μεταξύ τους, ενώ ο αλγόριθμος ορίζει και τα κέντρα των ομάδων ως «αντιπροσώπους» της ομάδας, χωρίς αυτά να ανήκουν απαραίτητα στα αρχικά δεδομένα, ανήκουν όμως στον ίδιο χώρο με αυτά. Αποτελεί έναν απλό, γρήγορο και εύρωστο αλγόριθμο που λειτουργεί αρκετά αποδοτικά για μεγάλο αριθμό δειγμάτων. Μερικές από τις αδυναμίες του όμως, είναι ότι δεν αποδίδει το ίδιο καλά σε όλων των ειδών τα δεδομένα (π.χ. αποδίδει καλύτερα σε πιο απομακρυσμένα μεταξύ τους δεδομένα σε σχέση με πιο πυκνά), χρειάζεται να οριστούν εξ αρχής οι ζητούμενες ομάδες, πράγμα που μπορεί να μην είναι εφικτό ή πολύ αυθαίρετο να γίνει, ενώ είναι συχνό φαινόμενο να τερματίζει σε τοπικό ελάχιστο της αντικειμενικής συνάρτησης που προσπαθεί να ελαχιστοποιήσει, χωρίς να καταλήγει σε βέλτιστη λύση και έτσι να απαιτούνται πολλαπλές επαναλήψεις του αλγορίθμου ώστε να είμαστε βέβαιοι ότι έχει βρεθεί μια βέλτιστη λύση.

Τα βήματα του αλγορίθμου είναι τα εξής:

1) Αρχικά, επιλέγονται τυχαία Κ κέντρα για τις συστάδες. Η αρχική επιλογή κέντρων μπορεί να γίνει με διάφορους τρόπους, ανάλογα την εφαρμογή και να βελτιώσουν την απόδοση του αλγορίθμου, αλλά εδώ περιγράφουμε την απλή περίπτωση.
2) Υπολογίζεται η Ευκλείδεια απόσταση κάθε σημείου των δεδομένων από κάθε ένα από τα κέντρα που επιλέχθηκαν.
3) Αντιστοιχίζεται κάθε σημείο στην αντίστοιχη ομάδα του κέντρου από το οποίο έχει τη μικρότερη απόσταση.
4) Υπολογίζονται τα νέα κέντρα των ομάδων που διαμορφώθηκαν από την παραπάνω αντιστοίχιση. Τα κέντρα αυτά αποτελούν το γεωμετρικό κέντρο βάρους της ομάδας των σημείων και υπολογίζονται με την εξίσωση:

$\boldsymbol{v_i} = (\mathbf{1}/\boldsymbol{c_i}) \sum_{j=1}^{c_i} \boldsymbol{x_i}$, όπου $\boldsymbol{c_i}$ είναι ο αριθμός των στοιχείων της ομάδας $\boldsymbol{i}$ και $\boldsymbol{x_i}$ τα σημεία της ομάδας.

5) Υπολογίζεται ξανά η Ευκλείδεια απόσταση των σημείων από όλα τα καινούρια κέντρα
6) Αντιστοιχίζονται ξανά τα σημεία στις ομάδες που έχουν την μικρότερη απόσταση από τα κέντρα τους. Αν δεν υπάρχει αλλαγή αντιστοίχισης για κάποιο σημείο, τότε ο αλγόριθμος τερματίζει. Διαφορετικά, υπολογίζονται εκ νέου τα κέντρα των ομάδων, όπως στο βήμα 4, και η διαδικασία συνεχίζει επαναληπτικά μέχρι να μην υπάρχουν αλλαγές στις ομάδες.

Η αντικειμενική συνάρτηση που προσπαθεί να ελαχιστοποιήσει ο αλγόριθμος είναι το άθροισμα τετραγωνικού σφάλματος των αποστάσεων των σημείων από τα κέντρα: $\boldsymbol{J} = \sum_{j=1}^{k} \sum_{i=1}^{c_j} \left\| \boldsymbol{x_i^{(j)}} - \boldsymbol{\mu_j} \right\|^2$. Αυτό αποτελεί και ένα μέτρο ποιότητας της συσταδοποίησης που έχει επιφέρει ο αλγόριθμος. Αν χρειάζεται να μειώσουμε ακόμα περισσότερο την αντικειμενική συνάρτηση, μπορούμε να το κάνουμε με 2 τρόπους:

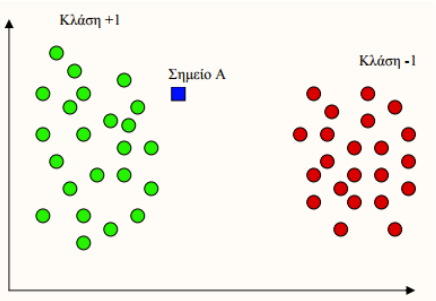
- **<u>Μεταβάλλοντας το Κ</u>**: Η επιλογή του αριθμού των συστάδων μπορεί να αλλάξει, καθώς δεν γνωρίζουμε από πριν τη μορφή των δεδομένων στο χώρο. Κατά γενική ομολογία, αυξάνοντας το Κ μειώνεται η αντικειμενική συνάρτηση. Όμως, όπως είναι εμφανές, το σφάλμα θα μηδενίζεται όταν το Κ γίνει ίσο με το πλήθος των στοιχείων. Έτσι, δεν μπορούμε να αυξάνουμε συνεχώς το Κ, μπορούμε όμως να δοκιμάσουμε ένα εύρος διαφορετικών τιμών και να επιλέξουμε προσεγγιστικά αυτή στην οποία μειώνεται σημαντικά ο ρυθμός μείωσης του σφάλματος. Στην πράξη, το Κ συνήθως επιλέγεται με βάση την εμπειρία σε παρόμοια προβλήματα με αυτό που αντιμετωπίζουμε.
- **<u>Επιλογή αρχικών κέντρων</u>**: Ο τρόπος που θα επιλέξουμε τα αρχικά κέντρα των ομάδων είναι και ο δραστικότερος στη μείωση του σφάλματος, καθώς ακόμα και με αύξηση του Κ, μια κακή επιλογή αρχικών κέντρων μπορεί να χειροτερέψει τα αποτελέσματα. Μια απλή προσέγγιση στο πρόβλημα είναι η επανάληψη του αλγορίθμου πολλές φορές με τυχαία επιλογή αρχικών κέντρων, ώστε να επιλέξουμε τελικά αυτή με το μικρότερο σφάλμα. Μια άλλη προσέγγιση είναι μια «σχεδόν» τυχαία επιλογή των αρχικών σημείων με μέριμνα τα σημεία αυτά να είναι όσο γίνεται πιο απομακρυσμένα μεταξύ τους, ώστε να χουν περισσότερες πιθανότητες να δημιουργήσουν καλά διαχωρισμένες και συμπαγής ομάδες που θα μειώνουν το σφάλμα. Συνήθως και οι δύο τρόποι χρησιμοποιούνται και βελτιώνουν σημαντικά το σφάλμα του αλγορίθμου.

Αναλύοντας τον τρόπο που λειτουργεί ο αλγόριθμος, παρατηρούμε ότι σημαντικό ρόλο σε όλη τη διαδικασία παίζουν τόσο η Ευκλείδεια απόσταση των σημείων μεταξύ τους, όσο και η διασπορά τους στο χώρο. Ο αλγόριθμος δείχνει να προϋποθέτει ότι οι ομάδες των δεδομένων έχουν κυκλικά σχήματα (αφού ορίζονται από την Ευκλείδεια απόστασή τους από σταθερό σημείο), ενώ πολύ πυκνά σημεία μπορεί να είναι δυσκολότερο να ομαδοποιηθούν σωστά. Αυτές οι παρατηρήσεις μας δείχνουν καθαρότερα και τη σημασία της προεπεξεργασίας των δεδομένων με την Ανάλυση Κύριων Συνιστωσών πριν

εφαρμοστεί ο αλγόριθμος Κ-μέσων. Οι κύριες συνιστώσες εμπεριέχουν τη μεγαλύτερη δυνατή διασπορά των δεδομένων, αποτελώντας μια πιο ομοιόμορφη αναπαράστασή τους, ενώ η μείωση διαστάσεων που πραγματοποιείται αποτελεί μια προσέγγιση που αναδεικνύει τις ομοιότητες και διαφορές μεταξύ των διαφορετικών ομάδων δεδομένων, μιας και η Ευκλείδεια απόσταση δεν είναι τόσο ενδεικτική μετρική σε περιπτώσεις πολλών διαστάσεων.

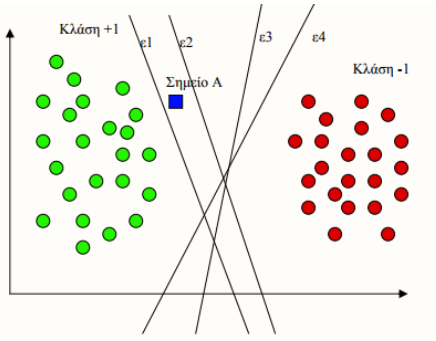
4.2.2 Support Vector Machines (SVM)

Οι Μηχανές Διανυσμάτων Υποστήριξης (Vapnik-Chervonenkis, 1963) [21] αποτελούν έναν διαδεδομένο αλγόριθμο επιβλεπόμενης μηχανικής μάθησης που χρησιμοποιείται τόσο σε προβλήματα ταξινόμησης, όσο και σε προβλήματα παλινδρόμησης. Σκοπός μιας τέτοιας μηχανής εκμάθησης είναι να αντιστοιχίζει μια τιμή $\boldsymbol{y_i}$ μιας, άγνωστης σε εμάς, συνάρτησης με ένα δοσμένο σημείο $\boldsymbol{x_i}$. Κατά τη διαδικασία της εκπαίδευσης η μηχανή παίρνει ένα σύνολο σημείων εκπαίδευσης $\boldsymbol{x_i} \in \boldsymbol{\Re^n}$ και τις αντίστοιχες τιμές τους $\boldsymbol{y_i} \in \boldsymbol{\Re}$ και εκπαιδεύεται στο να μάθει τη σχέση που συνδέει τα δύο αυτά σύνολα, ώστε να μπορεί στη συνέχεια, δοθέντος ενός σημείου διαφορετικού από αυτά του συνόλου εκπαίδευσης να δώσει την αντίστοιχη τιμή.



Εικόνα 10

Στην περίπτωση των SVM τα δεδομένα εκπαίδευσης χωρίζονται σε δύο υποσύνολα καθένα από τα οποία αποτελεί μια κατηγορία (κλάση), με τις αντίστοιχες τιμές τους $\boldsymbol{y_i}$ να είναι +1 ή -1 ανάλογα την κατηγορία. Έτσι, για να αποφασιστεί η κατηγορία στην οποία ανήκει ένα νέο Σημείο Α (όπως φαίνεται στην Εικ.10) που θα εισαχθεί στη μηχανή πρέπει να βρεθεί μια ευθεία (στην απλή περίπτωση αναφερόμαστε σε δυσδιάστατο χώρο) που να διαχωρίζει τις δύο κατηγορίες και από τη σχετική θέση του σημείου από αυτή τη διαχωριστική ευθεία θα συμπεράνουμε και την κατηγορία που ανήκει. Άρα, το πρόβλημα κατηγοριοποίησης περιορίζεται στην επιλογή αυτής της ευθείας διαχωρισμού των δεδομένων. Όμως, όπως φαίνεται και στην Εικ.11, άπειρες είναι οι υποψήφιες ευθείες διαχωρισμού και όπως είναι φυσικό δεν κατηγοριοποιούν με τον ίδιο τρόπο τα νέα σημεία που εξετάζουμε. Φαίνεται λοιπόν, πως στόχος του αλγορίθμου είναι να βρει τη βέλτιστη ευθεία διαχωρισμού των κατηγοριών η οποία θα ελαχιστοποιεί το σφάλμα κατάταξης και θα κατηγοριοποιεί σωστά όσο το δυνατό περισσότερα σημεία. Μια τέτοια ευθεία φαίνεται λογικό ότι θα απέχει όσο το δυνατό περισσότερο και από τις δύο κλάσεις εξίσου.



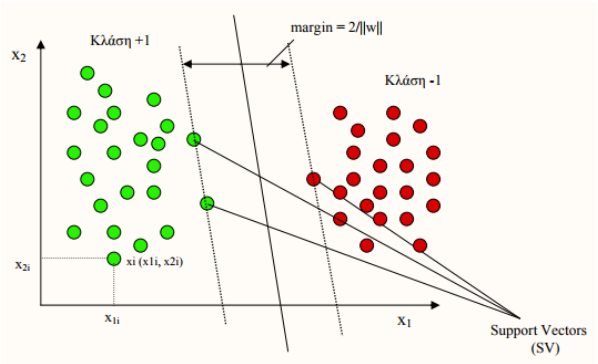
Εικόνα 11

Ζητούμενο λοιπόν των SVM είναι να βρουν αυτή την ευθεία (ή υπερεπίπεδο αν μιλάμε για μεγαλύτερο από 2Δ χώρο) που διαχωρίζει βέλτιστα τις κλάσεις. Με βάση τα χαρακτηριστικά των κλάσεων, τα προβλήματα που αντιμετωπίζουν οι SVM χωρίζονται σε:

1) Πλήρως γραμμικά διαχωρίσιμες κλάσεις
2) Μη απόλυτα γραμμικά διαχωρίσιμες κλάσεις
3) Μη γραμμικά διαχωρίσιμες κλάσεις

Παρακάτω, αναλύουμε τη διαδικασία κατηγοριοποίησης σε κάθε περίπτωση.

4.2.2.1 Πλήρως γραμμικά διαχωρίσιμες κλάσεις



Εικόνα 12

Σε αυτό το πρόβλημα, όλα δεδομένα της μιας κλάσης διαχωρίζονται από αυτά της άλλης με κατάλληλο υπερεπίπεδο. Η εξίσωση κάθε τέτοιου υπερεπιπέδου είναι της μορφής: $\boldsymbol{w^T x + b = 0}$, με $\boldsymbol{w}$ να είναι ένα διάνυσμα βαρών, $\boldsymbol{x}$ ένα διάνυσμα δεδομένων εισόδου και $\boldsymbol{b}$ μια τιμή πόλωσης. Αν συμβολίσουμε με $\boldsymbol{d_+}$ την απόσταση του υπερεπιπέδου διαχωρισμού από το πρώτο στοιχείο της κλάσης +1 και $\boldsymbol{d_-}$ την αντίστοιχη απόσταση από την κλάση -1, τότε το άθροισμα $(\boldsymbol{d_+ + d_-})$ ονομάζεται διάκενο (margin) και δείχνει το περιθώριο διαχωρισμού του υπερεπιπέδου. Στόχος του αλγορίθμου είναι να μεγιστοποιήσει την ποσότητα αυτή, βρίσκοντας το βέλτιστο υπερεπίπεδο διαχωρισμού.

Έστω τώρα, κάποιο $\boldsymbol{b}$ για το οποίο το διάκενο είναι ισομοιρασμένο στις δύο κλάσεις. Χωρίς βλάβη της γενικότητας, μπορούμε να εκφράσουμε τις εξισώσεις των παράλληλων υπερεπιπέδων εκατέρωθεν της διαχωριστικής που κείτονται στα πρώτα στοιχεία κάθε κλάσης ως εξής: $\boldsymbol{w^T x + b = +1}$ για την κλάση +1 και $\boldsymbol{w^T x + b = -1}$ για την κλάση -1 και έτσι τα σημεία της κλάσης +1 θα ικανοποιούν την $\boldsymbol{w^T x + b \geq +1}$ και την -1 την $\boldsymbol{w^T x + b \leq -1}$. Τα σημεία που καθορίζουν τις ευθείες αυτές ονομάζονται Διανύσματα Υποστήριξης (Support Vectors). Αποδεικνύεται ότι το διάκενο μεταξύ των κλάσεων στην περίπτωση αυτή ισούται με $\boldsymbol{(d_+ + d_-) = \frac{2}{\|w\|}}$.

Άρα, το ζητούμενο είναι η μεγιστοποίηση της ποσότητας αυτής που ισοδυναμεί με το: $\boldsymbol{\min_w J(w) = \frac{1}{2} w^T w}$ υπό τον περιορισμό $\boldsymbol{y_i(w^T x_i + b) \geq 1, \quad i = 1, 2, \dots, N}$. Το πρόβλημα αυτό αποτελεί χαρακτηριστικό παράδειγμα προβλήματος τετραγωνικού προγραμματισμού (Quadratic Programming Problem) και για την επίλυσή του

χρησιμοποιούμε το μετασχηματισμό Lagrange του προβλήματος, ενώ για να είναι βέλτιστη η λύση απαιτούμε να ικανοποιούνται οι συνθήκες Karush-Kuhn-Tucker (KKT). Από αυτές προκύπτει ότι για το ζητούμενο βέλτιστο $\boldsymbol{w}$ συμβάλλουν μόνο οι πολλαπλασιαστές Lagrange που αντιστοιχούν στα διανύσματα υποστήριξης και άρα το $\boldsymbol{w}$ αποτελεί έναν γραμμικό συνδυασμό τους. Έτσι καταλήγουμε στο δυϊκό του προβλήματος προς επίλυση, που παίρνει τη μορφή (1):

$$\max_{\boldsymbol{a}} \boldsymbol{L_d} = -\frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{a^T H a} + \boldsymbol{f^T a}$$

$$\boldsymbol{y^T a} = \mathbf{0}$$

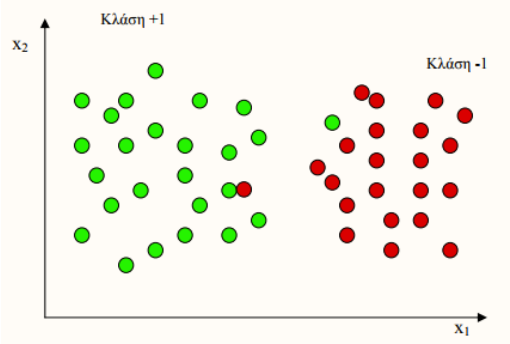
$$\boldsymbol{a} \geq \mathbf{0}$$

Όπου $\boldsymbol{a}$ το διάνυσμα των πολλαπλασιαστών Lagrange, $\boldsymbol{H}$ ο Εσσιανός πίνακας με $\boldsymbol{H_{ij}} = \boldsymbol{y_i y_j x_i^T x_j}$ και $\boldsymbol{f} = [\mathbf{1}\mathbf{1} \ldots \mathbf{1}]^{\boldsymbol{T}}$. Από τη λύση του προβλήματος το βέλτιστο υπερεπίπεδο είναι της μορφής:

$$\boldsymbol{w^*} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{a_i y_i x_i}$$

4.2.2.2 Μη απόλυτα γραμμικά διαχωρίσιμες κλάσεις

Στην περίπτωση αυτή, όπως βλέπουμε και στην Εικ.13 μπορεί να υπάρχουν περιοχές αλληλοεπικάλυψης στοιχείων διαφορετικών κλάσεων με αποτέλεσμα να μην μπορούν να διαχωριστούν απόλυτα από μια ευθεία. Μια πιθανή αντιμετώπιση του προβλήματος θα ήταν να χαράσσαμε μια γραμμή και όχι ευθεία για το διαχωρισμό των κλάσεων, ώστε να όλα τα στοιχεία να αντιστοιχίζονται σωστά. Κάτι τέτοιο όμως, πέραν της πολυπλοκότητας του, θα μείωνε αισθητά και το περιθώριο διαχωρισμού του μοντέλου και άρα τη δυνατότητα γενίκευσής του. Αντί αυτού, επιλέγεται η λύση της εισαγωγής μεταβλητών «χαλάρωσης» (slack variables) $\boldsymbol{\xi_{-1}, \xi_{+1}}$ στις ανισοτικές εξισώσεις που ορίζουν τις δύο κλάσεις, οι οποίες παίρνουν τη μορφή: $\boldsymbol{w^T x + b \geq +1 - \xi_{+1}}$ και $\boldsymbol{w^T x + b = -1 + \xi_{-1}}$. Έτσι, η προς ελαχιστοποίηση ποσότητα είναι η $\boldsymbol{\min_w J(w) = \frac{1}{2} w^T w + C(\sum_i \xi_i)}$ με το $\boldsymbol{\sum_i \xi_i}$ να αποτελεί ένα άνω φράγμα του αριθμού λάθος ταξινομήσεων και το $\boldsymbol{C}$ μια παράμετρος «ποινής» σε περίπτωση λάθους που ορίζεται από το χρήστη.



Εικόνα 13

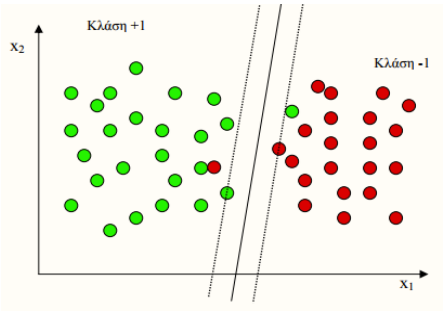
Η τιμή του $\boldsymbol{C}$ δεν μπορεί να είναι γνωστή στο χρήστη από πριν και η επιλογή της γίνεται είτε με εξαντλητική αναζήτηση είτε με βάση την εμπειρία. Συνήθως, μεγάλες τιμές του $\boldsymbol{C}$ αφορούν αξιόπιστα δεδομένα εκπαίδευσης όπου υπάρχει σχετικά μικρή πιθανότητα λάθος ταξινόμησης, ενώ μικρότερες τιμές επιλέγονται για θορυβώδη δεδομένα.

Το πρόβλημα προς επίλυση τελικά έχει αρκετές ομοιότητες με την προηγούμενη περίπτωση (1) και παίρνει τη μορφή (2):

$$\max_{\boldsymbol{a}} \boldsymbol{L_d} = -\frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{a^T H a} + \boldsymbol{f^T a}$$

$$\boldsymbol{y^T a} = \mathbf{0}$$

$$\mathbf{0} \leq \boldsymbol{a_i} \leq \boldsymbol{C}$$

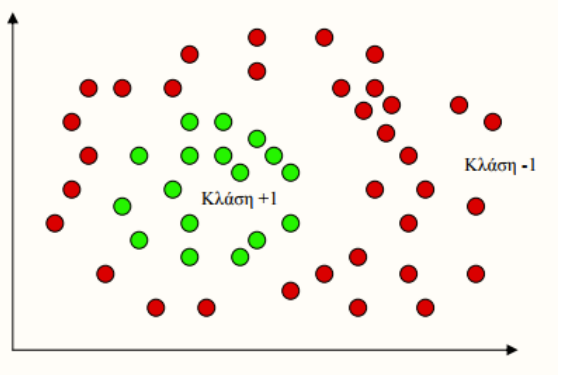


Εικόνα 14

Βλέπουμε ότι σε σχέση με τη μορφή (1) η διαφορά έγκειται στην ύπαρξη ενός άνω ορίου για τους πολλαπλασιαστές Lagrange που είναι η παράμετρος ποινής που επιλέγεται από τον χρήστη. Η ευθεία διαχωρισμού, όπως φαίνεται καθαρά στην Εικ.14 είναι παρόμοια με αυτή των πλήρως διαχωρίσιμων κλάσεων, καθώς με τους συντελεστές χαλάρωσης το μοντέλο μαθαίνει να «αγνοεί» τα λίγα στοιχεία που ταξινομούνται λάθος.

4.2.2.3 Μη γραμμικά διαχωρίσιμες κλάσεις

Αποτελούν την πιο συχνή περίπτωση των πραγματικών προβλημάτων που αντιμετωπίζει ο αλγόριθμος και ένα παράδειγμα τέτοιας περίπτωσης είναι αυτό στην Εικ.15. Βλέπουμε εδώ ότι δεν μπορεί να υπάρξει γραμμή που να διαχωρίζει τις δύο αυτές κλάσεις χωρίς να έχει μη αποδεκτό σφάλμα ταξινόμησης. Φαίνεται λοιπόν επιβεβλημένη η χρήση μιας καμπύλης γραμμής για το διαχωρισμό τους, όπως όμως είπαμε και προηγουμένως κάτι τέτοιο θα αύξανε την πολυπλοκότητα. Η λύση που προκρίνεται εδώ βασίζεται στην παρατήρηση ότι στη σχέση της μεθόδου ((1) ή (2)) τα δεδομένα εκπαίδευσης παρουσιάζονται μόνο με τη μορφή εσωτερικού γινομένου και έτσι αν χρησιμοποιήσουμε ένα μετασχηματισμό $\boldsymbol{x} \rightarrow \boldsymbol{\Phi(x)}$ μεγαλύτερων διαστάσεων θα εμφανίζεται και πάλι μόνο το εσωτερικό γινόμενο. Την ανάγκη του μετασχηματισμού αυτού μας την υπαγορεύει το Θεώρημα Cover που αναφέρει πως «κάθε πολυδιάστατος χώρος με μη γραμμικά διαχωρίσιμα πρότυπα, μπορεί να μετασχηματιστεί σε νέο χώρο που τα πρότυπα πιθανότατα είναι γραμμικά διαχωρίσιμα, αρκεί ο μετασχηματισμός να είναι μη γραμμικός και ο νέος χώρος να έχει την απαραίτητη διάσταση».



Εικόνα 15

Έτσι, αρκεί να βρούμε μια συνάρτηση στο χώρο μεγαλύτερων διαστάσεων της μορφής $\boldsymbol{K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j)}$ που θα εκφράζει το εσωτερικό γινόμενο των δεδομένων εκπαίδευσης. Μια τέτοια συνάρτηση, σύμφωνα με το Θεώρημα Mercer οφείλει να ικανοποιεί τη συνθήκη:

Για κάθε συνάρτηση $\boldsymbol{g(x)}$ τέτοια ώστε $\boldsymbol{\int g(x)^2 dx \in \Re}$ ισχύει $\boldsymbol{\int K(x_i, x_j)\, g(x_i) g(x_j) dx_i dx_j \geq 0}$. Οι συναρτήσεις αυτές ονομάζονται συναρτήσεις πυρήνα (kernel functions) και μερικές από τις πρώτες και πιο διαδεδομένες που χρησιμοποιήθηκαν είναι οι:

- **Γραμμικός πυρήνας**: $\boldsymbol{K(x_i, x_j) = x_i x_j}$ (η αρχική περίπτωση χωρίς μετασχηματισμό)
- **Πολυωνυμικός πυρήνας**: $\boldsymbol{K(x_i, x_j) = (k(x_i, x_j) + n)^p}$
- **RBF (Γκαουσιανός) πυρήνας**: $\boldsymbol{K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2 / 2\sigma^2}}$
- **Σιγμοειδής πυρήνας**: $\boldsymbol{K(x_i, x_j) = \tanh(k(x_i, x_j) - \delta)}$

Οι διάφορες παράμετροι που εμπεριέχουν οι πυρήνες πρέπει να επιλεγούν κατάλληλα από το χρήστη είτε από την εμπειρία είτε με εξαντλητική αναζήτηση, όπως αναφέραμε και για την παράμετρο $\boldsymbol{C}$ προηγουμένως. Η ακριβής μέθοδος προσέγγισης τέτοιων παραμέτρων σε μοντέλα μηχανικής μάθησης θα αναλυθεί στη συνέχεια.

Οι συναρτήσεις που περιγράφουν το μοντέλο για μη γραμμικά διαχωρίσιμες κλάσεις αποτελούν ουσιαστικά γενίκευση των αντίστοιχων των προηγούμενων περιπτώσεων (με επιλογή Γραμμικού πυρήνα). Η συνάρτηση απόφασης για την κατηγοριοποίηση ενός στοιχείου στη γενική περίπτωση είναι η:

$$\boldsymbol{f = sign(\sum_{i=1}^{l} y_i a_i K(x, x_i) + b^*)}$$

με το $\boldsymbol{b^* = 1 - w^{*T} \Phi(x) = 1 - \sum_{i=1}^{l} a_i y_i K(x, x_i)}$ να είναι η τιμή της πόλωσης που υπολογίζεται και αυτή με βάση τη συνάρτηση πυρήνα.

4.2.2.4 Πρόβλημα πολλών κλάσεων (multiclassification)

Παραπάνω αναλύσαμε τη λειτουργία των Διανυσμάτων Υποστήριξης σε διαφορετικές περιπτώσεις προβλημάτων και πάντα χρησιμοποιούσαμε παραδείγματα κατηγοριοποίησης δύο κλάσεων. Αυτό συμβαίνει επειδή η Μηχανή Διανυσμάτων Υποστήριξης είναι από τη φύση του δυαδικός ταξινομητής, καθώς αναζητεί ένα υπερεπίπεδο για να χωρίσει δύο ομάδες στοιχείων. Στην πλειοψηφία των προβλημάτων που αντιμετωπίζουμε όμως, οι κατηγορίες των στοιχείων που πρέπει να διαχωρίσουμε είναι πολύ περισσότερες. Υπάρχουν διάφορες μέθοδοι για την επίλυση προβλημάτων

πολλών κλάσεων με Μηχανές Διανυσμάτων Υποστήριξης και παρακάτω περιγράφουμε δύο από τις πιο διαδεδωμένες:

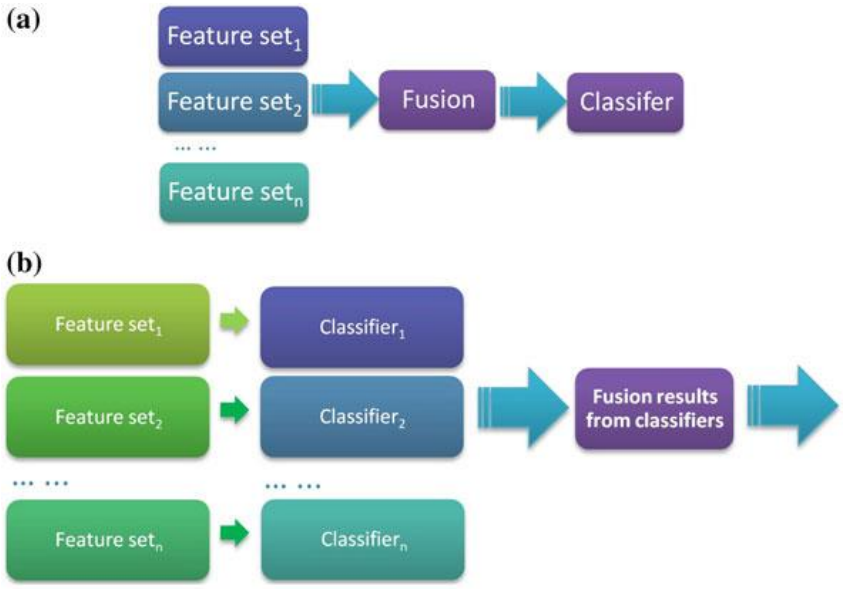
- **One versus All (OvA)**: Σε αυτή τη μέθοδο εκπαιδεύουμε ένα μοντέλο SVM για κάθε κλάση του προβλήματος, δηλαδή για $\boldsymbol{k}$ κλάσεις, $\boldsymbol{k}$ μοντέλα. Το μοντέλο i για παράδειγμα εκπαιδεύεται αντιστοιχώντας την ετικέτα +1 σε στοιχεία της κλάσης i και -1 στα στοιχεία όλων των υπόλοιπων κλάσεων. Έτσι τελικά δημιουργούνται $\boldsymbol{k}$ διαχωριστικά υπερεπίπεδα και $\boldsymbol{k}$ συναρτήσεις απόφασης και με συνδυασμό των αποτελεσμάτων τους καταλήγουμε στα όρια κάθε κλάσης. Αφού λοιπόν εκπαιδευτούν τα μοντέλα, όταν δεχθούμε ένα στοιχείο εισόδου προς εξέταση υπολογίζουμε την τιμή των $\boldsymbol{k}$ συναρτήσεων απόφασης (όχι μόνο το πρόσημο της όπως δείξαμε παραπάνω) όλων των μοντέλων και κατηγοριοποιούμε το στοιχείο στην αντίστοιχη κλάση της συνάρτηση με τη μεγαλύτερη τιμή συνάρτησης απόφασης (winner-takes-all).
- **One versus One (OvO)**: Σε αυτή τη μέθοδο εκπαιδεύουμε μοντέλα για κάθε πιθανό ζεύγος κλάσεων και έτσι για $\boldsymbol{k}$ κλάσεις έχουμε $\boldsymbol{k!/[2(k-2)!]}$ μοντέλα και άρα και τόσα διαχωριστικά υπερεπίπεδα και συναρτήσεις απόφασης. Εκπαιδεύουμε κάθε μοντέλο με την κλασσική μέθοδο που αναλύθηκε παραπάνω στην περίπτωση δύο κλάσεων, για τις δύο κλάσεις του ζεύγους που αντιστοιχεί στο μοντέλο. Όταν εξετάζουμε την κατηγοριοποίηση ενός στοιχείου εισόδου συνήθως επιλέγουμε την τελική κατηγοριοποίηση με τη μέθοδο των max-wins. Υπολογίζουμε δηλαδή τη συνάρτηση απόφασης για το στοιχείο αυτό σε όλα τα μοντέλα και το κατηγοριοποιούμε στην κλάση που κατατάχθηκε τις περισσότερες φορές από τα μοντέλα.

4.3 Τεχνικές Συγχώνευσης Χαρακτηριστικών

Στην παρούσα εργασία, διερευνούμε πέραν από την αποδοτικότητα διαφορετικών μεμονωμένων χαρακτηριστικών για την αναπαράσταση των βίντεο εκπαίδευσης και την τροφοδοσία μετά του ταξινομητή και τις δυνατότητες της συγχώνευσης αυτών των επί μέρους χαρακτηριστικών με σκοπό να αξιοποιήσουμε τη συμπληρωματικότητα μεταξύ τους και να μειώσουμε τα λάθη ταξινόμησης. Αυτή η μεθοδολογία χρησιμοποιείται συχνά σε εργασίες που ανιχνεύουν πιο «υψηλού επιπέδου» νοήματα [26], [47], [48], [49] και στοχεύουν στο να τα μάθουν τα μοντέλα (π.χ. ανίχνευση γεγονότων σε βίντεο) κάτι που για να επιτευχθεί απαιτεί την εξαγωγή πολυεπίπεδης και σε μεγάλη ποσότητα πληροφορίας από τα δεδομένα εκπαίδευσης ώστε να είναι ευδιάκριτες οι διαφορές μεταξύ των δειγμάτων και να καλυφθεί το «νοηματικό κενό» (semantic gap) ανάμεσα στην ωμή αναπαράσταση των βίντεο (δηλαδή ροών από bit) και τις έννοιες και τις σχέσεις μεταξύ τους που έχουν νόημα για τον άνθρωπο. Η συγχώνευση χαρακτηριστικών (feature fusion) υλοποιείται στις περισσότερες περιπτώσεις με τεχνικές συνολικής μάθησης (ensemble learning) που λαμβάνουν τα αποτελέσματα των επί μέρους

ταξινομητών του συστήματος και με βάση κάποιο κριτήριο εξάγουν την τελική τιμή πρόβλεψης για το προς εξέταση δείγμα.

Οι τεχνικές που θα χρησιμοποιήσουμε μπορούν να χωριστούν σε δύο κατηγορίες, αυτές της πρώιμης συγχώνευσης (early fusion) και της όψιμης συγχώνευσης (late fusion).



16 Σχεδιάγραμμα (a) πρώιμης και (b) όψιμης συγχώνευσης

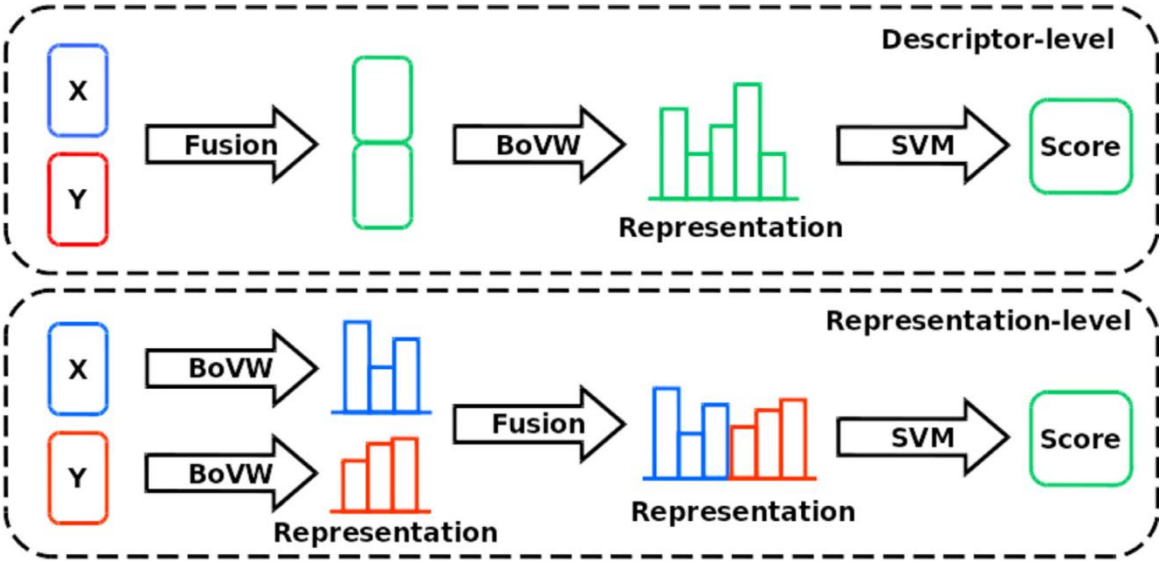
4.2.1 Early Fusion

Σε αυτή την κατηγορία τεχνικών, η συγχώνευση των χαρακτηριστικών συμβαίνει στο στάδιο της αναπαράστασης των δεδομένων εισόδου. Στη γενική περίπτωση, αντί να έχουμε πολλά διαφορετικά διανύσματα για την αναπαράσταση των βίντεο, ανάλογα να εξαγμένα χαρακτηριστικά, τα συνενώνουμε όλα σε ένα διάνυσμα (με διαστάσεις ίσες με το άθροισμα των επί μέρους) ώστε αυτή να είναι η συνολική αναπαράσταση του βίντεο που θα τροφοδοτήσει τον ταξινομητή του συστήματος. Μια επιλογή που υπάρχει σε αυτό το στάδιο συγχώνευσης είναι το αν η συνένωση των διανυσμάτων θα γίνει σε επίπεδο περιγραφέων (descriptor-level) ή σε επίπεδο αναπαράστασης (representation-level).

- **Descriptor-level**: Η πρώτη περίπτωση μπορεί να υλοποιηθεί μόνο για χαρακτηριστικά τα οποία εξάγονται από τις ίδιες γειτονιές του βίντεο ώστε να μπορούν να συνενωθούν τα διανύσματα που περιγράφουν σημεία ενδιαφέροντος από το ίδιο «κυβοειδές» (cuboid) του βίντεο. Έτσι, για παράδειγμα κάτι τέτοιο δεν μπορεί να υλοποιηθεί για χαρακτηριστικά εικόνας και ήχου. Ένα χαρακτηριστικό παράδειγμα τέτοιας συγχώνευσης αποτελούν εξ ορισμού τα STIP χαρακτηριστικά που παρουσιάσαμε παραπάνω, καθώς συνδυάζουν δύο μεθόδους περιγραφής των σημείων ενδιαφέροντος (Ιστογράμματα Προσανατολισμένων Κλίσεων και Οπτικής Ροής) σε ένα ενιαίο διάνυσμα για να περιγράψουν περιοχές

του βίντεο. Είναι προφανές, ότι αυτή η διαδικασία συγχώνευσης εφαρμόζεται πριν τη μεθοδολογία του σάκου λέξεων και την κωδικοποίηση τελικά του βίντεο σε ένα διάνυσμα.

- **Representation-level**: Στην περίπτωση του επιπέδου αναπαράστασης δεν έχουμε περιορισμό στο είδος των χαρακτηριστικών που στοχεύουμε να συγχωνεύσουμε. Αφού έχουμε εφαρμόσει το σάκο λέξεων και έχουμε κωδικοποιήσει το σύνολο σημείων ενδιαφέροντος του βίντεο σε ένα διάνυσμα ανά είδος χαρακτηριστικών μπορούμε να συνενώσουμε τα διανύσματα αυτά και έτσι να έχουμε μια καθολική αναπαράσταση του βίντεο (όμως με μεγάλη διαστατικότητα) που θα τροφοδοτήσει στη συνέχεια των ταξινομητή για την εκπαίδευση. Αυτή αποτελεί και την πιο απλή, αλλά και αποδοτική, εκδοχή συγχώνευσης χαρακτηριστικών και όπως θα φανεί και στα πειράματα που παρουσιάζουμε στη συνέχεια επιφέρει καλύτερη απόδοση στην ταξινόμηση από ότι κάθε ένα από τα επί μέρους χαρακτηριστικά μόνα τους.



17 Σχεδιάγραμμα επιπέδου περιγραφέων και αναπαράστασης

4.2.1 Late Fusion

Στην όψιμη συγχώνευση εκπαιδεύονται ταξινομητές μεμονωμένα για κάθε είδος χαρακτηριστικών και γίνεται προσπάθεια να συνδυαστούν τα αποτελέσματά τους για την τελική απόφαση ταξινόμησης ενός δείγματος εισόδου. Αυτή η κατηγορία μεθόδων έχει το πλεονέκτημα ότι δεν χρειάζεται να δημιουργήσει ένα διάνυσμα πολλών διαστάσεων για τα δεδομένα εισόδου με αποτέλεσμα η εκπαίδευση των ταξινομητών να γίνεται γρηγορότερα. Συνήθως ονομάζεται και συγχώνευση σε επίπεδο σκορ (score-level) καθώς όλη η διαδικασία γίνεται χρησιμοποιώντας είτε τις εξόδους των ταξινομητών, είτε τις πιθανότητες που εξάγουν για κάθε κατηγορία για κάθε δείγμα εισόδου, είτε και το σκορ απόδοσης που πετυχαίνουν στα δεδομένα ελέγχου (test data set). Υπάρχουν πολλοί

διαφορετικοί τρόποι να αξιοποιηθούν αυτά τα δεδομένα και η επιλογή του ποιος θα χρησιμοποιηθεί βασίζεται στο πλήθος των ταξινομητών που έχουμε, τις διαφορές μεταξύ των ειδών των χαρακτηριστικών που χρησιμοποιούνται, αλλά και στο είδος του προβλήματος που αντιμετωπίζουμε. Το κατά πόσο οι τεχνικές αυτές θα βελτιώσουν την απόδοση ενός συστήματος εξαρτάται κατά κύριο λόγο από το πόσο συμπληρωματικά είναι τα διαφορετικά χαρακτηριστικά που χρησιμοποιούμε μεταξύ τους. Θα μπορούσαμε να χωρίσουμε τις μεθόδους με τις οποίες πειραματιστήκαμε στην παρούσα εργασία σε δύο κατηγορίες: αυτές που δεν χρειάζονται εκπαίδευση και αυτές που χρειάζονται.

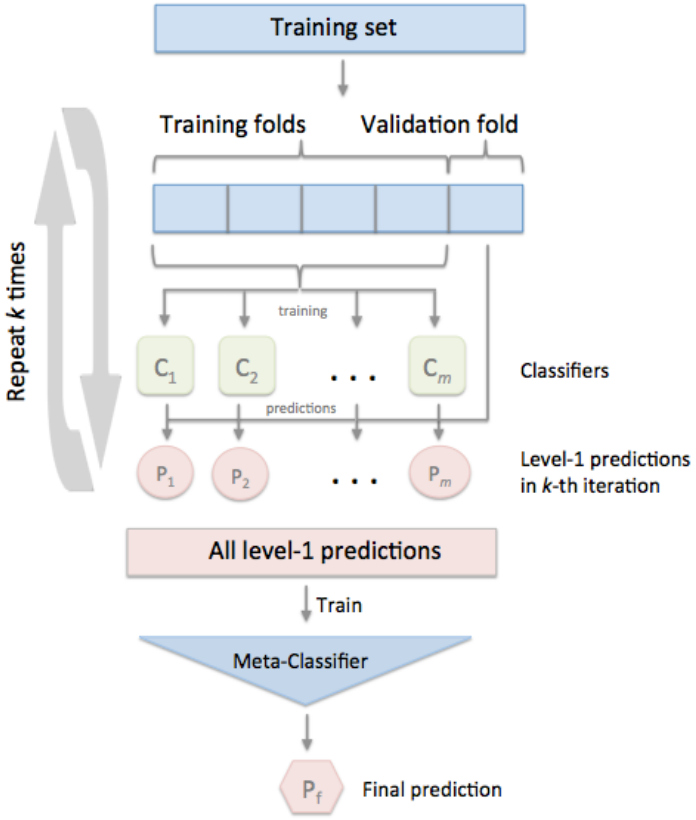
4.2.1.1 Μέθοδοι χωρίς εκπαίδευση

- **Ψηφοφορία Πλειοψηφίας (Majority Voting)**: Αποτελεί μια από τις απλούστερες μεθόδους. Σε αυτή, οι εκπαιδευμένοι ταξινομητές ταξινομούν κάθε δείγμα εισόδου σε μια κλάση και οι ταξινομήσεις αυτές προσμετρούνται ως «ψήφοι» για την ταξινόμηση του δείγματος στην κλάση αυτή. Η κλάση που θα έχει τις περισσότερες ψήφους θα είναι και η πρόβλεψη του συστήματος για το δείγμα εισόδου. Προφανώς, η μέθοδος αυτή δεν μπορεί να εφαρμοστεί για περιττό πλήθος ταξινομητών, καθώς υπάρχει κίνδυνος ισοπαλίας.
- **Ψηφοφορία Πλειοψηφίας με βάρη (Weighted Majority Voting)**: Η διαφορά με την προηγούμενη μέθοδο είναι ότι τώρα αντιστοιχούμε ένα βάρος στην ψήφο κάθε ταξινομητή, ώστε οι ταξινομήσεις τους να έχουν διαφορετική βαρύτητα. Η τελική πρόβλεψη θα είναι η κλάση με το μεγαλύτερο άθροισμα βαρών. Συνήθως, τα βάρη προκύπτουν από κάποια μετρική απόδοσης του ταξινομητή που υπολογίζεται κατά τον έλεγχο του και δείχνει την αξιοπιστία των αποτελεσμάτων του. Εδώ, δεν έχουμε περιορισμό στο πλήθος των ταξινομητών.
- **Κανόνας Αθροίσματος (Sum Rule)**: Στην περίπτωση αυτή, λαμβάνουμε στην έξοδο των ταξινομητών μια πιθανότητα κάθε δεδομένο εισόδου να ανήκει σε κάθε μια από τις κλάσεις. Οι πιθανότητες αυτές για τις Μηχανές Διανυσμάτων Υποστήριξης, προκύπτουν με βάση τις τιμές της συνάρτησης απόφασης κάθε μοντέλου (όπως περιγράψαμε στην περίπτωση OneVersusAll), δηλαδή της απόστασης κάθε δείγματος εισόδου από κάθε αντίστοιχο υπερπεπίπεδο διαχωρισμού και εφαρμόζοντας την κλιμάκωση Plat [50]. Έπειτα, αθροίζουμε τις πιθανότητες που δίνει κάθε ταξινομητής στο δείγμα εισόδου και η τελική πρόβλεψη είναι αυτή που συγκεντρώνει το μεγαλύτερο άθροισμα. Το ίδιο αποτέλεσμα θα προέκυπτε και αν υπολογίζαμε το μέσο όρο αντί για το άθροισμα.
- **Κανόνας Γινομένου (Product Rule)**: Εδώ, υπολογίζουμε το γινόμενο των εξαγμένων πιθανοτήτων για τους ταξινομητές και τελική πρόβλεψη είναι αυτή με το μέγιστο γινόμενο. Νοηματικά, ο κανόνας αυτός εκφράζει ουσιαστικά το γεωμετρικό μέσο των πιθανοτήτων.

- **<u>Κανόνας Διάμεσου (Median Rule)</u>**: Αντίστοιχα με παραπάνω, εδώ αποφασίζουμε την τελική πρόβλεψη με βάση τον διάμεσο των πιθανοτήτων από κάθε ταξινομητή.
- **<u>Κανόνας Μεγίστου (Max Rule)</u>**: Αντίστοιχα με παραπάνω, εδώ υπολογίζουμε τη μέγιστη πιθανότητα μεταξύ των ταξινομητών και από αυτές επιλέγουμε την κλάση με τη μεγαλύτερη από αυτές.

4.2.1.2 Μέθοδοι με εκπαίδευση

Αυτή η κατηγορία μεθόδων προσθέτει μια επιπλέον πολυπλοκότητα στο σύστημα, σε σχέση με την προηγούμενη. Στη γενική περίπτωση οι εκπαιδευμένοι ταξινομητές εξάγουν τις προβλέψεις τους ή τις πιθανότητες τους για τα δεδομένα εισόδου και αυτές αξιοποιούνται σε δεύτερο επίπεδο ως δεδομένα εισόδου σε έναν ταξινομητή ο οποίος εκπαιδεύεται με αυτά. Έτσι, στο τέλος ο ταξινομητής δεύτερου επιπέδου μαθαίνει τη σχέση που συνδέει τις εξόδους των ταξινομητών πρώτου επιπέδου με τις αντίστοιχες πραγματικές ταξινομήσεις των αρχικών δεδομένων εισόδου και έτσι παίρνει την τελική απόφαση για την πρόβλεψη του συστήματος. Στην αγγλική βιβλιογραφία, η μέθοδος αυτή ονομάζεται Stacking Classifier [51].



18 Σχεδιάγραμμα Stacking Classifier

Κατά την εκπαίδευση ενός τέτοιου ταξινομητή, βρισκόμαστε αντιμέτωποι με ένα πρόβλημα. Οι ταξινομητές $1^{ου}$ επιπέδου πρέπει όχι μόνο να εκπαιδευτούν με κάποια από τα δεδομένα εκπαίδευσης, αλλά πρέπει και στη συνέχεια να εξάγουν προβλέψεις που θα χρησιμοποιηθούν για την εκπαίδευση του ταξινομητή $2^{ου}$ επιπέδου. Αν χρησιμοποιήσουμε το ίδιο σύνολο δεδομένων για την εκπαίδευση των ταξινομητών και την εξαγωγή των προβλέψεων τους, τότε διατρέχουμε τον κίνδυνο της υπερεκπαίδευσης (over-fitting) του ταξινομητή $2^{ου}$ επιπέδου, καθώς οι προβλέψεις του $1^{ου}$ επιπέδου θα προέρχονται από δεδομένα που έχουν ήδη «δει» οι ταξινομητές κατά την εκπαίδευση τους. Έτσι, θα έχουμε ένα πλασματικό μεγάλο ποσοστό ορθών προβλέψεων που όμως θα στερούν από τους ταξινομητές τη δυνατότητα γενίκευσης που είναι από τα βασικά ζητούμενα για καλή απόδοση προβλέψεων σε άγνωστα δεδομένα. Για να αποφευχθεί αυτός ο κίνδυνος, χρησιμοποιούμε τη διαδεδομένη τεχνική Κ-στρώσεων διασταυρωμένη επικύρωση (K-fold cross validation). Χωρίζουμε δηλαδή τα δεδομένα εκπαίδευσης σε Κ μη επικαλυπτόμενα υποσύνολα και επαναληπτικά εκπαιδεύουμε τους ταξινομητές $1^{ου}$ επιπέδου με Κ-1 από τα υποσύνολα αυτά σε κάθε επανάληψη. Το εναπομείναν υποσύνολο κάθε φορά, χρησιμοποιείται για την επικύρωση της απόδοσης των ταξινομητών και δημιουργίας προβλέψεων από «άγνωστα» αυτή τη φορά δεδομένα. Από αυτό το σύνολο προβλέψεων, που προκύπτει από την επανάληψη για κάθε υποσύνολο από τα Κ, εκπαιδεύεται τελικά ο ταξινομητής $2^{ου}$ επιπέδου. Στη συνέχεια, μπορούμε να εκπαιδεύσουμε εκ νέου τους ταξινομητές $1^{ου}$ επιπέδου, χρησιμοποιώντας αυτή τη φορά όλα τα δεδομένα εκπαίδευσης.

Σημαντικό παράγοντα στην αποδοτικότητα αυτής της μεθόδου αποτελεί και η επιλογή κατάλληλου ταξινομητή στο $2^{ο}$ επίπεδο. Στη βιβλιογραφία δεν υπάρχουν κάποια σαφή κριτήρια για την επιλογή αυτή, με αποτέλεσμα η τελική απόφαση να προκύπτει από δοκιμές. Στην παρούσα εργασία πειραματιστήκαμε με την επιλογή της Λογιστικής Παλινδρόμησης (Logistic Regression) για την ταξινόμηση στο $2^{ο}$ επίπεδο, καθώς αποτελεί μια απλή περίπτωση ταξινομητή που δεν επηρεάζεται τόσο από τις συσχετίσεις μεταξύ των χαρακτηριστικών.

- **<u>Λογιστική Παλινδρόμηση</u>**: Ο αλγόριθμος Λογιστικής Παλινδρόμησης [21] (ή αλλιώς Μέγιστης Εντροπίας – Maximum Entropy) αποτελεί ένα γραμμικό μοντέλο ταξινόμησης και όχι παλινδρόμησης όπως μπορεί να υπονοεί το όνομά του. Σκοπός του είναι η μοντελοποίηση των πιθανοτήτων εξόδου με χρήση μιας λογιστικής συνάρτησης της μορφής: $\boldsymbol{f(x) = \frac{1}{1+e^{-k(x-x_0)}}}$ που έχει σιγμοειδή μορφή. Ο αλγόριθμος βασίζεται στην αρχή της Μέγιστης Εντροπίας, σύμφωνα με την οποία από όλα τα μοντέλα που ταιριάζουν στα δεδομένα εκπαίδευσης, επιλέγεται αυτό που δεν κάνει καμιά άλλη υπόθεση πέραν των περιορισμών των ίδιων των δεδομένων με αποτέλεσμα η κατανομή να είναι όσο το δυνατόν πιο ομοιόμορφη. Τελικά, το πρόβλημα που επιλύει ο αλγόριθμος είναι ένα πρόβλημα βελτιστοποίησης της μορφής:

$$\boldsymbol{P = \min_{w,c}(1/2 w^T w + C \sum_{i=1}^{n} \log(e^{-y_i(X_i^T w + c)} + 1))}$$

Με $\boldsymbol{w}$ το διάνυσμα συντελεστών της έκφρασης της εξόδου $\boldsymbol{y_i}$ σε μορφή γραμμικού συνδυασμού των χαρακτηριστικών του δείγματος εισόδου $\boldsymbol{x_i}$, όπως συμβαίνει και γενικότερα στα μοντέλα παλινδρόμησης. Το $\boldsymbol{C}$ αποτελεί έναν παράγοντα ομαλοποίησης που η τιμή του επιλέγεται με εξαντλητική αναζήτηση. Το πρόβλημα στο τέλος, επιλύεται με έναν επαναληπτικό αλγόριθμο κλιμάκωσης.

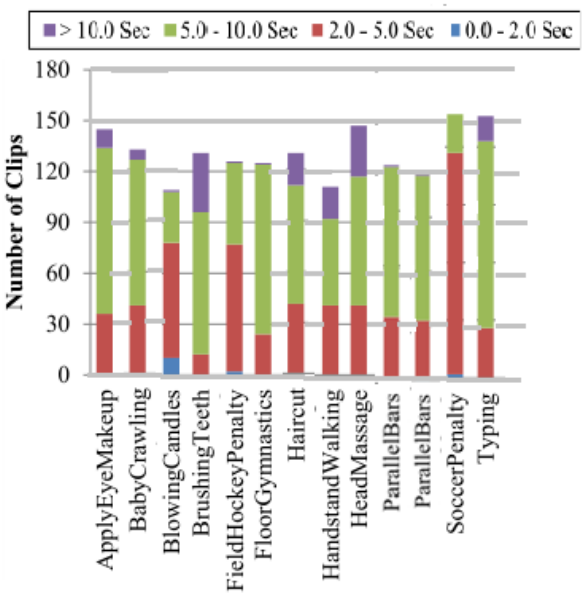
Κεφάλαιο 5 Παρουσίαση συστήματος και πειραματικών αποτελεσμάτων

Στο παρόν κεφάλαιο θα παρουσιάσουμε αναλυτικά την υλοποίηση του συστήματός μας για ανίχνευση ενεργειών σε βίντεο. Θα δείξουμε τα βήματα της ανάπτυξής του και τις σχεδιαστικές αποφάσεις που πήραμε κατά τη διάρκειά της και τα πειραματικά αποτελέσματα που αιτιολογούν κάποιες από αυτές. Θα παρουσιάσουμε συγκριτικούς πίνακες που δείχνουν τις διαφορές απόδοσης του συστήματος, ανάλογα με τις αλλαγές στο σχεδιασμό του.

5.1 Εισαγωγή

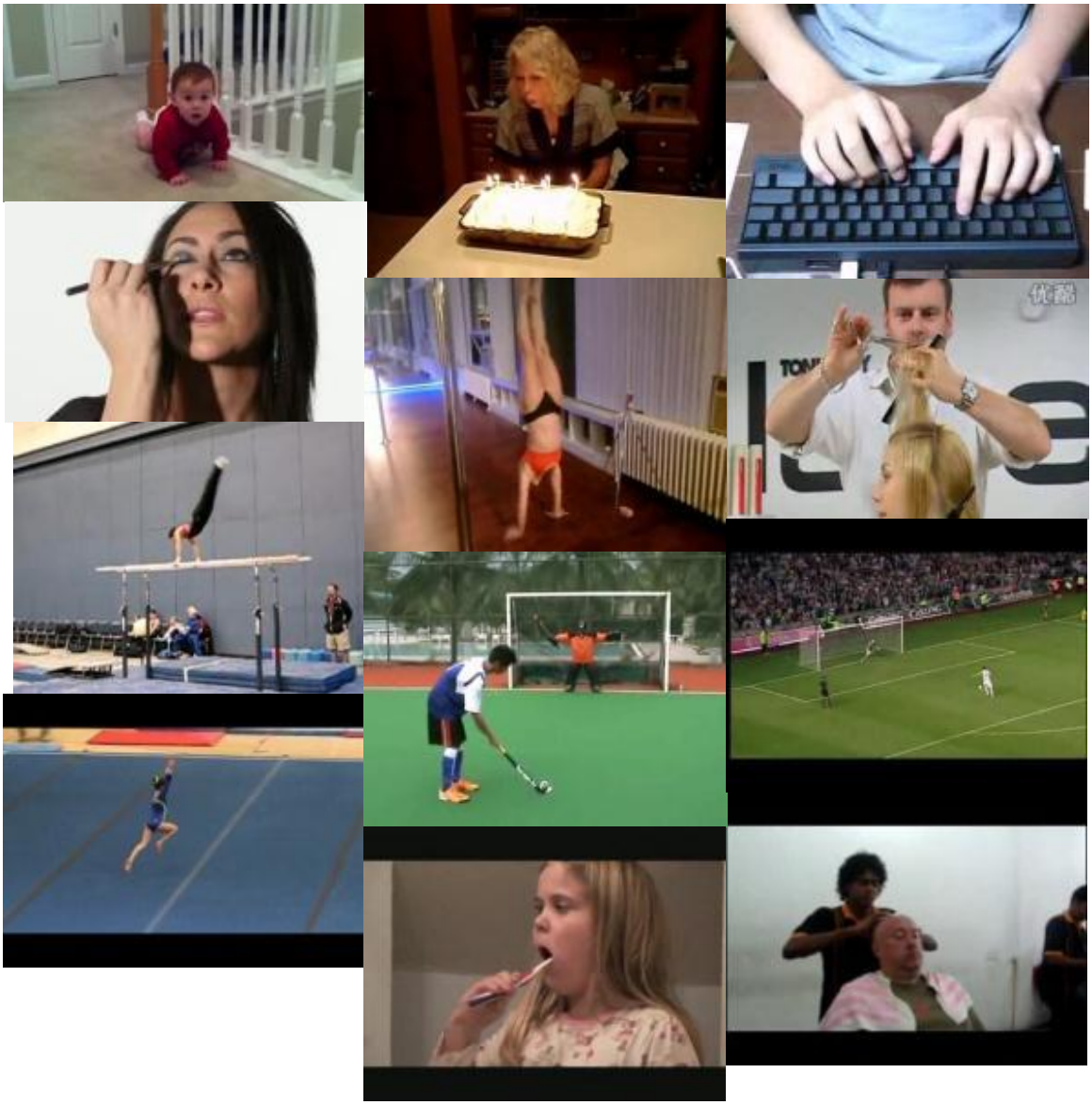
5.1.1 Παρουσίαση συνόλου δεδομένων (UCF101 Dataset)

Τα δεδομένα που χρησιμοποιήσαμε για την εκπαίδευση και τον έλεγχο του συστήματός μας προέρχονται από το διαδεδομένο dataset UCF101 [52] που δημιουργήθηκε από το University of Central Florida. Αποτελεί αυτή τη στιγμή το μεγαλύτερο dataset ανθρωπίνων ενεργειών και περιλαμβάνει 101 κατηγορίες ενεργειών, πάνω από 13.000 κλιπ και 27 ώρες δεδομένων βίντεο, ενώ είναι επέκταση του UCF50 που περιείχε 50 κατηγορίες. Τα βίντεο προέρχονται από ρεαλιστικά βίντεο ανεβασμένα διαδικτυακά από χρήστες και συνήθως περιέχουν κίνηση της κάμερας και ακατάστατο φόντο, κάνοντας το dataset ένα από προκλητικότερα που κυκλοφορούν. Το dataset συνοδεύουν και τα εξαγμένα STIP χαρακτηριστικά των βίντεο όλων των κατηγοριών.



19 Διάγραμμα διάρκειας βίντεο των 12 κατηγοριών που χρησιμοποιήσαμε

Οι 101 κατηγορίες χωρίζονται σε 5 υποκατηγορίες και συγκεκριμένα: Αλληλεπίδραση Ανθρώπου – Αντικειμένου, Κίνηση – Σώματος Μόνο, Αλληλεπίδραση Ανθρώπου – Ανθρώπου, Παίξιμο Μουσικών Οργάνων και Αθλήματα. Κάποιες από τις ενέργειες της ίδιας υποκατηγορίας έχουν αυξημένοι πιθανότητα να έχουν κοινά στοιχεία μεταξύ τους, κάνοντας τες δυσκολότερες στην ταξινόμηση από το σύστημα.



20 Παραδείγματα καρέ των 12 κλάσεων του UCF101 που χρησιμοποιήσαμε

Κάθε κατηγορία έχει 25 βίντεο τα οποία είναι κομμένα σε 4 με 7 κλιπ λίγων δευτερολέπτων το καθένα για τη διευκόλυνση της εκπαίδευσης των ταξινομητών. Κάθε βίντεο έχει 25 fps (frames per second), ενώ η ανάλυσή του είναι 320x240. Μόνο 51 από τις κατηγορίες αυτές έχουν βίντεο με ήχο, γι' αυτό και η επιλογή μας έγινε από αυτές.

Στην παρούσα εργασία, χρησιμοποιήσαμε 12 από αυτές τις κατηγορίες και συγκεκριμένα: Apply Eye Makeup, Baby Crawling, Blowing Candles, Brushing Teeth, Field Hockey Penalty, Floor Gymnastics, Haircut, Handstand Walking, Head Massage, Parallel Bars, Soccer Penalty, Typing. Η επιλογή δεν είναι τυχαία, καθώς έχει σημασία για να εκτιμήσουμε όσο το δυνατό πιο πειστική απόδοση του συστήματός μας όχι μόνο να δούμε τις σωστές ταξινομήσεις σε κάθε κατηγορία ξεχωριστά, αλλά και τι γίνεται μεταξύ παρόμοιων κατηγοριών (όπως π.χ. Haircut-Head Massage) που ο διαχωρισμός θα είναι δυσκολότερος.

5.1.2 Εργαλεία υλοποίησης

Το σύστημά μας υλοποιήθηκε εξ ολοκλήρου στη γλώσσα προγραμματισμού Python3. Συγκεκριμένα χρησιμοποιήσαμε πολλές από τις χρήσιμες βιβλιοθήκες της γλώσσας, όπως: OpenCV για την ανάγνωση των βίντεο και εξαγωγή χαρακτηριστικών, scikit-learn για τη μηχανική μάθηση και την επεξεργασία των δεδομένων, python-speech-features[1] για την εξαγωγή των χαρακτηριστικών ήχου. Επίσης, χρησιμοποιήσαμε και τις πολύ διαδεδομένες στην Python numpy, scipy, wav, pickle και matplotlib. Το σύστημα αναπτύχθηκε σε περιβάλλον PyCharm.

5.1.3 Μετρικές και μέθοδος αξιολόγησης

Τα δεδομένα βίντεο που έχουμε στη διάθεσή μας είναι 1543 από τις 12 κατηγορίες που επιλέξαμε. Η μέθοδος επικύρωσης που επιλέχθηκε σε όλα τα πειράματα είναι αυτή των K-στρώσεων διασταυρωμένη επικύρωση με K=5. Έτσι, για κάθε αξιολόγηση με τις μετρικές απόδοσης, χωρίζουμε το σύνολο δεδομένων σε 5 ίσα κομμάτια (χωρίς επικάλυψη μεταξύ τους και προσέχοντας να μην συμπεριληφθούν κλιπ του ίδιου βίντεο και για την εκπαίδευση και για την επικύρωση, γιατί τότε υπάρχει κίνδυνος υπερεκπαίδευσης) και χρησιμοποιούμε περίπου 1200 βίντεο για την εκπαίδευση του συστήματος και περίπου 300 για την επικύρωση. Στο τέλος της διαδικασίας αυτής, έχουμε 5 τιμές της μετρικής για το σύστημα και υπολογίζουμε το μέσο όρο τους για τελική τιμή. Με τον τρόπο αυτό, διασφαλίζεται ότι τα αποτελέσματά μας θα είναι αντιπροσωπευτικά για το σύνολο των δεδομένων, αφού οι μετρικές προκύπτουν κάθε φορά από δεδομένα που δεν έχει δει ο ταξινομητής κατά τη διάρκεια της εκπαίδευσης.

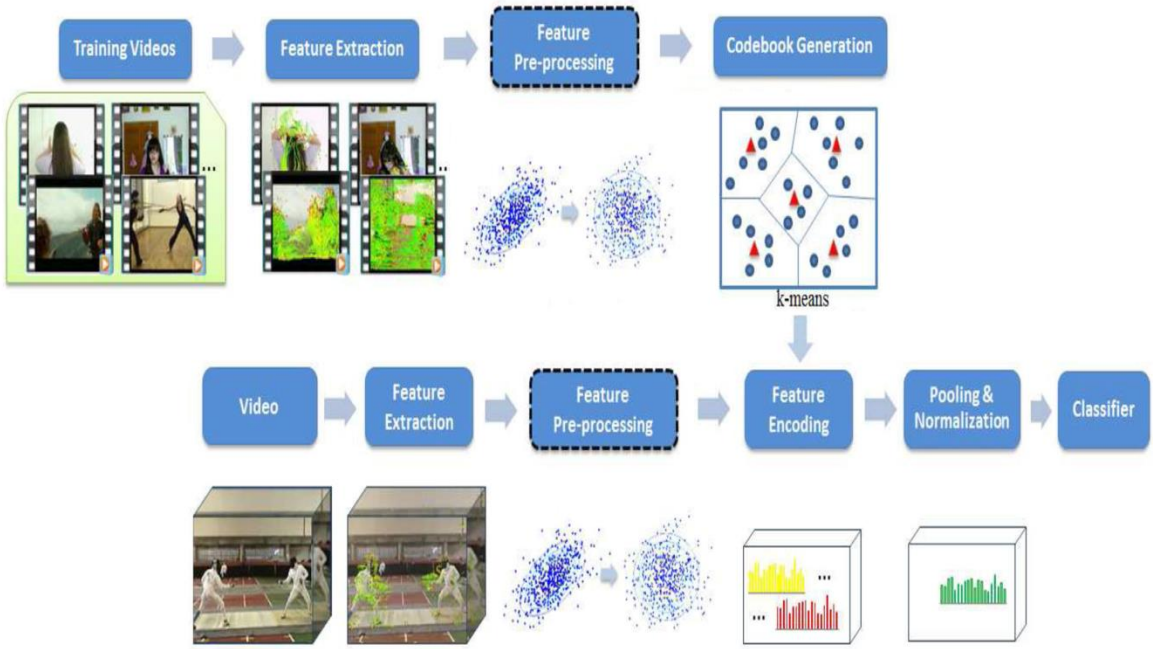
Οι μετρικές που χρησιμοποιούμε για την αξιολόγηση του συστήματος είναι οι εξής:

- **Απόδοση (Accuracy)**: $\boldsymbol{Ac = \frac{Σωστές\ προβλέψεις}{Σύνολο\ προβλέψεων}}$: Είναι ο λόγος των σωστών προβλέψεων του ταξινομητή προς το σύνολο των προβλέψεων του. Μας βοηθά να έχουμε μια καλή πρώτη εικόνα των δυνατοτήτων του συστήματός μας. Στη βιβλιογραφία εργασιών ανίχνευσης ενεργειών σε βίντεο, συνήθως χρησιμοποιείται αυτή η μετρική αξιολόγησης.
- **Ακρίβεια (Precision)**: $\boldsymbol{Pr = \frac{True\ Positives}{True + False\ Positives}}$: Αφορά την ικανότητα του ταξινομητή να μην ταξινομεί ως θετικά τα αρνητικά δείγματα. Μας δείχνει το κατά πόσο όσα βίντεο ταξινομήθηκαν σε μια κατηγορία, ανήκουν πράγματι σε αυτή.
- **Ανάκληση (Recall)**: $\boldsymbol{Re = \frac{True\ Positives}{True\ Positives + False\ Negatives}}$: Αφορά την ικανότητα του ταξινομητή να βρει όλα τα θετικά δείγματα και έτσι μας δείχνει αν μπόρεσε το σύστημα να βρει όλα τα βίντεο της κατηγορίας.

[1] https://github.com/jameslyons/python_speech_features

- **F1-score**: $F1 = \frac{2*Pr*Re}{Pr+Re}$: Ο αρμονικός μέσος ακρίβειας και ανάκλησης.
- **Πίνακας Σύγχυσης (Confusion Matrix)**: Είναι ένας τετραγωνικός πίνακας με διαστάσεις όσες και οι κατηγορίες μας που στη διαγώνιό του έχει το πλήθος των ορθών προβλέψεων ταξινόμησης για κάθε κατηγορία και στις υπόλοιπες θέσεις το πλήθος των λάθος ταξινομημένων βίντεο ανάλογα με το που ταξινομήθηκαν και που θα έπρεπε. Παράδειγμα τέτοιου για πολλές κατηγορίες θα δείξουμε παρακάτω στην παρουσίαση των αποτελεσμάτων.

5.2 Παρουσίαση του Συστήματος



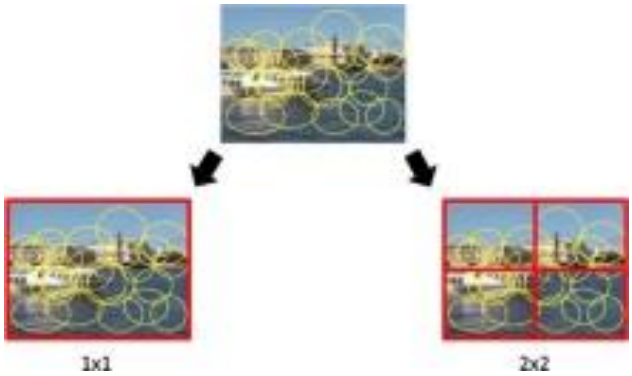
21 Σχεδιάγραμμα συστήματος με "Σάκο λέξεων"

Στο παραπάνω σχεδιάγραμμα παρουσιάζεται συνοπτικά η αρχιτεκτονική του συστήματός μας που υλοποιεί τη μεθοδολογία του Σάκου λέξεων (όπως παρουσιάστηκε στο 3.2), ενώ στα επόμενα σχεδιαγράμματα παρουσιάζονται η πρώιμη και όψιμη συγχώνευση των χαρακτηριστικών, που εφαρμόζονται στη συνέχεια.

Τα βήματα που ακολουθήθηκαν είναι τα εξής, ξεκινώντας με τη διαδικασία του σάκου λέξεων (Εικ.21):

- **Εξαγωγή Χαρακτηριστικών**: Αρχικά, για τα SIFT χαρακτηριστικά, κάνουμε δειγματοληψία ενός καρέ κάθε 2 sec από το οποίο εξάγουμε τα χαρακτηριστικά και παράλληλα το χωρίζουμε στα 4 με 2x2 υποπεριοχές (όπως στην Εικ.22) και εξάγουμε εκ νέου χαρακτηριστικά για κάθε μία από αυτές. Η διαδικασία αυτή

ονομάζεται χωρική πυραμίδα (spatial pyramid [53]) και βοηθάει κατά τη διαδικασία του σάκου λέξεων να διατηρείται η χωρική πληροφορία των σημείων της εικόνας, που χωρίς την επεξεργασία αυτή θα χανόταν. Έτσι συνολικά εξάγουμε 5 διανύσματα των 128 διαστάσεων το καθένα για κάθε καρέ που δειγματοληπτούμε. Στη συνέχεια, δημιουργούμε τα wav αρχεία ήχου των βίντεο για να μπορέσουμε να εξάγουμε τα MFCC χαρακτηριστικά. Για την εξαγωγή τους επιλέγουμε κυλιόμενο παράθυρο δειγματοληψίας 32 msec και βήμα 16 msec, δηλαδή έχουμε 16 msec επικάλυψη, ενώ επιλέγουμε να κρατήσουμε 13 Cepstral συντελεστές που είναι η συνηθισμένη επιλογή. Μαζί με αυτούς, υπολογίζουμε και τις πρώτες και δεύτερες παραγώγους τους (όπως περιγράψαμε στο 3.1.2) και συνολικά εξάγουμε διανύσματα 39 διαστάσεων για κάθε παράθυρο δειγματοληψίας. Τα STIP χαρακτηριστικά υπάρχουν ήδη και συνοδεύουν το dataset. Για κάθε βίντεο λαμβάνουμε από αυτά ένα σύνολο διανυσμάτων 162 διαστάσεων (από τους HOG/HOF περιγραφείς που αναφέρονται στο 3.1.3).



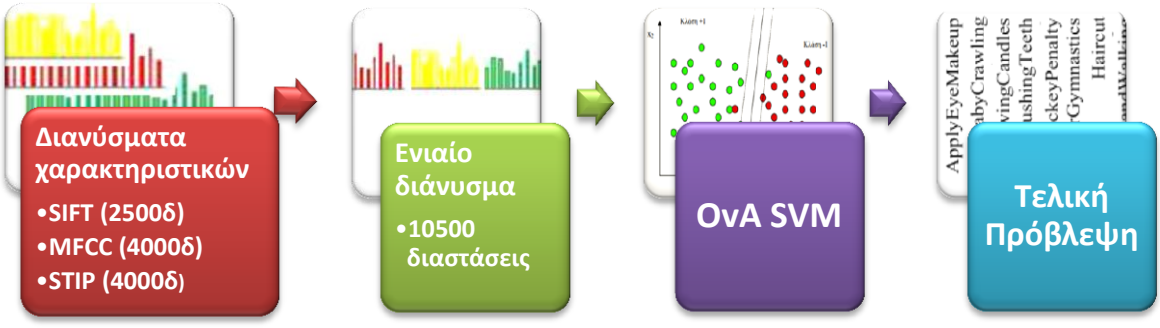
22 **Παράδειγμα χωρικής πυραμίδας 1x1-2x2**

- **Προεπεξεργασία Χαρακτηριστικών**: Εδώ εφαρμόζουμε ξεχωριστά σε κάθε σύνολο δεδομένων που εξάγαμε προηγουμένως την Ανάλυση Κύριων Συνιστωσών (4.1) και στη συνέχεια το Άσπρισμα των δεδομένων. Υπολογίζουμε τις κύριες συνιστώσες για τα χαρακτηριστικά των κλιπ των 15 πρώτων βίντεο κάθε κατηγορίας (που στη συνέχεια θα χρησιμοποιήσουμε και για τη δημιουργία του λεξικού) και έπειτα εφαρμόζουμε το μετασχηματισμό με βάση τις συνιστώσες αυτές στο σύνολο των δεδομένων. Για τα δεδομένα SIFT εφαρμόζουμε τη διαδικασία αυτή ξεχωριστά για κάθε μία από τις υποπεριοχές που χωρίσαμε προηγουμένως. Επιλέγουμε να κρατήσουμε το 99% της διακύμανσης των δεδομένων από τις κύριες συνιστώσες και έτσι καταλήγουμε να μειώνουμε τις διαστάσεις των διανυσμάτων από 128 σε 110 για κάθε περιοχή των SIFT χαρακτηριστικών, από 39 σε 13 για τα MFCC και από 162 σε 97 για τα STIP. Πετύχαμε έτσι μια σημαντική μείωση διαστατικότητας χωρίς να χάσουμε πρακτικά πληροφορία, πράγμα ιδιαίτερα χρήσιμο για τη δημιουργία του λεξικού στη συνέχεια. Στην παρουσίαση των αποτελεσμάτων στη συνέχεια, πειραματιστήκαμε και με την αποφυγή αυτού του βήματος για να αναδειχθεί η σημασία του στη συνολική απόδοση του συστήματος.
- **Δημιουργία λεξικού**: Εφαρμόζουμε τον αλγόριθμο K-μέσων για κάθε ένα από τα μετασχηματισμένα υποσύνολα δεδομένων (5 σύνολα από τα SIFT, 1 MFCC και 1 STIP). Με βάση παρόμοια παραδείγματα εφαρμογών στη βιβλιογραφία επιλέγουμε το K, που αποτελεί το μέγεθος του λεξικού που θα δημιουργήσουμε, να είναι 500 για τα SIFT χαρακτηριστικά, 4000 για τα MFCC και 4000 για τα STIP. Επιλέξαμε μικρότερο αριθμό για τα SIFT, καθώς θα δημιουργηθεί

ξεχωριστό λεξικό για κάθε περιοχή της πυραμίδας και έτσι οι συνολικές διαστάσεις θα φτάσουν στις 2500. Ο αλγόριθμος εφαρμόζεται με εκκίνηση τυχαίων κέντρων και σε 10 επαναλήψεις, ώστε να αποφευχθεί όσο είναι δυνατόν η σύγκλιση σε τοπικό ελάχιστο.

- **Κωδικοποίηση Χαρακτηριστικών**: Αφού έχουμε δημιουργήσει το λεξικό στο προηγούμενο βήμα που αποτελείται από τα σημεία – πρότυπα των συστάδων που δημιουργήθηκαν, κωδικοποιούμε τα διανύσματα των σημείων των συνόλων δεδομένων (για όλα τα βίντεο κάθε κατηγορίας αυτή τη φορά) θέτοντας σε ένα νέο διάνυσμα 1 στη θέση που αντιστοιχεί στην κοντινότερη στο σημείο συστάδα και 0 σε όλες τις υπόλοιπες.
- **Συγκέντρωση και Κανονικοποίηση**: Δημιουργούμε τα τελικά ιστογράμματα αναπαράστασης των βίντεο, αθροίζοντας τα κωδικοποιημένα διανύσματα των σημείων τους που δημιουργήσαμε στο προηγούμενο βήμα. Για τα δεδομένα SIFT, εδώ συνενώνουμε στον οριζόντιο άξονα και τις αναπαραστάσεις των επί μέρους περιοχών της πυραμίδας, αφού υπολογίσουμε τα αθροίσματα ξεχωριστά στην καθεμία, καταλήγοντας σε ένα διάνυσμα 2500 διαστάσεων για κάθε βίντεο. Αντίστοιχα, με βάση το μέγεθος των λεξικών που επιλέξαμε, καταλήγουμε σε διάνυσμα 4000 διαστάσεων για κάθε βίντεο για τα MFCC χαρακτηριστικά, το ίδιο και για τα STIP. Τέλος, κανονικοποιούμε τα διανύσματα διαιρώντας το καθένα με την Ευκλείδεια νόρμα του.

Στο τέλος της διαδικασίας του σάκου λέξεων, συνενώνουμε στον οριζόντιο άξονα τους πίνακες με τα διανύσματα που δημιουργήθηκαν, υλοποιώντας τη συγχώνευση σε επίπεδο αναπαράστασης (όπως περιγράψαμε στο 4.2.1) (Εικ.23). Τα διανύσματα που προκύπτουν έχουν 10500 διαστάσεις.



23 Σχεδιάγραμμα early representation-level fusion

Σε αυτό το σημείο, όλα μας τα δεδομένα είναι έτοιμα να τροφοδοτήσουν τους ταξινομητές. Έχουμε επιλέξει Μηχανές Διανυσμάτων Υποστήριξης με χρήση

συναρτήσεων πυρήνα, καθώς το πρόβλημα που αντιμετωπίζουμε αφορά μη γραμμικά διαχωρίσιμες κλάσεις. Η συνάρτηση πυρήνα που επιλέγουμε για όλα τα είδη χαρακτηριστικών είναι η $\boldsymbol{\chi^2}$ που προτείνεται από τη βιβλιογραφία της Ανίχνευσης Ενεργειών σε βίντεο και έχει τη μορφή:

$$\boldsymbol{K(x_i, x_j) = \exp(-\gamma \sum((x_i - x_j)^2/(x_i + x_j)))}$$

Η μέθοδος που επιλέγεται για την επίλυση του προβλήματος πολλών κλάσεων που αντιμετωπίζουμε με τα διανύσματα υποστήριξης είναι η One-Versus-All, εκπαιδεύοντας ένα SVM μοντέλο για κάθε κατηγορία που έχουμε. Κατά την παρουσίαση των αποτελεσμάτων θα δείξουμε και τα αποτελέσματα πειραμάτων που έγιναν και με άλλες συναρτήσεις πυρήνα (συγκεκριμένα με τη Γκαουσιανή και τη Γραμμική).

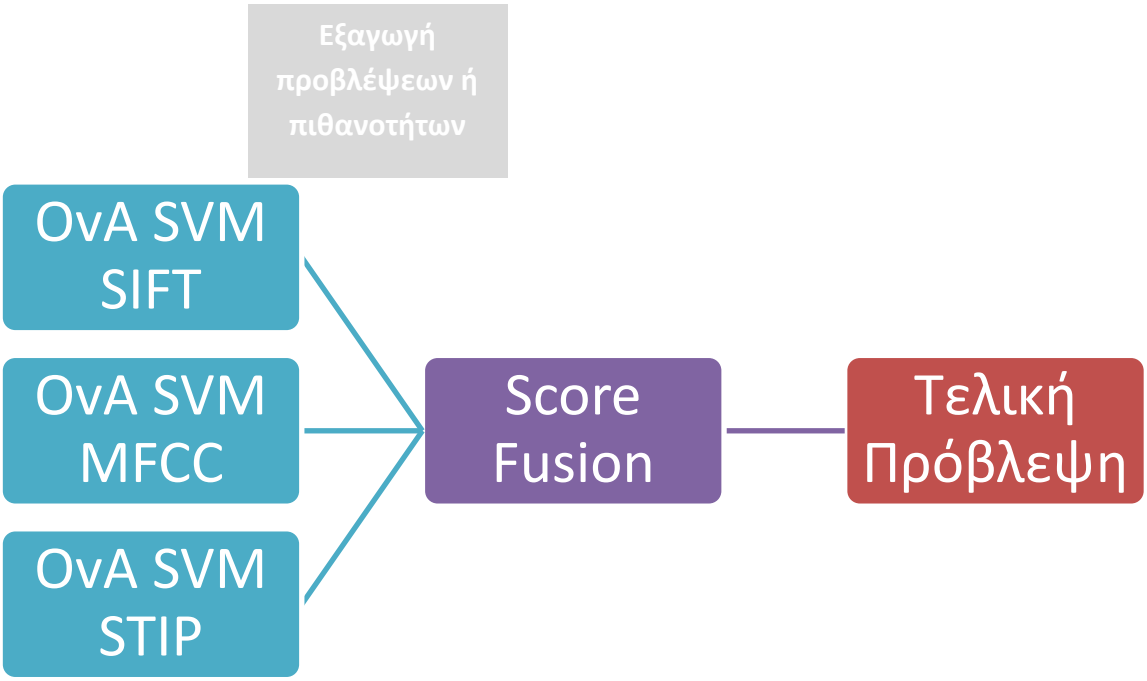
Για τις παραμέτρους των ταξινομητών που πρέπει να οριστούν από το χρήστη (συγκεκριμένα τον παράγοντα C για τα SVM μοντέλα και τον παράγοντα $\boldsymbol{\gamma}$ τόσο για τον πυρήνα $\boldsymbol{\chi^2}$, όσο και για τον Γκαουσιανό) διενεργήσαμε εξαντλητική αναζήτηση δοκιμάζοντας κάθε φορά άλλη τιμή από το εύρος [1, 1000] για τον C και [0.0001, 1] για τον $\boldsymbol{\gamma}$. Τα αποτελέσματα με τις τιμές που επιλέχθηκαν φαίνονται στον παρακάτω πίνακα:

Πίνακας 1 Τιμές παραμέτρων για τα SVM μοντέλα

	χ^2 πυρήνας	Γκαουσιανός πυρήνας	Γραμμικός πυρήνας
SVM (SIFT)	C= 100, γ= 0.001	C= 10, γ= 0.1	C= 10
SVM (MFCC)	C= 100, γ= 0.0001	C= 10, γ= 0.001	C= 1
SVM (STIP)	C= 100, γ= 0.01	C= 10, γ= 0.1	C= 1
SVM (representation fusion)	C= 1000, γ= 0.0001	C= 10, γ= 0.01	C= 1

Με βάση τη μέθοδο K-στρώσεων διασταυρωμένης επικύρωσης για K=5 που αναλύσαμε παραπάνω, για 5 επαναλήψεις εκπαιδεύουμε τα SVM μοντέλα χρησιμοποιώντας κάθε φορά 4 από τα 5 υποσύνολα των δεδομένων για εκπαίδευση (περίπου 12000 κλιπ βίντεο) και χρησιμοποιούμε το $5^{ο}$ υποσύνολο για τον υπολογισμό των μετρικών αξιολόγησης.

Στο τελευταίο στάδιο του συστήματος εφαρμόζουμε την όψιμη συγχώνευση των χαρακτηριστικών (Εικ.24). Αρχικά, για τις μεθόδους χωρίς εκπαίδευση, εξάγουμε από την παραπάνω διαδικασία τις πιθανότητες ταξινόμησης από καθέναν από τους ταξινομητές των επιμέρους χαρακτηριστικών (για την περίπτωση της Ψηφοφορίας Πλειοψηφίας, εξάγουμε τις τελικές προβλέψεις ταξινόμησης) για τα δεδομένα επικύρωσης. Στη συνέχεια, σε κάθε επανάληψη των K-στρώσεων υπολογίζουμε με τις μεθόδους συγχώνευσης χωρίς εκπαίδευση (όπως αναλύθηκαν στο 4.2.1.1) τις τελικές προβλέψεις του κάθε συστήματος και από αυτές τις μετρικές αξιολόγησης για το κάθε σύστημα. Οι τιμές των μετρικών αξιολόγησης που παρουσιάζουμε στα αποτελέσματα, αφορούν των μέσο όρων των μετρικών από τις 5 επαναλήψεις.



24 Σχεδιάγραμμα late score-level fusion

Για την όψιμη συγχώνευση χαρακτηριστικών με εκπαίδευση η διαδικασία είναι λίγο πιο πολύπλοκη. Για κάθε μια από τις επαναλήψεις των K-στρώσεων έχουμε περίπου 1200 βίντεο για την εκπαίδευση των ταξινομητών (K-1 υποσύνολα) και περίπου 300 για επικύρωση. Αυτά τα δεδομένα εκπαίδευση κάθε φορά τα διαχωρίζουμε ξανά σε 5 στρώσεις ($1200/5 = 240$ η κάθε στρώση τώρα) ώστε να εκπαιδεύουμε επαναληπτικά τους ταξινομητές και με τις προβλέψεις τους από τα δεδομένα επικύρωσης να εκπαιδεύουμε τον ταξινομητή Λογιστικής Παλινδρόμησης στο $2^{ο}$ επίπεδο (μετά από τη διαδικασία της εξαντλητικής αναζήτησης ορίζουμε τον παράγοντα C της Λογιστικής Παλινδρόμησης ίσο με 0.01). Μετά από την εκπαίδευση του ταξινομητή $2^{ου}$ επιπέδου, εκπαιδεύουμε εκ νέου τους ταξινομητές $1^{ου}$ επιπέδου, αυτή τη φορά με όλα τα δεδομένα εκπαίδευσης για τη συγκεκριμένη επανάληψη των K-στρώσεων (τα 1200 βίντεο του αρχικού διαχωρισμού) και στη συνέχεια τους τροφοδοτούμε με τα δεδομένα επικύρωσης (300 βίντεο του αρχικού διαχωρισμού) και με τις προβλέψεις τους τροφοδοτούμε και τον ήδη εκπαιδευμένο ταξινομητή του $2^{ου}$ επιπέδου και με τις προβλέψεις του υπολογίζουμε τις μετρικές αξιολόγησης. Έτσι, τα δεδομένα με τα οποία αξιολογούμε το σύστημά μας είναι «άγνωστα» στους ταξινομητές και των δύο επιπέδων. Στην παρουσίαση των αποτελεσμάτων πειραματιστήκαμε τόσο με τις προβλέψεις ταξινόμησης, όσο και με τις πιθανότητες ταξινόμησης για να εκπαιδεύσουμε τον ταξινομητή $2^{ου}$ επιπέδου.

5.3 Παρουσίαση Αποτελεσμάτων

Η παρουσίαση των πειραματικών μας αποτελεσμάτων πραγματοποιείται με τρόπο τέτοιο, ώστε να φανούν οι διαφορές και η βελτίωση που επιτυγχάνεται στην απόδοση του συνολικού συστήματος αξιοποιώντας με διαφορετικούς συνδυασμούς τα εξαγμένα χαρακτηριστικά που έχουμε στη διάθεσή μας για την αναπαράσταση των βίντεο καθώς και τις διαφορετικές μεθόδους που μπορούμε να αξιοποιήσουμε για να τα συγχωνεύσουμε.

5.3.1 Εκπαίδευση με μεμονωμένα χαρακτηριστικά

Στην απλούστερη εκδοχή του συστήματός μας πειραματιζόμαστε εκπαιδεύοντας ένα SVM μοντέλο, με One Versus All μέθοδο για το πρόβλημα πολλών κλάσεων, χρησιμοποιώντας κάθε φορά διαφορετικού είδους χαρακτηριστικά για τα δεδομένα εκπαίδευσης. Για τα STIP δεδομένα δοκιμάσαμε και μια υλοποίηση στην οποία παραλείψαμε το στάδιο της προεπεξεργασίας των δεδομένων, πριν τη δημιουργία του λεξικού, με την Ανάλυση Κύριων Συνιστωσών. Από τα αποτελέσματα φαίνεται η σημαντική βελτίωση στην απόδοση που επιτυγχάνεται με την προσθήκη της προεπεξεργασίας, ακόμα και στα δεδομένα STIP που από τα πειράματα φαίνεται ότι ήταν τα αποδοτικότερα στην αναπαράσταση των βίντεο σε σχέση με τα SIFT και MFCC.

Επίσης, πειραματιστήκαμε με τρεις διαφορετικές συναρτήσεις πυρήνα για την εκπαίδευση των SVM μοντέλων και με βάση τα αποτελέσματα μπορούμε να επιβεβαιώσουμε την καταλληλότητα της $\boldsymbol{\chi^2}$ για το πρόβλημα που αντιμετωπίζουμε, σε σχέση με παραδοσιακές επιλογές, όπως η Γκαουσιανή και η Γραμμική συνάρτηση, όπως προτεινόταν και από την βιβλιογραφία της Αναγνώρισης Ενεργειών. Τα αποτελέσματα της Απόδοσης του συστήματος στις εκδοχές αυτές παρουσιάζονται στον παρακάτω πίνακα, ενώ στη συνέχεια παραθέτουμε και τις μετρικές Ακρίβειας, Ανάκλησης και F1 σκορ για κάθε κατηγορία βίντεο για τα STIP χαρακτηριστικά. Με τον τρόπο αυτό λαμβάνουμε και μια πρώτη εικόνα για τις διαφορές μεταξύ των κατηγοριών σε επίπεδο δυσκολίας ταξινόμησή τους.

Πίνακας 2 Μέσω απόδοση ταξινομητών με μεμονωμένα χαρακτηριστικά

	SIFT	MFCC	STIP
NO_PCA χ^2	-	-	51.95%
PCA χ^2	53.19%	32.70%	61.34%
PCA RBF	50.78%	28.69%	57.37%
PCA LINEAR	47.73%	28.17%	56.99%

Πίνακας 3 Ακρίβεια, Ανάκληση και F1 για τα STIP χαρακτηριστικά

	precision	recall	f1-score	support
ApplyEyeMakeup	0.63	0.77	0.69	145
SoccerPenalty	0.76	0.91	0.82	137
Typing	0.57	0.59	0.58	136
Haircut	0.83	0.76	0.80	130
FloorGymnastics	0.49	0.58	0.53	125
BrushingTeeth	0.45	0.34	0.39	131
FieldHockeyPenalty	0.71	0.75	0.73	126
HeadMassage	0.50	0.40	0.44	147
ParallelBars	0.43	0.47	0.45	114
HandstandWalking	0.43	0.36	0.39	111
BabyCrawling	0.83	0.81	0.82	132
BlowingCandles	0.65	0.55	0.59	109
avg / total	0.61	0.61	0.61	1543

5.3.2 Εκπαίδευση με πρώιμη συγχώνευση

Στη συνέχεια πειραματιστήκαμε με την εφαρμογή της πρώιμης συγχώνευσης των χαρακτηριστικών του συστήματος. Για συνάρτηση πυρήνα στο SVM μοντέλο που εκπαιδεύσαμε, χρησιμοποιήσαμε τη $\boldsymbol{\chi^2}$ που διαπιστώσαμε την αποτελεσματικότητά της στο προηγούμενο πείραμα. Από την επικύρωση προκύπτουν αισθητά βελτιωμένα αποτελέσματα, τόσο σε Απόδοση, όσο και σε Ακρίβεια, Ανάκληση και F1 σκορ, σε σχέση με τα αποτελέσματα των μεμονωμένων χαρακτηριστικών και η μέση απόδοση που υπολογίστηκε έφτασε στο **74.53%**. Μάλιστα, όπως θα φανεί και στη συνέχεια, αυτή είναι και μια από τις μεγαλύτερες τιμές Απόδοσης που πέτυχε το σύστημά μας. Επιβεβαιώνεται έτσι η επιλογή για αναπαράσταση με πολλούς τρόπους των βίντεο εκπαίδευσης, αφού και με μια εύκολη στην υλοποίησή της συγχώνευσή τους έχουμε σημαντικό κέρδος στην απόδοση του συστήματος.

Πίνακας 4 Ακρίβεια, Ανάκληση και F1 για την πρώιμη συγχώνευση

	precision	recall	f1-score	support
ApplyEyeMakeup	0.83	0.83	0.83	145
SoccerPenalty	0.98	0.93	0.95	137
Typing	0.93	0.70	0.80	136
Haircut	0.83	0.69	0.76	130
FloorGymnastics	0.61	0.84	0.70	125
BrushingTeeth	0.57	0.53	0.55	131
FieldHockeyPenalty	0.86	0.87	0.86	126
HeadMassage	0.62	0.51	0.56	147
ParallelBars	0.89	0.85	0.87	114
HandstandWalking	0.58	0.51	0.55	111
BabyCrawling	0.69	0.84	0.76	132
BlowingCandles	0.60	0.83	0.70	109
avg / total	0.75	0.74	0.74	1543

5.3.3 Εκπαίδευση με όψιμη, χωρίς εκπαίδευση συγχώνευση

Στο επόμενο πείραμα, εφαρμόσαμε τις μεθόδους όψιμης συγχώνευσης χωρίς εκπαίδευση, τόσο για τις προβλέψεις ταξινόμησης, όσο και για τις πιθανότητες που εξήγαγαν οι ταξινομητές των επί μέρους χαρακτηριστικών από το πρώτο πείραμα. Για τα SVM μοντέλα επιλέξαμε τη συνάρτηση πυρήνα $\boldsymbol{\chi^2}$ που έδωσε τα καλύτερα αποτελέσματα. Αρχικά, δοκιμάσαμε όλους τους πιθανούς συνδυασμούς ανά δύο εκ των τριών χαρακτηριστικών, συνδυάζοντάς τα με τον κανόνα Αθροίσματος και τον κανόνα Γινομένου, για να φανεί καλύτερα η συσχέτιση μεταξύ τους και το κατά πόσο συμβάλλει το καθένα από αυτά στη βελτίωση της απόδοσης που επιτυγχάνεται. Στη συνέχεια, συνδυάσαμε και τα τρία είδη χαρακτηριστικών με όλες τις μεθόδους συγχώνευσης χωρίς εκπαίδευση, όπως παρουσιάστηκαν στο 4.2.1.1. Τα αποτελέσματα για τη μέση Απόδοση των διαφορετικών εκδοχών του συστήματος, φαίνονται στον παρακάτω πίνακα, με τις πρώτες γραμμές να είναι οι μετρήσεις των μεμονωμένων χαρακτηριστικών από το πρώτο πείραμα, ώστε να μπορούν να συγκριθούν οι αλλαγές:

Πίνακας 5 Αποτελέσματα Απόδοσης όψιμης συγχώνευσης χωρίς εκπαίδευση

SIFT	MFCC	STIP	MAJ	WEI_MAJ	SUM	PROD	MED	MAX
+			53.19%	53.19%	53.19%	53.19%	53.19%	53.19%
	+		32.70%	32.70%	32.70%	32.70%	32.70%	32.70%
		+	61.34%	61.34%	61.34%	61.34%	61.34%	61.34%
+	+		-	-	53.14%	56.25%	-	-
+		+	-	-	69.66%	74.59%	-	-
	+	+	-	-	59.17%	59.94%	-	-
+	+	+	57.20%	64.19%	69.34%	72.13%	66.55%	62.54%

Από τα αποτελέσματα μπορούμε να συμπεράνουμε ότι σχεδόν σε όλες τις περιπτώσεις (εκτός από κάποιες από αυτές που συμπεριλαμβάνονται και τα MFCC χαρακτηριστικά) τα αποτελέσματα βελτιώνονται λιγότερο ή περισσότερο με κάθε μέθοδο συγχώνευσης. Ο ταξινομητής των MFCC, που είχε πετύχει πολύ χαμηλή απόδοση (**32.70%**) μόνος του, βλέπουμε ότι εκτός από την περίπτωση του κανόνα Γινομένου με τον SIFT (όπου επιτυγχάνεται μια μικρή βελτίωση) λειτουργεί ανασταλτικά στην απόδοση του συστήματος με αποτέλεσμα οι συγχωνεύσεις που τον συμπεριλαμβάνουν να έχουν χαμηλότερα αποτελέσματα από ότι χωρίς αυτόν. Έτσι προκύπτει και το ότι την καλύτερη απόδοση σε αυτό το πείραμα πέτυχε ο συνδυασμός SIFT – STIP με τον κανόνα Γινομένου (**74.59%**), που σε όλες τις περιπτώσεις φαίνεται να είναι και η αποδοτικότερη μέθοδος συγχώνευσης από αυτές που εξετάσαμε εδώ. Αξίζει επίσης να παρατηρήσουμε ότι, μόνο ο συνδυασμός αυτός επέφερε αποτέλεσμα ανάλογο με αυτό της πρώιμης συγχώνευσης που εξετάσαμε στο προηγούμενο πείραμα, αναδεικνύοντας την αποδοτικότητα της συγχώνευσης σε επίπεδο αναπαράστασης σε σχέση με τη συγχώνευση σκορ. Ακόμη, σημειώνουμε και τη σημαντική διαφορά σε απόδοση που επιτυγχάνεται με

την προσθήκη βαρών στη μέθοδο Ψηφοφορίας Πλειοψηφίας, με τη σημασία της να αναδεικνύεται ακόμα περισσότερο στο επόμενο πείραμα.

Μπορούμε τελικά να συμπεράνουμε ότι, ενώ μεν η συγχώνευση διαφορετικών χαρακτηριστικών για την αναπαράσταση των δεδομένων μας είναι ευεργετική για την απόδοση του συστήματος, το γεγονός αυτό δεν μπορεί να θεωρείται κανόνας για όλες τις περιπτώσεις, καθώς είναι άμεσα εξαρτημένο και από την ποιότητα των χαρακτηριστικών που χρησιμοποιούμε και άρα υπάρχει κίνδυνος προσθέτοντας χαρακτηριστικά με χαμηλή απόδοση να καταλήξουμε σε μείωση της αποτελεσματικότητας του συστήματος. Κατά την ανάπτυξη ενός τέτοιου συστήματος λοιπόν, οφείλουμε πέραν από την επιλογή κατάλληλης μεθόδου συγχώνευσης, να δίνουμε έμφαση και στο πλήθος και την ποιότητα των διαφορετικών χαρακτηριστικών που χρησιμοποιούμε για να επιτύχουμε εύρωστα συστήματα.

5.3.4 Εκπαίδευση με συνδυασμό πρώιμης και όψιμης χωρίς εκπαίδευση συγχώνευσης

Σε αυτό το πείραμα, εξετάζουμε τη δυνατότητα συνδυασμού των δύο προηγούμενων κατηγοριών μεθόδων συγχώνευσης που παρουσιάσαμε. Έτσι, συμπεριλάβαμε και τον ταξινομητή που εκπαιδεύτηκε με τα διανύσματα της πρώιμης συγχώνευσης στις μεθόδους όψιμης συγχώνευσης χωρίς εκπαίδευση που εξετάσαμε. Αρχικά, εξετάζουμε τη συγχώνευση της πρώιμης συγχώνευσης με δύο από τα τρία κάθε φορά μεμονωμένα χαρακτηριστικά, με Ψηφοφορία Πλειοψηφίας, Ψηφοφορία Πλειοψηφίας με βάρη και κανόνα Γινομένου (που στο προηγούμενο πείραμα είχε την καλύτερη απόδοση). Στη συνέχεια, συγχωνεύουμε και τους τέσσερις ταξινομητές με όλες τις μεθόδους συγχώνευσης που παρουσιάσαμε (εκτός της Ψηφοφορίας Πλειοψηφίας λόγω ζυγού αριθμού ταξινομητών). Τα αποτελέσματα για τη μέση Απόδοση φαίνονται στον πίνακα:

Πίνακας 6 Αποτελέσματα Απόδοσης συνδυασμού πρώιμης και όψιμης συγχώνευσης

SIFT	MFCC	STIP	EARLY	MAJ	WEI_MAJ	SUM	PROD	MED	MAX
+				53.19%	53.19%	53.19%	53.19%	53.19%	53.19%
	+			32.70%	32.70%	32.70%	32.70%	32.70%	32.70%
		+		61.34%	61.34%	61.34%	61.34%	61.34%	61.34%
			+	74.53%	74.53%	74.53%	74.53%	74.53%	74.53%
+	+		+	64.58%	73.56%	-	69.15%	-	-
+		+	+	72.86%	74.79%	-	75.56%	-	-
	+	+	+	66.05%	73.56%	-	73.62%	-	-
+	+	+	+	-	74.66%	74.20%	74.27%	74.14%	70.18%

Και σε αυτό το πείραμα, βλέπουμε ότι η χαμηλή απόδοση του ταξινομητή των MFCC, επηρεάζει αρνητικά τα αποτελέσματα των συγχωνεύσεων που τον συμπεριλαμβάνουν. Αυτή η αρνητική επιρροή όμως, μετριάζεται σημαντικά στην περίπτωση της Ψηφοφορίας Πλειοψηφίας με βάρη, όπου όλοι οι συνδυασμοί που δοκιμάστηκαν παρέμειναν στα ίδια επίπεδα με αυτά της πρώιμης συγχώνευσης, ενώ μάλιστα παρατηρείται μια μικρή αλλά μετρήσιμη βελτίωση στον συνδυασμό SIFT – STIP και πρώιμης συγχώνευσης (**74.79%**). Με βάση αυτά, μπορούμε να βγάλουμε κάποια χρήσιμα συμπεράσματα.

Φαίνεται ότι η προσθήκη βαρών στην επιρροή που έχει ο κάθε ταξινομητής στην τελική πρόβλεψη, με βάση την ατομική του απόδοση, λειτουργεί σταθεροποιητικά για το σύστημα, αφού το «προστατεύει» από μια πτώση της απόδοσης λόγω κάποιου κακής ποιότητας ταξινομητή που συμπεριλήφθηκε. Μάλιστα, το γεγονός αυτό φαίνεται και από το ότι η Ψηφοφορία Πλειοψηφίας με βάρη καταφέρνει τα καλύτερα αποτελέσματα σχεδόν σε όλους τους συνδυασμούς που συμπεριλαμβάνουν τα MFCC χαρακτηριστικά σε σχέση με τις άλλες μεθόδους συγχώνευσης που εξετάσαμε εδώ. Συμπεραίνουμε ότι για το αποτέλεσμα αυτό ευθύνεται σε μεγάλο βαθμό, τόσο η ατομική απόδοση του ταξινομητή της πρώιμης συγχώνευσης που προστέθηκε στο σύστημα, όσο και το γεγονός ότι τη λαμβάνουμε υπόψη ως βάρος στη διαδικασία της Ψηφοφορίας. Βλέπουμε λοιπόν ότι, ενώ και σε αυτό και στο προηγούμενο πείραμα η Ψηφοφορία Πλειοψηφίας με βάρη επέφερε μικρή αύξηση στην απόδοση σε σχέση με την απόδοση του καλύτερου μεμονωμένου ταξινομητή, εξασφάλισε παράλληλα ότι το συνολικό σύστημα θα έχει απόδοση τουλάχιστον συγκρίσιμη με αυτή του καλύτερου ταξινομητή, πράγμα που στην περίπτωση εισαγωγής ενός ταξινομητή ήδη αρκετά καλής ποιότητας οδηγεί σε υψηλή απόδοση. Αντίθετα, οι μέθοδοι όψιμης συγχώνευσης που συνδυάζουν τις εξαγμένες πιθανότητες των ταξινομητών (κανόνας Γινομένου, Αθροίσματος κ.λπ.) μπορούν να καταγράψουν σημαντικά καλύτερη απόδοση από τους μεμονωμένους ταξινομητές (ο κανόνας Γινομένου με συνδυασμό SIFT – STIP και πρώιμης συγχώνευσης πετυχαίνει τη μεγαλύτερη απόδοση εδώ, αλλά και μεγαλύτερη συνολικά για όλα τα πειράματα: **75.56%**), είναι όμως ευάλωτοι στην προσθήκη ενός κακής ποιότητας ταξινομητή, η οποία μπορεί τελικά να οδηγήσει και σε μικρή μείωση της συνολικής απόδοσης σε σχέση με τον καλύτερο ταξινομητή. Στο πείραμα αυτό λοιπόν, αναδεικνύεται η σημασία της συμπερίληψης της ποιότητας του κάθε ταξινομητή ως βάρος στην εξίσωση συγχώνευσης των χαρακτηριστικών και η σημασία της απόρριψης (που ισοδυναμεί με μηδενικό βάρος) των κακής ποιότητας χαρακτηριστικών από την εξίσωση για τη δημιουργία αποδοτικότερων συστημάτων.

5.3.5 Εκπαίδευση με όψιμη συγχώνευση με εκπαίδευση

Στο τελευταίο μας πείραμα, εξετάζουμε τις δυνατότητες της όψιμης συγχώνευσης με εκπαίδευση ενός ταξινομητή 2ου επιπέδου με τις εξαγμένες προβλέψεις ή πιθανότητες των ταξινομητών του 1ου επιπέδου. Εδώ, δοκιμάσαμε να χρησιμοποιήσουμε για ταξινομητές 1ου επιπέδου αρχικά τους ταξινομητές των μεμονωμένων χαρακτηριστικών, στη συνέχεια προσθέσαμε και των ταξινομητή της πρώιμης συγχώνευσης, ενώ στο τέλος αφαιρέσαμε τον ταξινομητή των MFCC χαρακτηριστικών που έδειξε από τα προηγούμενα πειράματα ότι έχει αρνητική επίπτωση στη συνολική απόδοση του συστήματος. Τα αποτελέσματα που συλλέξαμε φαίνονται στον παρακάτω πίνακα:

Πίνακας 7 Αποτελέσματα απόδοσης όψιμης συγχώνευσης με εκπαίδευση

	Προβλέψεις	Πιθανότητες
Log_Regression Stacking (SIFT, MFCC, STIP)	17.28%	70.29%
Log_Regression Stacking (SIFT, MFCC, STIP, EARLY_FUSION)	19.24%	75.25%
Log_Regression Stacking (SIFT, STIP, EARLY_FUSION)	20.79%	74.93%

Μια από τις πρώτες προφανείς παρατηρήσεις που κάνουμε στο πείραμα αυτό είναι η τεράστια πτώση σε απόδοση που επιφέρει η χρήση των προβλέψεων των ταξινομητών 1ου επιπέδου για την εκπαίδευση του 2ου. Μια τέτοια επιλογή είναι εμφανές ότι αποτελεί λάθος και ίσως να μπορεί να εξηγηθεί λόγω της πολύ μικρής ποσότητας δεδομένων που προκύπτουν (3-4 προβλέψεις για κάθε δείγμα εκπαίδευσης) για την εκπαίδευση του ταξινομητή. Στην περίπτωση των πιθανοτήτων τα πράγματα είναι διαφορετικά και τα αποτελέσματα είναι αρκετά υψηλά για να επιβεβαιώσουν τη δύναμη της μεθόδου αυτής για αποδοτική συγχώνευση χαρακτηριστικών. Παρόλα αυτά, βλέπουμε πως το μεγαλύτερο αποτέλεσμα στο πείραμα αυτό (**75.25%** με συνδυασμό και των τεσσάρων ταξινομητών) δεν κατάφερε να ξεπεράσει την απόδοση του κανόνα Γινομένου στο προηγούμενο πείραμα. Από αυτό συμπεραίνουμε ότι, ενώ η όψιμη συγχώνευση με εκπαίδευση ταξινομητή 2ου επιπέδου μπορεί να βελτιώσει την απόδοση του συστήματος, συγκρίσιμη και μάλιστα μεγαλύτερη βελτίωση μπορεί να επιτευχθεί και με μεθόδους χωρίς εκπαίδευση που είναι και προτιμότερες ώστε να μην αυξάνεται κι άλλο ο χρόνος εκπαίδευσης του συνολικού συστήματος που απαιτείται.

Κεφάλαιο 6 Σύνοψη αποτελεσμάτων και μελλοντικές επεκτάσεις

6.1 Σύνοψη Αποτελεσμάτων

Με την παραπάνω διαδικασία πειραμάτων που διενεργήσαμε, μπορέσαμε να εμβαθύνουμε στη μεθοδολογία του σάκου λέξεων για την κωδικοποίηση εξαγμένων χαρακτηριστικών και να συγκρίνουμε μεθόδους συγχώνευσης των χαρακτηριστικών αυτών ως προς την απόδοση που επιτυγχάνουν. Στην πρώτη φάση των πειραμάτων, καταφέραμε να πετύχουμε 10% αύξηση της απόδοσης του συστήματος μέσω της προεπεξεργασίας των δεδομένων με Ανάλυση Κύριων Συνιστωσών και χρησιμοποιώντας STIP χαρακτηριστικά. Στη συνέχεια, είδαμε και πάλι μια αύξηση κοντά στο 13% της απόδοσης με την πρώιμη συγχώνευση όλων των διαθέσιμων χαρακτηριστικών, επιβεβαιώνοντας την αρχική μας υπόθεση ότι η συγχώνευση διαφορετικών χαρακτηριστικών, συμπληρωματικών μεταξύ τους, ακόμα και στην απλή της μορφή μπορεί να επιφέρει σημαντική βελτίωση στα συστήματά μας. Ακολούθως, αντιμετωπίσαμε το πρόβλημα συμπερίληψης ενός κακής ποιότητας ταξινομητή (MFCC) για τα δεδομένα μας και είδαμε πως η προσθήκη βαρών στην επιρροή των ταξινομητών στο σύστημα μπορεί να περιορίσει σημαντικά την πτώση της απόδοσης από αυτό. Τέλος, είδαμε πως είναι εφικτό να συνδυάσουμε μεθόδους πρώιμης και όψιμης συγχώνευσης στο σύστημά μας και μάλιστα με αυτό πετύχαμε και τη μεγαλύτερη τιμή απόδοσης (**75.56%**) χωρίς να απαιτείται κάποια επιπλέον εκπαίδευση, όπως στην περίπτωση του τελευταίου πειράματος.

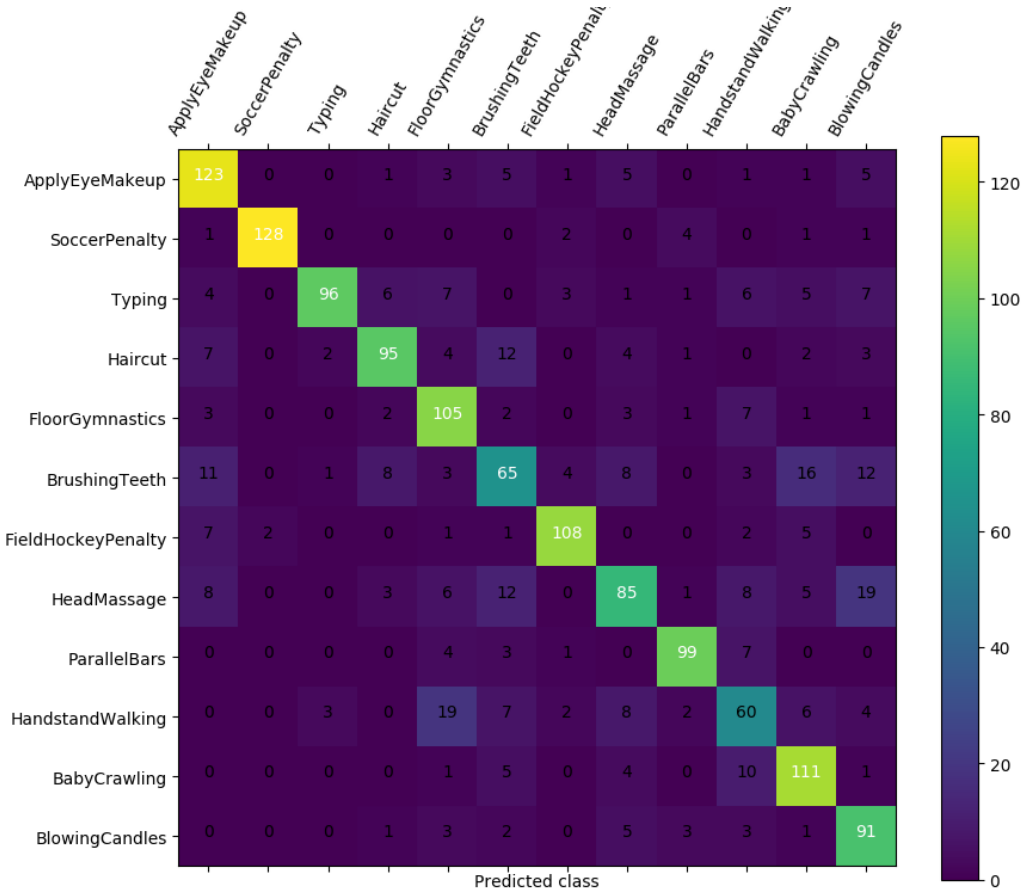
Στο σημείο αυτό, παραθέτουμε τις μετρικές Ακρίβειας, Ανάκλησης και F1 σκορ για τις

Πίνακας 8 Αποτελέσματα Ακρίβειας, Ανάκλησης και F1 για την αποδοτικότερη εκδοχή του συστήματος

	precision	recall	f1-score	support
ApplyEyeMakeup	0.75	0.85	0.80	145
SoccerPenalty	0.98	0.93	0.96	137
Typing	0.94	0.71	0.81	136
Haircut	0.82	0.73	0.77	130
FloorGymnastics	0.67	0.84	0.75	125
BrushingTeeth	0.57	0.50	0.53	131
FieldHockeyPenalty	0.89	0.86	0.87	126
HeadMassage	0.69	0.58	0.63	147
ParallelBars	0.88	0.87	0.88	114
HandstandWalking	0.56	0.54	0.55	111
BabyCrawling	0.72	0.84	0.78	132
BlowingCandles	0.63	0.83	0.72	109
avg / total	0.76	0.76	0.75	1543

διάφορες κατηγορίες που είχαμε στη διάθεσή μας, όπως και τον Πίνακα Σύγχυσης που προέκυψαν από την καλύτερη εκδοχή του συστήματός μας, δηλαδή το συνδυασμό SIFT – STIP και πρώιμης συγχώνευσης με τον κανόνα Γινομένου.

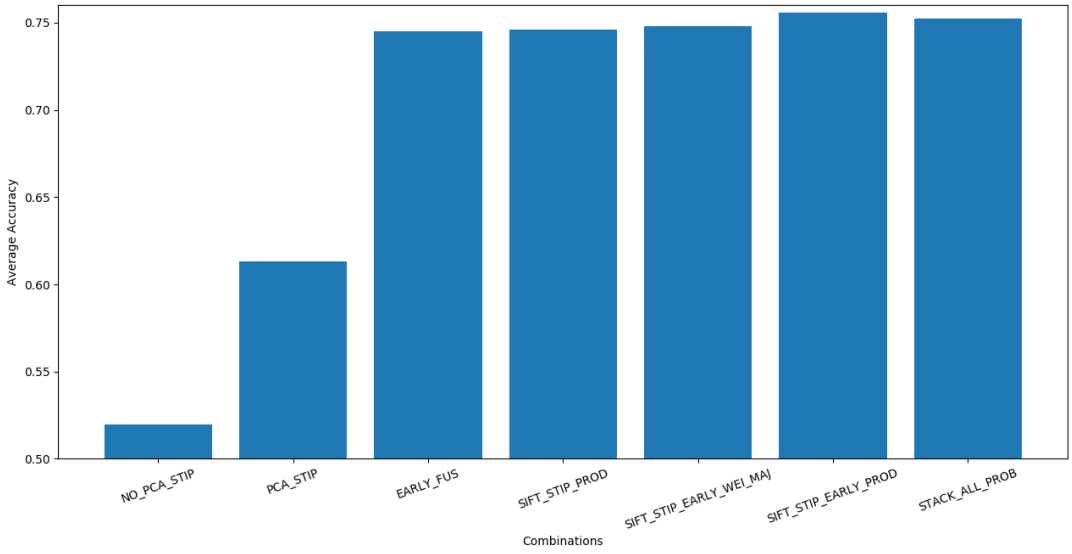
Πίνακας 9 Πίνακας Σύγχυσης για την αποδοτικότερη εκδοχή του συστήματος



Με βάση και την εικόνα των παραπάνω πινάκων, μπορούμε να συμπεράνουμε ότι το σύστημά μας επιτυγχάνει το σκοπό του, αφού μπορεί με μια ικανοποιητική βεβαιότητα να αναγνωρίσει την κατηγορία ενέργειας που περιέχεται σε ένα βίντεο που θα εισαχθεί. Από τον Πίνακα Σύγχυσης βλέπουμε ότι στις περισσότερες κατηγορίες έχουμε καταφέρει ικανοποιητικό αριθμό ορθών προβλέψεων, ώστε να μπορούμε να θεωρήσουμε ότι το σύστημα «έμαθε» την κατηγορία αυτή και μπορεί να την αναγνωρίσει. Από τις τιμές της Ανάκλησης των επί μέρους κατηγοριών, φαίνεται ότι οι κατηγορίες Brushing Teeth, Head Massage και Handstand Walking δυσκόλεψαν περισσότερο το σύστημα, οδηγώντας σε οριακά μη αποδεκτό ποσοστό βεβαιότητας. Ένας παράγοντας που ευθύνεται για το αποτέλεσμα αυτό, είναι οι μεγάλες παραλλαγές που υπάρχουν στο εσωτερικό των κατηγοριών αυτών, με διαφοροποιήσεις στη γωνία λήψης της κάμερας, στην πυκνότητα στο φόντο κ.ά. Η ιδιαιτερότητα αυτή, σε συνδυασμό και με την ύπαρξη κατηγοριών όπως το Floor Gymnastics για το Handstand Walking ή το Apply Eye Makeup για το Brushing Teeth, μπορούν να δημιουργήσουν σύγχυση στον ταξινομητή με αποτέλεσμα αρκετά από τα θετικά δείγματα της μιας να ταξινομούνται λανθασμένα στην άλλη.

Δεν συμβαίνει το ίδιο όμως και σε ζεύγη κατηγοριών όπως το Soccer Penalty και το Field Hockey Penalty, που αρχικά εκτιμούσαμε ότι μπορεί να συγχέονται μεταξύ τους, αλλά φαίνεται πως οι μικρές αποκλίσεις μεταξύ των δειγμάτων της ίδιας κατηγορίας κατέστησαν δυνατό τον αποτελεσματικό εντοπισμό της. Μάλιστα, η κατηγορία Soccer Penalty έφερε και το μεγαλύτερο αποτέλεσμα Ακρίβειας και Ανάκλησης, προσεγγίζοντας σχεδόν το 100%, πράγμα που μπορούσε σε ένα βαθμό να προβλεφθεί λόγω των κοινών χαρακτηριστικών των δειγμάτων της κατηγορίας (π.χ. ποδοσφαιριστής στα δεξιά της εικόνας, τέρμα στα αριστερά κ.ά.) που την έκαναν εύκολα αναγνωρίσιμη με τα χαρακτηριστικά που εξάγαμε.

Στη συνέχεια, παρουσιάζουμε ένα συγκεντρωτικό διάγραμμα με κάποια ενδεικτικά αποτελέσματα Μέσης Απόδοσης από τα πειράματα που διενεργήθηκαν. Βλέπουμε εδώ, ότι επιτύχαμε μια αύξηση κοντά στο 20% στην απόδοση του συστήματος σε σχέση με την αρχική εκδοχή με χρήση των δεδομένων STIP χωρίς προεπεξεργασία με Ανάλυση Κύριων Συνιστωσών, ενώ το μεγάλο άλμα στην απόδοση προέκυψε από την πρώιμη συγχώνευση των χαρακτηριστικών. Τέλος, καταφέραμε μια ακόμα μικρή αύξηση της απόδοσης του συστήματος, αξιοποιώντας τόσο την πρώιμη όσο και την όψιμη συγχώνευση.



25 Σύγκριση αποτελεσμάτων των διαφορετικών πειραμάτων

Αξίζει να σημειωθεί εδώ ότι αυτή η μεγάλη βελτίωση των αποτελεσμάτων επετεύχθη αποκλειστικά με χρήση μεθόδων συγχώνευσης των εξαγμένων χαρακτηριστικών, χωρίς να χρειαστεί να γίνουν αλλαγές ούτε στα είδη των χαρακτηριστικών, ούτε στο μοντέλο των ταξινομητών που χρησιμοποιήθηκαν, γεγονός πολύ σημαντικό για πραγματικές εφαρμογές πάνω στο πρόβλημα, όπου η αναθεώρηση της αρχιτεκτονικής ενός συστήματος συνολικά για τη βελτίωση της απόδοσής του μπορεί να μην είναι εφικτή και σίγουρα θα απαιτεί εκτεταμένη χρήση των υπολογιστικών πόρων που διαθέτουμε.

6.2 Μελλοντικές Επεκτάσεις

Με βάση τις ενδείξεις που παίρνουμε από τα αποτελέσματα και αναλογιζόμενοι τις προκλήσεις που αντιμετωπίσαμε κατά την ανάπτυξη του συστήματος αυτού, προτείνουμε κάποιες κατευθύνσεις για μελλοντική έρευνα πάνω στο πρόβλημα:

- Να αξιοποιηθούν όλες οι κατηγορίες που εμπεριέχονται στο UCF101 dataset, ώστε να διερευνηθούν όλες οι πιθανές συσχετίσεις μεταξύ τους και το κατά πόσο η προσθήκη μεγάλου αριθμού κατηγοριών επηρεάζει την απόδοση του συνολικού συστήματος.
- Να χρησιμοποιηθούν υλοποιήσεις των αλγορίθμων εκπαίδευσης που να μπορούν να παραλληλοποιηθούν και να αξιοποιήσουν τις υπολογιστικές δυνατότητες των καρτών γραφικών. Στην κατεύθυνση αυτή, μπορούν να αξιοποιηθούν και μέθοδοι Συνελικτικών Νευρωνικών Δικτύων για την αναπαράσταση και ταξινόμηση των δειγμάτων.
- Να δοθεί έμφαση στην όψιμη συγχώνευση με εκπαίδευση και να διερευνηθεί η αποδοτικότητα διαφορετικών ειδών ταξινομητών για το 2^{o} επίπεδο, ώστε να μπορούμε να προσεγγίσουμε καλύτερα τα κριτήρια με τα οποία κάποιος μπορεί να επιλέξει κατάλληλο ταξινομητή, ανάλογα το πρόβλημα που αντιμετωπίζει.
- Να γίνει χρήση διαφορετικών και πυκνότερων χαρακτηριστικών για την αναπαράσταση των βίντεο, όπως είναι οι βελτιωμένες Πυκνές Τροχιές (improved Dense Trajectories) και να διερευνηθούν άλλες επιλογές για τη δημιουργία λεξικού και την κωδικοποίηση των χαρακτηριστικών στη μεθοδολογία του σάκου λέξεων, όπως για παράδειγμα τα Μοντέλα Μείγματος Γκαουσιανών Κατανομών (Gaussian Mixture Models) και τα διανύσματα Fisher.
- Το επόμενο βήμα για την αναγνώριση ενεργειών σε βίντεο, είναι το αν θα μπορεί ένα σύστημα να αναγνωρίζει πάνω από μία ενέργειες που μπορεί να περιέχονται στο ίδιο βίντεο και μάλιστα να μπορεί να εντοπίζει τη χρονική θέση που έχουν οι ενέργειες αυτές στο βίντεο.

Βιβλιογραφία

[1] X. Peng et al., Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice, Computer Vision and Image Understanding (2016)

[2] Campbell, L.W. , Bobick, A.F. , 1995. Recognition of human body motion using phase space constraints. In: ICCV, pp. 624–630 .

[3] Niyogi, S.A. , Adelson, E.H. , 1994. Analyzing and recognizing walking figures in XYT. In: CVPR, pp. 469–474 .

[4] Webb, J.A. , Aggarwal, J.K. , 1981. Structure from motion of rigid and jointed objects. In: IJCAI, pp. 686–691 .

[5] Yacoob, Y. , Black, M.J. , 1999. Parameterized modeling and recognition of activities. Comput. Vis. Image Understand. 73 (2), 232–247 .

[6] Laptev, I. , 2005. On space-time interest points. Int. J. Comput. Vis. 64 (2-3), 107–123.

[7] Wang, H. , Schmid, C. , 2013a. Action recognition with improved trajectories. In: ICCV, pp. 3551–3558 .

[8] Krizhevsky, A. , Sutskever, I. , Hinton, G.E. , 2012. Imagenet classification with deep convolutional neural networks. In: NIPS, pp. 1097–1105 .

[9] Simonyan, K. , Zisserman, A. , 2014. Two-stream convolutional networks for action recognition in videos. In: NIPS, pp. 568–576 .

[10] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1–8, June 2008

[11] Samitha Herath, Mehrtash Harandi, and Fatih Porikli. Going deeper into action recognition: A survey. arXiv preprint arXiv:1605.04988, 2016.

[12] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, ACM Transactions on Intelligent Systems and Technology 2 (3)(2011) 27

[13] David G. Lowe. 2004. Distinctive Image Features from Scale-Invariant Keypoints. Int. J. Comput. Vision 60, 2 (November 2004), 91-110.

[14] vandeSande,K.E.A.,Gevers,T.,Snoek,C.G.M.:Evaluatingcolor descriptors for object and scene recognition. PAMI 32(9), 1582–1596 (2010)

[15] Paul Scovanner, Saad Ali, and Mubarak Shah. A 3-dimensional sift descriptor and its application to action recognition. In Proceedings of the 15th ACM international conference on Multimedia, pages 357–360. ACM, 2007.

[16] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), volume 1, pages 886–893 vol. 1, 2005

[17] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1–8, June 2008

[18] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Dense trajectories and motion boundary descriptors for action recognition. International journal of computer vision, 103(1):60–79, 2013.

[19] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In Proceedings of the IEEE International Conference on Computer Vision, pages 3551–3558, 2013.

[20] N. Dalal, B. Triggs, C. Schmid, Human detection using oriented histograms of flow and appearance, in: European Conference on Computer Vision, Springer, 2006.

[21] Bishop, C.M. , 2006. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag New York, Inc., Secaucus, NJ, USA .

[22] Perronnin, F. , Sánchez, J. , Mensink, T. , 2010. Improving the fisher kernel for large-scale image classification. In: ECCV, pp. 143–156 .

[23] K. Lee and D. Ellis, "Audio-based semantic concept classification for consumer video," IEEE Trans. ASLP, vol. 18, no. 6, pp. 1406 –1416, 2010.

[24] Janin A, Stolcke A, Anguera X, Boakye K, Çetin Ö, Frankel J, Zheng J (2006) The ICSI-SRI spring 2006 meeting recognition system. In: MLMI'06 proceedings of the third international conference on machine learning for multimodal, interaction, pp 444–456

[25] Liu N, Dellandréa E, Chen L, Zhu C, Zhang Y, Bichot CE, Bres S, Tellez B (2013) Multimodal recognition of visual concepts using histograms of textual concepts and selective weighted late fusion scheme. Computer Vision and Image Understanding 117:493–512

[26] Bogdan Ionescu, Jenny Benois-Pineau, Tomas Piatrik, and Georges Qunot. 2014. Fusion in Computer Vision: Understanding Complex Visual Content. Springer Publishing Company, Incorporated.

[27] Sangmin Oh, Scott Mccloskey, Ilseo Kim, Arash Vahdat, Kevin J. Cannons, Hossein Hajimirsadeghi, Greg Mori, A. G. Perera, Megha Pandey, and Jason J. Corso. 2014. Multimedia event detection with multimodal feature fusion and temporal concept localization. Mach. Vision Appl. 25, 1 (January 2014), 49-69.

[28] Moreno-Seco F., Iñesta J.M., de León P.J.P., Micó L. (2006) Comparison of Classifier Fusion Methods for Classification in Pattern Recognition Tasks. In: Yeung DY., Kwok J.T., Fred A., Roli F., de Ridder D. (eds) Structural, Syntactic, and Statistical Pattern Recognition. SSPR /SPR 2006. Lecture Notes in Computer Science, vol 4109. Springer, Berlin, Heidelberg.

[29] Y. Le Cun, B. Boser, J. S. Denker, R. E. Howard, W. Habbard, L. D. Jackel, and D. Henderson. 1990. Handwritten digit recognition with a back-propagation network. In Advances in neural information processing systems 2, David S. Touretzky (Ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA 396-404.

[30] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.

[31] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. IEEE Trans. Pattern Analysis and Machine Intelligence, 35(1):221–231, Jan 2013. ISSN 0162-8828

[32] Joe Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4694–4702, 2015

[33] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1725–1732, 2014

[34] Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia, and Atilla Baskurt. Sequential deep learning for human action recognition. In Proceedings of the Second International Conference on Human Behavior Unterstanding, HBU'11, pages 29–39, 2011

[35] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2625–2634, 2015

[36] A. J. Robinson and F. Fallside. Static and dynamic error propagation networks with application to speech coding. In Proc. Advances in Neural Information Processing Systems (NIPS), pages 632–641, 1988

[37] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural Comput., 9(8):1735–1780, 1997.

[38] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In Proc. Advances in Neural Information Processing Systems (NIPS), pages 568–576, 2014.

[39] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In Proc. Int. Conference on Computer Vision (ICCV), pages 4489–4497, Dec 2015.

[40] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1933–1941, 6 2016

[41] David G Lowe. Object recognition from local scale-invariant features. In Computer vision, 1999. The proceedings of the seventh IEEE international conference on, volume 2, pages 1150–1157. Ieee, 1999.

[42] P. Mermelstein, "Distance Measures for Speech Recognition – Psychological and Instrumental", Pattern Recognition and Artificial Intelligence, pp. 374–388, 1976.

[43] Harris, C. and Stephens, M. 1988. A combined corner and edge detector. Alvey Vision Conference, pp. 147–152.

[44] R. Baeza-Yates and B. Ribeiro-Neto. Modern Information Retrieval. ACM Press, ISBN: 020139829, 1999.

[45] Sivic J, Zisserman A (2003) Video google: a text retrieval approach to object matching in videos. In: Proceedings of the IEEE international conference on computer vision (ICCV)

[46] Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, and Cdric Bray. Visual categorization with bags of keypoints. In In Workshop on Statistical Learning in Computer Vision, ECCV, pages 1–22, 2004

[47] Tamrakar, A., Ali, S., Yu, Q., Liu, J., Javed, O., Divakaran, A., Cheng, H., Sawhney, H.S.: Evaluation of low-level features and their combinations for complex event detection in open source videos. In: CVPR (2012)

[48] Sangmin Oh, Scott Mccloskey, Ilseo Kim, Arash Vahdat, Kevin J. Cannons, Hossein Hajimirsadeghi, Greg Mori, A. G. Perera, Megha Pandey, and Jason J. Corso. 2014. Multimedia event detection with multimodal feature fusion and temporal concept localization. Mach. Vision Appl. 25, 1 (January 2014), 49-69.

[49] Lan, Z.Z., Bao, L., Yu, S.I., Liu, W., Hauptmann, A.G.: Double fusion for multimedia event detection. In: ICME (2012)

[50] Platt, J.: Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In: Advances in Large Margin Classifiers, 2000

[51] Tang, J., S. Alelyani, and H. Liu. "Data Classification: Algorithms and Applications." Data Mining and Knowledge Discovery Series, CRC Press (2015): pp. 498-500.

[52] Khurram Soomro, Amir Roshan Zamir and Mubarak Shah, UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild., CRCV-TR-12-01, November, 2012

[53] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. 2006. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR '06), Vol. 2. IEEE Computer Society, Washington, DC, USA, 2169-2178