Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής
Εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων

Αυτόματη Αναγνώριση Ανθρώπινων Δράσεων με Εμπλουτισμένες Αναπαραστάσεις Βίντεο

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

Ευφροσύνης Α. Μαυρουδή

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2015

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής
Εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων

Αυτόματη Αναγνώριση Ανθρώπινων Δράσεων με Βελτιωμένες Αναπαραστάσεις Βίντεο

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

Ευφροσύνης Α. Μαυρουδή

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 16η Ιουλίου 2015.

.....	.....	.....
Πέτρος Μαραγκός Καθηγητής Ε.Μ.Π	Κωνσταντίνος Τζαφέστας Επίκουρος Καθηγητής Ε.Μ.Π	Γεράσιμος Ποταμιάνος Αναπληρωτής Καθηγητής Παν/μίου Θεσσαλίας

Αθήνα, Ιούλιος 2015.

.....
Ευφροσύνη Α. Μαυρουδή
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Ευχαριστίες

Αρχικά θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντά μου, Καθ. Πέτρο Μαραγκό, για την εμπιστοσύνη που μου έδειξε, δίνοντάς μου την ευκαιρία να ασχοληθώ ερευνητικά με το πρόβλημα της Αναγνώρισης Δράσεων. Μέσω των διαλέξεών του ήρθα σε επαφή για πρώτη φορά με τα πεδία της Όρασης Υπολογιστών και Αναγνώρισης Προτύπων και συνέβαλε με αυτό τον τρόπο στη διαμόρφωση των ερευνητικών μου ενδιαφερόντων. Ο ενθουσιασμός του και η αφοσίωσή του στην έρευνα αποτελούν πηγές έμπνευσης για κάθε φοιτητή, ενώ χωρίς τη στήριξη και την καθοδήγησή του θα ήταν αδύνατη η ολοκλήρωση αυτής της εργασίας. Επίσης, ευχαριστώ όλα τα μέλη του εργαστηρίου Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων και του εργαστηρίου Ρομποτικής και Αυτοματισμού για τη βοήθειά τους. Ιδιαίτερα θα ήθελα να ευχαριστήσω το Βασίλη Πιτσικάλη για τη στενή καθοδήγησή του και τις πολύωρες ερευνητικές συζητήσεις μας. Ακόμα, ευχαριστώ τον Κέβη Μανίνη για την παράχωρηση του κώδικά του, με τον οποίο ξεκίνησα τον πειραματισμό μου στα πλαίσια αυτής της διπλωματικής, καθώς και το Νίκο Κάρδαρη για την ευχάριστη συνεργασία μας και τη βοήθειά του.

Επίσης, θα ήθελα να ευχαριστήσω όλους τους ανθρώπους που με στήριξαν και αποτέλεσαν πρότυπα για μένα όλα αυτά τα χρόνια, τους καθηγητές μου και τους ερευνητές με τους οποίους είχα τη χαρά να συνεργαστώ. Ευχαριστώ τους φίλους μου για τα ευχάριστα φοιτητικά χρόνια που περάσαμε μαζί, την υπομονή και τη συμπαράστασή τους κατά τη διάρκεια εκπόνησης αυτής της εργασίας.

Τέλος, ευχαριστώ την οικογένειά μου για την αγάπη, τη φροντίδα, την υποστήριξη και την υπομονή τους καθ'όλη τη διάρκεια των προπτυχιακών σπουδών μου και αφιερώνω την παρούσα διπλωματική εργασία στον αδερφό μου Βασίλη.

Περίληψη

Η παρούσα διπλωματική εργασία ασχολείται με το πρόβλημα της αυτόματης αναγνώρισης ανθρώπινων δράσεων σε ρεαλιστικά βίντεο, εστιάζοντας σε μεθόδους αναπαράστασης των βίντεο. Για την εξαγωγή χαρακτηριστικών εκμεταλλευόμαστε την πλούσια πληροφορία κίνησης που μας προσφέρουν τα διαδεδομένα χαρακτηριστικά "Πυκνών Τροχιών". Σημαντικό μέρος της εργασίας αφιερώνεται στην ανάλυση των μεθόδων που χρησιμοποιούνται για την αναγνώριση των δράσεων, με ιδιαίτερη έμφαση σε επιτυχημένες σύγχρονες μεθόδους αναπαράστασης βίντεο, όπως οι Bag-Of-Visual-Words και VLAD. Αρχικά, πραγματοποιείται εκτενής πειραματισμός με διάφορες γνωστές μεθόδους εξαγωγής χαρακτηριστικών και υπολογισμού αναπαραστάσεων για την επίλυση του προβλήματος της αναγνώρισης συνεχόμενων δράσεων σε RGB-D βίντεο, τα οποία περιέχουν δράσεις που εκτελούνται από ηλικιωμένα άτομα. Στη συνέχεια, προτείνουμε δύο νέες μεθόδους αναπαράστασης βίντεο. Η πρώτη μέθοδος μοντελοποιεί την αλληλεπίδραση μεταξύ των συστάδων οπτικών χαρακτηριστικών (τροχιών) ποσοτικοποιώντας την κατευθυνόμενη ομοιότητα μεταξύ των συστάδων με το συνδυασμό εργαλείων όπως η Ανάλυση σε Κύριες Συνιστώσες και η απόκλιση Kullback-Leibler. Η δεύτερη μέθοδος αναπαριστά τα βίντεο ως χρονικές ακολουθίες συχνά εμφανιζόμενων οπτικών λέξεων, αποσκοπώντας στην μοντελοποίηση της εγγενούς χρονικής διάταξης των κινήσεων που αποτελούν μια δράση. Επιπρόσθετα, προτείνεται μέθοδος υπολογισμού της απόστασης μεταξύ αυτών των ακολουθιών οπτικών λέξεων με χρήση αλγορίθμου τοπικής στοίχισης συμβολικών ακολουθιών, που μας επιτρέπει την ταξινόμησή τους με χρήση SVMs. Η πειραματική αξιολόγηση των μεθόδων μας σε απαιτητικές βάσεις ανθρώπινων δράσεων επιβεβαιώνει την αποτελεσματικότητά τους, καθώς επιτυγχάνουν επιδόσεις που ξεπερνούν αυτές αρκετών γνωστών μεθόδων και είναι συγκρίσιμες με αυτές των καλύτερων σύγχρονων μεθόδων αναπαράστασης βίντεο της διεθνούς βιβλιογραφίας.

Λέξεις κλειδιά: αναγνώριση ανθρώπινων δράσεων, αναπαράσταση βίντεο, πυκνές τροχιές, Bag-Of-Visual-Words, Ανάλυση σε Κύριες Συνιστώσες, στοίχιση ακολουθιών, Μηχανές Διανυσματικής Υποστήριξης, συσταδοποίηση

Abstract

This thesis deals with the problem of automatic human action recognition in realistic videos, focusing on video representation methods. For feature extraction, we exploit the rich motion information captured in the state-of-the-art "Dense Trajectories" features. A significant part of this work is devoted to the analysis of action recognition methods, with a special focus on successful modern video representations, such as Bag-Of-Visual-Words and VLAD. We experiment with various popular feature extraction methods and video representations in the context of action classification and temporal localization in continuous RGB-D videos, which contain actions performed by elderly people. Furthermore, we propose two novel video representation methods. The first method models the interaction between clusters of visual features, quantifying the directional similarity between clusters, combining tools such as the Principal Component Analysis and the Kullback-Leibler divergence. The other method represents videos as temporal sequences of frequently occurring visual words, aiming at the modelling of the inherent temporal order of motions constituting an action. We also propose a method for the computation of distances between these visual word sequences, using a local sequence alignment algorithm, which enables their classification with Support Vector Machines. The experimental evaluation of our methods in demanding human action datasets confirms their efficacy, since they achieve high action recognition accuracy, outperforming many popular video representations and they are comparable with recently published top-performing video representations.

Keywords: human action recognition, video representation, dense trajectories, Bag-Of-Visual-Words, Principal Component Analysis, sequence alignment, Support Vector Machines, clustering

Περιεχόμενα

Κατάλογος σχημάτων

Κατάλογος πινάκων

List of Algorithms

Κεφάλαιο 1

Εισαγωγή

1.1 Γενικά για την Όραση Υπολογιστών

Η Όραση Υπολογιστών είναι ο τομέας της επιστήμης και της τεχνολογίας που μελετά μεθόδους για την ανάλυση και κατανόηση μιας εικόνας ή ακολουθίας εικόνων με στόχο την εξαγωγή συμβολικής πληροφορίας. Η είσοδος των αλγορίθμων της Όρασης Υπολογιστών είναι αριθμητικά δεδομένα, π.χ. σήματα εικόνων, ενώ η έξοδος είναι συμβολική, π.χ. εντοπισμός και περιγραφή αντικειμένων, δράσεων. Σε αντίθεση με τον τομέα των Γραφικών Υπολογιστών, που προσπαθεί να επιλύσει το πρόβλημα της μετατροπής συμβόλων σε εικόνα, η Όραση Υπολογιστών προσπαθεί να λύσει το αντίστροφο πρόβλημα και από εικόνες να εξάγει συμπεράσματα για το είδος των μεταβλητών ενδιαφέροντος που υπάρχουν σε μια σκηνή (σχήμα, μέγεθος, υφή, χρώμα, ταυτότητα), τη θέση τους (τοποθεσία, κίνηση) καθώς και τις σχέσεις και ομοιότητες μεταξύ τους. Παραδείγματα μεταβλητών ενδιαφέροντος είναι τα αντικείμενα, οι άνθρωποι, τα πρόσωπα, οι χειρονομίες ή οι ανθρώπινες δράσεις.

Οι απαρχές της Όρασης Υπολογιστών μπορούν να εντοπιστούν στη δεκαετία του 1960 και στις προσπάθειες ερευνητών από το χώρο της Τεχνητής Νοημοσύνης να κατασκευάσουν υπολογιστές που αντιλαμβάνονται τον ορατό κόσμο όπως οι άνθρωποι. Παρά την πρόοδο που έχει συντελεστεί τα προηγούμενα χρόνια στον τομέα, απέχουμε πολύ από την κατασκευή ενός συστήματος που θα μπορεί να κατονομάσει και να σκιαγραφήσει όλα τα αντικείμενα σε μια φωτογραφία με την άνεση ενός μικρού παιδιού.

Η Όραση Υπολογιστών είναι ένα διεπιστημονικό πεδίο που συνδυάζει μεθόδους και αρχές από περιοχές όπως η Επεξεργασία Σημάτων, η Αναγνώριση Προτύπων, η Τεχνητή Νοημοσύνη, τα Εφαρμοσμένα Μαθηματικά και Φυσική αλλά και η Νευροβιολογία και Ψυχολογία. Η Όραση Υπολογιστών βρίθει εφαρμογών. Ενδεικτικά βρίσκει εφαρμογή στην επεξεργασία εικόνων,

τη ρομποτική, βιοϊατρική τεχνολογία και ιατρική απεικόνιση, την επικοινωνία ανθρώπου-υπολογιστή, την τηλεπισκόπιση, τα ευφυή συστηματα, τον κινηματογράφο και την τεχνολογία βίντεο.

Κάποια ενδεικτικά πρόβλήματα που επιλύονται με μεθόδους της όρασης υπολογιστών είναι:

- η αναγνώριση αντικειμένων
- η ανίχνευση και εκτίμηση οπτικής κίνησης
- η ανακατασκευή τρισδιάστατης δομής
- η κατάτμηση εικόνων
- η ανίχνευση προσώπου
- η ανίχνευση και κατηγοριοποίηση ανθρώπινων δράσεων
- η μοντελοποίηση και ανάλυση υφής, χρώματος και σχήματος.

1.2 Το πρόβλημα της Αναγνώρισης Ανθρώπινων Δράσεων

Οι ανθρώπινες δράσεις μπορεί να θεωρηθεί ότι αποτελούνται από διαδοχικές ή επαναλαμβανόμενες πρωταρχικές κινήσεις (motion primitives), δηλαδή απλές, βασικές κινήσεις οι οποίες μπορούν να συνδυαστούν και να διαταχθούν σχηματίζοντας μια δράση. Αυτή η θεώρηση ενισχύεται από ευρήματα της Νευροβιολογίας και συγκεκριμένα από το σύστημα των νευρώνων-καθρεπτών (mirror- neuron system), όπου ο κάθε νευρώνας ενεργοποιείται όταν εκτελείται ή παρατηρείται κάποια συγκεκριμένη πρωταρχική κίνηση και όλοι μαζί συνδυάζονται για να σχηματίσουν μια αλυσίδα ενεργοποιήσεων που αναπαριστά όλη τη δράση [1]. Παραδείγματος χάριν, ως motion primitive μπορεί να θεωρηθεί η κίνηση "πιάνω", η οποία μπορεί να οδηγήσει σε διαφορετικές ενεργοποιήσεις όταν αποτελεί μέρος διαφορετικών δράσεων (π.χ. "τρώω", "τοποθετώ αντικείμενο").

Το πρόβλημα της αναγνώρισης δράσεων σε βίντεο μπορεί να αφορά την απλή ταξινόμηση ενός βίντεο, δηλαδή στην απόδοση μιας ετικέτας σχετικά με τη δράση που εκτελείται στο βίντεο ή μπορεί να αφορά τον εντοπισμό των χρονικών διαστημάτων του βίντεο στα οποία εκτελείται κάποια δράση και την ταξινόμηση αυτής της δράσης στη σωστή κατηγορία ή ακόμα και το χωρικό εντοπισμό των δράσεων στα πλαίσια (frames) του βίντεο. Στην πρώτη περίπτωση η έξοδος του συστήματος είναι το όνομα της δράσης (π.χ.

το ρήμα *Περπατάει - Walk*), ενώ στη δεύτερη, όπου έχουμε συνεχόμενες δράσεις, είναι μια λίστα που περιέχει τα ονόματα των δράσεων με τη σειρά που αυτές εμφανίζονται στο βίντεο. Στο Σχήμα 1.1 απεικονίζονται μερικά ενδεικτικά frames των δράσεων *Κάθομαι - Sit*, *Τρώω - Eat* και *Αγκαλιάζω - Hug*.

(α′) Δράση *Sit*.

(β′) Δράση *Eat*.

(γ′) Δράση *Hug*.

Σχήμα 1.1: Ενδεικτικά frames από βίντεο ανθρώπινων δράσεων.

Εφαρμογές της αναγνώρισης δράσεων

Με νέα βίντεο διάρκειας 300 ωρών να ανεβαίνουν κατά μέσο όρο κάθε ώρα στο YouTube ανά τον κόσμο και με πληθώρα βίντεο, που αφορούν ταινίες, αθλητικά γεγονότα και τηλεοπτικά προγράμματα, να παράγονται καθημερινά, η ανάγκη αυτόματης επεξεργασίας τους με ακριβείς και αποδοτικούς αλγορίθμους για την εξαγωγή συμβολικών πληροφοριών (όπως τα ονόματα των δράσεων που εκτελούνται) είναι αδήριτη.

Ενδεικτικές εφαρμογές της ανάλυσης βίντεο για την αναγνώριση δράσεων αποτελούν η αλληλεπίδραση ανθρώπου-υπολογιστή, τα συστήματα παρακολούθησης (video surveillance) και η αυτόματη ανάκτηση βίντεο (video retrieval) από βάσεις δεδομένων [2]. Για παράδειγμα, μια εφαρμογή που μας

ενδιαφέρει είναι η ανίχνευση εγκληματικών ή ασυνήθιστων δράσεων σε περιβάλλοντα παρακολούθησης με κάμερες, το οποίο επιτρέπει την αυτόματη ειδοποίηση των αρμόδιων υπηρεσιών. Επίσης, βρίσκει μεγαλή εφαρμογή σε ανάλυση αθλητικών βίντεο, όπου η αναγνώριση συνεχόμενων δράσεων μπορεί να συνεισφέρει στην κατασκευή αυτόματων περιγραφών των αγώνων.

Σημαντικά πεδία της εφαρμογής μεθόδων αναγνώρισης δράσεων είναι η παρακολούθηση της υγείας και καθημερινής δραστηριότητας (health/daily-activity monitoring) ηλικιωμένων ατόμων, καθώς και η αποκατάσταση. Μάλιστα, τα τελευταία χρόνια με την εμφάνιση οπτικών αισθητήρων μικρού κόστους και υψηλής ευκρίνειας, όπως ο αισθητήρας Kinect, ο οποίος παρέχει πληροφορία RGB (έγχρωμες εικόνες), αλλά και πληροφορία βάθους (θέσεις των εικονιζόμενων αντικειμένων στον τρισδιάστατο κόσμο), όλο και περισσότερες εφαρμογές προσανατολίζονται στην παροχή βοήθειας στο σπίτι.

Η παρακολούθηση των καθημερινών δραστηριοτήτων των ηλικιωμένων έχει ως σκοπό να τους δώσει τη δυνατότητα να ζουν με ασφάλεια, άνεση και αυτονομία. Για να το καταφέρουν αυτό, τα περισσότερα συστήματα καταγράφουν συνεχώς τις κινήσεις των ηλικιωμένων, αναγνωρίζοντας αυτόματα τις δραστηριότητές τους και ανιχνεύοντας τόσο σταδιακές αλαλγές σε βασικές δράσεις, που μπορεί να οφείλονται σε κινησιακά ή γνωσιακά προβλήματα [3], [4], όσο και ασυνήθιστες δράσεις, όπως η πτώση [5]. Στην παρούσα διπλωματική εργασία πειραματιστήκαμε σε βάση δράσεων, η οποία έχει ληφθεί από την παρακολούθηση της δραστηριότητας ηλικιωμένων ατόμων με κινητικές διαταραχές. Στη συγκεκριμένη εφαρμογή, ένας ρομποτικός βοηθός παρακολουθεί τις κινήσεις ενός ασθενή με τη χρήση οπτικών αισθητήρων Kinect, οι οποίοι είναι τοποθετημένοι πάνω στον κινούμενο ρομποτικό βοηθό, έτσι ώστε να καταλάβει τις ανάγκες του ανθρώπου [6]. Σκοπός είναι, εκτός των άλλων, η παροχή βοήθειας κατά τη βάδιση (π.χ. στήριξη για ισορροπία) και κατά το σήκωμα/κάθισμα σε καρέκλα. Επίσης, ο ρομποτικός βοηθός θα ήταν επιθυμητό να μπορεί να αναγνωρίζει αυτόματα παθολογίες κίνησης κατά τη διάρκεια της βάδισης.

Η αυτόματη αναγνώριση δράσεων μπορεί να οδηγήσει στη μείωση του κόστους και στη βελτίωση της αποτελεσματικότητας της αποκατάστασης από τραυματισμούς ή άλλες ασθένειες, καθώς δίνεται η δυνατότητα εκτέλεσης των ασκήσεων της θεραπείας στο σπίτι. Αναγνωρίζοντας τις ασκήσεις και καταγράφοντας τις κινήσεις του ασθενούς, ο υπολογιστής μπορεί να προσφέρει feedback σε πραγματικό χρόνο και μπορεί να υπολογίζει και να στέλνει στους αρμόδιους φορείς τις τιμές διάφορων παραμέτρων της κίνησης, διευκολύνοντας την αξιολόγηση της προόδου της αποκατάστασης καθώς και την έγκαιρη διάγνωση παθολογιών [7]–[9].

Δυσκολίες/προκλήσεις της αναγνώρισης δράσεων

Όπως είναι αναμενόμενο, λόγω των πολλαπλών τεχνολογικών εφαρμογών του, το πρόβλημα της αναγνώρισης δράσεων είναι από τα πιο ενεργά ερευνητικά πεδία της Όρασης Υπολογιστών και της Αναγνώρισης Προτύπων. Ωστόσο, παρά τις πολυάριθμες ερευνητικές προσπάθειες και καινοτομίες, το πρόβλημα παραμένει ακόμα ανοιχτό. Αυτό εδράζεται στην ιδιαιτερότητα των δεδομένων και στην πολυπλοκότητα των δράσεων. Ενδεικτικά, μερικές από τις προκλήσεις που καλούνται να αντιμετωπίσουν οι αλγόριθμοι αναγνώρισης δράσεων είναι:

- Μεταβλητότητα του περιβάλλοντος (environment variation): συχνά στις σκηνές που εκτελούνται οι δράσεις συντελούνται άλλες κινήσεις στο υπόβαθρο. Π.χ. μπορεί να κινούνται αυτοκίνητα ή άλλοι άνθρωποι πίσω από τον άνθρωπο που εκτελεί τη δράση προς αναγνώριση. Επίσης, συχνές είναι και επικαλύψεις του ανθρώπου από άλλα αντικείμενα που τον καθιστούν μερικώς ή και καθόλου ορατό σε κάποια frames του βίντεο.

- Γωνίες Λήψης: η ίδια δράση μπορεί να καταγραφεί από διαφορετικές γωνίες στα διάφορα βίντεο, οδηγώντας σε διαφορετικές εικόνες σε κάθε χρονική στιγμή και συνεπώς διαφορετικές εκφάνσεις της ίδιας δράσης. Μάλιστα, η κίνηση της κάμερας κατά τη λήψη των βίντεο δυσκολεύει σημαντικά την αναγνώριση δράσεων, καθώς η κίνηση του ανθρώπου που εκτελεί την κίνηση αναμιγνύεται με αυτή της κάμερας.

- Μεταβλητότητα τρόπου ή/και διάρκειας εκτέλεσης: κάθε άνθρωπος εκτελεί μία δράση με το δικό του τρόπο και με διαφορετική ταχύτητα. Επομένως, τα στιγμιότυπα βίντεο μιας κατηγορίας δράσης εμφανίζουν έντονες διαφορές μεταξύ τους (intra-class variation).

- Πολλαπλά είδη δράσεων: οι δράσεις που θέλουμε να αναγνωρίσουμε καλύπτουν ένα μεγάλο φάσμα κινήσεων από κινήσεις των άνω άκρων (χειρονομίες - gestures) (π.χ. *Χαιρετώ*), μέχρι συνηθισμένες δράσεις που εκτελούνται από ένα άτομο (π.χ. *Σηκώνομαι*), δράσεις που περιλαμβάνουν αλληλεπίδραση ανθρώπου με άνθρωπο (π.χ. *Χειραψία*) ή με αντικείμενα (π.χ. *Κλωτσάω μπάλα* ή *Τρώω*) και ομαδικές δράσεις.

- Χωροχρονική επικάλυψη δράσεων: πολύ συχνά μπορεί να συμβαίνουν δράσεις ταυτόχρονα στον ίδιο χώρο (π.χ. καθώς εκτελείται χειραψία μεταξύ δύο ανθρώπων, ένας τρίτος άνθρωπος να σηκώνεται από την καρέκλα).

Μέθοδοι αναγνώρισης δράσεων

Για την επίλυση του δύσκολου προβλήματος της ταξινόμησης δράσεων ακολουθούνται ποικίλες προσεγγίσεις, οι οποίες μπορούν να κατηγοριοποιηθούν σε δύο μεγάλες οικογένειες μεθόδων [10]:

1. Χρήση διαχωριστικών ταξινομητών (discriminative classifiers): σε αυτή την προσέγγιση, δίνεται έμφαση στο διαχωρισμό μεταξύ δύο ή περισσότερων κλάσεων και όχι στη μοντελοποίηση τους. Ένας από τους πιο διαδεδομένους διαχωριστικούς ταξινομητές είναι ο αλγόριθμκος των Μηχανών Διανυσματικής Υποστήριξης (Support Vector Machines - SVMs), οι οποίες μαθαίνουν ένα υπερεπίπεδο στο χώρο των χαρακτηριστικών, το οποίο διαχωρίζει τα στιγμιότυπα δύο διαφορετικών κλάσεων. Επίσης, τα Βαθιά Νευρωνικά Δίκτυα (Deep Neural Nets) έχουν οδηγήσει σε ιδιαίτερα υψηλές επιδόσεις σε εφαρμογές ταξινόμησης εικόνων και αναγνώρισης αντικειμένων.

2. Χρήση μοντέλων του χρονικού χώρου καταστάσεων (temporal state-space models): Τα μοντέλα χώρου καταστάσεων αποτελούνται από καταστάσεις (states) που ενώνονται με ακμές. Αυτές οι ακμές μοντελοποιούν τις πιθανότητες μετάβασης από κατάσταση σε κατάσταση και τις πιθανότητες μεταξύ καταστάσεων και παρατηρήσεων. Στο συγκεκριμένο πρόβλημα, κάθε κατάσταση αντιστοιχεί σε μία φάση δράσης σε μία συγκεκριμένη χρονική στιγμή. Τα μοντέλα αυτά μπορεί να είναι διαχωριστικά (discriminative) ή αναγεννητικά (generative). Τα generative μοντέλα μαθαίνουν την από κοινού κατανομή πάνω από τις παρατηρήσεις και τις ετικέτες των δράσεων. Έτσι μαθαίνουν να μοντελοποιούν μία συγκεκριμένη δράση με όλες τις παραλλαγές της. Αντίθετα, τα διαχωριστικά μοντέλα μαθαίνουν την πιθανότητα των δράσεων δεδομένων των παρατηρήσεων. Έτσι, δε μοντελοποιούν μία συγκεκριμένη δράση, αλλά εστιάζουν στις διαφορές μεταξύ των δράσεων. Ένα από τα πιο διαδεδομένα αναγεννητικά μοντέλα είναι τα Κρυφά Μαρκοβιανά Μοντέλα (Hidden Markov Models - HMMs), τα οποία έχουν χρησιμοποιηθεί και στον τομέα της αναγνώρισης δράσεων. Τα Conditional Random Fields (CRFs) είναι ένα παράδειγμα διαχωριστικών μοντέλων που έχουν δοκιμαστεί σε εφαρμογές αναγνώρισης δράσεων [11].

Παρά τις πολλά υποσχόμενες μεθόδους που χρησιμοποιούν μοντέλα του χώρου καταστάσεων, τα καλύτερα αποτελέσματα αναγνώρισης δράσεων σε διεθνείς, απαιτητικές βάσεις δράσεων έχουν επιτευχθεί από μεθόδους που χρησιμοποιούν SVMs [12] ή Βαθιά Νευρωνικά Δίκτυα [13]. Στην παρούσα εργασία θα βασιστούμε στο ισχυρό εργαλείο των Μηχανών Διανυσματικής Υποστήριξης.

Για την ταξινόμηση βίντεο με χρήση SVM ταξινομητών, εκπαιδεύουμε τα διαχωριστικά μοντέλα σε ένα σύνολο βίντεο εκπαίδευσης (training set), καθένα από τα οποία συνοδεύεται από την επισημείωση της δράσης που περιέχει. Ο σκοπός της εκπαίδευσης είναι να μπορούν τα μοντέλα να διαχωρίσουν τις διαφορετικές κλάσεις αλλά και να έχουν την ικανότητα γενίκευσης (generalization), δηλαδή να μπορούν να ταξινομήσουν σωστά εκτελέσεις της ίδιας δράσης με μεγάλη μεταβλητότητα. Για να αξιολογήσουμε την ικανότητα γενίκευσης των μοντέλων μας, ελέγχουμε κατά πόσο μπορούν να ταξινομήσουν σωστά τα βίντεο ενός συνόλου αξιολόγησης (testing set), τα οποία είναι διαφορετικά από αυτά πάνω στα οποία εκπαιδεύτηκαν.

Φυσικά, για να είναι σε θέση τα διαχωριστικά μοντέλα να ταξινομήσουν σωστά ένα βίντεο, πρέπει να λάβουν στην είσοδο τους μια αναπαράσταση του βίντεο, η οποία να έχει ενσωματωμένη όσο το δυνατόν περισσότερη πληροφορία σχετικά με τη δράση που εκτελείται. Το πρώτο βήμα για την απόκτηση αυτής της πληροφορίας είναι η ανίχνευση των σημείων/περιοχών ενδιαφέροντος του βίντεο, οι οποίες περιέχουν χρήσιμη πληροφορία (feature detection). Για παράδειγμα, μας ενδιαφέρουν τα σημεία που ανήκουν στα κινούμενα μέλη των ανθρώπων. Στη συνέχεια, πρέπει να περιγράψουμε κατάλληλα τις ιδιότητες των χωροχρονικών περιοχών γύρω από τα σημεία ενδιαφέροντος. Ιδιαίτερα σημαντικές είναι οι ιδιότητες της στατικής εμφάνισης και της κίνησης κάθε χωροχρονικής γειτονιάς. Αυτές οι ιδιότητες κωδικοποιούνται με τη χρήση ειδικών *περιγραφητών - descriptors*. Με τον υπολογισμό των περιγραφητών ολοκληρώνεται η *διαδικασία εξαγωγής χαρακτηριστικών* (feature extraction). Εντούτοις, τα χαρακτηριστικά σε αυτή τη μορφή τους δεν είναι κατάλληλες είσοδοι για τον αλγόριθμο ταξινόμησης, μιας και λόγω του διαφορετικού αριθμού χαρακτηριστικών που εξάγονται από κάθε βίντεο και των συνήθως μεγάλων διαστάσεων τους, είναι αρκετά δύσκολο να συγκριθούν άμεσα μεταξύ τους. Γι'αυτό είναι απαραίτητος ο υπολογισμός μιας *αναπαράστασης βίντεο* (video representation). Για τον υπολογισμό αυτού του διανύσματος αναπαράστασης κάθε βίντεο χρειάζεται εν γένει να προηγηθεί η κατασκευή ενός λεξικού "οπτικών λέξεων" από τη συσταδοποίηση (clustering) των χαρακτηριστικών. Ένα παράδειγμα μιας τέτοιας αναπαράστασης είναι το ιστόγραμμα Bag-Of-Visual-Words [14], το οποίο είναι ένα ιστόγραμμα συχνοτήτων εμφάνισης των οπτικών λέξεων στο βίντεο.

1.3 Διάρθρωση της Διπλωματικής Εργασίας

Στην παρούσα διπλωματική εργασία μελετάμε το πρόβλημα της αναγνώρισης δράσεων υπό το πρίσμα των αναπαραστάσεων βίντεο, αναλύοντας κάποιες από τις πιο δημοφιλείς υπάρχουσες μεθόδους αναπαράστασης βίντεο, συγκρίνοντας τις επιδόσεις τους στο πρόβλημα του χρονικού εντοπισμού και ταξινόμησης ανθρώπινων δράσεων σε μια νέα πολυαισθητηριακή βάση δεδομένων και τέλος, προτείνοντας δύο νέες μεθόδους αναπαράστασης που έρχονται να συμπληρώσουν τις υπάρχουσες μεθόδους αξιοποιώντας συμπληρωματική πληροφορία σε σχέση με αυτές.

Πιο συγκεκριμένα, το περιεχόμενο της διπλωματικής είναι οργανωμένο σε κεφάλαια ως εξής:

Το Κεφάλαιο 2 αφορά μεθόδους εξαγωγής χαρακτηριστικών (ανιχνευτές και περιγραφητές). Μετά από μια συνοπτική αναφορά στις διάφορες μεθόδους που έχουν προταθεί στη βιβλιογραφία, αναλύονται διεξοδικά τα δημοφιλή χαρακτηριστικά *πυκνών τροχιών* (dense trajectories), καθώς και η βελτιωμένη παραλλαγή αυτών που είναι εύρωστη στην κίνηση της κάμερας. Τα χαρακτηριστικά αυτά χρησιμοποιήθηκαν στα πειράματά μας και σε αυτά βασίσαμε τις νέες μεθόδους αναπαράστασης βίντεο που προτείνουμε.

Στο Κεφάλαιο 3 μελετάμε υπάρχουσες μεθόδους αναπαράστασης βίντεο. Αρχικά, παρουσιάζουμε δύο ισχυρά εργαλεία για την κατασκευή του "οπτικού λεξικού": τους αλγορίθμους συσταδοποίησης K-μέσων (K-means clustering) και Μοντέλου Μείγματος Γκαουσιανών (Gaussian Mixture Model clustering). Στη συνέχεια, αναλύονται και συγκρίνονται θεωρητικά τρεις αναπαραστάσεις βίντεο: η απλή μέθοδος Bag-Of-Visual Words (BoVW) και οι state-of-the-art μέθοδοι: Vector of Locally Aggregated Descriptors (VLAD) και διάνυσμα Fisher (FV). Επίσης, αναφερόμαστε σε μεθόδους προεπεξεργασίας των χαρακτηριστικών που βελτιώνουν τις προκύπτουσες αναπαραστάσεις. Τέλος, θίγουμε συνοπτικά κάποιες αδυναμίες αυτών των μεθόδων, τις οποίες θα προσπαθήσουμε να διορθώσουμε προτείνοντας νέες συμπληρωματικές μεθόδους αναπαράστασης δράσεων.

Στο Κεφάλαιο 4 αναλύεται ο τρόπος λειτουργίας των Μηχανών Διανυσματικής Υποστήριξης (SVMs), οι οποίες χρησιμοποιούνται στην εργασία μας ως αλγόριθμος ταξινόμησης των βίντεο σε κατηγορίες δράσεων. Ιδιαίτερη έμφαση δίνεται σε συναρτήσεις πυρήνα, οι οποίες χρησιμοποιούνται ευρέως στη βιβλιογραφία για την ταξινόμηση αναπαραστάσεων BoVW, VLAD και FV. Τέλος, αναφερόμαστε σε τρόπους σύμμειξης ροών πληροφορίας, όπως πολλαπλών περιγραφητών ή αναπαραστάσεων.

Στο Κεφάλαιο 5 ασχολούμαστε με το πρόβλημα της αναγνώρισης συνεχό-

μενων ανθρώπινων δράσεων. Εστιάζουμε κυρίως στον τρόπο που επεκτείναμε το σύστημα ταξινόμησης δράσεων. Παρατίθενται τα αποτελέσματα από τον πειραματισμό μας με διαφορετικές μεθόδους εξαγωγής χαρακτηριστικών καθώς και αναπαραστάσεων σε μία νέα πολυαισθητηριακή βάση δράσεων με δεδομένα από ηλικιωμένους, τη βάση MOBOT [1]. Τέλος, για να διευκολύνουμε την αναγνώριση, συνδυάζουμε τις πυκνές τροχιές με την πληροφορία βάθους (depth), εξάγοντας έναν κατάλληλο περιγραφητή εμφάνισης από το διαθέσιμο κανάλι βάθους.

Στο Κεφάλαιο 6 προτείνεται μια βελτίωση της μέθοδου αναπαράστασης BoVW, η οποία κωδικοποιεί τις σχέσεις μεταξύ των συστάδων οπτικών χαρακτηριστικών, χρησιμοποιώντας μια κατευθυνόμενη μετρική της ομοιότητας μεταξύ των clusters που βασίζεται στη μέθοδο Ανάλυσης Κύριων Συνιστωσών (Principal Component Analysis - PCA). Εξετάζουμε δύο παραλλαγές της μεθόδου που αξιοποιούν δύο διαφορετικούς αλγορίθμους συσταδοποίησης. Στο τέλος του κεφαλαίου παρατίθενται τα πειραματικά αποτελέσματα στη δημοφιλή βάση KTH και η μέθοδος συγκρίνεται με γνωστές αναπαραστάσεις.

Στο Κεφάλαιο 7 προτείνουμε μια νέα μέθοδο αναπαράστασης βίντεο, η οποία αξιοποιεί τη χρονική αλληλουχία των οπτικών λέξεων. Κατασκευάζουμε υπο-ακολουθίες από συχνά εμφανιζόμενες οπτικές λέξεις, οι οποίες εν τέλει συνενώνονται σε μια τελική ακολουθία οπτικών λέξεων που αναπαριστά μια δράση. Επίσης προτείνουμε και μια μετρική της ομοιότητας μεταξύ των διανυσμάτων που προκύπτουν με αυτή τη μέθοδο αναπαράστασης, βασισμένη σε αλγόριθμο τοπικής στοίχισης των ακολουθιών, έτσι ώστε να γίνει δυνατή η αξιοποίησή της από μηχανές διανυσματικής υποστήριξης. Τέλος, εξετάζουμε τη σύμμειξη της νέας μεθόδου αναπαράστασης με τη μέθοδο BoVW. Τα αποτελέσματα των πειραμάτων μας στην ευρέως χρησιμοποιούμενη βάση δεδομένων δράσεων KTH και την απαιτητική βάση μεγάλης κλίμακας HMDB51 δείχνουν ότι η προτεινόμενη αναπαράσταση είναι συμπληρωματική της αναπαράστασης BoVW και ο συνδυασμός τους οδηγεί σε αποτελέσματα συγκρίσιμα με αυτά των καλύτερων μεθόδων της τρέχουσας διεθνούς βιβλιογραφίας, πολλές εκ των οποίων χρησιμοποιούν πιο πολύπλοκα μοντέλα.

Στο Κεφάλαιο 8 συνοψίζονται οι κυριότερες συνεισφορές και συμπεράσματα της παρούσας διπλωματικής καθώς και οι βασικές κατευθύνσεις για μελλοντική έρευνα.

[1] `http://www.mobot-project.eu/`

Κεφάλαιο 2

Εξαγωγή Χαρακτηριστικών

Στο κεφάλαιο αυτό παρουσιάζονται μέθοδοι της βιβλιογραφίας που χρησιμοποιούνται για την εξαγωγή χαρακτηριστικών εμφάνισης και κίνησης από τα βίντεο ανθρώπινων δράσεων. Θα αναλυθούν διεξοδικά οι μέθοδοι εξαγωγής χαρακτηριστικών "πυκνών τροχιών" και "βελτιωμένων πυκνών τροχιών", οι οποίες έχουν διαδεχτεί στη διεθνή βιβλιογραφία τις μεθόδους των χωροχρονικών σημείων ενδιαφέροντος (Spatio-temporal Interest Points - STIPs). Στο κεφάλαιο 5 θα παρουσιαστούν τα πειραματικά αποτελέσματα που προκύπτουν από τη σύγκριση των μεθόδων Dense Trajectories και improved Dense Trajectories στην απαιτητική βάση MOBOT [1], σε βίντεο με έντονη κίνηση κάμερας. Επίσης, οι πυκνές τροχιές χρησιμοποιούνται και σε όλα τα υπόλοιπα πειράματα αυτής της διπλωματικής εργασίας, ως μία αξιόπιστη μέθοδος εξαγωγής χαρακτηριστικών που ενσωματώνουν πλούσια πληροφορία.

2.1 Εισαγωγή

Ο στόχος της εξαγωγής χαρακτηριστικών είναι η εξαγωγή χρήσιμης πληροφορίας από το βίντεο, το οποίο δεν είναι άλλο παρά μια αλληλουχία από εικόνες (πίνακες φωτεινοτήτων), σχετικά με τη δράση που εκτελείται, η οποία μπορεί να αφορά την πόζα του ανθρώπου/κινούμενων αντικειμένων ή την κίνηση (ταχύτητα).

Για την εξαγωγή χαρακτηριστικών χρειάζονται δύο βήματα:

1. Ανίχνευση χαρακτηριστικών: Πρέπει να προσδιοριστούν τα χωροχρονικά σημεία ενδιαφέροντος γύρω από τα οποία μας ενδιαφέρει να παρατηρήσουμε τη στατική εμφάνιση και την κίνηση. Για παράδειγμα, ιδανικά θα θέλαμε να συλλέγουμε πληροφορία που να σχετίζεται με την κίνηση του

[1] `http://www.mobot-project.eu/`

ανθρώπου που πραγματοποιεί τη δράση, και επομένως να περιοριστούμε σε σημεία ενδιαφέροντος της εικόνας όπου βρίσκεται ο άνθρωπος. Επίσης, επειδή συχνά οι δράσεις περιλαμβάνουν την αλληλεπίδραση με αντικείμενα, όπως για παράδειγμα η δράση *Πίνω* συνδέεται με το ποτήρι, θα μας ενδιέφερε να καταγράψουμε πληροφορία σχετικά με την εμφάνιση των αντικειμένων κ.ο.κ.

2. Υπολογισμός περιγραφητών: Για να κωδικοποιήσουμε π.χ. την πληροφορία στατικής εμφάνισης γύρω από ένα σημείο ενδιαφέροντος, χρειαζόμαστε κάποιο κατάλληλο περιγραφητή, ο οποίος να ενσωματώνει αυτή την πληροφορία σε ένα διάνυσμα συγκεκριμένου μήκους. Οι τοπικοί περιγραφητές συνήθως εξάγονται σε χωροχρονικές γειτονιές που περιέχουν ανιχνευθέντα σημεία ενδιαφέροντος. Συνήθως κωδικοποιούν πληροφορία σχετικά με τα 2D/3D gradients ή/και την οπτική ροή των χωροχρονικών όγκων γύρω από κάθε σημείο ενδιαφέροντος.

Στην επόμενη ενότητα, θα αναφερθούμε σε μια πληθώρα μεθόδων της βιβλιογραφίας για ανίχνευση χαρακτηριστικών και εξαγωγή περιγραφητών, ενώ θα εστιάσουμε στα χαρακτηριστικά των "πυκνών τροχιών", τα οποία χρησιμοποιήσαμε ως μέθοδο εξαγωγής χαρακτηριστικών στα πειράματά μας και την πληροφορία των οποίων προσπαθήσαμε να αξιοποιήσουμε πιο αποδοτικά από τις υπάρχουσες μεθόδους αναπαράστασης βίντεο της βιβλιογραφίας.

2.2 Σχετική βιβλιογραφία

Την τελευταία δεκαετία οι προσπάθειες για την κατασκευή τοπικών χωροχρονικών περιγραφών κινήθηκαν από τη χρήση μεθόδων αραιών τοπικών χωροχρονικών σημείων ενδιαφέροντος στην πυκνή δειγματοληψία και πρόσφατα στη χρήση πυκνών τροχιών. Όσον αφορά τους περιγραφητές, χρησιμοποιήθηκαν κυρίως περιγραφητές βασισμένοι στις κλίσεις (gradients) και στις ιδιότητες της κίνησης.

Οι ανιχνευτές τοπικών χωροχρονικών σημείων ενδιαφέροντος (spatio-temporal interest points - STIPs) βρίσκουν θέσεις και κλίμακες όπου είναι επιθυμητό να εξαχθούν χαρακτηριστικά μεγιστοποιώντας συγκεκριμένες συναρτήσεις σημαντικότητας (saliency functions). Για παράδειγμα, ο Harris3D ανιχνευτής [15] επεκτείνει τον δισδιάστατο Harris ανιχνευτή, ο οποίος ανιχνεύει γωνίες στις εικόνες. Υπολογίζεται ένας χωροχρονικός πίνακας ροπών δεύτερης τάξης σε κάθε σημείο και αναζητούνται σημεία τα οποία έχουν μεγάλες ιδιοτιμές του πίνακα. Τα σημεία ενδιαφέροντος λοιπόν είναι τοπικά μέγιστα ενός κριτηρίου γωνιότητας που βασίζεται στον πίνακα ροπών δεύτερης τάξης. Τα σημεία αυτά διαισθητικά είναι χωρικά σημεία ενδιαφέροντος με συγκεκριμένη θέση

στο χρόνο που αντιστοιχεί στις στιγμές με μη σταθερή κίνηση της εικόνας σε μία τοπική χωροχρονική γειτονιά. Στους αλγορίθμους ανίχνευσης STIPs συγκαταλέγεται και ο ανιχνευτής Cuboids που προτάθηκε από τους Dollar et al. [16] το 2005 και αντικαθιστά το κριτήριο γωνιότητας της μεθόδου Harris3D με το κριτήριο $R = (I * g * h_{ev})^2 + (I * g * h_{od})^2$, όπου $g(x, y; \sigma)$ είναι δισδιάστο Γκαουσιανό φίλτρο ομαλοποίησης, εφαρμοσμένο στο πεδίο του χώρου (space domain), ενώ h_{ev} και h_{od} είναι ένα quadrature pair μονοδιάστατων φίλτρων Gabor που εφαρμόζονται στο πεδίο του χρόνου (time domain). Ο ανιχνευτής Hessian προτάθηκε από τους Willems et al. [17] ως μία χωροχρονική επέκταση του μέτρου σημαντικότητας Hessian που χρησιμοποιείται για ανίχνευση blobs (κηλίδων) στις εικόνες. Ο ανιχνευτής μετρά τη σημαντικότητα κάθε σημείου χρησιμοποιώντας την ορίζουσα του 3D Hessian πίνακα. Άλλος ένας ανιχνευτής είναι ο Gabor3D [18], ο οποίος φιλτράρει το βίντεο με συστοιχία Gabor φίλτρων, ξεχωριστά για κάθε διάσταση. Τα σημεία ενδιαφέροντος επιλέγονται ως τα τοπικά μέγιστα στις τρεις διαστάσεις της ενέργειας που προκύπτει.

Όσον αφορά τους περιγραφητές που υπολογίζονται σε όγκους κατάλληλης χωροχρονικής κλίμακας κεντραρισμένους γύρω από τα τοπικά χωροχρονικά σημεία ενδιαφέροντος (x, y, t) έχουν προταθεί πληθώρα μεθόδων υπολογισμού τους. Έχουν χρησιμοποιηθεί παράγωγοι υψηλής τάξης (local jets) [14] και περιγραφητές βασισμένοι σε πληροφορία φωτεινοτήτων, κλίσεων και οπτικής ροής [16]. Ο περιγραφητής HOG3D που αναπτύχθηκε από τους Klaeser et al. [19] βασίζεται σε ιστογράμματα των κατευθύνσεων 3D κλίσεων και μπορεί να ειδωθεί ως μια επέκταση του περιγραφητή SIFT σε βίντεο. Κανονικά πολύεδρα χρησιμοποιούνται για να κβαντιστούν ομοιόμορφα οι κατευθύνσεις των χωροχρονικών κλίσεων. Έτσι, ο περιγραφητής συνδυάζει ταυτόχρονα πληροφορία σχήματος και κίνησης. Οι Willems et al. [17] πρότειναν τη χρήση του ESURF περιγραφητή, ο οποίος επεκτείνει τον περιγραφητή εικόνων SURF [20] σε βίντεο. Δύο άλλοι δημοφιλείς περιγραφητές είναι οι HOG/HOF [21], [22], οι οποίοι για να περιγράψουν την τοπική κίνηση και εμφάνιση βασίζονται στον υπολογισμό ιστογραμμάτων χωρικών κλίσεων και οπτικής ροής σε χωροχρονικές γειτονιές γύρω από τα ανιχνευμένα σημεία ενδιαφέροντος.

Οι Wang et al. συνέκριναν το 2013 [23] κάποιους από τους πιο διαδεδομένους ανιχνευτές τοπικών χωροχρονικών σημείων ενδιαφέροντος, καταλήγοντας στο συμπέρασμα ότι η καταλληλότητα κάθε ανιχνευτή εξαρτάται από τη βάση δεδομένων και τον περιγραφητή που χρησιμοποιείται και ότι εν γένει παρουσιάζουν παραπλήσιες επιδόσεις, με τον Cuboids ανιχνευτή να οδηγεί σε υψηλότερες ακρίβειες αναγνώρισης δράσεων σε απαιτητικές βάσεις όπως η UCFSports και η Hollywood2D. Ωστόσο, η πιο χρήσιμη παρατήρηση που έκαναν είναι ότι η χρήση απλής πυκνής δειγματοληψίας, δηλαδή η χρήση των pixels των εικόνων που ανήκουν σε ένα σταθερό χωροχρονικό πλέγμα οδηγεί σε καλύτερα αποτελέσματα σε σχέση με την επιλογή σημείων ενδιαφέροντος

βάσει πολύπλοκων συναρτήσεων οπτικής σημαντικότητας (saliency), και επομένως με πολύ μεγαλύτερη υπολογιστική ταχύτητα.

Παράλληλα με την ιδέα της πυκνής δειγματοληψίας, πολλές εργασίες ασχολήθηκαν με τη χρήση "τροχιών" ως χαρακτηριστικών. Αντί να ανιχνεύουμε σημεία στο χωροχρονικό όγκο που ορίζει ένα βίντεο, μπορούμε να ανιχνεύουμε σημεία ενδιαφέροντος στις εικόνες και να τα παρακολουθούμε στο χρόνο (tracking). Εξάλλου διαισθητικά οι ανθρώπινες δράσεις μπορούν να αναπαρασταθούν ως χωροχρονικές τροχιές. Διαφορετικές δράσεις έχουν διαφορετικά μοτίβα τροχιών. Κατά τη διάρκεια ενός βίντεο, ένα σημείο, το οποίο ανήκει π.χ. στο πόδι ενός ανθρώπου που περπατάει, διαγράφει μια τροχιά στο χρόνο. Η χρήση αυτών των τροχιών επιτρέπει το διαχωρισμό της επιλογής σημείων στο δισδιάστατο πεδίο χώρου (space domain) από το πεδίο του χρόνου (time domain), καθώς τα πεδία αυτά έχουν προφανώς διαφορετικές ιδιότητες. Έτσι, οι Matikainen et al. [24] και οι Messing et al. [25] χρησιμοποίησαν την οπτική ροή Lucas-Kanade για να παρακολουθήσουν στο χρόνο σημεία ενδιαφέροντος που είχαν ανιχνευθεί από τον "good features to track" detector [26] και από τον Harris3D detector, ενώ οι Sun et al. [27] χρησιμοποίησαν τις αντιστοιχίες ανάμεσα σε SIFT περιγραφητές που είχαν εξαχθεί ανάμεσα σε διαδοχικά frames για να σχηματίσουν τις τροχιές. Οι Sun et al. το 2010 [28] συνδύασαν τις δύο προηγούμενες προσεγγίσεις. Εκτός από την παρακολούθηση αραιών σημείων, οι Brox και Malik [29] εξήγαγαν τροχιές από όλο το βίντεο χρησιμοποιώντας πυκνή οπτική ροή για να διαχωρίσουν τα διάφορα αντικείμενα αντικείμενα σε ένα βίντεο.

Συνδυάζοντας τις ιδέες της πυκνής δειγματοληψίας και των τροχιών, οι Wang et al. εισήγαγαν το 2011 [30] τη μέθοδο των Πυκνών Τροχιών (Dense Trajectories), οι οποίες βασίζονται στην παρακολούθηση στο χρόνο σημείων από ένα πυκνό πλέγμα με χρήση της πυκνής οπτικής ροής. Στη συνέχεια, εξάγονται περιγραφητές στατικής εμφάνισης και κίνησης, όπως HOG, HOF, κατά μήκος κάθε τροχιάς. Η μέθοδός τους οδήγησε σε σημαντικές βελτιώσεις της ακρίβειας αναγνώρισης σε πολλές απαιτητικές βάσεις. Το 2013 οι Wang et al. [31] πρότειναν μία βελτιωμένη έκδοση της μεθόδου τους (improved Dense Trajectories) κατά την οποία αφαιρούνται τροχιές οι οποίες οφείλονται στην κίνηση της κάμερας και διορθώνεται η οπτική ροή. Με το πρόβλημα της αντιμετώπισης της κίνησης της κάμερας ασχολήθηκαν και οι Jain et al. [32], οι οποίοι προτείνουν μια αποσύνθεση της οπτικής κίνησης σε κυρίαρχες (dominant) και δευτερεύουσες (residual) κινήσεις.

Η χρήση των Συνελικτικών Νευρωνικών Δικτύων (Convolutional Neural Networks) ως ταξινομητή ή ως μεθόδου εξαγωγής χαρακτηριστικών έχει οδηγήσει σε ραγδαία πρόοδο στο πεδίο της ταξινόμησης εικόνων [33], χωρίς να έχει οδηγήσει ωστόσο σε αντίστοιχα μεγάλες βελτιώσεις στον τομέα της αναγνώρισης δράσεων [13], [34].

2.3 Πυκνές τροχιές (Dense Trajectories)

Οι Wang et al. ανέπτυξαν τη μέθοδο των πυκνών τροχιών (dense trajectories), κατά την οποία παρακολουθούν με χρήση πυκνής οπτικής ροής την κίνηση σημείων της εικόνας στο χρόνο, εξάγοντας τροχιές και υπολογίζοντας περιγραφητές εμφάνισης και ταχύτητας στο χωροχρονικό όγκο κατά μηκος της τροχιάς. Επίσης εισήγαγαν ένα νέο περιγραφητή κίνησης, ο οποίος είναι εύρωστος στην κίνηση της κάμερας μιας και βασίζεται στην παράγωγο της οπτικής ροής. Πριν παρουσιαστούν τα βασικά βήματα της μεθόδου, είναι χρήσιμο να αναφερθούμε στην οπτική ροή και σε τρόπους υπολογισμού της, καθώς αποτελεί βασική συνιστώσα του αλγορίθμου.

2.3.1 Οπτική Ροή

Η 3D κίνηση των αντικειμένων της σκηνής του πραγματικού κόσμου που καταγράφεται προκαλεί μια 2D κίνηση των αντίστοιχων μοτίβων φωτεινότητας στο επίπεδο της εικόνας. Αυτή η φαινομενική 2D κίνηση ονομάζεται οπτική ροή (optical flow) [35].

Ένας από τους πιο γνωστούς αλγορίθμους υπολογισμού της οπτικής ροής είναι ο αλγόριθμος Lucas-Kanade [36], ο οποίος βασίζεται στις μερικές χωρικές και χρονικές παραγώγους του σήματος εικόνας και μπορεί να υπολογίσει την οπτική ροή για ένα αραιό σύνολο σημείων. Ο στόχος του αλγορίθμου είναι να αντιστοιχιστεί μια εικόνα-τεμπλέτα $T(\mathbf{x})$ με μία εικόνα εισόδου $I(\mathbf{x})$, όπου $\mathbf{x}$ είναι το διάνυσμα-στήλη που περιέχει τις συντεταγμένες των pixels. Αν ο αλγόριθμος των Lucas-Kanade χρησιμοποιείται για να υπολογίσει τη μετατόπιση ενός τμήματος μιας εικόνας μεταξύ δύο διαδοχικών frames, η τεμπλέτα $T(\mathbf{x})$ είναι η το τμήμα της εικόνας στο $n-1$-οστό frame $I_{n-1}(\mathbf{x})$ και η εικόνα $I(\mathbf{x})$ είναι το ίδιο τμήμα της εικόνας στο επόμενο frame $I(\mathbf{x}) = I_n(\mathbf{x})$. Αναζητούμε το διάνυσμα απόστασης $\mathbf{d}$ το οποίο ελαχιστοποιεί κάποιο μέτρο της διαφοράς μεταξύ της μετατόπισης της I_n ($I_n(\mathbf{x}+\mathbf{d})$) και της τεμπλέτας I_{n-1}, για τα $\mathbf{x}$ εντός της περιοχής ενδιαφέροντος. Αυτό μπορούμε να το επιτύχουμε ελαχιστοποιώντας κάποιο κριτήριο σφάλματος, όπως την L_2 νόρμα του σφάλματος αντιστοίχισης, αφού πρώτα εφαρμόσουμε γκαουσιανό παράθυρο στάθμισης, με τυπική απόκλιση ρ που δίνει μεγαλύτερη έμφαση στα κεντρικά pixel της περιοχής και έτσι έχουμε:

$$J_{\mathbf{x}}(\mathbf{d}) = \int\limits_{\mathbf{x}' \in \mathbb{R}^2} G_\rho(\mathbf{x}-\mathbf{x}') \left[I_n(\mathbf{x}'+\mathbf{d}) - I_{n-1}(\mathbf{x}')\right]^2 d\mathbf{x} \tag{2.1}$$

Ο αλγόριθμος κάνει την υπόθεση ότι υπάρχει μια τρέχουσα εκτίμηση του διανύσματος οπτικής ροής $\mathbf{d}_i$ και μετέπειτα σε κάθε επανάληψη προσπαθεί να

εκτιμήσει τις μεταβολές $\mathbf{u}$ στη μετατόπιση. Αναπτύσσοντας κατά Taylor την έκφραση $I_{n-1}(\mathbf{x}+\mathbf{d}_i+\mathbf{u})$ γύρω από το σημείο $\mathbf{x}+\mathbf{d}_i$ προκύπτει ότι:

$$I_{n-1}(\mathbf{x}+\mathbf{d}) \approx I_{n-1}(\mathbf{x}+\mathbf{d}_i) + \nabla I_{n-1}(\mathbf{x}+\mathbf{d}_i)^T\mathbf{u} \tag{2.2}$$

και επομένως συμπεραίνουμε ότι η μέθοδος Lucas-Kanade μπορεί να εφαρμοστεί για μικρές κινήσεις/μετατοπίσεις, έτσι ώστε να ισχύει η προσέγγιση. Αντικαθιστώντας αυτή την προσέγγιση στο κριτήριο σφάλματος και ελαχιστοποιώντας το, προκύπτει η εξίσωση που δίνει τη μεταβολή $\mathbf{u}$ και την οποία εφαρμόζουμε επαναληπτικά έως ότου συγκλίνει για κάθε ζεύγος διαδοχικών εικόνων του βίντεο.

Η μέθοδος Lucas-Kanade υπολογίζει την οπτική ροή για ένα αραιό σύνολο σημείων ενδιαφέροντος (π.χ. τις γωνίες που εμφανίζονται στις εικόνες). Στην περίπτωση των πυκνών τροχιών επιθυμούμε να υπολογίσουμε την οπτική ροή για όλα τα pixels της εικόνας. Πιο συγκεκριμένα, δεδομένων δύο διαδοχικών frames που έχουν ληφθεί με μικρή χρονική διαφορά, θέλουμε να βρούμε την αλλαγή της θέσης κάθε pixel του πρώτου frame στο δεύτερο. Ένας δημοφιλής αλγόριθμος υπολογισμού πυκνής οπτικής ροής είναι ο αλγόριθμος του Gunner Farneback [37]. Η μέθοδος εκμεταλλεύεται την προσέγγιση των φωτεινοτήτων της χωρικής γειτονιάς κάθε pixel των δύο γειτονικών frames με τη χρήση πολυωνυμικών αναπτυγμάτων (polynomial expansion). Στη συνέχεια, λαμβάνοντας υπόψη την επίδραση μιας μετατόπισης σε ένα πολυώνυμο, αναπτύσσουν μια μέθοδο, η οποία μπορεί να εκτιμά τα πεδία μετατόπισης (displacement fields) από τους συντελεστές των πολυωνυμικών αναπτυγμάτων. Με άλλα λόγια, δεδομένου ενός αρχικού πολυωνύμου $f(\mathbf{x})$ κατασκευάζουν ένα νέο πολυώνυμο, το οποίο έχει προέλθει από τη μετατόπιση του αρχικού $f(\mathbf{x}-\mathbf{d})$ και εξισώνοντας τους συντελεστές των δύο πολυωνύμων βρίσκουν τη μετατόπιση $\mathbf{d}$.

2.3.2 Ανίχνευση χαρακτηριστικών

Για την ανίχνευση των χαρακτηριστικών, δηλαδή την κατασκευή των τροχιών, οι Wang et al. προβαίνουν σε πυκνή δειγματοληψία της εικόνας, η οποία ακολουθείται από την παρακολούθηση της τροχιάς των σημείων στο χρόνο βάσει της πυκνής οπτικής ροής.

Πυκνή δειγματοληψία Η μέθοδος ξεκινά δειγματοληπτώντας σημεία ενδιαφέροντος σε ένα πυκνό πλέγμα σε κάθε frame, με βήμα δειγματοληψίας W pixels. Το βήμα δειγματοληψίας καθορίζει πόσο πυκνές θα είναι οι τροχιές και διαπιστώθηκε πειραματικά ότι η αποτελεσματικότητα της μεθόδου μειώνεται σημαντικά όσο αυξάνεται το W. Η δειγματοληψία εκτελείται σε

πολλαπλές χωρικές κλίμακες, το πολύ 8 ανάλογα με την ανάλυση του βίντεο, με την κλίμακα να αυξάνεται κάθε φορά με παράγοντα $\frac{1}{\sqrt{2}}$. Η πυκνή δειγματοληψία αφορά τις μη ομοιογενείς περιοχές του βίντεο, έτσι ώστε η οπτική ροή να είναι μη μηδενική και να μπορούμε να παρακολουθήσουμε την κίνηση των σημείων στο χρόνο. Τα σημεία που τελικά παρακολουθούνται στο χρόνο υπολογίζονται με βάση το κριτήριο Shi & Tomasi [26]. Το κριτήριο αυτό ανιχνεύει τις προεξέχουσες γωνίες σε μία εικόνα ακολουθώντας τα εξής βήματα:

1. Υπολογίζει ένα κριτήριο γωνιότητας σε κάθε pixel της εικόνας:

$$R(x, y) = \min(\lambda_+, \lambda_-) \tag{2.3}$$

   όπου λ_+, λ_- είναι οι ιδιοτιμές του πίνακα $G_\rho * \left\{ \nabla I_\sigma (\nabla I_\sigma)^T \right\}$ με $I_\sigma = G_\sigma * I$ και G_σ, G_ρ δισδιάστατοι Γκαουσιανοί πυρήνες με τυπικές αποκλίσεις σ και ρ αντίστοιχα.

2. Χρησιμοποιώντας το κριτήριο αυτό, δηλαδή την τιμή του σε κάθε σημείο της εικόνας, επιλέγουμε τα pixels τα οποία ικανοποιούν δύο κριτήρια: α) έχουν τη μέγιστη τιμή του R εντός των 3×3 παραθύρων που τα περιβάλλουν και β) για ένα κατώφλι q, αντιστοιχούν σε τιμή του R μεγαλύτερη της ποσότητας $q \max_{x,y} R(x, y)$.

3. Αυτές οι γωνίες ταξινομούνται σε φθίνουσα σειρά βάσει του κριτηρίου γωνιότητας.

4. Απορρίπτονται οι γωνίες των οποίων η Ευκλείδεια απόσταση από "ισχυρότερες" γωνίες είναι μικρότερη από ένα κατώφλι. Εν προκειμένω, αυτό το κατώφλι W ορίζει το βήμα της δειγματοληψίας.

Τροχιές Αφού επιλεχθούν τα σημεία, παρακολουθούμε την κίνηση τους σε κάθε χωρική κλίμακα ξεχωριστά. Για κάθε frame I_t υπολογίζουμε το πεδίο της οπτικής ροής $\mathbf{d}_t = (d_x, d_y)$ ως προς το επόμενο frame I_{t+1}, όπου d_x, d_y είναι η οριζόντια και κάθετη συνιστώσα της οπτικής ροής αντίστοιχα. Έστω σημείο $P_t = (x_t, y_t)$ στο frame I_t. Η προβλεπόμενη θέση του σημείου στο frame I_{t+1} υπολογίζεται από το πεδίο οπτικής ροής που έχει ομαλοποιηθεί με median φιλτράρισμα:

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + \text{med}_W(\mathbf{d}_t)|_{(x_t, y_t)} \tag{2.4}$$

όπου W είναι ένα 3×3 τετραγωνικό παράθυρο.

Μια *τροχιά* (trajectory) σχηματίζεται από τη συνένωση των θέσεων ενός pixel που παρακολουθείται σε L διαδοχικά frames: $(P_t, P_{t+1}, \ldots, P_{t+L-1})$. Ο

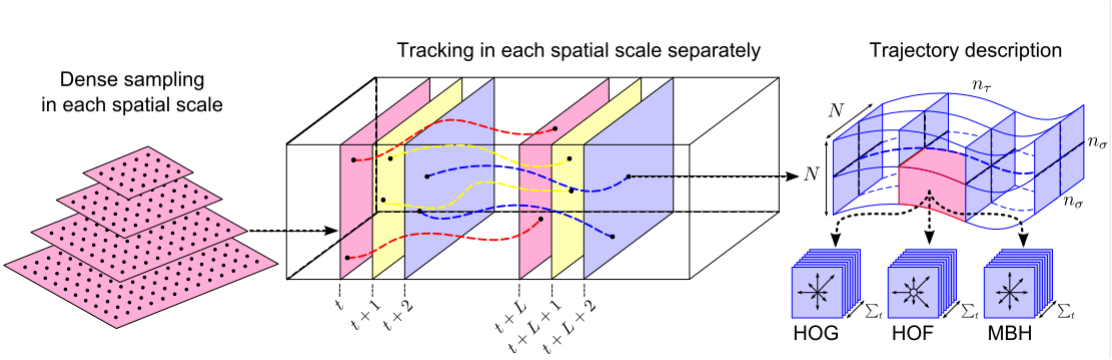
λόγος που περιορίζουν οι συγγραφείς το μήκος της τροχιάς σε L σημεία, είναι η τάση των τροχιών να ολισθαίνουν από τις αρχικές τους θέσεις κατά τη διάρκεια του tracking. Μια ενδεικτική τιμή του μήκους των τροχιών είναι 15 frames. Για κάθε frame, αν δεν υπάρχει κάποιο σημείο που παρακολουθείται σε γειτονιά $W \times W$, διαλέγουμε ένα νέο σημείο και αρχίζουμε να το παρακολουθούμε, εξασφαλίζοντας ένα πυκνό σύνολο τροχιών. Επίσης, τροχιές με πολύ μικρή μετατόπιση (στατικές) ή πολύ μεγάλη μετατόπιση αφαιρούνται.

2.3.3 Περιγραφητές

Αφού συλλέχθηκαν οι τροχιές, πρέπει να περιγράψουμε κατάλληλα τις ιδιότητές τους και της γειτονιάς τους. Για αυτό το σκοπό εξάγονται πληθώρα περιγραφητών.

Περιγραφητής Trajectory Ο περιγραφητής αυτός περιγράφει το σχήμα μιας τροχιάς, το οποίο σχετίζεται προφανώς με το μοτίβο της κίνησης. Υπολογίζεται ως το κανονικοποιημένο διάνυσμα των διανυσμάτων μετατόπισης $\Delta P_t = (P_{t-1} - P_t) = (x_{t+1} - x_t, y_{t+1} - y_t)$.

$$T = \frac{(\Delta P_t, \ldots, \Delta P_{t+L-1})}{\sum_{j=1}^{t+L-1} \|\Delta P_j\|} \tag{2.5}$$



Σχήμα 2.1: Διαδικασία εξαγωγής οπτικών χαρακτηριστικών πυκνών τροχιών [38].

Εκτός από τον περιγραφητή Trajectory, υπολογίζονται δισδιάστατοι περιγραφητές κίνησης και δομής εντός ενός όγκου, ο οποίος είναι ευθυγραμμισμένος με την τροχιά, σε αντίθεση με παλαιότερες μεθόδους που υπολόγιζαν τοπικούς περιγραφητές σε 3D όγκους γύρω από τα σημεία ενδιαφέροντος.

Για κάθε τροχιά, υπολογίζονται περιγραφητές μέσα στον όγκο διάστασης $N \times N$ pixels και L frames, ο οποίος διαιρείται σε ένα χωροχρονικό πλέγμα με $n_\sigma \times n_\sigma \times n_\tau$ κελιά, όπως απεικονίζεται και στο Σχήμα 2.1. Πιο συγκεκριμένα, για κάθε τροχιά υπολογίζονται οι περιγραφητές μέσα σε υποδιαιρέσεις του αρχικού όγκου, οι οποίες έχουν η κάθεμια διάσταση $\frac{N}{n_\sigma} \times \frac{N}{n_\sigma}$ pixels και $\frac{L}{n_\tau}$ frames. Η τελική περιγραφή της τροχιάς προκύπτει από τη συνένωση αυτών των περιγραφητών όλων των υποδιαιρέσεων. Εφόσον χρησιμοποιούνται 2D περιγραφητές, αυτοί υπολογίζονται για κάθε υποδιαίρεση στις γειτονιές $\frac{N}{n_\sigma} \times \frac{N}{n_\sigma}$ pixels και μετέπειτα αθροίζονται στη διάσταση του χρόνου. Οι παράμετροι N, n_σ και n_τ ρυθμίζουν το μέγεθος της χωρικής γειτονιάς γύρω από κάθε σημείο της τροχιάς όπου εξάγονται οι περιγραφητές και τον τρόπο που διαμερίζεται ο όγκος χωρικά και χρονικά αντίστοιχα. Ενδεικτικές τιμές είναι $N = 32$ pixels, $n_\sigma = 2$ και $n_\tau = 3$. Επειδή οι όγκοι που εξάγονται οι περιγραφητές γύρω από τις τροχιές είναι συχνά αλληλοεπικαλυπτόμενοι, το μέγεθος της γειτονιάς δεν επηρεάζει σημαντικά το αποτέλεσμα.

Περιγραφητής HOG Ένας από τους πιο δημοφιλείς περιγραφητές εμφάνισης είναι το Ιστόγραμμα προσανατολισμένης κλίσης (Histogram of Oriented Gradients - HOG) που εισήχθη από τους Dalal και Triggs [21] για την ανίχνευση ανθρώπων. Το σχήμα των αντικειμένων και οι σιλουέτες των ανθρώπων μπορούν να μοντελοποιηθούν από τις τιμές των τοπικών gradients, δηλαδή από τις ακμές της εικόνας. Ο περιγραφητής HOG είναι ένα ιστόγραμμα τοπικών κλίσεων. Για το υπολογισμό του πρέπει αρχικά να υπολογιστούν οι κλίσεις ως προς τους άξονες x και y με κέντρο κάθε pixel της γειτονιάς που υπολογίζεται ο περιγραφητής. Αν και μπορούν να χρησιμοποιηθούν μονοδιάστατες μάσκες για τον υπολογισμό των κλίσεων, όπως η $[-1, 0, 1]$, εν προκειμένω χρησιμοποιούνται 3×3 μάσκες Sobel:

$$I_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} * I, I_y = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} * I \tag{2.6}$$

Βάσει των προσανατολισμένων κλίσεων I_x και I_y μπορούμε να υπολογίσουμε το μέτρο m και την κατεύθυνση (γωνία) θ της κλίσης σε κάθε pixel:

$$\begin{aligned} m(x, y) &= \sqrt{I_x^2(x, y) + I_y^2(x, y)} \\ \theta(x, y) &= \arctan(\frac{I_y(x, y)}{I_x(x, y)}) \end{aligned} \tag{2.7}$$

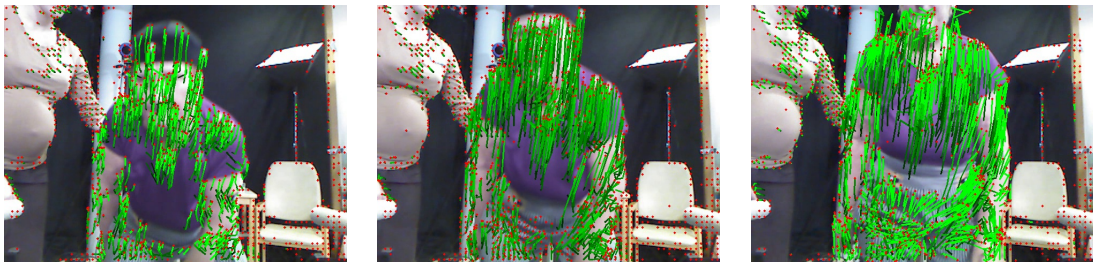
Στη συνέχεια, θέλουμε να κατασκευάσουμε ένα ιστόγραμμα που θα απεικονίζει την κατανομή των κλίσεων των pixels μιας γειτονιάς στις διάφορες γωνίες.

Κάθε ράβδος/bin του ιστογράμματος αντιστοιχεί σε ένα εύρος γωνιών, π.χ. το πρώτο σε 0°- 22.5°, το δεύτερο σε 22.5°- 45°κ.ο.κ. Κάθε pixel υπολογίζει μια σταθμισμένη ψήφο για τη ράβδο του ιστογράμματος στην οποία ανήκει ο προσανατολισμός της κλίσης που είναι κεντραρισμένη στο pixel. Το βάρος της ψήφου ισούται με το μέτρο της κλίσης m. Συχνά χρησιμοποιείται και διγραμμική παρεμβολή έτσι ώστε το κάθε pixel να μπορεί να ψηφίζει σε δύο ράβδους. Οι ψήφοι από τα pixels μιας γειτονιάς συσσωρεύονται στα bins του ιστογράμματος. Εν προκειμένω κατασκευάζεται ιστόγραμμα 8 bins. Τέλος, είναι πολύ σημαντικό να κανονικοποιηθεί το ιστόγραμμα. Η αρχική κανονικοποίηση που είχε προταθεί από τους Dalal και Triggs ήταν η κανονικοποίηση με την ℓ_2 νόρμα του ιστογράμματος: $\mathbf{h}' = \frac{\mathbf{h}}{\sqrt{\|\mathbf{h}\|^2+\epsilon}}$, όπου ϵ μικρή θετική σταθερά. Ωστόσο, οι Wang et al. χρησιμοποιούν μια πρόσφατη προσέγγιση, την κανονικοποίηση RootSIFT [39], που πρώτα πραγματοποιεί κανονικοποίηση με την ℓ_1 νόρμα και στη συνέχεια υπολογισμό της τετραγωνικής ρίζας κάθε στοιχείου του ιστογράμματος, $\mathbf{h}' = \sqrt{\frac{\mathbf{h}}{\sum_{i=1}^{8} |h_i|}}$ και η οποία αποδείχθηκε ότι βελτιώνει σημαντικά την επίδοση του περιγραφητή.

Όπως προαναφέραμε, ο περιγραφητής HOG, όπως και οι υπόλοιποι περιγραφητές εμφάνισης και κίνησης, υπολογίζεται σε $n_\sigma \times n_\sigma \times n_\tau$ υποδιαιρέσεις του όγκου που περιβάλλει την τροχιά και στη συνέχεια τα αποτελέσματα συνενώνονται σε έναν τελικό περιγραφητή διάστασης $n_\sigma \times n_\sigma \times n_\tau \times 8 = 96$ στοιχείων.

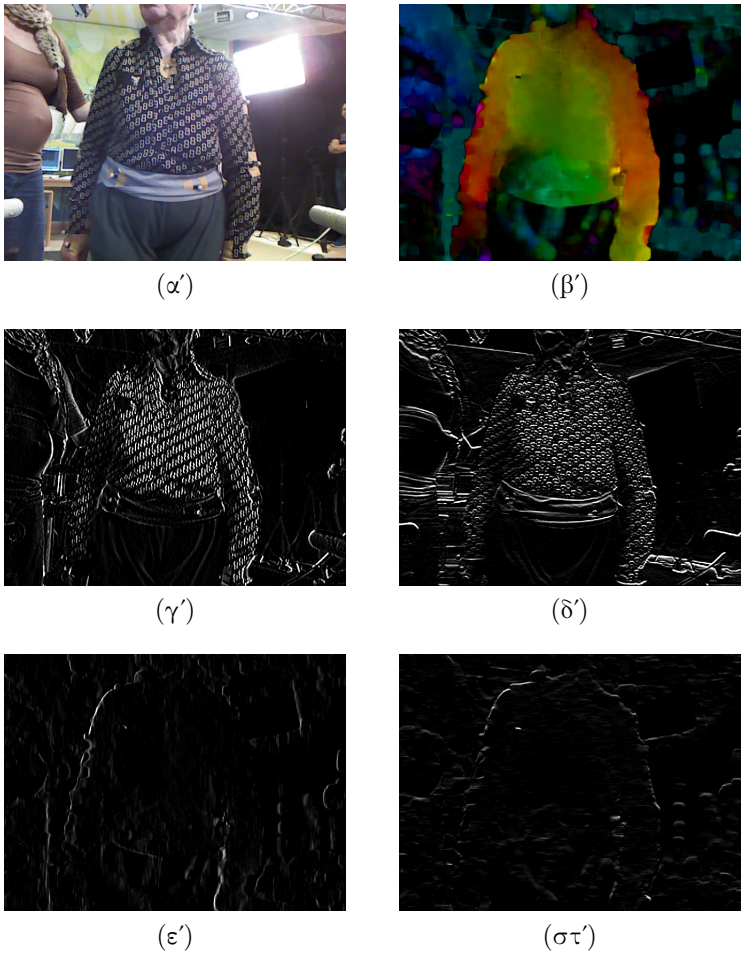
Περιγραφητής HOF Ο περιγραφητής HOF (Histogram Of Optical Flow ή Ιστόγραμμα Οπτικής Ροής) [22] εκμεταλλεύεται την οπτική ροή για να περιγράψει την τοπική κίνηση σε ένα μικρό χωρίο. Αφού υπολογιστούν οι δύο συνιστώσες της οπτικής ροής σε κάθε pixel, από τις κατευθύνσεις τους και τα πλάτη τους κατασκευάζουμε το ιστόγραμμα οπτικής ροής, με τον ίδιο τρόπο που κατασκευάστηκε στον περιγραφητή HOG. Η μόνη διαφορά είναι ότι στον HOF χρησιμοποιούμε ένα ακόμη bin, δηλαδή συνολικά 9, στο οποίο ανατίθενται τα pixels των οποίων τα πλάτη της οπτικής ροής είναι μικρότερα από ένα κατώφλι. Για τον περιγραφητή HOF χρησιμοποιείται η μέθοδος κανονικοποίησης RootSIFT και η τελική διάσταση του περιγραφητή για κάθε τροχιά είναι $n_\sigma \times n_\sigma \times n_\tau \times 9 = 108$ στοιχεία.

Περιγραφητής MBH Ο περιγραφητής MBH (Motion Boundary Histogram ή Ιστόγραμμα Ορίων Κίνησης) ήταν μια προσπάθεια των Dalal et al. [40] να βελτιώσουν τον περιγραφητή HOF, έτσι ώστε να λαμβάνει υπόψη τη σχετική κίνηση μεταξύ των pixels και να μην επηρεάζεται από την κίνηση μεταξύ των frames που οφείλεται σε άλλες πηγές, όπως η κίνηση της κάμερας. Αρχικά



Σχήμα 2.2: Παραδείγματα πυκνών τροχιών εξαγμένα από βίντεο της δράσης "Σηκώνομαι".

υπολογίζονται οι δύο συνιστώσες της οπτικής ροής σε κάθε pixel, έχουμε δηλαδή στη διάθεση μας για κάθε frame τις εικόνες οι οποίες περιέχουν τις x (οριζόντια) και y (κάθετη) συνιστώσες. Στη συνέχεια, υπολογίζουμε τις χωρικές παραγώγους αυτών των δύο εικόνων ως προς x και y και από τις κατευθύνσεις και το πλάτος των gradients κατασκευάζουμε κατά τα γνωστά τα ιστογράμματα MBHx και MBHy, δηλαδή ένα για κάθε συνιστώσα. Εφόσον ο MBH περιγραφητής αναπαριστά την κλίση της οπτικής ροής, οποιαδήποτε τοπικά σταθερή κίνηση της κάμερας αφαιρείται και διατηρείται μόνο η πληροφορία σχετικά με τις αλλαγές στο πεδίο ροής. Συνεπώς, αυτός ο περιγραφητής είναι πιο εύρωστος στην κίνηση της κάμερας και έτσι περιγράφει καλύτερα τις κινήσεις που απαρτίζουν μια δράση.



(α′) (β′) (γ′) (δ′) (ε′) (στ′)

Σχήμα 2.3: Απεικόνιση της πληροφορίας που ενσωματώνουν οι περιγραφητές HOG, HOF και MBH. (α′) Frame ενός βίντεο. (β′) Πυκνή οπτική ροή στην οποία βασίζεται ο περιγραφητής HOF. (γ′), (δ′) Μερικές παράγωγοι ενός frame ως προς τις κατευθύνσεις x και y στις οποίες βασίζεται ο περιγραφητής HOG. (ε′),(στ′) Μερικές παράγωγοι της οπτικής ροής ως προς x και y στις οποίες βασίζονται οι περιγραφητές MBHx και MBHy αντίστοιχα.

2.4 Βελτιωμένες πυκνές τροχιές (improved Dense Trajectories)

Παρά την ευρωστία του περιγραφητή MBH στην κίνηση της κάμερας, το πρόβλημα της αλλοίωσης της οπτικής ροής από την κίνηση της κάμερας συνε-

χίζει να επιδρά αρνητικά τόσο στην επίδοση των περιγραφητών κίνησης, όσο και στην επιλογή των τροχιών, οι οποίες θα θέλαμε να είναι συγκεντρωμένες στις περιοχές της εικόνας που αφορούν κινήσεις ανθρώπων και αντικειμένων σχετικών με τη δράση. Οι Wang et al. το 2013 πρότειναν μια βελτιωμένη εκδοχή των πυκνών τροχιών, η οποία βασίζεται στην εκτίμηση της κίνησης της κάμερας και την αφαίρεσή της από την οπτική ροή. Για να εκτιμήσουν την κίνηση του υποβάθρου (background) που οφείλεται στην κίνηση της κάμερας, οι Wang et al. κάνουν την υπόθεση ότι κάθε ζεύγος διαδοχικών frames συνδέεται από μία ομογραφία. Η ομογραφία είναι ένας προβολικός μετασχηματισμός, ο οποίος αναπαρίσταται ως ένας 3×3 πίνακας $\mathbf{H}$. Αν $\mathbf{x} = (u, v, 1)$ και $\mathbf{x}' = (u', v, 1)$ είναι προβολές του ίδιου σημείου του 3D κόσμου που ανήκουν σε ένα επίπεδο, εκφρασμένες σε ομογενείς συντεταγμένες, τότε συνδέονται από έναν πίνακα $\mathbf{H}$ που αντιστοιχεί σε αυτό το επίπεδο:

$$\mathbf{x}' = \mathbf{H}\mathbf{x} \tag{2.8}$$

Στα βίντεο ανθρώπινων δράσεων, η υπόθεση ότι τα διαδοχικά frames συνδέονται με ομογραφία ισχύει στις περισσότερες περιπτώσεις, αφού η κίνηση ανάμεσα σε δύο frames είναι συνήθως μικρή. Φυσικά, τα σημεία που ανήκουν σε ανεξάρτητα κινούμενα αντικείμενα, όπως ανθρώπους και αντικείμενα, δε συμφωνούν με την ομογραφία. Όταν μερικά pixels από το επίπεδο που έχει ειδωθεί από δύο διαφορετικές γωνίες είναι διαθέσιμα, τότε ο $\mathbf{H}$ μπορεί να υπολογιστεί. Για κάθε ζευγάρι αντιστοιχισμένων σημείων $\mathbf{x}, \mathbf{x}'$, η εξίσωση (2.8) δίνει δύο ανεξάρτητες γραμμικές εξισώσεις που μπορούν να γραφτούν και ως [41]:

$$\mathbf{x}' \times \mathbf{H}\mathbf{x} = 0 \tag{2.9}$$

Εφόσον ο πίνακας $\mathbf{H}$ έχει 8 βαθμούς ελευθερίας, αρκούν 4 αντιστοιχίες σημείων για να προσδιοριστεί. Όπως είναι προφανές, σε ρεαλιστικές εφαρμογές η χρήση μόνο 4 τυχαία επιλεγμένων ζευγαριών σημείων δεν είναι αρκετή για μια καλή προσέγγιση της ομογραφίας που πραγματικά συνδέει τα frames. Επομένως, μπορούν να χρησιμοποιηθούν N_{cor} αντιστοιχίες σημείων (point correspondences) που οδηγούν σε $2N_{cor}$ γραμμικούς περιορισμούς μέσω της εξίσωσης 2.9. Αυτό καταλήγει σε ένα σύστημα της μορφής $\mathbf{B}\mathbf{h} = 0$, όπου $\mathbf{h} = (h_{11}, h_{12}, h_{13}, h_{21}, h_{22}, h_{23}, h_{31}, h_{32}, h_{33})^T$. Άρα πρέπει να λυθεί το πρόβλημα ελαχιστοποίησης:

$$\min_{\mathbf{h}} \|\mathbf{B}\mathbf{h}\|^2 \quad s.t. \ \|\mathbf{h}\| = 1 \tag{2.10}$$

Η λύση του είναι κατά τα γνωστά τα ιδιοδιανύσματα του $\mathbf{B}^T\mathbf{B}$ που αντιστοιχούν στη μικρότερη ιδιοτιμή.

Μια εναλλακτική λύση για την εύρεση της ομογραφίας είναι η χρήση του γρήγορου και εύρωστου επαναληπτικού αλγόριθμου RANSAC (Random

Sample Consensus) [42]. Η βασική ιδέα του αλγορίθμου είναι ότι τα δεδομένα εισόδου (εν προκειμένω τα ζεύγη αντιστοιχισμένων σημείων) μπορούν να διαχωριστούν σε inliers, δηλαδή δεδομένα που η κατανομή τους μπορεί να εξηγηθεί βάσει κάποιων παραμέτρων ενός μοντέλου (εν προκειμένω ζεύγη τα οποία συμφωνούν με την ομογραφία) και σε outliers, δηλαδή δεδομένα που δεν ταιριάζουν στο συγκεκριμένο μοντέλο, είτε επειδή είναι θορυβώδη, είτε επειδή έχουν γίνει λανθασμένες υποθέσεις για την ερμηνεία των δεδομένων. Ο RANSAC είναι αρκετά εύρωστος, έτσι ώστε να μπορεί να υπολογίζει τις παραμέτρους ενός μοντέλου στο οποίο μπορεί να ταιριάζουν ακόμα και μικρός αριθμός inliers, λόγω πολύ θορυβώδων δεδομένων.

Για την εκτίμηση της με τον αλγόριθμο RANSAC, δεδομένου ενός συνόλου από υποψήφια ζεύγη εκτελούνται τα ακόλουθα βήματα:

1. Επιλέγονται 4 ζευγάρια σημείων από το σύνολο των υποψήφιων αντιστοιχίσεων και υπολογίζεται μια ομογραφία μέσω της (2.9).

2. Τα ζεύγη που συμφωνούν με την ομογραφία που υπολογίσθηκε (inliers) αποθηκεύονται. Ένα ζεύγος $(\mathbf{x}, \mathbf{x}')$ θεωρούμε ότι συμφωνεί με την ομογραφία αν για κάποια μικρή σταθερά ϵ ισχύει: $dist(\mathbf{Hx}, \mathbf{x}') < \epsilon$.

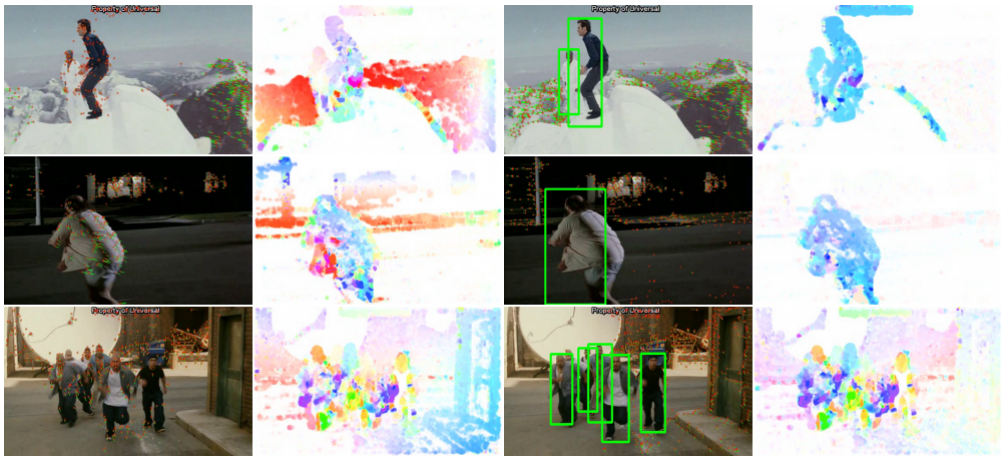
3. Τα δύο προηγούμενα βήματα επαναλαμβάνονται μέχρι να βρεθεί μια ομογραφία με ικανοποιητικό αριθμό από inlier δεδομένα, τα οποία είναι συνεπή με την υπολογισμένη ομογραφία.

4. Χρησιμοποιώντας το μεγαλύτερο σύνολο συνεπών αντιστοιχίσεων (inlier matches), υπολογίζεται η τελική ομογραφία με χρήση της (2.10).

Για την εκτίμηση της ομογραφίας χρειάζεται να βρούμε λοιπόν αντιστοιχίες σημείων μεταξύ κάθε ζεύγους συνεχόμενων frames. Οι συγγραφείς συνδυάζουν δύο μεθόδους για να παράξουν αρκετές και συμπληρωματικές αντιστοιχίες. Εξάγουν SURF [20] χαρακτηριστικά από τα frames και τα αντιστοιχίζουν βάσει της μεθόδου του κοντινότερου γείτονα (nearest neighbor rule). Τα SURF χαρακτηριστικά επιλέχθηκαν λόγω της ευρωστίας τους στη θόλωση της εικόνας εξαιτίας της κίνησης (motion blur). Εκτός από αυτά τα χαρακτηριστικά, επιλέγονται κάποια διανύσματα οπτικής ροής, έτσι ώστε να βρεθούν πυκνές αντιστοιχίες ανάμεσα στα frames. Τα σημεία των οποίων τη μετατόπιση καταγράφουν τα επιλεγμένα διανύσματα οπτικής ορής επιλέγονται με χρήση του κριτηρίου "good features to track" [26], το οποίο βρίσκει ισχυρές γωνίες σε μια εικόνα. Άρα τα χαρακτηριστικά που αντιστοιχίζονται είναι συμπληρωματικά, αφού όπως είδαμε τα διανύσματα οπτικής ροής επιλέγονται βάσει των γωνιών, ενώ ο ανιχνευτής SURF εστιάζει σε δομές τύπου blob (κηλίδων). Έτσι συλλέγεται μια ισορροπημένη κατανομή αντιστοιχισμένων

σημείων, που είναι ιδιαίτερα σημαντική για μια καλή εκτίμηση της ομογραφίας.

Στο Σχήμα 2.4 απεικονίζονται τα inlier matches για κάποια ενδεικτικά frames από βίντεο δράσεων και η στρεβλωμένη (warped) οπτική ροή, όπου έχει γίνει προσπάθεια αντιστάθμισης της κίνησης της κάμερας. Παρατηρούμε ότι η κίνηση του υποβάθρου έχει περιοριστεί σε αρκετά μεγάλο βαθμό, ενώ έχει τονιστεί η κίνηση του ανθρώπου. Στη στήλη 2 του σχήματος παρατηρούμε τη διορθωμένη οπτική ροή, όπου όμως βλέπουμε ότι δεν έχει αφαιρεθεί τελείως η κίνηση της κάμερας από την οπτική ροή, δηλαδή η εκτίμηση της ομογραφίας βάσει των inlier matches δεν ήταν βέλτιστη. Αυτό μπορεί να εξηγηθεί αν κανείς παρατηρήσει τα inlier matches των εικόνων. Τότε θα δει ότι πολλά από τα ζεύγη σημείων ανήκουν στις σιλουέτες των κινούμενων ανθρώπων, και επομένως η εκτιμούμενη ορθογραφία είναι λανθασμένη. Γι'αυτό οι συγγραφείς προτείνουν τη χρήση ενός ανιχνευτή ανθρώπων, του οποίου η έξοδος είναι ορθογώνια πλαίσια (bounding boxes) που περιβάλλουν τους ανθρώπους και μας επιτρέπουν να αγνοήσουμε τα inlier matches σημείων που βρίσκονται εντός τους. Τα αποτελέσματα με τον ανιχνευτή ανθρώπων είναι σαφώς βελτιωμένα, όπως βλέπουμε στις δύο τελευταίες στήλες του σχήματος.



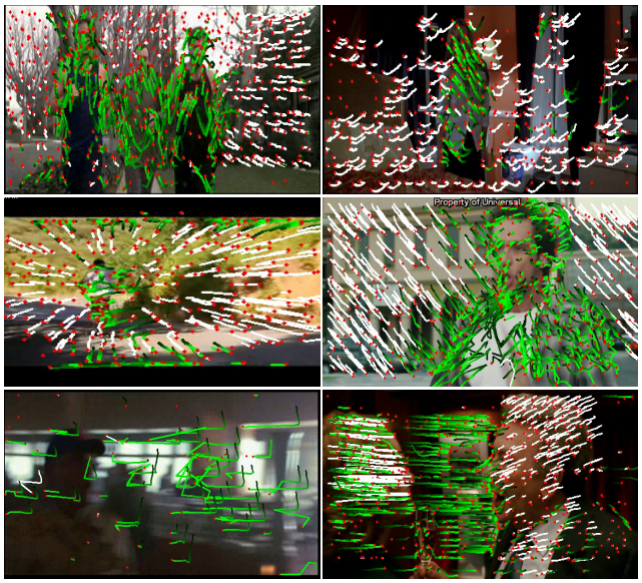
Σχήμα 2.4: Εκτίμηση ομογραφίας χωρίς της χρήση ανιχνευτή ανθρώπων (αριστερά) και με χρήση ανιχνευτή ανθρώπων (δεξιά). Τα inlier matches του αλγορίθμου RANSAC απεικονίζονται στην πρώτη και τρίτη στήλη. Η οπτική ροή (δεύτερη και τέταρτη στήλη) έχει στρεβλωθεί βάσει της εκτιμούμενης ομογραφίας. [38]

Αφού υπολογιστεί η όσο το δυνατόν περισσότερο απαλλαγμένη από την κίνηση της κάμερας οπτική ροή, ο αλγόριθμος των βελτιωμένων πυκνών τροχιών ανιχνεύει κατά τα γνωστά τροχιές, παρακολουθώντας τα σημεία στο χρόνο με

χρήση της διορθωμένης οπτικής ροής. Το όφελος από τη διορθωμένη οπτική ροή είναι διπλό:

- Οι περιγραφητές που βασίζονται στην οπτική ροή (HOF και MBH) περιγράφουν καλύτερα την κίνηση των ανθρώπων στο προσκήνιο. Ενώ ο MBH ήταν ήδη κάπως πιο εύρωστος στην κίνηση της κάμερας, καθώς χρησιμοποιεί την παράγωγο της οπτικής ροής, ο HOF επηρεαζόταν σημαντικά από αυτή και έτσι βελτιώθηκε.

- Οι τροχιές που οφείλονται στην κίνηση της κάμερας μπορούν να αφαιρεθούν. Αυτό επιτυγχάνεται με την κατωφλίωση των διανυσμάτων μετατόπισης στο διορθωμένο πεδίο οπτικής ροής. Αν η μετατόπιση των σημείων είναι πάρα πολύ μικρή, η τροχιά θεωρείται παρόμοια με την κίνηση της κάμερας και απορρίπτεται. Στο Σχήμα 2.5 βλέπουμε κάποια ενδεικτικά αποτελέσματα που όντως οι τροχιές εστιάζουν στις περιοχές του προσκηνίου, οι οποίες έχουν μεγάλη σημαντικότητα (motion saliency).



Σχήμα 2.5: Παραδείγματα τροχιών που αφαιρούνται μετά τη διόρθωση τις οπτικής ροής. Οι λευκές τροχιές θεωρείται ότι αντιστοιχούν στην κίνηση της κάμερας και γι'αυτό αφαιρούνται. Τα κόκκινα σημεία είναι οι θέσεις των τροχιών στο τρέχον frame. Η τελευταία γραμμή δείχνει δύο περιπτώσεις αποτυχίας. Η αριστερή οφείλεται σε έντονη θόλωση της εικόνας λόγω κίνησης, ενώ η δεύτερη σε λάθος εκτίμηση της ομογραφίας λόγω των inlier matches που αφορούν τον κινούμενο άνθρωπο, ο οποίος κυριαρχεί στην εικόνα. [38]

Κεφάλαιο 3

Αναπαραστάσεις video

Στο κεφάλαιο αυτό παρουσιάζονται οι κυριότερες μέθοδοι της βιβλιογραφίας που χρησιμοποιούνται για την αναπαράσταση των βίντεο με συμπαγή τρόπο, δεδομένων των χαρακτηριστικών που έχουν εξαχθεί κατά το προηγούμενο στάδιο. Οι μέθοδοι αυτές έχουν ως σκοπό τη συγκέντρωση στατιστικών που αφορούν την κατανομή των χαρακτηριστικών στο χώρο χαρακτηριστικών και οδηγούν σε ένα διάνυσμα αναπαράστασης του βίντεο, το οποίο στη συνέχεια θα αποτελέσει την είσοδο της μεθόδου ταξινόμησης. Θα αναλυθούν οι μέθοδοι Bag-Of-Visual-Words (BoVW) [14], Vector of Locally Aggregated Descriptors (VLAD) [43] και Fisher Vector (FV) [44] καθώς και τα βασικά μαθηματικά εργαλεία που χρησιμοποιούνται για την υλοποίησή τους. Στο κεφάλαιο 5 θα παρουσιαστούν τα πειραματικά αποτελέσματα που προκύπτουν με τη χρήση των μεθόδων BoVW και VLAD και των διάφορων κανονικοποιήσεών για την αναγνώριση ανθρώπινων δράσεων στην πολυ-αισθητηριακή βάση MOBOT [1]. Επιπρόσθετα, με αυτές τις μεθόδους συγκρίθηκαν οι προτεινόμενες μέθοδοι αναπαράστασης video της παρούσας διπλωματικής εργασίας (κεφάλαια 6, 7).

3.1 Εισαγωγή

Έχοντας εξαγάγει ένα σύνολο τοπικών περιγραφητών (στους οποίους θα αναφερόμαστε και με το γενικότερο όρο "χαρακτηριστικά" στο παρόν κεφάλαιο) (π.χ. HOF περιγραφητές των πυκνών τροχιών), $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}, \mathbf{x}_i \in \mathbb{R}^L$ για κάθε βίντεο, όπου N είναι ο αριθμός των διανυσμάτων χαρακτηριστικών and L είναι η διάσταση του περιγραφητή, πρέπει να συγκεντρώσουμε στατιστικά από αυτά τα χαμηλού επιπέδου χαρακτηριστικά. Αυτό είναι αναγκαίο αν λάβει κανείς υπόψη το πλήθος αυτών των χαρακτηριστικών, που μπορεί να

[1] `http://www.mobot-project.eu/`

φτάσει μέχρι και τα $N = 50000$ σε ένα βίντεο που περιέχει μία δράση. Είναι σαφές ότι το σύνολο αυτών των χαρακτηριστικών, του οποίου τα στοιχεία έχουν μεταβαλλόμενο αριθμό από βίντεο σε βίντεο και δεν έχουν κάποια διάταξη με σημασία, δεν μπορεί να δοθεί ως είσοδος σε μεθόδους ταξινόμησης, τόσο λόγω της υψηλής υπολογιστικής πολυπλοκότητας που επιβάλλει, όσο και λόγω της περιορισμένης διακριτικής του ικανότητας και ευρωστίας στη μεταβλητότητα της διάρκειας, της γωνίας λήψης και των συνθηκών φωτισμού των βίντεο. Επομένως χρειάζεται να παραχθεί μια αναπαράσταση του βίντεο, με τη μορφή ενός διανύσματος, η οποία να είναι απλή, αποτελεσματική και με χαμηλή υπολογιστική πολυπλοκότητα.

Ο τρόπος που υπολογίζονται τέτοιες αναπαραστάσεις βασίζεται στην κβαντοποίηση των χαρακτηριστικών με τη βοήθεια ενός "οπτικού λεξικού" (visual dictionary) που περιέχει "οπτικές λέξεις" (visual words) προερχόμενες από κάποιο αλγόριθμο ομαδοποίησης (clustering algorithm) των χαρακτηριστικών. Υπάρχουν πολλοί δημοφιλείς αλγόριθμοι ομαδοποίησης, εκ των οποίων δύο χρησιμοποιούνται κατά κόρον σε εφαρμογές αναγνώρισης δράσεων, ο αλγόριθμος των K-μέσων (K-means clustering) και ο αλγόριθμος ομαδοποίησης με Μείγμα Γκαουσιάνων Κατανομών (GMM clustering). Κατά το στάδιο της εκπαίδευσης, παράγεται με τη βοήθεια αυτών των αλγορίθμων παράγεται από ένα υποσύνολο των δεδομένων εκπαίδευσης ένα λεξικό K οπτικών λέξεων $\mathcal{D} = \{\mathbf{d}_1, \ldots, \mathbf{d}_K\}, \mathbf{d}_i \in \mathbb{R}^L$. Ο στόχος των μεθόδων αναπαράστασης των βίντεο είναι η εύρεση ενός κώδικα (code) $\mathbf{s}_n$ για κάθε χαρακτηριστικό $\mathbf{x}_n$ του σύνολου χαρακτηριστικών και η συσσώρευση αυτών των codes για την εξαγωγή μιας τελικής καθολικής αναπαράστασης του βίντεο, δοθέντος του οπτικού λεξικού $\mathcal{D}$ [12]. Για αυτό το λόγο, οι συγκεκριμένες μέθοδοι είναι ευρέως γνωστές και με την ονομασία "μέθοδοι κωδικοποίησης" (encoding methods). Στο αυτό το κεφάλαιο θα αναφερθούμε μόνο στις μεθόδους αναπαράστασης βίντεο που χρησιμοποιούνται στην παρούσα διπλωματική εργασία. Για μια οργανωμένη επισκόπηση των διαφορετικών κλασικών ειδών αναπαράστασης βίντεο, ο αναγνώστης μπορεί να συμβουλευτεί τη δημοσίευση των [12].

3.2 Κατασκευή οπτικού λεξικού

Η κατασκευή του οπτικού λεξικού μπορεί να γίνει με τη χρήση των αλγορίθμων K-means ή GMM, οι οποίοι ομαδοποιούν ένα υποσύνολο των χαρακτηριστικών που έχουν εξαχθεί από τα βίντεο που ανήκουν στο σύνολο εκπαίδευσης. Ο μεν αλγόριθμος K-means είναι ένας αλγόριθμος διανυσματικής κβαντοποίησης, ο οποίος διαχωρίζει το χώρο των χαρακτηριστικών σε K περιοχές, καθεμία από τις οποίες αναπαρίσταται από το κέντρο της, ενώ ο

αλγόριθμος GMM χρησιμοποιεί ένα αναγεννητικό (generative) μοντέλο έτσι ώστε να προσεγγίσει την κατανομή πιθανότητας των χαρακτηριστικών.

3.2.1 Αλγόριθμος K-means

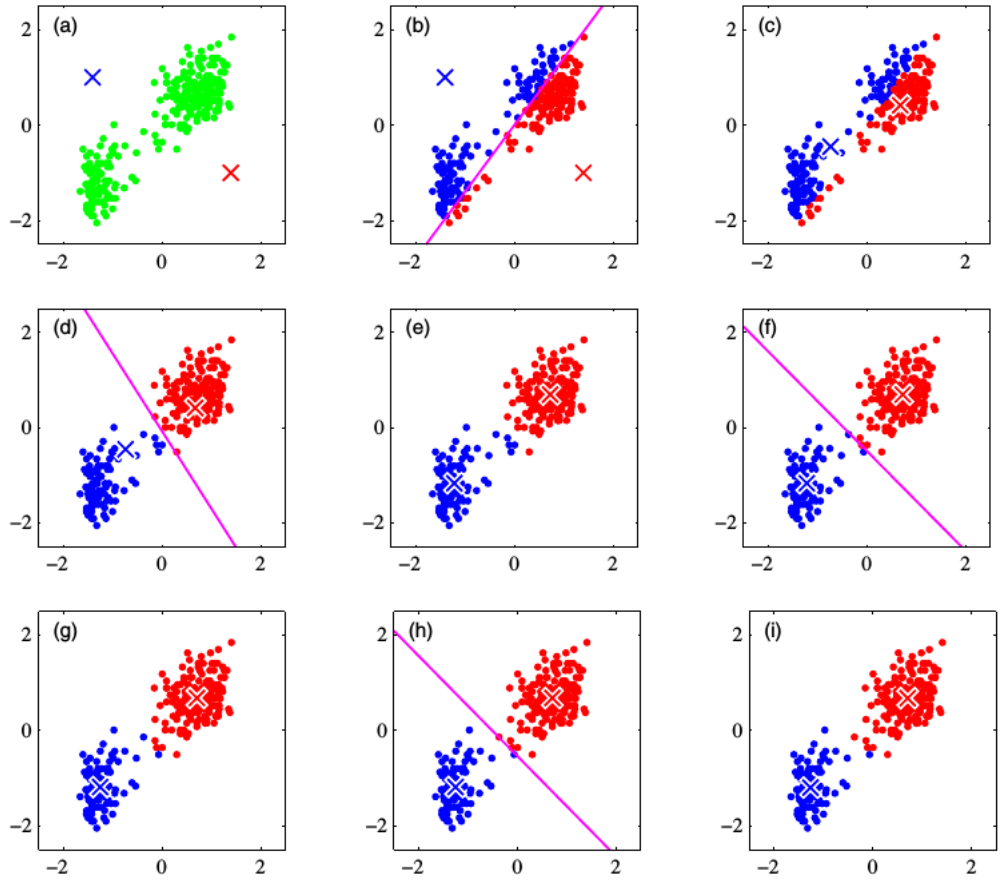
Ο K-means είναι ένας επαναληπτικός αλγόριθμος μη επιβλεπόμενης μάθησης. Πιο συγκεκριμένα, κατατάσσει το σύνολο των δεδομένων μας (χαρακτηριστικά) σε K συστάδες (clusters) χωρίς να έχει προηγηθεί εκπαίδευση με επισημειωμένα δεδομένα. Κάθε συστάδα μπορεί να θεωρηθεί ως ένα σύνολο δεδομένων, κάθε στοιχείο του οποίου απέχει πολύ λιγότερο από τα υπόλοιπα στοιχεία της ίδιας συστάδας σε σχέση με τα δεδομένα των άλλων. Κάθε συστάδα εκπροσωπείται από το κέντρο της. Σε κάθε επανάληψη ο αλγόριθμος ανανεώνει τα κέντρα των συστάδων και επανυπολογιζει τις αποστάσεις των χαρακτηριστικών από αυτά, ανανεώνοντας τελικά και την ετικέτα τους, δηλαδή τη συστάδα στην οποία ανατίθενται. Τα βήματα του αλγορίθμου είναι συνοπτικά τα ακόλουθα 1:

Αλγόριθμος 1 Αλγόριθμος K-means

1: Αρχικοποίηση των κέντρων των ομάδων με τυχαία επιλογή σημείων στο χώρο χαρακτηριστικών.
2: Επιλογή μετρικού απόστασης (εν προκειμένω χρησιμοποιείται η Ευκλείδεια νόρμα) και υπολογισμός της απόστασης κάθε χαρακτηριστικού $\mathbf{x}_n$ από τα κέντρα των K ομάδων. Κάθε $\mathbf{x}_n$ ανατίθεται στην ομάδα με το κεντροειδές $\boldsymbol{\mu}_k$ με τη μικρότερη απόσταση από αυτές που υπολογίστηκαν για το συγκεκριμένο $\mathbf{x}_n$, ελαχιστοποιώντας το μέτρο παραμόρφωσης:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \left|\left|\mathbf{x}_n - \boldsymbol{\mu}_k\right|\right|^2 \tag{3.1}$$

όπου $r_{nk} = 1$ αν το $\mathbf{x}_n$ ανήκει στην ομάδα k και $r_{nk} = 0$ διαφορετικά, δηλαδή $r_{nk} = 1$, αν $k = argmin_j \left|\left|\mathbf{x}_n - \boldsymbol{\mu}_j\right|\right|^2$.
3: Ανανέωση των κέντρων των ομάδων με επιλογή ως νέου κέντρου για κάθε ομάδα, του σημείου εκείνου που αποτελεί τον αριθμητικό μέσο των δεδομένων της ομάδας: $\boldsymbol{\mu}_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}}$
4: Έλεγχος συνθήκης τερματισμού. Ως τέτοια συνθήκη μπορεί να επιλεγεί ο μέγιστος αριθμός επαναλήψεων, ή ένα κατώφλι για το μέτρο παραμόρφωσης. Υπάρχουν επίσης και πιο σύνθετες συνθήκες.



Σχήμα 3.1: Απεικόνιση των βημάτων του K-means αλγορίθμου [45]. (a) Τα πράσινα σημεία σηματοδοτούν ένα σύνολο δεδομένων στο δισδιάστατο Ευκλείδιο χώρο. Η αρχικοποίηση των κέντρων $\boldsymbol{\mu}_1$ και $\boldsymbol{\mu}_2$ απεικονίζεται με τον κόκκινο και μπλε σταυρό, αντίστοιχα. (b) Στο Βήμα 1, κάθε σημείο/δεδομένο ανατίθεται είτε στην κόκκινη είτε στην μπλε ομάδα, ανάλογα με το ποιο κέντρο είναι κοντινότερο. (Ελαχιστοποίηση του μέτρου παραμόρφωσης J ως προς τα r_{nk}, κρατώντας τα $\boldsymbol{\mu}_k$ σταθερά.) (c) Στο Βήμα 2, ανανεώνονται τα κέντρα κάθε ομάδας έτσι ώστε να είναι ο αριθμητικός μέσος των σημείων που έχουν ανατεθεί στην αντίστοιχη ομάδα. (Ελαχιστοποίηση του μέτρου παραμόρφωσης J ως προς τα $\boldsymbol{\mu}_k$, κρατώντας τα r_{nk} σταθερά.) (d)-(i) τα βήματα 1 και 2 επαναλαμβάνονται μέχρι τη σύγκλιση του αλγορίθμου.

Η επαναληπτική διαδικασία ομαδοποίησης των δεδομένων με χρήση K-means περιγράφεται και εποπτικά στο Σχήμα 3.1. Επειδή σε κάθε επανάληψη μειώνεται η τιμή της αντικειμενικής συνάρτησης J, η σύγκλιση του αλγορίθμου είνα εγγυημένη. Ωστόσο, υπάρχει η πιθανότητα να συγκλίνει σε κάποιο τοπικό

ελάχιστο της J, αντί για το επιθυμητό ολικό ελάχιστο. Αυτό εξαρτάται από την αρχική επιλογή των κέντρων των ομάδων. Για αυτό το λόγο, συνίσταται η πολλαπλή εκτέλεση του αλγορίθμου με τυχαία αρχικοποίηση των κεντροειδών και η διατήρηση της λύσης με το μικρότερο τελικό μέτρο παραμορφωσης. Μετά το τέλος του αλγορίθμου καταλήγουμε σε ένα σύνολο K κέντρων (οπτικών λέξεων) που εκπροσωπούν τις ομάδες που δημιουργήθηκαν.

3.2.2 Ομαδοποίηση με Μοντέλο Μείγματος Γκαουσιανών

Σε αυτή τη μέθοδο ομαδοποίησης θέλουμε να μοντελοποιήσουμε την κατανομή πιθανότητας των δεδομένων (χαρακτηριστικών). Αυτό επιτυγχάνεται προσεγγίζοντας την συνάρτηση πυκνότητας πιθανότητας που ακολουθεί το διάνυσμα χαρακτηριστικών με ένα γραμμικό συνδυασμό από Γκαουσιανές συναρτήσεις:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{3.2}$$

όπου $\mathbf{x}$ είναι το διάνυσμα χαρακτηριστικών που μοντελοποιείται, K είναι ο αριθμός των Γκαουσιανών, π_k είναι μη αρνητικά βάρη που αθροίζουν στη μονάδα και δίνουν τη συνεισφορά κάθε Γκαουσιανής και $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ είναι η k-οστή πολυδιάστατη γκαουσιανή κατανομή με διάνυσμα μέσης τιμής $\boldsymbol{\mu}_k$ και πίνακα συνδιακύμανσης $\boldsymbol{\Sigma}_k$.

Αρχικά επιθυμούμε να προσδιορίσουμε τις παραμέτρους $\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}$ του μοντέλου μείγματος Γκαουσιανών που θα περιγράφει βέλτιστα τα διανύσματα κάποιων τυχαία επιλεγμένων δεδομένων εκπαίδευσης. Για να μειώσουμε την πολυπλοκότητα του προβλήματος θεωρούμε διαγώνιους πίνακες συνδιακύμανσης, δηλαδή ότι τα χαρακτηριστικά εκπαίδευσης είναι ασυσχέτιστα μεταξύ τους. Έτσι αντί για $\frac{L(L+1)}{2}$ αγνώστους για κάθε πίνακα $\boldsymbol{\Sigma}_k$ των Γκαουσιανών του μείγματος, τελικά χρειάζεται να προσδιοριστούν μόνο τα στοιχεία της διαγωνίου. Οι βέλτιστες παράμετροι προκύπτουν από τη μεγιστοποίηση της παρακάτω λογαριθμικής πιθανοφάνειας (θεωρώντας ανεξάρτητα δείγματα) :

$$\ln p(\mathbf{x}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}. \tag{3.3}$$

Η μεγιστοποίηση αυτής της παράστασης είναι αρκετά δύσκολη λόγω του αθροίσματος που βρίσκεται εντός του λογαρίθμου και δεν υπάρχει κλειστή λύση της. Γι'αυτό θα εισάγουμε μια κρυφή τυχαία μεταβλητή στο μοντέλο μας και θα χρησιμοποιήσουμε τον ιδιαίτερα δημοφιλή και ισχυρό επαναληπτικό

αλγόριθμο Expectation Maximization (EM) που βρίσκει λύσεις μέγιστης πιθανοφάνειας για μοντέλα με κρυφές μεταβλητές.

Εισάγουμε μια τυχαία μεταβλητή για κάθε δείγμα $\mathbf{x}_n$ η οποία προσδιορίζει ποια γκαουσιανή συνιστώσα παρήγαγε το δείγμα. Έστω $\mathbf{z}_n = (z_{n1}, \ldots, z_{nK})$ μια K-διάστατη δυαδική τυχαία μεταβλητή με $z_{nk} \in \{0, 1\}$ και $\sum_k z_{nk} = 1$, η οποία προσδιορίζει την Γκαουσιανή συνιστώσα (από τις K) που παρήγαγε το δείγμα $\mathbf{x}_n$.

$$z_{nk} = \begin{cases} 1 & \text{αν η Γκαουσιανή } k \text{ παρήγαγε το δείγμα } \mathbf{x}_n \\ 0 & \text{διαφορετικά .} \end{cases} \tag{3.4}$$

Τότε τα βάρη π_k δεν είναι παρά η περιθώρια συνάρτηση πιθανότητας της μεταβλητής $\mathbf{z}_n$:

$$p(\mathbf{z}_n) = \prod_{k=1}^{K} \pi_k^{z_{nk}} \tag{3.5}$$

και ο συντελεστής π_k ερμηνεύεται ως η πρότερη πιθανότητα η παρατήρηση $\mathbf{x}_n$ να προέρχεται από την k-οστή γκαουσιανή συνιστώσα του μοντέλου. Επίσης ισχύει ότι:

$$p(\mathbf{x}_n|\mathbf{z}_n) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}} \tag{3.6}$$

Επομένως, η περιθώρια συνάρτηση πιθανότητας $p(\mathbf{x}_n)$ παίρνει τη μορφή:

$$p(\mathbf{x}_n) = \sum_{\mathbf{z}_n} p(\mathbf{z}_n) p(\mathbf{x}_n|\mathbf{z}_n) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{3.7}$$

Καταφέραμε, λοιπόν, να εξάγουμε μια ισοδύναμη παράσταση για το μείγμα Γκαουσιανών εισάγοντας μια κρυφή μεταβλητή $\mathbf{z}_n$ για κάθε δείγμα $\mathbf{x}_n$.

Μια άλλη χρήσιμη ποσότητα που θα εμφανιστεί στον αλγόριθμο EM είναι η "ευθύνη" (responsibility):

$$\gamma_{nk} = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \tag{3.8}$$

η οποία προκύπτει από το θεμελιώδες θεώρημα του Bayes:

$$p(z_{nk} = 1|\mathbf{x}_n) = \frac{p(\mathbf{x}_n|z_{nk} = 1) p(z_{nk} = 1)}{p(\mathbf{x}_n)}$$

Η ποσότητα γ_{nk} είναι η ύστερη πιθανότητα το δείγμα $\mathbf{x}_n$ να έχει παραχθεί από την γκαουσιανή συνιστώσα k, ή σε όρους ομαδοποίησης εκφράζει το βαθμό

συμμετοχής του (membership weight) στη συστάδα k. Ο ενεργός αριθμός δειγμάτων που ανατίθενται τελικά στη συστάδα ισούται με:

$$N_k = \sum_{n=1}^{N} \gamma_{nk} \tag{3.9}$$

Αξίζει να υπενθυμιστεί ότι στο μοντέλο μας υποθέτουμε ότι το καθε δείγμα έχει παραχθεί μόνο από μία γκαουσιανή συνιστώσα και έτσι οι responsibilities αντικατοπτρίζουν την αβεβαιότητά μας σχετικά με το από ποια συνιστώσα προήλθε το δείγμα.

Αφού ορίσαμε τις παραπάνω μεταβλητές, θα παρουσιάσουμε τον EM αλγόριθμο με τη μορφή που παίρνει στο πρόβλημα της μεγιστοποίησης της συνάρτησης λογαριθμικής πιθανότητας ως προς τις παραμέτρους (μέση τιμή, πίνακας συνδιακύμανσης και βάρος κάθε συνιστώσας) ενός μοντέλου GMM. Ο γενικευμένος EM αλγόριθμος και η αναλυτική εφαρμογή του μπορεί να βρεθεί στο Κεφάλαιο 9 του [45].

Αλγόριθμος 2 Αλγόριθμος EM

1: Αρχικοποίηση των παραμέτρων $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\pi}_k$ και υπολογισμός της αρχικής τιμής της λογαριθμικής πιθανοφάνειας (εξ. 3.3).
2: Ε-βήμα: Υπολογισμός των "ευθυνών" γ_{nk} (εξ. 3.8) (ευθύνη της k-οστής γκαουσιανής να δικαιολογήσει το n-οστό δείγμα) για $n = 1, \ldots, N$ και $k = 1, \ldots, K$.
3: Μ-βήμα: Επανυπολογισμός των τιμών των παραμέτρων του μοντέλου χρησιμοποιώντας τα δεδομένα και τις responsibilities που υπολογίστηκαν στο Ε-βήμα:

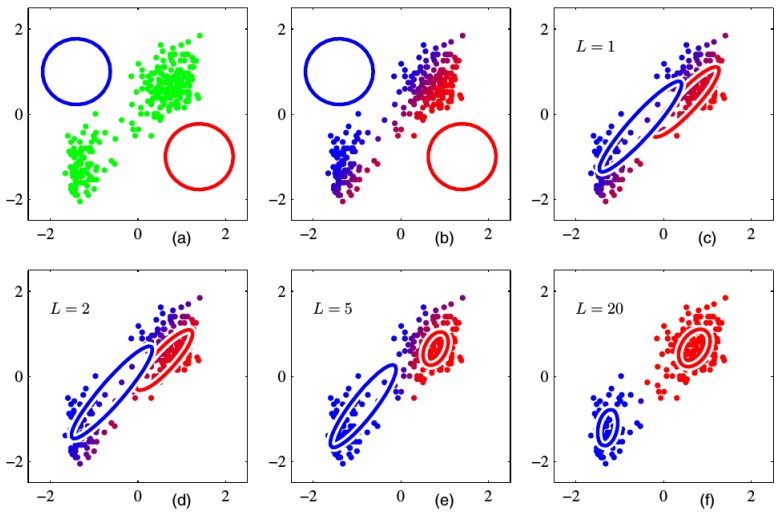
$$\boldsymbol{\mu}_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \tag{3.10}$$

$$\boldsymbol{\Sigma}_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (\mathbf{x}_n - \boldsymbol{\mu}_k^{new})(\mathbf{x}_n - \boldsymbol{\mu}_k^{new})^T \tag{3.11}$$

$$\pi_k^{new} = \frac{N_k}{N} \tag{3.12}$$

όπου N_k ο ενεργός αριθμός δεδομένων που ανατίθενται στη συνιστώσα k (εξ. 3.9).
4: Υπολογισμός της λογαριθμικής πιθανοφάνειας (εξ. 3.3) και έλεγχος σύγκλισης των παραμέτρων ή της λογαριθμικής πιθανοφάνειας. Αν δεν ικανοποιείται το κριτήριο σύγκλισης, επιστροφή στο Βήμα 2.

Όπως και ο αλγόριθμος K-means, ο αλγόριθμος EM εγγυάται τη σύγκλιση, αυξάνοντας τη λογαριθμική πιθανότητα σε κάθε επανάληψη, αλλά είναι ιδιαίτερα ευαίσθητος στην αρχικοποίηση των παραμέτρων και μπορεί να συγκλίνει σε κάποιο τοπικό μέγιστο. Γι'αυτό συχνά χρησιμοποιούνται ως αρχικοποίηση των $\boldsymbol{\mu}_k$ τα κέντρα των συστάδων που έχουν προκύψει από τον K-means, ως αρχικοποίηση των $\boldsymbol{\Sigma}_k$ οι συνδιακυμάνσεις δείγματος των δεδομένων που έχουν ομαδοποιηθεί στη συστάδα k και ως αρχικοποίηση των βαρών pi_k το κλάσμα των δεδομένων που ανατέθηκαν στη συστάδα.



Σχήμα 3.2: Απεικόνιση των βημάτων του αλγορίθμου EM [45]. (a) Τα πράσινα σημεία σηματοδοτούν ένα σύνολο δεδομένων στο δισδιάστατο Ευκλείδιο χώρο. Η αρχικοποίηση των $K = 2$ γκαουσιανών μοναδιαίας τυπικής απόκλισης απεικονίζεται με τον κόκκινο και μπλε κύκλο, αντίστοιχα. (b) Στο E-Βήμα κάθε σημείο απεικονίζεται με μπλε απόχρωση που αντιστοιχεί στην ύστερη πιθανότητα να έχει παραχθεί από την μπλε συνιστώσα και με κόκκινη απόχρωση που αντιστοιχεί στην ύστερη πιθανότητα να έχει παραχθεί από την κόκκινη συνιστώσα. Έτσι τα σημεία που έχουν μεγάλη πιθανότητα να ανήκουν στη μία όσο και στην άλλη ομάδα φαίνονται μωβ. (c) Στο M-Βήμα ανανεώνονται οι παράμετροι κάθε γκαουσιανής, έτσι ώστε η μέση τιμή της μπλε γκαουσιανής να είναι το κέντρο μάζας των σημείων που έχουν μπλε απόχρωση και η συνδιακύμανσή της να είναι ίση με με τη συνδιακύμανση δείγματος των σημείων με μπλε απόχρωση. Ομοίως και για την κόκκινη. (d)-(f) Αποτελέσματα μετά από 2, 5 και 20 πλήρεις επαναλήψεις του EM αλγορίθμου.

Όπως παρατηρούμε, το GMM clustering οδηγεί σε μια "χαλαρή" ανάθεση (soft assignment) των χαρακτηριστικών στις ομάδες, εκφρασμένη μέσω των

εκ των υστέρων πιθανοτήτων, σε αντίθεση με τον αλγόριθμο K-means όπου υπάρχει "αυστηρή" ανάθεση (hard assignment) κάθε δεδομένου σε μια μοναδική ομάδα, αυτή της οποίας το κέντρο απέχει λιγότερο. Είναι διαισθητικά κατανοητό ότι στη δεύτερη περίπτωση χάνεται περισσότερη πληροφορία λόγω της κβάντισης των χαρακτηριστικών. Επιπρόσθετα, ενώ ο K-means μας δίνει πληροφορία μόνο για τα κέντρα των συστάδων, ο GMM αποκαλύπτει και το σχήμα της κατανομής των χαρακτηριστικών εντός της κάθε συστάδας. Τέλος, μπορεί να παρατηρήσει κανείς την αντιστοιχία του βήματος 2 του K-means αλγορίθμου με το E-βήμα του EM αλγορίθμου και την αντιστοιχία του βήματος 3 του K-means αλγορίθμου, με το M-βήμα του EM αλγορίθμου και μάλιστα αποδεικνύεται ότι ο K-means είναι ειδική περίπτωση του EM αλγορίθμου.

3.3 Σύνολα Οπτικών Λέξεων - Bag of Visual Words (BoVW)

Η μέθοδος Bag-Of-Visual-Words [14] είναι μια δημοφιλής μέθοδος αναπαράστασης εικόνων και εν γένει βίντεο, η οποία βασίζεται στη συχνότητα εμφάνισης των "οπτικών" λέξεων σε ένα βίντεο. Πιο συγκεκριμένα, δοθέντος ενός λεξικού $\mathcal{D}$ οπτικών λέξεων η BoVW αναπαράσταση είναι το ιστόγραμμα των συχνοτήτων εμφάνισης των οπτικών λέξεων (Σχήμα 3.4).

Η μέθοδος είναι εμπνευσμένη από την επεξεργασία κειμένου και θεωρεί ότι όπως τα κείμενα είναι συλλογές από λέξεις ενός προκαθορισμένου λεξιλογίου, έτσι και οι εικόνες/βίντεο αποτελούνται από οπτικές λέξεις ενός οπτικού λεξικού και μπορούν να αναπαρασταθούν από τη συχνότητα εμφάνισης αυτών των λέξεων. Για να υπολογιστεί η BoVW αναπαράσταση ενός video αρχικά χρειάζεται το οπτικό λεξικό, το οποίο έχει υπολογιστεί μέσω του αλγορίθμου K-means από κάποια τυχαία επιλεγμένα χαρακτηριστικά εκπαίδευσης. Τα κέντρα των ομάδων που προκύπτουν από την ομαδοποίηση είναι οι οπτικές λέξεις. Στη συνέχεια, κάθε χαρακτηριστικό που έχει εξαχθεί από το video ανατίθεται στην κοντινότερή του οπτική λέξη με βάση την Ευκλείδεια απόσταση. Τέλος, έχοντας αντιστοιχίσει κάθε περιγραφητή σε μία οπτική λέξη, υπολογίζεται η συχνότητα εμφάνισης της κάθε οπτικής λέξης στο βίντεο. Το ιστόγραμμα που προκύπτει ονομάζεται Bag-Of-Visual-Words (BoVW) ιστόγραμμα.

Πιο τυπικά, αυτή η μέθοδος αναπαράστασης, όπως και οι υπόλοιπες που ακολουθούν, μπορεί να θεωρηθεί ότι αποτελείται από δύο στάδια: το στάδιο της *κωδικοποίησης* και το στάδιο της *συσσώρευσης*. Στο στάδιο της κωδικοποίησης παράγεται ένας κώδικας $\mathbf{s}_n$ για κάθε χαρακτηριστικό $\mathbf{x}_n$ με χρήση μιας συνάρτησης κωδικοποίησης. Στο στάδιο της συσσώρευσης εφαρμόζεται

ένας τελεστής συσσώρευσης (pooling operator) ο οποίος δίνει την τελική καθολική αναπαράσταση p του βίντεο.

Στα πλαίσια της αναπαράστασης BoVW, η συνάρτηση κωδικοποίησης παίρνει τη μορφή της "αυστηρής" ανάθεσης (hard assignment) κάθε χαρακτηριστικού στην κοντινότερη οπτική λέξη από τις K λέξεις του λεξικού $\mathcal{D}$. Για το χαρακτηριστικό $\mathbf{x}_n$ ο κώδικας $\mathbf{s}_n$ είναι ένα διάνυσμα στοιχείων του οποίου το i-οστό στοιχείο δίνεται από την παρακάτω εξίσωση:

$$\mathbf{s}_n(i) = \begin{cases} 1, & \text{αν } i = \underset{j}{\operatorname{argmin}} \left\| \mathbf{x}_n - \mathbf{d}_j \right\|_2, \\ 0, & \text{διαφορετικά.} \end{cases} \tag{3.13}$$

Όπως είναι προφανές, ο κάθε κώδικας είναι ένα διάνυσμα μήκους K του οποίου όλα τα στοιχεία είναι ίσα με μηδέν εκτός από αυτό που αντιστοιχεί στην κοντινότερη οπτική λέξη. Έχοντας υπολογίσει τους κώδικες $\mathbf{s}_n$ για κάθε χαρακτηριστικό $\mathbf{x}_n$, η τελική αναπαράσταση του video προκύπτει από τη συσσώρευση των $\mathbf{s}_n$ με τη χρήση ενός τελεστή μέσου όρου (averaging operator) και είναι ένα διάνυσμα μήκους με i-οστό στοιχείο:

$$p_i = \frac{1}{N} \sum_{n=1}^{N} \mathbf{s}_n(i) \tag{3.14}$$

Βλέπουμε ότι και με αυτή την πιο τυπική διατύπωση καταλήγουμε σε ένα ιστόγραμμα εμφανίσεων των οπτικών λέξεων μήκους K.

Κάποιες μέθοδοι κανονικοποίησης που χρησιμοποιούνται συχνά με αυτή τη μέθοδο είναι:

- ℓ_1-κανονικοποίηση: το ιστόγραμμα BoVW διαιρείται με την ℓ_1 νόρμα του $(\frac{\mathbf{p}}{\|\mathbf{p}\|_1})$.
- ℓ_2-κανονικοποίηση: το ιστόγραμμα BoVW διαιρείται με την ℓ_2 νόρμα του $(\frac{\mathbf{p}}{\|\mathbf{p}\|_2})$.

Αξίζει να σημειωθεί ότι αυτή η μέθοδος απλώς αποθηκεύει τις συχνότητες εμφάνισης των λέξεων, χωρίς να κρατάει κάποια πληροφορία για τη χωροχρονική τους θέση μέσα στο video. Στο Κεφάλαιο 7 θα παρουσιάσουμε μια νέα μέθοδο, η οποία λαμβάνει υπόψη τη χρονική ακολουθία των οπτικών λέξεων και έρχεται να προστεθεί στην ομάδα μεθόδων της βιβλιογραφίας που προσπαθούν να εμπλουτίσουν το BoVW με χωροχρονική πληροφορία.

3.4 Διάνυσμα Τοπικά Συσσωρευμένων Περιγραφητών - Vector of Locally Aggregated Descriptors (VLAD)

Αυτή η μέθοδος αναπαράστασης, που εισήγαγαν οι Jegou et al. [43], περιλαμβάνει, όπως και η μέθοδος BoVW, την κατασκευή ενός οπτικού λεξικού χρησιμοποιώντας τον αλγόριθμο K-means. Ωστόσο, αντί να καταγράφει για κάθε κεντροειδές τον αριθμό των περιγραφητών που έχουν ανατεθεί σε αυτό, καταγράφει το άθροισμα των διαφορών αυτών των περιγραφητών από το κεντροειδές. Τα διανύσματα που προκύπτουν με αυτό τον τρόπο για κάθε κεντροειδές συνενώνονται, έτσι ώστε να παραχθεί ένα τελικό διάνυσμα (super-vector). Η αναπαράσταση VLAD βελτιώνει την αναπαράσταση BoVW εκμεταλλευόμενη στατιστικά πρώτης τάξης.

Πιο συγκεκριμένα, για κάθε οπτική λέξη $\mathbf{d}_i$, αθροίζονται οι διαφορές $\mathbf{x}_n - \mathbf{d}_i$ των διανυσμάτων $\mathbf{x}_n$ που έχουν ανατεθεί στη λέξη $\mathbf{d}_i$ (δηλαδή $\text{NearestNeighbor}(\mathbf{x}_n) = \mathbf{d}_i$).

$$\mathbf{u}_i = \sum_{\mathbf{x}_n : \underset{j}{\text{argmin}} \|\mathbf{x}_n - \mathbf{d}_j\|_2 = i} \mathbf{x}_n - \mathbf{d}_i \tag{3.15}$$

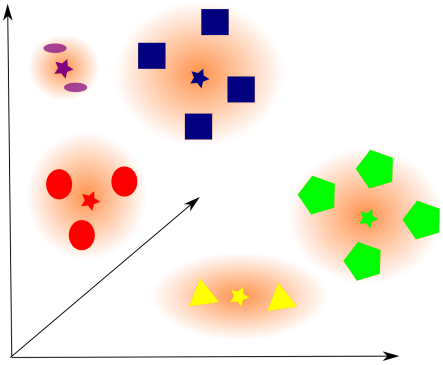
Η αναπαράσταση VLAD είναι η συνένωση (concatenation) αυτών των L-διάστατων διανυσμάτων $\mathbf{u}_i$, όπου L είναι η διάσταση του περιγραφητή, και η διάστασή του είναι συνεπώς ίση με $K \cdot L$.

Αναλύοντας την αναπαράσταση VLAD ως το συνδυασμό μιας συνάρτησης κωδικοποίησης και ενός pooling τελεστή έχουμε:

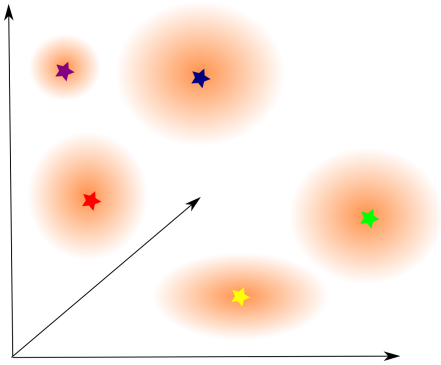
$$\mathbf{s}_n(i) = \begin{cases} \mathbf{x}_n - \mathbf{d}_i, & \text{αν } i = \underset{j}{\text{argmin}} \|\mathbf{x}_n - \mathbf{d}_j\|_2, \\ 0, & \text{διαφορετικά.} \end{cases} \tag{3.16}$$

$$\mathbf{s}_n = [0, \ldots, 1 \cdot (\mathbf{x}_n - \mathbf{d}_i), \ldots, 0] \tag{3.17}$$

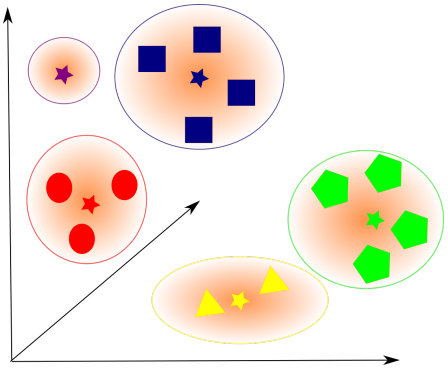
$$\mathbf{p}_i = \sum_{n=1}^{N} \mathbf{s}_n(i) \tag{3.18}$$



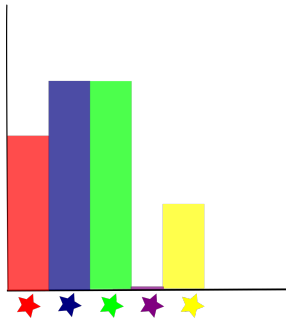
(α′) Ομαδοποίηση K-means



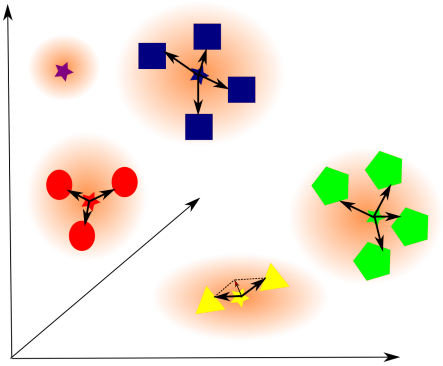
(β′) Λεξικό Οπτικών Λέξεων



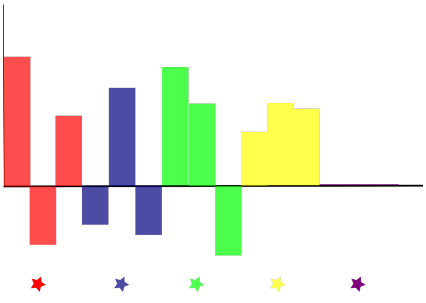
(γ′) Καταμέτρηση αριθμού χαρακτηριστικών που έχουν ανατεθεί σε οπτικές λέξεις



(δ′) Διάνυσμα BoVW



(ε′) Διανυσματικές διαφορές των χαρακτηριστικών από τις οπτικές λέξεις που έχουν ανατεθεί



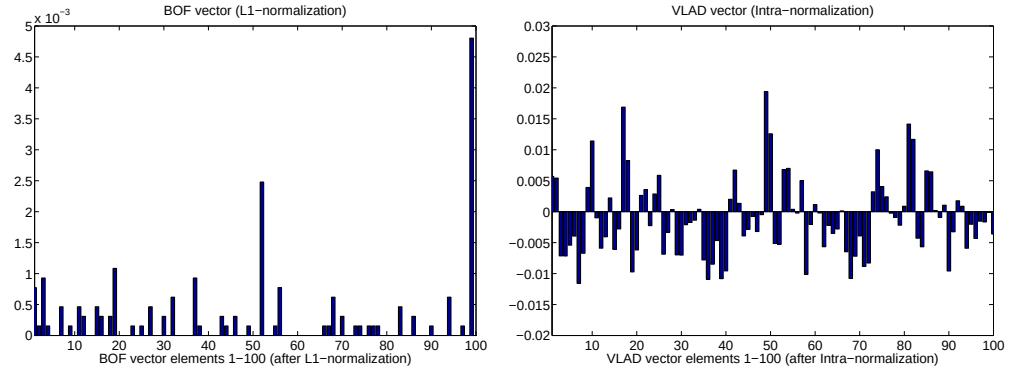
(στ′) Διάνυσμα VLAD

Σχήμα 3.3: Βήματα κατασκευής αναπαραστάσεων BoVW και VLAD.

Το σχήμα 3.3 συνοψίζει τα βασικά βήματα των μεθόδων αναπαράστασης

BoVW και VLAD. Αρχικά, χρησιμοποιούνται κάποια τυχαία επιλεγμένα χαρακτηριστικά από τα video εκπαίδευσης, τα οποία ανήκουν στο τρισδιάστατο Ευκλείδειο χώρο ($L = 3$), και ομαδοποιούνται με χρήση του K-means αλγορίθμου (3.3α′) για την εύρεση των οπτικών λέξεων που θα αποτελέσουν το οπτικό λεξικό (3.3β′). Τα χαρακτηριστικά που αντιστοιχούν σε κάθε ομάδα αναπαρίστανται με διαφορετικό χρώμα, ενώ τα κεντροειδή (οπτικές λέξεις) απεικονίζονται με αστεράκια. Στη συνέχεια επιθυμούμε να εξάγουμε BoVW και VLAD αναπαραστάσεις κάποιο video εισόδου, του οποίου έχουμε διαθέσιμα τα χαρακτηριστικά. Το κάθε χαρακτηριστικό ανατίθεται στην κοντινότερή του θέση. Στη μέθοδο BoVW μετράμε πόσα χαρακτηριστικά έχουν ανατεθεί στην κάθε οπτική λέξη (3.3γ′) και παράγουμε ένα ιστόγραμμα συχνότητας εμφάνισης των οπτικών λέξεων στο video (3.3δ′). Στη μέθοδο VLAD μας ενδιαφέρει η σχετική θέση των χαρακτηριστικών σε σχέση με τα κέντρα των clusters και για κάθε οπτική λέξη υπολογίζουμε τις διανυσματικές διαφορές των χαρακτηριστικών από την οπτική λέξη (βέλη με μαύρη κεφαλή στο) και το διανυσματικό άθροισμα αυτών (μαύρο βέλος στο κίτρινο cluster)(3.3ε′). Η τελική αναπαράσταση του video κατασκευάζεται από τη συνένωση αυτών των διανυσματικών αθροισμάτων που έχουν προκύψει για κάθε οπτική λέξη και έχει διάσταση $K \cdot D$ (3.3στ′).

Η εικόνα 3.4 απεικονίζει αποσπάσματα 100 στοιχείων από τα διανύσματα BoVW και VLAD που αναπαριστούν ένα video. Παρατηρούμε ότι το BoVW είναι ένα αραιό διάνυσμα, το οποίο περιέχει θετικά στοιχεία, που αντιστοιχούν στον κανονικοποιημένο αριθμό χαρακτηριστικών που έχουν ανατεθεί στην κάθε οπτική λέξη, ενώ το VLAD παράγει μια συμπαγή αναπαράσταση με θετικά και αρνητικά στοιχεία που προκύπτουν από διανυσματικά αθροίσματα διαφορών.



(α′) Διάνυσμα BoVW (β′) Διάνυσμα VLAD

Σχήμα 3.4: Αποσπάσματα από τα διανύσματα BoVW και VLAD που αναπαριστούν ένα video. Στο (α) απεικονίζονται μόνο τα 100 πρώτα στοιχεία του BoVW, που αντιστοιχούν στις πρώτες 100 από K_{BoVW} οπτικές λέξεις. Ομοίως, στο (β) διακρίνονται μόνο τα 100 πρώτα στοιχεία του VLAD διανύσματος, του οποίου η διάσταση είναι $K_{VLAD} \cdot L$. Ο αριθμός των οπτικών λέξεων που χρησιμοποιήθηκαν για τις BoVW και VLAD αναπαραστάσεις είναι $K_{BoVW} = 4000$ και $K_{VLAD} = 256$ αντίστοιχα.

Η αναπαράσταση VLAD οδηγεί σε μια συμπαγή αναπαράσταση του video, διάστασης $K \cdot L$, με λιγότερα μηδενικά στοιχεία σε σύγκριση με την BoVW αναπαράσταση και συνεπώς περισσότερη πληροφορία αποθηκευμένη ανά οπτική λέξη. Γι'αυτό η VLAD αναπαράσταση χρειάζεται σημαντικά μικρότερο αριθμό οπτικών λέξεων και είναι πιο αποδοτική για μεγάλες εφαρμογές. Ένα άλλο πλεονέκτημα της είναι ότι έχει δειχθεί ότι δουλεύει εξαιρετικά καλά με γραμμικούς SVM ταξινομητές (βλ. Κεφάλαιο 4), οι οποίοι υπολογίζονται αποδοτικά.

Όπως και στην περίπτωση του BoVW ιστογράμματος, τα διάνυσματα VLAD κανονικοποιούνται συχνά πριν τη χρήση τους για την ταξινόμηση των videos και έχει αποδειχθεί ότι η κανονικοποίηση επηρεάζει σημαντικά το τελικό αποτέλεσμα. Μερικές από τις κανονικοποιήσεις που προτιμούνται για την αναπαράσταση VLAD είναι:

- ℓ_2-κανονικοποίηση: το διάνυσμα VLAD διαιρείται με την ℓ_2 νόρμα του $(\frac{\mathbf{p}}{\|\mathbf{p}\|_2})$.
- Power-Normalization: εφαρμογή της συνάρτησης προσημασμένης τετραγωνικής ρίζας (Signed Squared Rooting - SSR) $(\text{sign}(p_k)|p_k|^{\frac{1}{2}})$ σε κάθε στοιχείο πριν την ℓ_2-κανονικοποίηση [44]. Συχνά κάποια χαρακτηριστικά του video εμφανίζονται πολύ συχνά ("bursty" features), π.χ. λόγω κάποιας επαναλαμβανόμενης δομής της εικόνας όπως ένα δάπεδο με πλακάκια. Όπως είναι προφανές, τα στοιχεία του VLAD που αντιστοιχούν

σε clusters αυτών των χαρακτηριστικών θα έχουν πολύ μεγαλύτερη τιμή από τα υπόλοιπα στοιχεία, και θα επηρεάσουν την αποτίμηση της ομοιότητας μεταξύ δύο videos. Η κανονικοποίηση με χρήση της συνάρτησης SSR μειώνει τις μεγάλες τιμές υπολογίζοντας τη τετραγωνική ρίζα κάθε στοιχείου και κανονικοποιώντας εκ των υστέρων το διάνυσμα με χρήση της ℓ_2 νόρμας.

- Intra-Normalization: κανονικοποίηση κάθε συνιστώσας του VLAD που αντιστοιχεί σε ένα cluster ξεχωριστά [46]. Η μέθοδος αυτή δεν μειώνει απλώς τις μεγάλες τιμές των στοιχείων του διανύσματος VLAD που αντιστοιχούν σε "bursty" χαρακτηριστικά, αλλά εξαλείφει τελείως το φαινόμενο.

$$\mathbf{p} = [\frac{\mathbf{p}^{(1)}}{\|\mathbf{p}^{(1)}\|}, \ldots, \frac{\mathbf{p}^{(K)}}{\|\mathbf{p}^{(K)}\|}] \tag{3.19}$$

3.5 Διάνυσμα Fisher - Fisher Vector (FV)

Η αναπαράσταση Fisher vector μοντελοποιεί την πιθανοτική κατανομή των οπτικών χαρακτηριστικών του βίντεο χρησιμοποιώντας ένα αναγεννητικό (generative) μοντέλο και υπολογίζει το διάνυσμα κλίσης της λογαριθμικής πιθανοφάνειας ως προς τις παραμέτρους του μοντέλου $\boldsymbol{\lambda}$, έτσι ώστε να αναπαραστήσει το βίντεο.

$$\mathcal{G}^{\mathbf{x}}_{\boldsymbol{\lambda}} = \nabla_{\boldsymbol{\lambda}} \log p\left(\mathbf{x}|\boldsymbol{\lambda}\right) \tag{3.20}$$

Χρησιμοποιώντας ένα μοντέλο μείγματος Γκαουσιανών (GMM) για να μοντελοποιήσουμε την κατανομή των χαρακτηριστικών, μπορούμε να γράψουμε τη συνάρτηση κωδικοποίησης ενός χαρακτηριστικού $\mathbf{x}_n$:

$$\mathcal{G}^{\mathbf{x}_n}_{\boldsymbol{\mu},k} = \frac{1}{\sqrt{\pi_k}}\gamma_{nk}(\frac{\mathbf{x}_n - \mu_k}{\sigma_k}) \tag{3.21}$$

$$\mathcal{G}^{\mathbf{x}_n}_{\boldsymbol{\sigma},k} = \frac{1}{\sqrt{2\pi_k}}\gamma_{nk}(\frac{(\mathbf{x}_n - \mu_k)^2}{\sigma_k^2} - 1) \tag{3.22}$$

Ως pooling τελεστής χρησιμοποιείται ο τελεστής μέσου όρου, οπότε έχουμε ότι $\mathcal{G}^{\mathbf{x}}_{\boldsymbol{\sigma},k} = \frac{1}{N}\sum_{n=1}^{N}\mathcal{G}^{\mathbf{x}_n}_{\boldsymbol{\sigma},k}$ και ομοίως $\mathcal{G}^{\mathbf{x}}_{\boldsymbol{\mu},k} = \frac{1}{N}\sum_{k=1}^{N}\mathcal{G}^{\mathbf{x}_n}_{\boldsymbol{\mu},k}$

Το τελικό διάνυσμα Fisher προκύπτει από τη συνένωση αυτών των διανυσμάτων:

$$\mathcal{G}^{\mathbf{x}}_{\lambda} = \left[\left(\mathcal{G}^{\mathbf{x}}_{\mu,1}\right) \quad \left(\mathcal{G}^{\mathbf{x}}_{\sigma,1}\right) \quad \cdots \quad \left(\mathcal{G}^{\mathbf{x}}_{\mu,K}\right) \quad \left(\mathcal{G}^{\mathbf{x}}_{\sigma,K}\right)\right], \tag{3.23}$$

όπου

Όπως φαίνεται από τις εξισώσεις, η διάσταση του διανύσματος Fisher είναι ίση με $2 \cdot K \cdot L$, δηλαδή διπλάσια από αυτή του VLAD, ενώ όσον αφορά τις μεθόδους κανονικοποίησής του, μπορούν να εφαρμοστούν οι ίδιες που εφαρμόζονται και στο VLAD vector. Η πιο διαδεδομένη κανονικοποίηση του FV είναι η power-normalization που αναλύσαμε παραπάνω.

Ωστόσο, ένα μειονέκτημα της μεθόδου Fisher Vector, όπως και της προαναφερθείσας μεθόδου VLAD, είναι ότι το μέγεθος της αναπαράστασης του video αυξάνεται γραμμικά με τη διάσταση του περιγραφητή L. Γι'αυτό συνηθίζεται να χρησιμοποιείται η μέθοδος Ανάλυσης Κύριων Συνιστωσών (PCA), έτσι ώστε να μειωθούν οι διαστάσεις των περιγραφητών. Η PCA έχει ευεργετική επίδραση και στην κωδικοποίηση FV, καθώς εκτός από μείωση της διάστασης του διανύσματος, τα ασυσχέτιστα δεδομένα μπορούν να μοντελοποιηθούν με μεγαλύτερη ακρίβεια από GMMs με διαγώνιους πίνακες συνδιακύμανσης [44]. Επιπρόσθετα, η εκτίμηση των παραμέτρων του μοντέλου GMM είναι θορυβώδης για τις λιγότερο σημαντικές συνιστώσες [47]. Σε αυτό το σημείο θα προχωρήσουμε σε μια αναλυτική περιγραφή της μεθόδου.

3.6 Ανάλυση σε Κύριες Συνιστώσες - Principal Component Analysis (PCA)

Η μέθοδος PCA είναι μια από τις πιο δημοφιλείς τεχνικές μείωσης των διαστάσεων ενός συνόλου δεδομένων. Η PCA πραγματοποιεί έναν ορθογώνιο μετασχηματισμό ενός συνόλου δεδομένων (εν προκειμένων περιγραφητών) σε ένα σύνολο ασυσχέτιστων μεταβλητών, που ανήκουν σε ένα γραμμικό χώρο μικρότερης ή ίσης διάστασης, γνωστό ως κύριο υποχώρο (principal subspace), και ονομάζονται Κύριες Συνιστώσες (principal components). Η πρώτη κύρια συνιστώσα αντιστοιχεί στο μεγαλύτερο δυνατό ποσοστό της συνολικής διακύμανσης (variance) των δεδομένων και κάθε ακόλουθη κύρια συνιστώσα έχει την υψηλότερη δυνατή διακύμανση υπό τον περιορισμό ότι είναι ορθογώνια ως προς τις προηγούμενες συνιστώσες. Συνήθως ένας μικρός αριθμός κύριων συνιστωσών ερμηνεύει ένα επαρκές ποσοστό της διακύμανσης των δεδομένων και έτσι επιτυγχάνουμε μείωση των διαστάσεων του συνόλου δεδομένων.

Έστω ότι το σύνολο δεδομένων απαρτίζεται από N διανύσματα $\mathbf{x}_n \in \mathbb{R}^L, n = 1, \ldots, N$ με μηδενική μέση τιμή. Επιθυμούμε να βρούμε ένα ορθογώνιο πίνακα, δηλαδή έναν τετραγωνικό πίνακα για τον οποίο ισχύει $\mathbf{A}\mathbf{A}^T = \mathbf{I}$ ($\mathbf{A}^{-1} = \mathbf{A}^T$), έτσι ώστε τα μετασχηματισμένα διανύσματα $\mathbf{y} = \mathbf{A}^T\mathbf{x}$ να έχουν δύο ιδιότητες:

1. Να έχουν ορθογώνιες (ή ασυσχέτιστες) συνιστώσες

2. Αν κρατήσουμε τις πρώτες $p < L$ συνιστώσες να έχουμε ελάχιστο μέσο τετραγωνικό σφάλμα (MSE).

Έστω ότι γνωρίζουμε τα ορθοκανονικά διανύσματα $\{\mathbf{e}_1, \ldots, \mathbf{e}_L\}$ ($\mathbf{e}_i\mathbf{e}_j = \delta_{ij}$), που σχηματίζουν τις στήλες του $\mathbf{A}$ και τα θεωρούμε σαν μια ορθοκανονική βάση όλου του χώρου. Τότε:

$$\mathbf{x}_n = \sum_{i=1}^{L} y_{in}\mathbf{e}_i \tag{3.24}$$

$$y_{in} =< \mathbf{x}_n, \mathbf{e}_i >= \mathbf{e}_i^T\mathbf{x}_n \tag{3.25}$$

Αν προβάλλουμε αυτό το διάνυσμα στον υποχώρο που ορίζεται από τη μικρότερη βάση $\{\mathbf{e}_1, \ldots, \mathbf{e}_p\}$, τότε η καλύτερη προσέγγιση $\hat{\mathbf{x}}_n$ είναι:

$$\hat{\mathbf{x}}_n = \sum_{k=1}^{p} y_{kn}\mathbf{e}_k \tag{3.26}$$

και το αντίστοιχο MSE J είναι:

$$\begin{aligned} J =& \frac{1}{N}\sum_{n=1}^{N} \|\mathbf{x}_n - \hat{\mathbf{x}}_n\| = \\ & \frac{1}{N}\sum_{n=1}^{N} \left\| \sum_{i=1}^{L} y_{in}\mathbf{e}_i - \sum_{k=1}^{p} y_{kn}\mathbf{e}_k \right\|^2 = \\ & \frac{1}{N}\sum_{n=1}^{N} \left\| \sum_{i=p+1}^{L} y_{in}\mathbf{e}_i \right\|^2 \end{aligned} \tag{3.27}$$

το οποίο λόγω της ορθογωνιότητας των διανυσμάτων βάσης και του μοναδιαίου μέτρου τους ισούται με:

$$\begin{aligned} J =& \frac{1}{N}\sum_{n=1}^{N}\sum_{i=p+1}^{L} \|y_{in}\mathbf{e}_i\|^2 = \frac{1}{N}\sum_{n=1}^{N}\sum_{i=p+1}^{L} \|y_{in}\|^2 = \\ & \frac{1}{N}\sum_{i=p+1}^{L}\sum_{n=1}^{N} \|y_{in}\|^2 = \sum_{i=p+1}^{L} \mathbf{e}_i^T\mathbf{C}\mathbf{e}_i \end{aligned} \tag{3.28}$$

όπου $\mathbf{C} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n\mathbf{x}_n^T$ είναι ο πίνακας συνδιακύμανσης (covariance matrix) των δεδομένων.

Επιστρέφοντας στο αρχικό μας στόχο, ο οποίος ήταν να βρούμε τα βέλτιστα διανύσματα βάσης $\{\mathbf{e}_1, \ldots, \mathbf{e}_L\}$, που ορίζουν το βέλτιστο πίνακα $\mathbf{A}$, και να διαλέξουμε τις p κύριες συνιστώσες έτσι ώστε να ελαχιστοποιήσουμε το MSE J. Ελαχιστοποιώντας το J ύπο τον περιορισμό $\mathbf{e}_i^T\mathbf{e}_i = 1$ $i = 1, \ldots, L$ εισάγοντας πολλαπλασιαστές Lagrange λ_i για κάθε i έχουμε:

$$\min \mathbf{e}_i^T\mathbf{C}\mathbf{e}_i + \lambda_i(1 - \mathbf{e}_i^T\mathbf{e}_i) \tag{3.29}$$

Θέτοντας την παράγωγο ως προς $\mathbf{e}_i$ ίση με 0 έχουμε:

$$\mathbf{C}\mathbf{e}_i = \lambda_i\mathbf{e}_i \ , i = 1, \ldots, L. \tag{3.30}$$

Επομένως, η βέλτιστη ορθοκανονική βάση $\{\mathbf{e}_1, \ldots, \mathbf{e}_L\}$ αποτελείται από τα ιδιοδιανύσματα του $\mathbf{C}$ και τα λ_i συνιστούν τις αντίστοιχες ιδιοτιμές. Άρα για να ελαχιστοποιήσουμε το $J = \sum_{i=p+1}^{L} \mathbf{e}_i^T\mathbf{C}\mathbf{e}_i = \sum_{i=p+1}^{L} \lambda_i$ αρκεί να ταξινομήσουμε τις ιδιοτιμές λ_i σε φθίνουσα σειρά ($\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_L$) και να διαλέξουμε ως κύριες κατευθύνσεις $\{\mathbf{e}_1, \ldots, \mathbf{e}_L\}$ αυτές που αντιστοιχούν στις p μεγαλύτερες ιδιοτιμές. Τα αντίστοιχα μετασχηματισμένα δεδομένα $\{y_k : k = 1, \ldots, p\}$ ονομάζονται κύριες συνιστώσες. Άλλη μια σημαντική παρατήρηση είναι ότι οι παραπάνω επιλογές διαγωνοποιούν τον πίνακα διακύμανσης των μετασχηματισμένων διανυσμάτων.

$$\mathbf{C}_y = \frac{1}{N}\sum_{i=1}^{N} \mathbf{y}\mathbf{y}^T = \mathbf{A}^T\mathbf{C}_x\mathbf{A} = \text{diag}\left[\lambda_1, \ldots, \lambda_L\right] \tag{3.31}$$

Επομένως, τα μετασχηματισμένα χαρακτηριστικά $\{\mathbf{y}_i\}$ είναι ορθογώνια και οι διακυμάνσεις τους (variances) ισούνται με τις ιδιοτιμές του $\mathbf{C}_x$. Συνεπώς, εφόσον η συνολική διακύμανση των διανυσμάτων $\mathbf{x}_n$ είναι ίση με το άθροισμα των ιδιοτιμών λ_i, η επιλογή των ιδιοδιανυσμάτων που αντιστοιχούν στις μεγαλύτερες ιδιοτιμές προσφέρει μέγιστη δυνατή διακύμανση για το εκάστοτε p.

Η PCA, λοιπόν, προβάλλει ένα σύνολο δεδομένων διάστασης d σε ένα νέο χώρο διάστασης $p \leq d$, μεγιστοποιώντας την ενέργεια/διακύμανση της προσέγγισης με τις p πρωτεύουσες συνιστώσες, ελαχιστοποιώντας το μέσο τετραγωνικό σφάλμα της προβολής και εξασφαλίζοντας πως τα μετασχηματισμένα δεδομένα είναι ορθογώνια (ασυσχέτιστα) μεταξύ τους. Αν επιθυμούμε τα μετασχηματισμένα ασυσχέτιστα δεδομένα να έχουν και την ίδια διακύμανση, τότε μπορούμε να πολλαπλασιάσουμε το διάνυσμα των μετασχηματισμένων δεδομένων με το διαγώνιο πίνακα $\mathbf{\Lambda}$, ο οποίος ορίζεται μέσω της σχέσης 3.32:

$$\mathbf{\Lambda} = \text{diag}(\frac{1}{\sqrt{\lambda_1 + \epsilon}}, \cdots, \frac{1}{\sqrt{\lambda_p + \epsilon}}) \tag{3.32}$$

όπου λ_i είναι η i-οστή μεγαλύτερη ιδιοτιμή του πίνακα συνδιακύμανσης και ϵ μια μικρή σταθερά regularization που μαας εξασφαλίζει αριθμητική ευστάθεια. Ο πίνακας διακύμανσης των δεδομένων μετά από αυτή τη διαδικασία, που ονομάζεται λευκοποίηση (whitening) θα είναι ένας μοναδιαίος πίνακας.

Συνοψίζοντας, η μαθηματική διατύπωση της μεθόδου PCA με χρήση whitening είναι:

$$\mathbf{Y}^T = \mathbf{X}^T \mathbf{A} \mathbf{\Lambda} \tag{3.33}$$

όπου $\mathbf{X} \in \mathbb{R}^{L \times N}$ είναι ο πίνακας των δεδομένων, μηδενικής μέσης τιμής και μοναδιαίας διακύμανσης, $\mathbf{Y} \in \mathbb{R}^{p \times N}$ είναι η αναπαράσταση του $\mathbf{X}$ μετά την εφαρμογή της PCA, όπου η διάσταση του εκάστοτε διανύσματος (π.χ. περιγραφητή) έχει μειωθεί από L σε p και $\mathbf{A} \in \mathbb{R}^{L \times p}$ είναι ο πίνακας των ιδιοδιανυσμάτων του $\mathbf{X}\mathbf{X}^T$. Επομένως η PCA είναι ουσιαστικά μια διαδικασία περιστροφής του αρχικού πίνακα δεδομένων σε ένα νέο σύστημα συντεταγμένων με κριτήριο τη μεγιστοποίηση της διακύμανσης. Για να υπολογίσουμε τα ιδιοδιανύσματα του πίνακα $\mathbf{X}\mathbf{X}^T$ μπορεί να χρησιμοποιηθεί η μέθοδος Διάσπασης Ιδιάζουσων Τιμών (Singular Value Decomposition - SVD).

Επιστρέφοντας στο πρόβλημα της αναγνώρισης δράσεων, για τις ανάγκες της μείωσης της διάστασης των περιγραφητών, οι p κύριες κατευθύνσεις (principal directions), δηλαδή ο πίνακας $\mathbf{A}$, υπολογίζονται για κάθε περιγραφητή με βάση ένα υποσύνολο των χαρακτηριστικών εκπαίδευσης (το ίδιο υποσύνολο που χρησιμοποιείται και για τον υπολογισμό των οπτικών λέξεων). Στη συνέχεια, κάθε χαρακτηριστικό προβάλλεται στο σύστημα συντεταγμένων που ορίζει ο $\mathbf{A}$. Η νέα διάσταση κάθε μετασχηματισμένου χαρακτηριστικού μπορεί να διαφέρει για διαφορετικούς περιγραφητές. Παραδείγματος χάριν, το μήκος του περιγραφητή Trajectory μπορεί να επιλεχθεί να μειωθεί από 30 σε 15, ενώ του περιγραφητή HOG από 96 σε 64.

Κεφάλαιο 4

Μηχανές Διανυσματικής Υποστήριξης (SVMs)

Σε αυτό το κεφάλαιο θα αναλύσουμε τον ταξινόμητη SVM (Support Vector Machine), ο οποίος χρησιμοποιείται ευρέως για την κατηγοριοποίηση ανθρώπινων δράσεων. Θα περιγράψουμε πώς τα γραμμικά SVMs μπορούν να ταξινομήσουν γραμμικώς ή μη γραμμικώς διαχωρίσιμα δεδομένα και θα αναφερθούμε στη σημασία των συναρτήσεων πυρήνα (kernel functions) για την κατασκευή μη γραμμικών ταξινόμητων, παραθέτοντας τις πιο δημοφιλείς συναρτήσεις πυρήνα που χρησιμοποιούνται στο πεδίο της αναγνώρισης δράσεων. Τέλος, θα σκιαγραφήσουμε κάποιες από τις μεθόδους της βιβλιογραφίας που επιτρέπουν το συνδυασμό πολλών περιγραφητών (και εν γένει τη σύμμειξη (fusion) πολλών ροών πληροφορίας) στο επίπεδο του ταξινομητή.

4.1 Εισαγωγή

Οι μηχανές διανυσματικής υποστήριξης (support vector machines) είναι μοντέλα μηχανικής μάθησης που χρησιμοποιούνται στην ταξινόμηση και τη γραμμική παλινδρόμηση. Εισήχθησαν από τους Cortes και Vapnik [48] το 1995 και βασίζονται στις έννοιες του μέγιστου περιθωρίου (maximum margin classifier) και των συναρτήσεων πυρήνα [45]. Ανήκει στην οικογένεια των αλγόριθμων ταξινόμησης, οι οποίοι δεδομένου ενός πεπερασμένου πλήθος διακριτών κατηγοριών έχουν σκοπό να προσδιορίσουν σε ποια κατηγορία ανήκει κάποιο διάνυσμα εισόδου (π.χ. μια Bag-Of-Visual-Words αναπαράσταση ενός βίντεο). Ο αλγόριθμος SVM ανήκει στους αλγορίθμους επιβλεπόμενης μάθησης (supervised learning), όπου το σύνολο των στιγμιοτύπων εκπαίδευσης (training set) αποτελείται από διανύσματα εισόδου συνοδευόμενα από την ετικέτα της κατηγορίας στην οποία ανήκουν.

4.2 Support Vector Machines: Η γραμμική περίπτωση

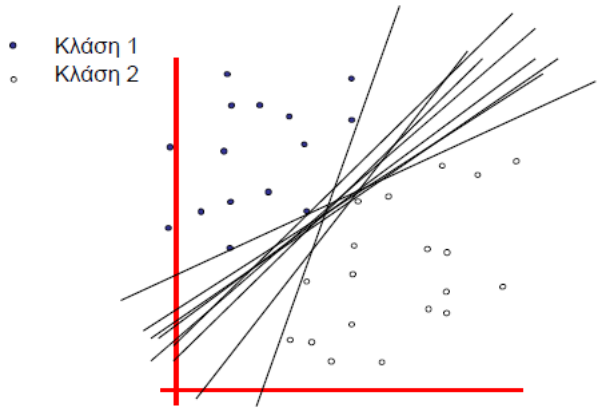
4.2.1 Γραμμικώς διαχωρίσιμα δεδομένα

Έστω ότι έχουμε ένα δυαδικό πρόβλημα ταξινόμησης και επιθυμούμε να διαχωρίσουμε τα δεδομένα σε δύο κλάσεις με ένα γραμμικό μοντέλο της μορφής:

$$y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b \tag{4.1}$$

Το σύνολο εκπαίδευσης αποτελείται από N διανύσματα εισόδου $\mathbf{x}_1, \ldots, \mathbf{x}_N$ [1] με αντίστοιχες ετικέτες κλάσεων $t_1, \ldots, t_N$, όπου $t_i \in \{-1, 1\}$, και ο αλγόριθμος ταξινομεί νέα δεδομένα $\mathbf{x}$ σύμφωνα με το πρόσημο του $y(\mathbf{x})$.

Εφόσον τα δεδομένα είναι γραμμικώς διαχωρίσιμα, υπάρχει τουλάχιστον ένας συνδυασμός $\mathbf{w}, b$ τέτοιος ώστε $y(\mathbf{x}_n) > 0$ για τα διανύσματα εισόδου με $t_n = 1$ και $y(\mathbf{x}_n) < 0$ για τα διανύσματα εισόδου με $t_n = -1$. Όμως η επιλογή της διαχωριστικής επιφάνειας δεν είναι μοναδική, όπως φαίνεται και στο Σχήμα 4.1.

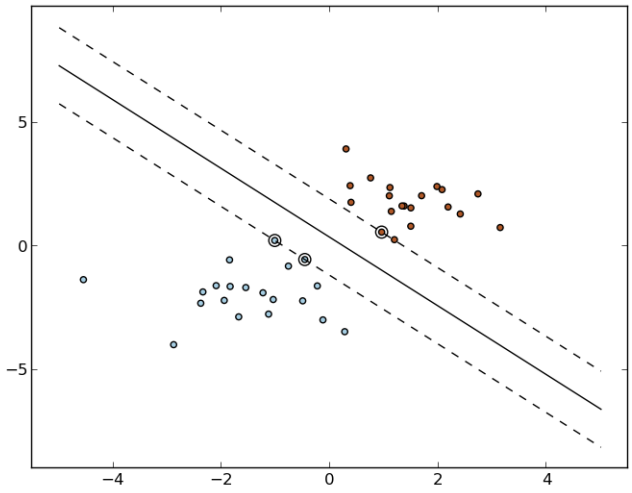


Σχήμα 4.1: Παραδείγματα διαχωριστικών γραμμών μεταξύ των δεδομένων δύο κλάσεων.

Ο αλγόριθμος SVM επιλέγει συστηματικά τον ταξινομητή που δίνει το μικρότερο σφάλμα, και αυτός είναι διαισθητικά ο ταξινομητής με το μεγαλύτερο περιθώριο, όπου περιθώριο ονομάζεται η κάθετη απόσταση ανάμεσα στο

[1] Τονίζουμε ότι το σύνολο διανυσμάτων εισόδου $\mathbf{x}_1, \ldots, \mathbf{x}_N$ δεν πρέπει να συγχέεται με το σύνολο χαρακτηριστικών που είχαν τον ίδιο συμβολισμό στα προηγούμενα κεφάλαια. Στο πλαίσιο της κατηγοριοποίησης των βίντεο με τον όρο διανύσματα εισόδου αναφερόμαστε στα διανύσματα αναπαράστασης (π.χ. BoVW, VLAD, FV) των N βίντεο εισόδου.

σύνορο απόφασης και στα πιο κοντινά του διανύσματα δεδομένων. Η θέση του συνόρου καθορίζεται μόνο από ένα υποσύνολο των διανυσμάτων εισόδου, γνωστά ως *διανύσματα υποστήριξης*, τα οποία βρίσκονται πάνω στο περιθώριο, όπως φαίνεται στην εικόνα 4.2.



Σχήμα 4.2: Περιθώριο ονομάζεται η κάθετη απόσταση μεταξύ του συνόρου απόφασης και των πιο κοντινών σε αυτό διανυσμάτων εισόδου. Η μεγιστοποίηση του περιθωρίου οδηγεί σε μια συγκεκριμένη επιλογή διαχωριστικής επιφάνειας, της οποίας η θέση ορίζεται από ένα υποσύνολο των δεδομένων εισόδου που απεικονίζονται κυκλωμένα.

Το ερώτημα που τίθεται λοιπόν είναι πώς μπορούμε να υπολογίσουμε τα $\mathbf{w}, b$ που μας εξασφαλίζουν σύνορο απόφασης με μέγιστο περιθώριο. Όπως βλέπουμε στο Σχήμα 4.3, το διάνυσμα $\mathbf{w}$ είναι κάθετο στο επίπεδο $\mathbf{w}^T\mathbf{x}+b = 0$ και σε όλα τα παράλληλα προς αυτό επίπεδα. Πράγματι, αν θεωρήσουμε $\mathbf{u}, \mathbf{v}$ διανύσματα που ανήκουν στο επίπεδο $\mathbf{w}^T\mathbf{x} + b = 1$ π.χ. τότε

$$\mathbf{w}^T\mathbf{u} + b = 1$$

και

$$\mathbf{w}^T\mathbf{v} + b = 1.$$

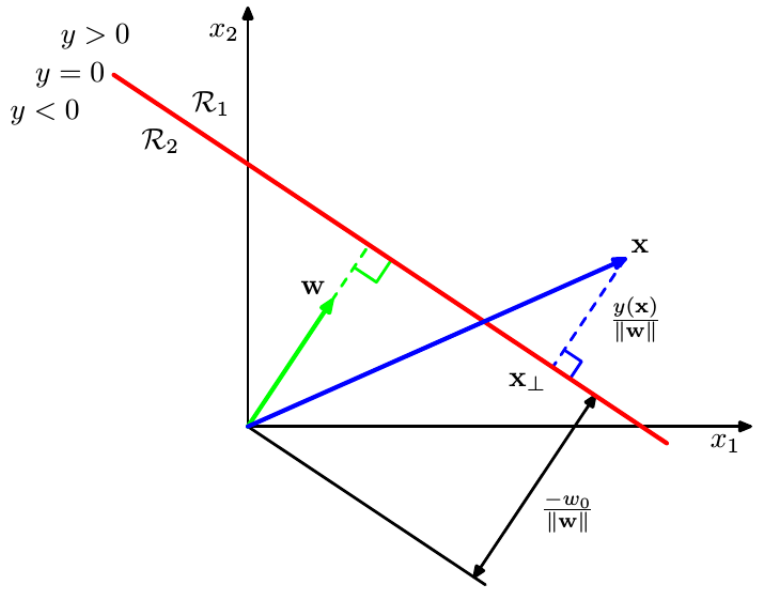
Αν αφαιρέσουμε κατά μέλη τις παραπάνω εξισώσεις παίρνουμε $\mathbf{w}^T(\mathbf{u}-\mathbf{v}) = 0$.

Άρα το $\mathbf{w}$ είναι η διεύθυνση της επιφάνειας απόφασης. Για τα διανύσματα που βρίσκονται πάνω στο σύνορο απόφασης ισχύει $y(\mathbf{x}) = 0$ και έτσι η κάθετη

απόσταση του από την αρχή των αξόνων ισούται με:

$$\frac{\mathbf{w}^T\mathbf{x}}{||\mathbf{w}||} = -\frac{b}{||\mathbf{w}||}$$

Όπως βλέπουμε και στο σχήμα 4.3 $\mathbf{x} = \mathbf{x}_{normal} + r\frac{\mathbf{w}}{||\mathbf{w}||}$, όπου $\mathbf{x}_{normal}$ είναι η προβολή του διανύσματος $\mathbf{x}$ πάνω στο σύνορο απόφασης, r η κάθετη απόσταση του διανύσματος $\mathbf{x}$ από το σύνορο και $\frac{\mathbf{w}}{||\mathbf{w}||}$ το μοναδιαίο διάνυσμα στη διεύθυνση του $\mathbf{w}$. Πολλαπλασιάζοντας τα δύο μέλη της παραπάνω σχέσης με $\mathbf{w}^T$ και προσθέτοντας το b , λαμβάνοντας υπόψη ότι $y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b = 0$ και $y(\mathbf{x}_{normal}) = \mathbf{w}^T\mathbf{x}_{normal} + b = 0$, παίρνουμε ότι $r = \frac{y(\mathbf{x})}{||\mathbf{w}||}$.



Σχήμα 4.3: Η διαχωριστική επιφάνεια που χωρίζει το χώρο των διανυσμάτων εισόδου σε δύο περιοχές $\mathcal{R}_1, \mathcal{R}_2$ απεικονίζεται με κόκκινο χρώμα και είναι κάθετη στο διάνυσμα $\mathbf{w}$. Τέλος, η μετατόπισή της από την αρχή των αξόνων ελέγχεται από το κατώφλι w_0 (ή b). [45]

Όπως αναλύσαμε, η λύση που ψάχνουμε είναι αυτή που μεγιστοποιεί την απόσταση του κοντινότερου σημείου από το επίπεδο, η οποία απόσταση δεν είναι άλλη από το r που μόλις προσδιορίσαμε. Άρα η απαίτησή μας μπορεί να διατυπωθεί ως:

$$\underset{\mathbf{w},b}{\operatorname{argmax}} \left\{ \frac{1}{||\mathbf{w}||} t_n(\mathbf{w}^T\mathbf{x}_n + b) \right\} \tag{4.2}$$

Έστω ότι η απόσταση των γραμμών του περιθωρίου από το σύνορο απόφασης είναι d. Τότε στην κλάση 1 ανήκουν όσα διανύσματα ικανοποιούν την ανίσωση:

$\mathbf{w}^T\mathbf{x}+b \geq d$, ενώ στην κλάση -1 όσα ικανοποιούν την ανίσωση: $\mathbf{w}^T\mathbf{x}+b \leq -d$. Η παράμετρος d μπορεί να λάβει οποιαδήποτε τιμή, που σημαίνει ότι τα δύο επίπεδα $\mathbf{w}^T\mathbf{x}+b = \pm d$ μπορούν να πλησιάσουν ή να απομακρυνθούν όσο θέλουν. Κρατώντας σταθερή την τιμή της παραμέτρου d και διαιρώντας και τα δύο μέλη της προηγούμενης εξίσωσης με d, λαμβάνουμε ± 1 στη δεξιά πλευρά της εξίσωσης. Όμως προφανώς δεν έχει αλλάξει η διεύθυνση και η θέση των δύο υπερεπιπέδων στο χώρο. Το ίδιο ισχύει και για το σύνορο απόφασης. Η κανονικοποίηση με μια σταθερά d δεν επηρεάζει τα σημεία που βρίσκονται πάνω σε ένα υπερεπίπεδο (και το ορίζουν) [49]. Επομένως μπορούμε να θεωρήσουμε το περιθώριο ως την περιοχή μεταξύ δύο παράλληλων υπερεπιπέδων $\mathbf{w}^T\mathbf{x}+b = \pm 1$. Άρα τώρα η Ευκλείδια απόσταση μεταξύ οποιουδήποτε σημείου πάνω σε αυτά τα υπερεπίπεδα και του υπερεπιπέδου του ταξινομητή (σύνορο απόφασης) είναι ίση με $\frac{1}{||\mathbf{w}||}$. Άρα θέλουμε να μεγιστοποιήσουμε το $\frac{1}{||\mathbf{w}||}$ ή ισοδύναμα να λύσουμε το ακόλουθο πρόβλημα ελαχιστοποιήσης:

$$\begin{aligned} &\underset{\mathbf{w},b}{\operatorname{argmin}} \quad \left\{ \frac{1}{2} ||\mathbf{w}||^2 \right\} \\ &\text{s.t.} \qquad t_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1,\ n = 1, \ldots, N. \end{aligned} \tag{K.K.T.1}$$

Πρόκειται για ένα πρόβλημα τετραγωνικού προγραμματισμού, όπου στόχος είναι η βελτιστοποίηση μιας τετραγωνικής συνάρτησης διάφορων μεταβλητών που υπόκεινται σε γραμμικούς περιορισμούς ανισοτήτων.

Για την επίλυση αυτού του προβλήματος εισάγουμε τους πολλαπλασιαστές Lagrange. Κατασκευάζουμε τη συνάρτηση Lagrange:

$$L(\mathbf{w}, b, \mathbf{a}) = \frac{1}{2} ||\mathbf{w}||^2 - \sum_{i=1}^{N} a_i \left[t_i(\mathbf{x}_i\mathbf{w} + b) - 1 \right] \tag{4.3}$$

όπου $\mathbf{a}^T = (a_1, \ldots, a_N)$ είναι ένα διάνυσμα μη αρνητικών πολλαπλασιαστών, ένας για κάθε περιορισμό.

$$a_n \geq 0,\ n = 1, \ldots, N \tag{K.K.T.2}$$

Αποδεικνύεται ότι η λύση στο δικό μας πρόβλημα βελτιστοποίησης καθορίζεται από το σαγματικό σημείο αυτής της Lagrangian συνάρτησης στον $2N+1$-διάστατο χώρο των $\mathbf{w}, b, \mathbf{a}$, όπου θέλουμε να την ελαχιστοποιήσουμε ως προς $\mathbf{w}$ και b και να την μεγιστοποιήσουμε ως προς a (εξού και το αρνητικό πρόσημο μπροστά από τον πολλαπλασιαστή). Άρα στο βέλτιστο σημείο $(\mathbf{w}_o, b_o)$ παίρνουμε:

$$\left.\frac{\partial L(\mathbf{w}, b, \mathbf{a})}{\partial x}\right|_{\mathbf{w}=\mathbf{w}_0} = \left(\mathbf{w}_0 - \sum_{i=1}^{N} a_i t_i \mathbf{x}_i \right) = 0 \rightarrow \mathbf{w}_0 = \sum_{i=1}^{N} a_i t_i \mathbf{x}_i \tag{K.K.T.3}$$

$$\left.\frac{\partial L(\mathbf{w}, b, \mathbf{a})}{\partial b}\right|_{b=b_0} = \left(\sum_{a_i}^{N} a_i t_i\right) = 0 \qquad \text{(K.K.T.4)}$$

Τα προβλήματα βελτιστοποίησης αυτής της μορφής ικανοποιούν τις συνθήκες Karush-Kuhn-Tucker (K.K.T), τέσσερις εκ των οποίων παρουσιάστηκαν παραπάνω και η πέμπτη δίνεται από το θεώρημα Kuhn-Tucket, το οποίο δηλώνει ότι στο σαγματικό σημείο $\mathbf{w}_0, b_0, \mathbf{a}_0$ κάθε πολλαπλασιαστής Lagrange και ο περιορισμός στον οποίο αντιστοιχεί, συνδέονται μέσω της εξίσωσης:

$$a_n(t_n(\mathbf{w}_0^T\mathbf{x}_n + b_0) - 1) = 0,\ n = 1, \ldots, N \qquad \text{(K.K.T.5)}$$

Από τις συνθήκες K.K.T.3 και K.K.T.5 προκύπτουν δύο θεμελιώδους σημασίας συμπεράσματα: 1) $\mathbf{w}_0 = \sum_{i=1}^{N} a_i t_i \mathbf{x}_i$ και άρα η λύση του βέλτιστου υπερεπιπέδου μπορεί να γραφεί ως γραμμικός συνδυασμός των διανυσμάτων εκπαίδευσης και 2) από την K.K.T.5 παρατηρούμε ότι $a_n \neq 0$ μόνο για τα διανύσματα $\mathbf{x}_n$ που ικανοποιούν τη σχέση $t_n(\mathbf{w}_0^T\mathbf{x}_n + b_0) = 1$. Αυτά όμως δεν είναι άλλα παρά τα διανύσματα υποστήριξης που προαναφέραμε. Επομένως αποδείξαμε ότι η θέση του συνόρου καθορίζεται μόνο από ένα υποσύνολο των διανυσμάτων εισόδου, τα διανύσματα υποστήριξης, και είναι αυτά που βρίσκονται πάνω στο περιθώριο.

Στη συνέχεια, αντικαθιστώντας τις βέλτιστες τιμές των $\mathbf{w}$ και b στην Lagrangian παίρνουμε:

$$L(\mathbf{w}_0, b_0, \mathbf{a}) = \ldots = \sum_{i=1}^{N} a_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j t_i t_j (\mathbf{x}_i)^T \mathbf{x}_j - b\left(\sum_{a_i} a_i t_i\right) \quad (4.4)$$

με τον τελευταίο όρο να μηδενίζεται λόγω της K.K.T.4 Άρα καταλήξαμε στο δυικό πρόβλημα:

$$\begin{aligned} \max_{\mathbf{a}} \quad & \sum_{i=1}^{N} a_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j t_i t_j < \mathbf{x}_i, \mathbf{x}_j > \\ \text{s.t.} \quad & a_n \geq 1,\ \forall n \\ & \left(\sum_{i=1}^{N} a_i t_i\right) = 0 \end{aligned} \qquad (4.5)$$

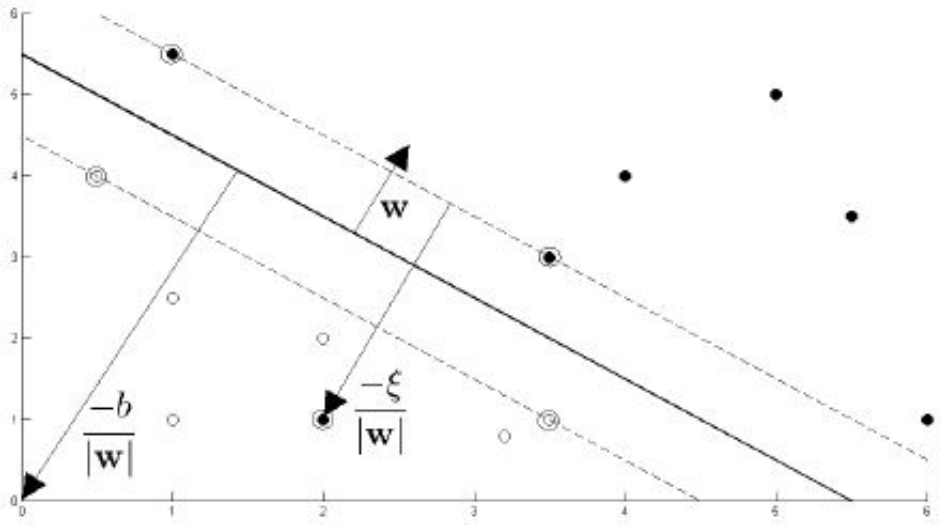
Το παραπάνω τετραγωνικό πρόβλημα βελτιστοποίησης είναι κυρτό και έτσι το πρόγραμμα επίλυσης βρίσκει μοναδική λύση [50]. Σε αυτό το σημείο πρέπει να σημειωθεί μια σημαντική λεπτομέρεια με σημαντικές προεκτάσεις που

θα αναλυθούν στη συνέχεια του κεφαλαίου: στο πρόβλημα βελτιστοποίησης εμφανίζεται μόνο το εσωτερικό γινόμενο $< \mathbf{x}_i, \mathbf{x}_j >$ ενώ τα διανύσματα εισόδου $\mathbf{x}_i$ δεν εμφανίζονται ποτέ μόνα τους. Λύνοντας το τετραγωνικό πρόβλημα θα υπολογιστούν το $\mathbf{a}^T = (a_1, \ldots, a_N)$ και μετά μέσω της K.K.T.3 το $\mathbf{w}$. Όπως προαναφέρθηκε, τα διανύσματα υποστήριξης είναι αυτά για τα οποία οι πολλαπλασιαστές Lagrange δεν είναι μηδενικοί. Έτσι βρίσκουμε λοιπόν και το σύνολο S των support vectors. Τέλος μας μένει να προσδιορίσουμε το b και αυτό γίνεται μέσω της σχέσης $b = \frac{1}{N_S} \sum_{s \in S} \left(t_s - \sum_{m \in S} a_m t_m \mathbf{x}_m \mathbf{x}_s \right)$.

Ο αλγόριθμός μας έχει δυνατότητα γενίκευσης, το οποίο είναι και το ζητούμενο, και μπορεί να ταξινομεί νέα διανύσματα εισόδου $\mathbf{x}'$ αποτιμώντας το $y' = \text{wgn}\left(\mathbf{w}^T \mathbf{x}' + b\right)$, όπως αναφέραμε και στην εισαγωγή.

4.2.2 Μη Γραμμικώς διαχωρίσιμα δεδομένα

Για να αντιμετωπίσουμε την περίπτωση των μη γραμμικώς διαχωρίσιμων δεδομένων θα χαλαρώσουμε τους περιορισμούς για να επιτρέπουν και κάποια λάθος ταξινομημένα διανύσματα. Αυτό το επιτυγχάνουμε εισάγοντας τις *μεταβλητές χαλαρότητας* (slack variables) $\xi_i \geq 0 \ i = 1, \ldots, N$, έτσι ώστε να επιτρέψουμε κάποια διανύσματα εκπαίδευσης να είναι στο λάθος ημιεπίπεδο, επιβάλλοντάς τους ποινή, η οποία αυξάνεται με την απόσταση από το σύνορο απόφασης (Σχήμα 4.4).



Σχήμα 4.4: Υπερεπίπεδο που προκύπτει από ένα γραμμικό μοντέλο SVM στην προσπάθεια διαχωρισμού δύο μη γραμμικώς διαχωρίσεων κλάσεων δεδομένων.

Άρα τώρα οι περιορισμοί του προβλήματος γίνονται:

$$\begin{aligned} \mathbf{w}^T\mathbf{x}_n + b &\geq 1 - \xi_n, \qquad t_n = 1 \\ \mathbf{w}^T\mathbf{x}_n + b &\leq -1 + \xi_n, \quad t_n = -1 \end{aligned} \tag{4.6}$$

Το πρόβλημα βελτιστοποίησης μπορεί να επαναδιατυπωθεί ως η εύρεση ενός υπερεπιπέδου ταξινόμησης που να ελαχιστοποιεί το άθροισμα των αποκλίσεων των σφαλμάτων εκπαίδευσης και μεγιστοποιεί το περιθώριο για τα ορθά ταξινομημένα διανύσματα εισόδου.

$$\begin{aligned} \min \quad & \frac{1}{2}\left|\left|\mathbf{w}\right|\right|^2 + C\sum_{i=1}^{N}\xi_i \\ s.t.: \quad & t_n(\mathbf{w}^T\mathbf{x}_n + b) - 1 + \xi_n \geq 0,\ n = 1, \ldots, N \end{aligned} \tag{4.7}$$

Η παράμετρος C ελέγχει το trade-off μεταξύ της ποινής των ξ_i και του μεγέθους του περιθωρίου και η επιλογή του είναι πολύ σημαντική για την ακρίβεια της ταξινόμησης και εξαρτάται κάθε φορά απο την εκάστοτε εφαρμογή και έτσι προσδιορίζεται ως επί το πλείστον πειραματικά. Για να επιλύσουμε αυτό το πρόβλημα εργαζόμαστε όπως προηγουμένως:

- Διατυπώνουμε τη συνάρτηση Lagrange:

$$L(\mathbf{w}, b, \mathbf{a}, \boldsymbol{\xi}, \mathbf{v}) = \frac{1}{2}\left|\left|\mathbf{w}\right|\right|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}a_i\left[t_i(\mathbf{x}_i\mathbf{w} + b) - 1 + \xi_i\right] - \sum_{i=1}^{N}v_i\xi_i \tag{4.8}$$

- Την ελαχιστοποιούμε ως προς $\mathbf{w}, b$ και ξ και τη μεγιστοποιούμε ως προς $\mathbf{a}$.

$$\begin{aligned} &\left.\frac{\partial L(\mathbf{w}, b, \mathbf{a}, \boldsymbol{\xi}, \mathbf{v})}{\partial x}\right|_{\mathbf{w}=\mathbf{w}_0} = \left(\mathbf{w}_0 - \sum_{i=1}^{N}a_i t_i \mathbf{x}_i\right) = 0 \rightarrow \mathbf{w}_0 = \sum_{i=1}^{N}a_i t_i \mathbf{x}_i \\ &\left.\frac{\partial L(\mathbf{w}, b, \mathbf{a}, \boldsymbol{\xi}, \mathbf{v})}{\partial b}\right|_{b=b_0} = \left(\sum_{a_i}^{N}a_i t_i\right) = 0 \\ &\frac{\partial L(\mathbf{w}, b, \mathbf{a}, \boldsymbol{\xi}, \mathbf{v})}{\partial \xi_i} = 0 \rightarrow C = a_i + v_i \end{aligned} \tag{4.9}$$

- Αντικαθιστούμε τις παραπάνω τιμές στη συνάρτηση Lagrange και βρίσκουμε κατά τα γνωστά το δυικό πρόβλημα βελτιστοποίησης και τις

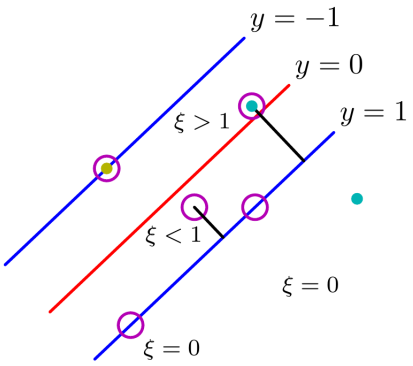
συνθήκες Karush-Kuhn-Tucker. Προκύπτει ίδιο προβλημα μεγιστοποίησης με πριν αλλά με διαφοροποιημένους περιορισμούς (επειδή $C = a_i + v_i$ και $v_i \geq 0 \rightarrow 0 \leq a_i \leq C$):

$$\begin{aligned} \max_{\mathbf{a}} \quad & \sum_{i=1}^{N} a_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j t_i t_j < \mathbf{x}_i, \mathbf{x}_j > \\ \text{s.t.} \quad & 0 \leq a_n \geq C, \ \forall n \\ & \left(\sum_{i=1}^{N} a_i t_i \right) = 0 \end{aligned} \tag{4.10}$$

και συνθήκες K.K.T.:

$$\begin{aligned} & C = a_n + v_n \\ & a_n \geq 0 \\ & t_n(\mathbf{w}^T \mathbf{x}_n + b) - 1 + \xi_n \geq 0 \\ & t_n(\mathbf{w}^T \mathbf{x}_n + b) - 1 + \xi_n = 0 \\ & v_n \geq 0 \\ & \xi_n \geq 0 \\ & v_n \xi_n = 0 \end{aligned} \tag{4.11}$$

Τα διανύσματα υποστήριξης είναι, όπως είδαμε και στην περίπτωση των γραμμικών διαχωρίσιμων δεδομένων, αυτά για τα οποία ισχύει $t_n(\mathbf{w}^T \mathbf{x}_n + b) - 1 + \xi_n > 0$. Αν για αυτά τα ισχύει ότι $a_n < C$ τότε πρέπει $v_n > 0$ και άρα $v_n \xi_n = 0 \rightarrow \xi_n = 0$ (δηλαδή έχουν μηδενική απόσταση από το περιθώριο, δηλαδή βρίσκονται πάνω σε αυτό). Διαφορετικά, αν $a_n = C$ τα σημεία βρίσκοντα εντός του περιθωρίου και αν $\xi_n \leq 1$ είναι σωστά ταξινομημένα, αλλιώς $\xi_n > 1$ και βρίσκονται στο λάθος ημιεπίπεδο (Σχήμα 4.5).



Σχήμα 4.5: Απεικόνιση των μεταβλητών χαλάρωσης $\xi_n \geq 0$. Τα κυκλωμένα σημεία (στιγμιότυπα εκπαίδευσης) είναι τα διανύσματα υποστήριξης. [45]

4.2.3 Πιθανοτική έξοδος ενός SVM

Σε αρκετές περιπτώσεις θέλουμε να δώσουμε ένα δεδομένο εισόδου σε έναν ταξινομητή και μας ενδιαφέρει ο βαθμός βεβαιότητας με τον οποίο ταξινομείται αυτό το δεδομένο στην κλάση $+1$. Τυπικά παραδείγματα αποτελούν ο συνδυασμός ατομικών προβλέψεων, δηλαδή προβλέψεων ένα δεδομένο να ανήκει ή όχι σε καθεμία από κλάσεις, και η επιλογή της "απόρριψης" του δεδομένου όταν δεν ανήκει σε καμία κλάση. Σε αυτές τις περιπτώσεις είναι χρήσιμο να υπολογίζεται η ύστερη πιθανότητα $P(t_n = 1|\mathbf{x}_n)$. Ωστόσο, οι μηχανές διανυσματικής υποστήριξης *δεν* παράγουν πιθανοτικές εξόδους αλλά αντιθέτως παίρνουν αποφάσεις ταξινόμησης αν ένα διάνυσμα εισόδου ανήκει στη μία ή στην άλλη κλάση.

Προσπαθώντας να δώσει μια λύση στο παραπάνω πρόβλημα, ο Platt [51] πρότεινε τη χρήση μιας παραμετρικής μορφής της σιγμοειδούς συνάρτησης:

$$P(t_n = 1|\mathbf{x}_n) = \frac{1}{1 + \exp(Ay(\mathbf{x}_n) + B)} \tag{4.12}$$

όπου το $y(\mathbf{x}_n)$ δίνεται από την εξίσωση 4.1 και αποτελεί την τιμή εξόδου του SVM ταξινομητή. Οι παράμετροι A, B προσδιορίζονται από την ελαχιστοποίηση της αρνητικής λογαριθμικής πιθανότητας των δεδομένων ενός συνόλου εκπαίδευσης:

$$\min -\sum_i t_i \log(p_i) + (1 - t_i) \log(1 - p_i) \tag{4.13}$$

όπου $p_i = p(t_i = 1|\mathbf{x}_i)$. Αν ως σύνολο εκπαίδευσης για τον προσδιορισμό της σιγμοειδούς χρησιμοποιηθούν τα ίδια δεδομένα που χρησιμοποιήθηκαν για την

εκπαίδευση του SVM μοντέλου, τότε ενέχει ο κίνδυνος της υπερπροσαρμογής (over-fitting). Γι'αυτό προτείνεται η χρήση της μεθόδου cross-validation, όπου το αρχικό σύνολο χωρίζεται σε 3 μέρη και εκπαιδεύεται ένα SVM μοντέλο για τα δεδομένα που ανήκουν σε 2 από τα 3 μέρη, για κάθε πιθανή δυάδα, και η σιγμοειδής προσαρμόζεται στο εκάστοτε τρίτο μέρος.

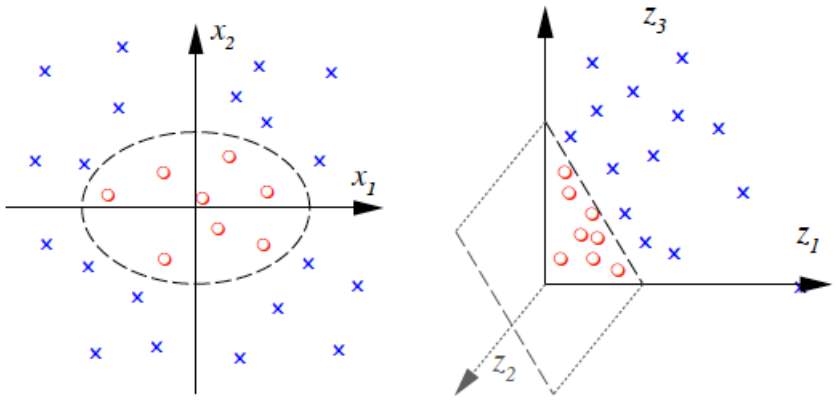
4.3 Support Vector Machines: Η μη γραμμική περίπτωση

Στην προηγούμενη ενότητα αναλύσαμε πώς οι μηχανές διανυσματικής υποστήριξης μπορούν να δημιουργήσουν *γραμμικές* διαχωριστικές επιφάνειες με σκοπό τον καλύτερο δυνατό διαχωρισμό δύο κλάσεων, που μπορεί να είναι γραμμικά ή μη γραμμικά διαχωρίσιμες. Εντούτοις σε πολλές περιπτώσεις οι κλάσεις είναι τέτοιες που καμία γραμμική διαχωριστική επιφάνεια δεν μπορεί εύλογα να τις διαχωρίσει, ενώ το μοντέλο SVM βρίσκει εξ ορισμού ένα $(N-1)$-διάστατο υπερεπίπεδο για ένα N-διάστατο σύνολο δεδομένων. Μια λύση για να παρακάμψουμε αυτό το πρόβλημα είναι να μετασχηματίσουμε τα διανύσματα εισόδου με κάποιο κατάλληλο (μη γραμμικό) μετασχηματισμό που να οδηγήσει σε αύξηση των διαστάσεων του αρχικού χώρου δεδομένων. Παραδείγματος χάριν, στο Σχήμα 4.6 απεικονίζεται ένα αρχικό σύνολο δισδιάστατων δεδομένων το οποίο δεν είναι εφικτό να διαχωριστεί με κάποια ευθεία σε δύο κλάσεις. Αν εφαρμόσουμε όμως έναν κατάλληλο μετασχηματισμό χώρου που απεικονίζει κάθε διάνυσμα δύο διαστάσεων σε ένα διάνυσμα τριών διαστάσεων $\Phi : \mathbb{R}^2 \to \mathbb{R}^3$:

$$\Phi(\mathbf{x}) = \begin{pmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{pmatrix} \tag{4.14}$$

παρατηρούμε ότι το μετασχηματισμένο σύνολο δεδομένων είναι γραμμικώς διαχωρίσιμο στο χώρο $\mathbb{R}^3$.

$$\Phi : R^2 \rightarrow R^3$$
$$(x_1, x_2) \mapsto (z_1, z_2, z_3) := (x_1^2, \sqrt{(2)}x_1x_2, x_2^2)$$



Σχήμα 4.6: Επίλυση μη γραμμικά διαχωρίσιμου προβλήματος με το μετασχηματισμό των δεδομένων σε έναν χώρο μεγαλύτερης διάστασης.

Με αυτή τη μέθοδο καταφέραμε να διαχωρίσουμε τα δεδομένα μας με ένα γραμμικό υπερεπίπεδο σε ένα χώρο μεγαλύτεων διαστάσεων, εφαρμόζοντας απλώς το μη γραμμικό μετασχηματισμό Φ σε κάθε διάνυσμα εισόδου και εκπαιδεύοντας κατά τα γνωστά ένα μοντέλο SVM. Το πρόβλημα με αυτή την προσέγγιση είναι ότι πολλές φορές τα δεδομένα πρέπει να μετασχηματιστούν σε ένα χώρο πολύ μεγάλης (ή ακόμα και άπειρης) διάστασης, με άμεση απόρροια ο υπολογισμός του μεταχηματισμού Φ για κάθε δεδομένο $\mathbf{x}_i$ να είναι υπολογιστικά δύσκολος έως και αδύνατος.

Το πρόβλημα αυτό μπορεί να ξεπεραστεί αν κανείς λάβει υπόψη τη διατύπωση του δυικού προβλήματος (Εξίσωση 4.10) που στην περίπτωση των μετασχηματισμένων δεδομένων η αντικειμενική συνάρτηση γίνεται:

$$\sum_{i=1}^{N} a_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j t_i t_j < \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) > \tag{4.15}$$

Παρατηρούμε ότι αρκεί να γνωρίζουμε το εσωτερικό γινόμενο $< \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) >$ για κάθε ζεύγος στιγμιοτύπων. Αυτό το εσωτερικό γινόμενο ονομάζεται *συνάρτηση πυρήνα* και μπορεί να υπολογιστεί πολύ πιο αποδοτικά:

$$k(\mathbf{x}_i, \mathbf{x}_j) =< \Phi(\mathbf{x}_i)\Phi(\mathbf{x}_j) > \tag{4.16}$$

Μάλιστα, οποιαδήποτε συνάρτηση μπορεί να γραφτεί με την παραπάνω μορφή, και είναι συμμετρική με τον πίνακα $\mathbf{K} = \{k(\mathbf{x}_i, \mathbf{x}_j)\}$ (kernel matrix) να είναι θετικά ημιορισμένος για κάθε πιθανό συνδυασμό διανυσμάτων στο χώρο των παρατηρήσεων, μπορεί να χρησιμοποιηθεί ως συνάρτηση πυρήνα χωρίς τη γνώση της συνάρτησης μετασχηματισμού $\Phi(\cdot)$. Οι συναρτήσεις πυρήνα μπορούν να ερμηνευτούν ως ένα μέτρο της ομοιότητας δύο διανυσμάτων στο χώρο εισόδου, όπως θα διαφανεί και στα παραδείγματα κάποιων δημοφιλών συναρτήσεων πυρήνα που ακολουθούν.

Υπάρχει πληθώρα συναρτήσεων πυρήνα για την κατασκευή μη γραμμικών μοντέλων SVM, όπως η Γκαουσιανή RBF, η Πολυωνυμική, η Σιγμοειδής και άλλες. Παρακάτω θα παραθέσουμε τρεις συναρτήσεις πυρήνα που χρησιμοποιούνται ευρέως σε συστήματα αναγνώρισης δράσεων και οδηγούν σε ικανοποιητικά αποτελέσματα όταν χρησιμοποιούνται στην ταξινόμηση διανυσμάτων εισόδου όπως τα BoVW, VLAD και Fisher.

1. Γραμμικός πυρήνας: Η πιο απλή συνάρτηση πυρήνα προκύπτει για τη μοναδιαία απεικόνιση $\Phi(\mathbf{x}) = \mathbf{x}$:

$$k(\mathbf{x}_i, \mathbf{x}_j) = {\mathbf{x}_i}^T \mathbf{x}_j \tag{4.17}$$

   Είναι δηλαδή το εσωτερικό γινόμενο μεταξύ δύο διανυσμάτων και οδηγεί στο γραμμικό SVM ταξινομητή της προηγούμενης ενότητας. Έχει διαπιστωθεί πειραματικά ότι οι αναπαραστάσεις VLAD και Fisher συνεργάζονται καλά με γραμμικούς ταξινομητές, οι οποίοι υπολογίζονται αποδοτικά.

2. Πυρήνας χ^2: Μη γραμμικές μηχανές διανυσματικής υποστήριξης με χ^2 πυρήνα [52] χρησιμοποιούνται κατά κόρον σε συνδυασμό με την BoVW αναπαράσταση. Έστω $\mathbf{x}_i = (u_1, \ldots, u_K)$ and $\mathbf{x}_j = (w_1, \ldots, w_K)$ δύο ιστογράμματα, π.χ. BoVW αναπαραστάσεις. Για να συγκρίνουμε αυτά τα ιστογράμματα χρησιμοποιούμε την χ^2 απόσταση που ορίζεται ως εξής:

$$D(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{2} \sum_{k=1}^{K} \frac{(u_k - w_k)^2}{u_k + w_k}$$

   Η αντίστοιχη χ^2 συνάρτηση πυρήνα είναι:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{1}{A} D(\mathbf{x}_i, \mathbf{x}_j))$$

   όπου A είναι η μέση τιμή των χ^2 αποστάσεων ανάμεσα στα στιγμιότυπα εκπαίδευσης και D η συνάρτηση απόστασης. Επειδή η χ^2 απόσταση είχε αναπτυχθεί ως απόσταση ανάμεσα σε διακριτές κατανομές πιθανότητας, συχνά το BoVW ιστόγραμμα κανονικοποιείται με τη χρήση της ℓ_1 νόρμας.

3. Πυρήνας Τομής Ιστογραμμάτων (Histogram Intersection Kernel): Ο πυρήνας τομής ιστογραμμάτων [53] (HIK) δίνεται από τη σχέση:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sum_{i=1}^{L} \min(u_i, w_i) \tag{4.18}$$

για δύο ιστογράμματα $\mathbf{x}_i = (u_1, \ldots, u_L)$ and $\mathbf{x}_j = (w_1, \ldots, w_L)$.

4.4 Σύμμειξη ροών πληροφορίας

Μέχρι τώρα είδαμε πώς δοθέντων των διανυσματικών αναπαραστάσεων των βίντεο (π.χ. των BoVW αναπαραστάσεων) μπορούμε να χρησιμοποιήσουμε μια μηχανή διανυσματικής υποστήριξης για να ταξινομήσουμε το κάθε βίντεο σε μια κατηγορία δράσης. Ωστόσο, όπως αναλύσαμε στο κεφάλαιο των διανυσματικών αναπαραστάσεων, κάθε βίντεο έχει πολλαπλές διανυσματικές αναπαραστάσεις, μία για κάθε τύπο περιγραφητή που έχει εξαχθεί από το βίντεο (Trajectory, HOG, HOF, MBH κλπ). Άρα με το παρόν πλαίσιο μπορούμε να ταξινομήσουμε το βίντεο σε μία κατηγορία για κάθε τύπο περιγραφητή, λαμβάνοντας υπόψιν τον κάθε περιγραφητή ανεξάρτητα από τους υπόλοιπους. Αν αναλογιστούμε όμως τη συμπληρωματικότητα των διάφορων περιγραφητών, που κωδικοποιούν διαφορετικές ροές πληροφορίας, όπως τη στατική εμφάνιση (HOG) ή την κίνηση (HOF), θα θέλαμε να είμαστε σε θέση να τους συνυπολογίσουμε όλους για την ταξινόμηση ενός video.

Η σύμμειξη (fusion) αυτών των ροών πληροφορίας μπορεί να γίνει σε διάφορες φάσεις της επεξεργασίας του βίντεο για την ταξινόμησή του.

- Μπορεί να επιτευχθεί με απλή συνένωση των περιγραφητών σε ένα διάνυσμα περιγραφητών πριν το στάδιο της αναπαράστασης του βίντεο (πρώιμη σύμμειξη - early fusion). Για παράδειγμα, οι Shu et al. [54] ενώνουν τους περιγραφητές κάθε τροχιάς σε ένα μεγάλο περιγραφητή και κατασκευάζουν το λεξικό οπτικών λέξεων βασισμένοι σε αυτόν.

- Η σύμμειξη τελικού σταδίου (late fusion) λαμβάνει μια απόφαση από τα μοντέλα SVM για κάθε περιγραφητή ξεχωριστά, και εξάγει μια τελική απόφαση μέσω της αξιολόγησης και του συνδυασμού των επιμέρους αποφάσεων. Οι Oneata et al. [55] ακολουθούν αυτή τη στρατηγική και συνδυάζουν γραμμικά τις εξόδους των SVM ταξινομητών που έχουν υπολογιστεί για κάθε περιγραφητή. Για την εύρεση των συντελεστών του γραμμικού συνδυασμού εκτελούν μια αναζήτηση σε ένα grid πιθανών συντελεστων, χρησιμοποιώντας cross-validation για την εύρεση των βέλτιστων συντελεστών ως προς μία μετρική αξιολόγησης.

- Οι διαφορετικοί περιγραφητές μπορούν να συνδυαστούν και στο επίπεδο του μοντέλου SVM (representation level fusion).

Σε αυτή τη διπλωματική ασχολούμαστε με την τελευταία μέθοδο σύμμειξης ροών πληροφορίας, όπου οι αναπαραστάσεις που έχουν εξαχθεί για κάθε περιγραφητή συνδυάζονται με απλή συνένωση ή με την κατασκευή μιας κατάλληλης συνάρτησης πυρήνα στο επίπεδο της SVM ταξινόμησης. Η μέθοδος αυτή χρησιμοποιείται κατά κόρον και στη βιβλιογραφία για το συνδυασμό των ροών πληροφορίας από διαφορετικούς descriptors, αν και μπορούν να χρησιμοποιηθούν και υβριδικές μέθοδοι [12]. Πώς μπορούμε να συνδυάσουμε όμως στο επίπεδο του SVM ταξινομητή τις επιμέρους ροές πληροφορίας;

Στην περίπτωση των VLAD και Fisher αναπαραστάσεων ο πιο συχνά χρησιμοποιούμενος τρόπος συνδυασμού διαφορετικών περιγραφητών είναι η συνένωση των διανυσματικών αναπαραστάσεων σε ένα μεγάλο διάνυσμα εισόδου του γραμμικού SVM ταξινομητή [31], [32].

Ένας τρόπος κατασκευής μιας συνάρτησης πυρήνα που να ενσωματώνει τις αποστάσεις όλων των περιγραφητών είναι η πολυκαναλική σύμμειξη, που χρησιμοποιήθηκε εκτός των άλλων και από τους Wang et al. για το συνδυασμό των χαρακτηριστικών που εξήγαγαν από τις πυκνές τροχιές. Ο χ^2 πυρήνας που χρησιμοποιείται στον ταξινομητή SVM ισούται με:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\sum_c \frac{1}{A^c} D(\mathbf{x}_i{}^c, \mathbf{x}_j{}^c)) \tag{4.19}$$

όπου c είναι το c-οστό κανάλι, δηλαδή $\mathbf{x}_i{}^c$ είναι το BoVW ιστόγραμμα ενός video υπολογισμένο βάσει του c-οστού περιγραφητή.

Μια σημαντική ιδιότητα των συναρτήσεων πυρήνα είναι ότι αν k_1, k_2 είναι θετικά ορισμένες συναρτήσεις πυρήνα, τότε η $k = ak_1 + bk_2$ είναι επίσης θετικά ορισμένη συνάρτηση πυρήνα για $a, b \geq 0$. Επομένως μια άλλη προσέγγιση συνδυασμού διαφορετικών περιγραφητών είναι ο υπολογισμός μιας συνάρτησης πυρήνα από την άθροιση των επιμέρους συναρτήσεων πυρήνα που έχουν υπολογιστεί για κάθε χαρακτηριστικό $K = \sum_c K_c$. Για παράδειγμα, στην περίπτωση του χ^2 πυρήνα:

$$\mathbf{K} = \left\{ \sum_c \exp(-\frac{1}{A^c} D(\mathbf{x}_i{}^c, \mathbf{x}_j{}^c)) \right\} \tag{4.20}$$

Επεκτείνοντας την παραπάνω μέθοδο, κανείς μπορεί να θεωρήσει οποιοδήποτε γραμμική συνδυασμό από επιμέρους συναρτήσεις πυρήνα. Μάλιστα, υπάρχει ένα σύνολο μεθόδων μηχανικής μάθησης, γνωστές με τον όρο Μάθηση Πολλαπλών Πυρήνων (Multiple Kernel Learning), οι οποίες έχουν ως σκοπό

να μάθουν από κοινού τις παραμέτρους του SVM και του γραμμικού συνδυασμού των πυρήνων από το σύνολο των δεδομένων εκπαίδευσης $\{(\mathbf{x}_i, t_i)\}$. Δοθέντος ενός συνόλου πυρήνων $\{K_c\}$ οι γραμμικές μέθοδοι MKL στοχεύουν στην εύρεση ενός γραμμικού συνδυασμού των επιμέρους kernels.

$$\mathbf{K} = \sum_c d_c \mathbf{K}_c \tag{4.21}$$

Στη μέθοδο Γενικευμένης Μάθησης Πολλαπλών Πυρήνων (GMKL) [56] αυτό το πρόβλημα διατυπώνεται ως ένα πρόβλημα βελτιστοποίησης, που περιέχει στην αντικειμένικη συνάρτηση δύο όρους κανονικοποίησης (regularization terms) και μια συνάρτηση απώλειας (loss function) (εξίσωση 4.22). Αυτή η διατύπωση κανονικοποιεί ταυτόχρονα τα βάρη του υπερεπιπέδου και τα βάρη του συνδυασμού των πυρήνων.

$$\begin{aligned} \min_{\mathbf{w},\boldsymbol{\xi}\geq 0,b,\mathbf{d}\geq 0} \quad & \frac{1}{2}\sum_k ||\mathbf{w}_k||_2^2 + C\sum_{i=1}^{N}\xi_i + r(\mathbf{d}) \\ \text{s.t.} \quad & t_i(\sum_k \sqrt{d_k}\mathbf{w}_k^T\Phi_k(\mathbf{x}_i) + b) \geq 1 - \xi_i \quad \forall i \end{aligned} \tag{4.22}$$

Ο regularizer $r(\mathbf{d})$ μπορεί να είναι οποιαδήποτε νόρμα (π.χ. L_1 ή L_2 νόρμα).

Η Μάθηση Πολλαπλών Πυρήνων μας επιτρέπει όχι μόνο να συνδυάζουμε διαφορετικούς πυρήνες, οι οποίοι μπορούν να αντιστοιχούν σε διαφορετικούς περιγραφητές και εν γένει σε διαφορετικές αναπαραστάσεις του βίντεο, αλλά και να διαπιστώνουμε με ένα διαισθητικό τρόπο ποιες ροές πληροφορίας είναι οι πιο σημαντικές για την ταξινόμηση του δείγματος σε μία κλάση. Τα βάρη που αντιστοιχίζονται σε κάθε πυρήνα σχετιζόμενο με έναν περιγραφητή μας δίνουν έμμεσα μια κατάταξη των περιγραφητών και μπορούν να μας βοηθήσουν στην επιλογή των καταλληλότερων.

4.5 Ταξινόμηση πολλαπλών κλάσεων

Όπως είδαμε και στη μαθηματική θεμελίωσή του, ο αλγόριθμος SVM εφαρμόζεται άμεσα μόνο σε προβλήματα δυαδικής ταξινόμησης. Εντούτοις, στις περισσότερες εφαρμογές επιθυμούμε την ταξινόμηση των δεδομένων σε περισσότερες από μία κλάσεις, για παράδειγμα σε περισσότερες από δύο δράσεις. Για να μπορέσει προσεγγιστικά να εφαρμοστεί ο αλγόριθμος SVM σε μία εφαρμογή ταξινόμησης πολλαπλών κλάσεων (multiclass classification) μπορεί να αναχθεί το πρόβλημα ταξινόμησης πολλαπλών κλάσεων σε πολλά προβλήματα δυαδικής ταξινόμησης. Θα αναφερθούμε συνοπτικά σε δύο από αυτές τις μεθόδους αναγωγής, οι οποίες εφαρμόζονται στη διεθνή βιβλιογραφία σε συστήματα αναγνώρισης δράσεων.

- Μέθοδος Ένας-Εναντίον-Όλων (One-Against-All ή OAA): Σύμφωνα με αυτή τη μέθοδο, εκπαιδεύουμε έναν ταξινομητή για κάθε κλάση του προβλήματός μας (άρα M μοντέλα SVM, όπου M ο αριθμός των κλάσεων). Το σύνολο εκπαίδευσης (training set) του i-οστού SVM αποτελείται από το σύνολο των δεδομένων της κλάσης i (ετικέτες $+1$) και το συμπλήρωμά του, δηλαδή τα στιγμιότυπα εκπαίδευσης που ανήκουν σε άλλες κλάσεις (ετικέτες -1). Για να ταξινομήσουμε ένα νέο διάνυσμα εισόδου $\mathbf{x}$, εφαρμόζουμε σε αυτό και τους M ταξινομητές ξεχωριστά και λέμε ότι το $\mathbf{x}$ ανήκει στην κλάση που έχει τη μεγαλύτερη τιμή εξόδου του SVM ταξινομητή.

- Μέθοδος Ένας-Εναντίον-Ενός (One-Against-One ή OAO): Εδώ εκπαιδεύουμε $\frac{M(M-1)}{2}$ ταξινομητές, ο καθένας εκ των οποίων εκπαιδεύεται σε δεδομένα από 2 κλάσεις i και j. Για να ταξινομήσουμε ένα νέο διάνυσμα εισόδου εφαρμόζουμε σε αυτό όλους τους ταξινομητές και ο καθένας δίνει μία ψήφο στην επικρατούσα κλάση. Το διάνυσμα εισόδου αποδίδεται στην κλάση που έλαβε τις περισσότερες ψήφους.

Κεφάλαιο 5

Χρονικός εντοπισμός και ταξινόμηση δράσεων σε συνεχή ροή βίντεο

5.1 Εισαγωγή

Στα προηγούμενα κεφάλαια παρουσιάσαμε μια πληθώρα μεθόδων εξαγωγής χαρακτηριστικών, αναπαράστασης βίντεο και ταξινόμησης, που μας επιτρέπουν να ταξινομήσουμε ένα βίντεο το οποίο περιέχει την εκτέλεση μιας δράσης σε μία από τις κατηγορίες ανθρώπινων δράσεων, τις οποίες έχουμε εκπαιδεύσει το σύστημα μας να αναγνωρίζει. Ωστόσο, η απλή ταξινόμηση βίντεο, τα οποία έχουν υποστεί χρονική κατάτμηση έτσι ώστε η αρχή και το τέλος τους να συμπίπουν με την αρχή και το τέλος μιας δράσης (presegmented videos), απέχει αρκετά από ρεαλιστικές εφαρμογές της αναγνώρισης δράσεων. Για παράδειγμα, ένα σύστημα παρακολούθησης θα πρέπει να μπορεί να ανιχνεύει τα χρονικά διαστήματα στα οποία λαμβάνει χώρα κάποια ασυνήθιστη δραστηριότητα, π.χ. όταν κάποιος άνθρωπος τρέχει ή μαλώνει με κάποιον άλλον [57]. Σε μια αθλητική εφαρμογή, ένα αυτόματο σύστημα αναγνώρισης δράσεων θα ήταν επιθυμητό να αναγνωρίζει ανά πάσα στιγμή ποια κίνηση αντισφαίρισης (tennis) εκτελείται (π.χ. service, backhand, smash). Τέλος, οι ταινίες είναι παραδείγματα μη χρονικά κομμένων βίντεο που περιέχουν συνεχή αλλαγή δράσεων.

Ο στόχος της αναγνώρισης δράσεων σε βίντεο, στα οποία δεν έχει εφαρμοστεί χρονική κατάτμηση έτσι ώστε να απομονωθεί μία μοναδική δράση, είναι:

1. ο χρονικός εντοπισμός (temporal localization) στο video των δράσεων,

2. η ταξινόμηση των δράσεων που ανιχνεύουμε σε μία από τις κατηγορίες ανθρώπινων δράσεων που επιθυμούμε να αναγνωρίζει το σύστημά μας.

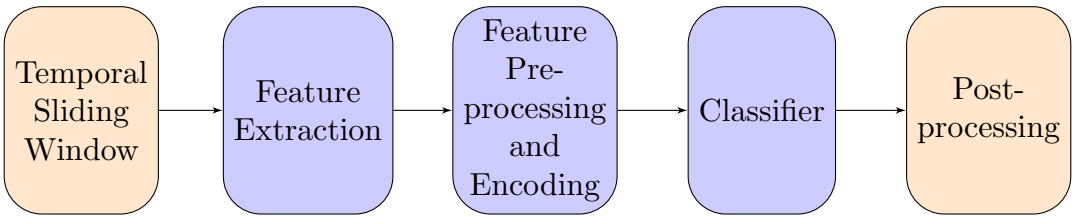
Από την παραπάνω διατύπωση γίνεται σαφές το γεγονός ότι η ταξινόμηση δράσεων, που παρουσιάστηκε στα προηγούμενα κεφάλαια, είναι υποπρόβλημα της αναγνώρισης δράσεων σε χρονικά "μη κομμένα" βίντεο (temporally untrimmed videos). Επομένως, αυτό το πρόβλημα συνιστά μια μεγαλύτερη πρόκληση για την ερευνητική κοινότητα στο πεδίο της αναγνώρισης ανθρώπινων δράσεων. Αξίζει να τονιστεί, ότι εκτός από το χρονικό εντοπισμό των δράσεων σε ένα βίντεο, ο χωρικός εντοπισμός (spatial localization) τους είναι επίσης βαρύνουσας σημασίας. Στο παράδειγμα του συστήματος αναγνώρισης δράσεων αντισφαίρισης, δε μας ενδιαφέρει μόνο *πότε* και *ποια* κίνηση εκτελείται, αλλά και *πού* εκτελείται στο χώρο, καθώς αυτή η πληροφορία μας επιτρέπει έμμεσα να συμπεράνουμε π.χ. ποιος παίκτης κάνει την κίνηση). Λόγω της σημασίας του, ο χωρικός εντοπισμός είναι ένα πρόβλημα που έχει απασχολήσει πολλούς ερευνητές και διαρκώς εξελίσσεται. Ωστόσο, στα πλαίσια αυτής της εργασίας θα ασχοληθούμε μόνο με το θέμα του χρονικού εντοπισμού και ταξινόμησης δράσεων, παρουσιάζοντας μια απλή επέκταση του συστήματος ταξινόμησης δράσεων των Κεφαλαίων 2 3, 4 με τη χρήση κυλιόμενου παραθύρου για την αναγνώριση συνεχόμενων δράσεων. Το προτεινόμενο σύστημα χρησιμοποιήθηκε σε ένα υποσύνολο της πολυτροπικής και πολυαισθητηριακής βάσης MOBOT για πειραματισμό και σύγκριση διαφορετικών ανιχνευτών χαρακτηριστικών (πυκνές τροχιές, βελτιωμένες πυκνές τροχιές), περιγραφητών (Trajectory, HOG, HOF, MBHx, MBHy) , αναπαραστάσεων (BoVW, VLAD) και κανονικοποιήσεων. Επίσης, εκμεταλλευόμενοι τη διαθέσιμη πληροφορία βάθους που προσφέρει η βάση MOBOT, πειραματιζόμαστε και με την εξαγωγή περιγραφητή εμφάνισης κατά μήκος των RGB τροχιών από το κανάλι βάθους.

5.2 Αναγνώριση συνεχόμενων ανθρώπινων δράσεων με τη χρήση πυκνών τροχιών

5.2.1 Επισκόπηση συστήματος

Για το χρονικό εντοπισμό ανθρώπινων δράσεων σε ρεαλιστικά βίντεο, τα οποία μπορεί να περιέχουν πολλαπλά στιγμιότυπα δράσεων από διαφορετικές κατηγορίες χρησιμοποιούμε ένα μη επικαλυπτόμενο κυλιόμενο παράθυρο (sliding window). Με τη βοήθεια του κυλιόμενου παραθύρου χωρίζουμε το βίντεο σε τμήματα (video segments) συγκεκριμένης χρονικής διάρκειας Δt πλαισίων (frames). Στη φάση της εκπαίδευσης, κατασκευάζουμε SVM μο-

ντέλα χρησιμοποιώντας αυτά τα video segments ως βίντεο εκπαίδευσης, αφού πρώτα έχουν εξαχθεί τα χαρακτηριστικά και έχουν υπολογιστεί οι αναπαραστάσεις του κάθε video segment. Κατά της φάση της αξιολόγησης, χωρίζουμε τα αντίστοιχα βίντεο σε τμήματα και ταξινομούμε με τον ίδιο τρόπο το κάθε τμήμα σε μία κατηγορία. Τέλος, αφού έχει ανατεθεί μια ετικέτα δράσης σε κάθε χρονικό τμήμα του βίντεο, μπορεί να ακολουθήσει μια μετεπεξεργασία (post-processing), που στόχο έχει να εκμεταλλευτεί τα αποτελέσματα πρόβλεψης γειτονικών video segments, έτσι ώστε να διορθώσει κάποιες λανθασμένες προβλέψεις. Η έξοδος του συστήματος για κάθε βίντεο αξιολόγησης είναι μια ακολουθία από κατηγορίες δράσεων στις οποίες ταξινομούνται τα τμήματα του βίντεο. Το συνολικό σύστημα απεικονίζεται στο Σχήμα 5.1. Όπως παρατηρούμε, για το χρονικό εντοπισμό και την αναγνώριση των δράσεων που εκτελούνται σε ένα βίντεο εισόδου, μια απλή μέθοδος που μπορεί να ακολουθηθεί είναι να χωρίσουμε το βίντεο σε χρονικά κομμάτια, να ταξινομήσουμε το καθένα ακολουθώντας τα γνωστά βήματα (εξαγωγή χαρακτηριστικών, κωδικοποίηση χαρακτηριστικών σε συμπαγή αναπαράσταση, χρήση SVM ταξινομητών) και προαιρετικά να επεξεργαστούμε τις εξόδους των ταξινομητών καταλήγοντας σε μια ακολουθία από αναγνωρισμένες δράσεις.



Σχήμα 5.1: Μπλοκ διάγραμμα του συστήματος χρονικού εντοπισμού και ταξινόμησης δράσεων. Αποτελείται κυρίως από πέντε στάδια: (α) ένα κυλιόμενο παράθυρο που χωρίζει το βίντεο εισόδου σε τμήματα τα οποία πρέπει να ταξινομηθούν (temporal sliding window) (β) εξαγωγή χαρακτηριστικών (feature extraction), (γ) προεπεξεργασία χαρακτηριστικών και κωδικοποίηση (feature pre-processing and encoding), (δ) χρήση ταξινομητών (classifiers) και (ε) επεξεργασία των πιθανοτικών εξόδων των SVM ταξινομητών (post-processing). (Τα γαλάζια blocks αντιστοιχούν στα βήματα ταξινόμησης ενός βίντεο που περιέχει μία δράση).

5.2.2 Κυλιόμενο παράθυρο

Στην υλοποίησή μας, εξάγουμε πυκνές τροχιές (dense trajectories ή improved dense trajectories) από όλο το video και έχουμε στη διάθεσή μας ένα σύνολο

από περιγραφητές (Trajectory, HOG, HOF, MBHx, MBHy) για κάθε τροχιά. Καθώς δουλεύουμε στο επίπεδο του χρονικού παραθύρου των Δt frames χρειάζεται να αναθέσουμε τροχιές στα τμήματα του βίντεο που δημιουργούναι με τη χρήση του κυλιόμενου παραθύρου. Για τα πειράματα που θα ακολουθήσουν, όσες τροχιές λήγουν σε frame που ανήκει σε ένα παράθυρο ανατίθενται στο παράθυρο αυτό. Εναλλακτικά θα μπορούσαμε να θεωρήσουμε ότι όσες τροχιές έχουν χρονική επικάλυψη με το παράθυρο μεγαλύτερη από το μισό μήκος της τροχιάς θα ανατίθενται σε αυτό.

Το μήκος του κυλιόμενου παραθύρου είναι μια πολύ σημαντική παράμετρος, η οποία καθορίζει την "διακριτική ικανότητα" του συστήματός μας. Πιο συγκεκριμένα, ένα πολύ μικρό παράθυρο θα οδηγήσει σε video segments με πολύ μικρή πληροφορία σχετικά με τη δράση που εκτελείται. Αντιθέτως, ένα υπερβολικά μεγάλο παράθυρο θα μπορούσε να συμπεριλαμβάνει την εκτέλεση δύο συνεχόμενων δράσεων, τις οποίες δε θα είμαστε σε θέση να διαχωρίσουμε, αφού το σύστημα προβλέπει μία ετικέτα για κάθε παράθυρο. Εν γένει, το μέγεθος του παραθύρου σχετίζεται με τη χρονική κλίμακα των χαρακτηριστικών που εξάγονται. Εν προκειμένω, στα χαρακτηριστικά τροχιών σχετίζεται με το μήκος των τροχιών.

5.2.3 Κατηγορία Background class

Προτού εστιάσουμε περισσότερο στις λεπτομέρειες υλοποίησης του κυλιόμενου παραθύρου και της μετεπεξεργασίας των εξόδων των ταξινομητών, χρειάζεται να τονιστεί μια ιδιαιτερότητα που εμφανίζεται στην ταξινόμηση τμημάτων από βίντεο συνεχόμενων δράσεων. Μια συνήθης υπόθεση στην περίπτωση της ταξινόμησης ενός βίντεο που περιέχει μόνο μία δράση είναι ότι το βίντεο ανήκει σε μία από τις κατηγορίες που το μοντέλο ταξινόμησης έχει εκπαιδευτεί να αναγνωρίζει. Όταν όμως το πρόβλημά μας είναι να ταξινομήσουμε ένα χρονικό κομμάτι ενός μεγάλου βίντεο, που περιέχει την εκτέλεση πολλαπλών δράσεων, ενδέχεται αυτό το τμήμα να μην περιέχει κάποια ανθρώπινη δράση ή να μην περιέχει καμία δράση που να ανήκει στις κατηγορίες που θέλουμε να ανιχνεύσουμε. Για παράδειγμα, σε ένα βίντεο ενός αγώνα αντισφαίρισης συνήθως υπάρχουν χρονικά διαστήματα στα οποία δεν εκτελείται από τους παίκτες κάποια κίνηση του αθλήματος (μπορεί να μην εκτυλίσσεται καμία ανθρώπινη δράση ή μπορεί κάποιος παίκτης να κάθεται ή να πίνει νερό, τα οποία δε συνιστούν δράση ενδιαφέροντος του υποθετικού συστήματος). Γι'αυτό χρησιμοποιούνται τέτοια video segments, που δεν περιέχουν κάποια δράση προς αναγνώριση, έτσι ώστε να εκπαιδευτεί ένα κατάλληλο μοντέλο SVM που θα αναγνωρίζει αυτή την ειδική κατηγορία που ονομάζουμε *Background class*.

5.2.4 Ομαλοποίηση Αποτελεσμάτων

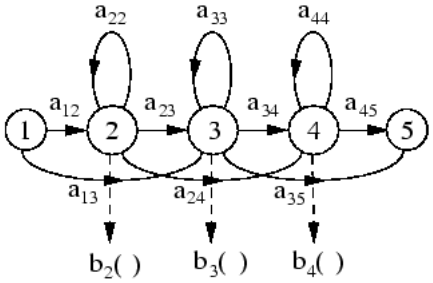
Αφού χωριστεί το βίντεο σε T τμήματα των Δt frames και προσδιοριστούν τα χαρακτηριστικά που ανήκουν σε κάθε τμήμα, ο αλγόριθμος ταξινόμησης αναλαμβάνει να ταξινομήσει το καθένα από αυτά τα τμήματα σε μία κατηγορία. Για την ταξινόμηση πολλαπλών κλάσεων με χρήση SVM χρησιμοποιούμε τη μέθοδο "ένας-εναντίον-όλων" (one-against-all), στην οποία, όπως αναλύθηκε και στο Κεφάλαιο 4, εκπαιδεύουμε M δυαδικούς SVM ταξινομητές που αναγνωρίζουν την κάθε κλάση, όπου M ο αριθμός των κλάσεων, και για την ταξινόμηση ενός στιγμιοτύπου παίρνουμε τις προβλέψεις αυτών των M ταξινομητών και κρατάμε την απόφαση αυτού που έχει το μεγαλύτερο score. Με άλλα λόγια κάθε δυαδικός ταξινομητής δίνει στην έξοδό του α) μια ετικέτα 1 ή -1, αν προβλέπει ότι το βίντεο ανήκει στην κατηγορία ή όχι και β) τις πιθανότητες να ανήκει στην κλάση ή όχι. Για κάθε τμήμα του βίντεο συγκρίνουμε, λοιπόν, τις πιθανότητες να ανήκει σε κάθε μία από τις κλάσεις, όπως αυτές προκύπτουν από τους ταξινομητές, και το ταξινομούμε στην κλάση με τη μεγαλύτερη πιθανότητα.

Μπορούμε να εκμεταλλευτούμε αυτές τις πιθανότητες για να λάβουμε υπόψη τις προβλέψεις γειτονικών παραθύρων, ώστε να διορθώσουμε την τελική ακολουθία δράσεων. Παραδείγματος χάριν, αν όλα τα χρονικά τμήματα πριν και μετά από ένα τμήμα, που έχει ταξινομηθεί στην κλάση c_i, έχουν ταξινομηθεί στην κλάση c_j είναι πιθανότερο το τμήμα αυτό να ανήκει επίσης στην κλάση c_j.

Γι'αυτό το σκοπό επεξεργαζόμαστε τις πιθανοτικές εξόδους των δυαδικών ταξινομητών του κάθε παραθύρου σε δύο βήματα. Αρχικά φιλτράρουμε τις πιθανοτικές εξόδους των SVMs που έχουν ληφθεί για κάθε παράθυρο και για κάθε κατηγορία δράσης, έτσι ώστε να τις ομαλοποιήσουμε. Το φιλτράρισμα ομαλοποιεί τις πιθανότητες έτσι ώστε να διορθώνονται λάθος ταξινομήσεις κάποιων παραθύρων βάσει των ταξινομήσεων των γειτονικών παραθύρων. Στη συνέχεια, μοντελοποιούμε το πρόβλημα αναγνώρισης συνεχόμενων δράσεων με χρήση ενός Κρυφού Μαρκοβιανού Μοντέλου (Hidden Markov Model - HMM).

Κρυφά Μαρκοβιανά Μοντέλα - Hidden Markov Models

Τα κρυφά Μαρκοβιανά μοντέλα HMM είναι στοχαστικές διαδικασίες που αποτελούνται από μια υποβόσκουσα στοχαστική διαδικασία, η οποία δεν είναι άμεσα παρατηρήσιμη (διαδικασία Markov με κρυφές καταστάσεις) παρά μόνο μέσω ενός άλλου συνόλου στοχαστικών διαδικασιών που παράγουν την ακολουθία των παρατηρήσεων.



Σχήμα 5.2: HMM μοντέλο.

Ένα μοντέλο HMM περιγράφεται λοιπόν από:

1. Το σύνολο των κρυφών καταστάσεων $\{1, 2, \ldots, N_s\}$

2. Την κατάσταση q_t του μοντέλου κάθε χρονική στιγμή t.

3. Το σύνολο των παρατηρήσιμων συμβόλων σε κάθε κατάσταση. Εν προκειμένω $V = \{v_1, v_2, ..., v_{M+1}\}$.

4. Την ακολουθία των παρατηρήσεων $\mathbf{O} = (\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T)$.

5. Την κατανομή πιθανότητας των παρατηρήσεων, $\mathbf{B} = \{b_j(k)\}$, όπου $b_j(k) = P[\mathbf{o}_t = \mathbf{v}_k | q_t = j], 1 \leq k \leq M+1$ είναι η κατανομή των συμβόλων στην κατάσταση j.

6. Ο πίνακας μεταβάσεων $\mathbf{A} = \{a_{ij}\}$, όπου a_{ij} είναι η πιθανότητα μετάβασης από την κατάσταση i στην κατάσταση j ($a_{ij} = P\left[q_{t+1} = j | q_t = i\right],\ 1 \leq i, j \leq N_s$) . Οι πιθανότητες a_{ij} θα πρέπει να ικανοποιούν τις σχέσεις $a_{ij} \geq 0$ και $\sum_{j=1}^{N_s} a_{ij} = 1$.

7. Τις πρότερες πιθανότητες $\pi_i = P\left[q_1 = i\right],\ 1 \leq i \leq N_s$ των καταστάσεων. Με άλλα λόγια, π_i είναι η πιθανότητα το HMM να ξεκινά από την κατάσταση i.

Τρία είναι τα βασικά προβλήματα που πρέπει να λυθούν, έτσι ώστε ένα μοντέλο HMM να είναι χρήσιμο σε εφαρμογές [58]:

- Πρόβλημα 1: Δεδομένης της ακολουθίας παρατηρήσεων $\mathbf{O} = (\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T)$ και ενός μοντέλου $\boldsymbol{\lambda} = (\mathbf{A}, \mathbf{B}, \pi)$, πώς υπολογίζουμε αποδοτικά την πιθανότητα η ακολουθία παρατηρήσεων να έχει παραχθεί από το μοντέλο ($P(\mathbf{O}|\lambda)$);

- Πρόβλημα 2: Δεδομένης της ακολουθίας παρατηρήσεων $\mathbf{O} = (\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T)$ και ενός μοντέλου $\boldsymbol{\lambda} = (\mathbf{A}, \mathbf{B}, \pi)$, πώς επιλέγουμε μια "βέλτιστη" ακολουθία κρυφών καταστάσεων $Q = q_1 q_2 \ldots q_T$;

- Προβλήμα 3: Πώς προσαρμόζουμε τις παραμέτρους του μοντέλου $\boldsymbol{\lambda} = (\mathbf{A}, \mathbf{B}, \pi)$ για να μεγιστοποιήσουμε την πιθανότητα $P(\mathbf{O}|\boldsymbol{\lambda})$; (Εκπαίδευση μοντέλου HMM).

Το πρόβλημα 1 λύνεται με τον αλγόριθμο forward-backward, το πρόβλημα 2 με τον αλγόριθμο Viterbi και το πρόβλημα 3 με τον αλγόριθμο Baum-Welch, που πρόκειται για ειδική περίπτωση του EM (Expectation-Maximization).

Επιστρέφοντας στο πρόβλημα της αναγνώρισης δράσεων, μοντελοποιούμε το πρόβλημα με ένα HMM με πρότερες πιθανότητες π_i και πιθανότητες μετάβασης από την κατάσταση i στην κατάσταση j, a_{ij}, του οποίου οι κρυφές κατάστασεις θεωρούμε οτι αντιστοιχούν στις δράσεις, ενώ ως παρατηρήσεις $\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T$ χρησιμοποιούμε τις πιθανοτικές εξόδους των SVMs, που μας δίνουν την πιθανότητα κάθε τμήμα του βίντεο να ανήκει σε καθεμία από τις $M + 1$ κλάσεις (κλάσεις ανθρώπινων δράσεων και η κατηγορία background class). Προσθέτοντας μια σταθερά στη διαγώνιο του πίνακα μετάβασης [59] (state transition penalty), μπορούμε να ρυθμίσουμε την ευελιξία του μοντέλου μας να αλλάζει καταστάσεις, αποτρέποντας συχνές μεταβάσεις από δράση σε δράση ανάμεσα σε γειτονικά τμήματα του βίντεο. Για να βρούμε την πιο πιθανή ακολουθία κρυφών καταστάσεων, εν προκειμένω δράσεων, που παρήγαγε την ακολουθία των πιθανοτικών εξόδων των SVMs που παρατηρήθηκε, πρέπει να χρησιμοποιήσουμε τον αλγόριθμο Viterbi.

Ο αλγόριθμος Viterbi

Το πρόβλημα της αποκωδικοποίησης (decoding) συνίσταται στην εύρεση μιας ακολουθίας κρυφών καταστάσεων που μεγιστοποιεί την πιθανότητα της ακολουθίας παρατηρήσεων. Θέλουμε να μεγιστοποιήσουμε, δηλαδή, την ύστερη πιθανότητα του να είμαστε σε μια κατάσταση i τη χρονική στιγμή t, δεδομένης της ακολουθίας παρατηρήσεων και του μοντέλου. Το πρόβλημα αυτό επιλύεται, όπως προαναφέραμε, με τον αλγόριθμο δυναμικού προγραμματισμού Viterbi.

Ο αλγόριθμος Viterbi ξεκινά με τον υπολογισμό των μερικών πιθανοτήτων $\delta_t(i),\ i = 1, \ldots, N_s$, όπου $\delta_t(i)$ είναι η πιθανότητα του βέλτιστου μονοπατιού που τελειώνει στην κατάσταση i στο χρόνο T.

$$\delta_t(i) = \max_{q_1 q_2 \cdots q_{t-1}} P\left[q_1 q_2 \ldots q_t = i, \mathbf{o}_1 \mathbf{o}_2 \ldots \mathbf{o}_t | \boldsymbol{\lambda}\right] \tag{5.1}$$

Επαγωγικά έχουμε:

$$\delta_{t+1}(j) = \left[\max_i \delta_t(i) a_{ij}\right] b_j(\mathbf{o}_{t+1}) \tag{5.2}$$

Για να είμαστε σε θέση να ανακτήσουμε τη βέλτιστη ακολουθία καταστάσεων που μας οδήγησε στην κατάσταση i, αποθηκεύουμε το όρισμα που μεγιστοποίησε την εξίσωση 5.2, για κάθε t και j, στον πίνακα με στοιχεία $\psi_t(j)$.

Αλγόριθμος 3 Αλγόριθμος Viterbi

1: Αρχικοποίηση.

$$\begin{aligned} \delta_1(i) &= \pi_i b_i(\mathbf{o}_1),\ 1 \leq N_s \\ \psi_1(i) &= 0 \end{aligned} \tag{5.3}$$

2: Αναδρομή.

$$\begin{aligned} \delta_t(j) &= \left[\max_{1\leq i\leq N_s} \delta_{t-1}(i) a_{ij}\right] b_j(\mathbf{o}_t),\ 2 \leq t \leq T \\ \psi_t(j) &= \operatorname*{argmax}_{1\leq i\leq N_s} \left[\delta_{t-1}(i) a_{ij}\right],\ 1 \leq j \leq N_s, 2 \leq t \leq T, 1 \leq j \leq N_s \end{aligned} \tag{5.4}$$

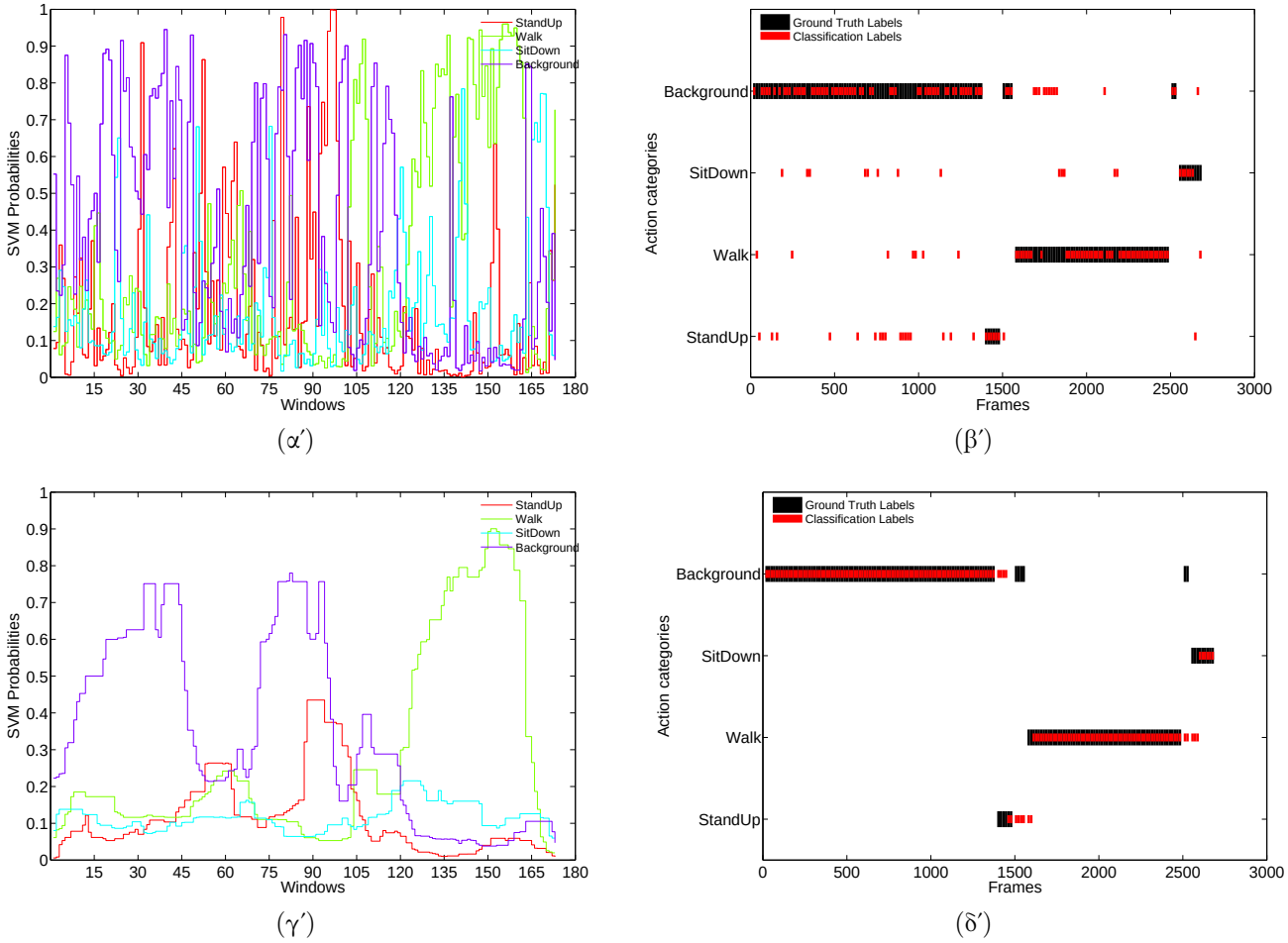
3: Τερματισμός

$$\begin{aligned} P^\star &= \max_{1\leq i\leq N_s} \left[\delta_T(i)\right] \\ q_T^\star &= \operatorname*{argmax}_{1\leq i\leq N_s} \left[\delta_T(i)\right] \end{aligned} \tag{5.5}$$

4: Οπισθοδρόμηση (backtracking) για την εύρεση βέλτιστου μονοπατιού.

$$q_t^\star = \psi_{t+1}(q_{t+1}^*),\ t = T-1, T-2, \ldots, 1 \tag{5.6}$$

Η έξοδος του αλγορίθμου μας δίνει μια τελική συλλογή ετικετών για τα παράθυρα του βίντεο εισόδου και, εφόσον αυτά είναι μη επικαλυπτόμενα, αυτή ισοδυναμεί με μια τελική ακολουθία για κάθε ετικετών για κάθε frame του βίντεο.



Σχήμα 5.3: Ομαλοποίηση πιθανοτικών εξόδων των δυαδικών SVM ταξινομητών. (α′) Πιθανοτικές Έξοδοι των SVM ταξινομητών για κάθε κλάση για κάθε τμήμα ενός βίντεο. (β′) Ετικέτες που έχουν αποδωθεί σε κάθε πλαίσιο του βίντεο από ένα σύνολο 4 κατηγοριών δράσεων (*StandUp*, *Walk*, *Sit Down*, *Background class*). (γ′) Φιλτραρισμένες πιθανοτικές έξοδοι SVM ταξινομητών. (δ′) Τελική ακολουθία ετικετών των πλαισίων του βίντεο, όπως προέκυψε από τον αλγόριθμο Viterbi.

Στο Σχήμα 5.3 απεικονίζεται το αποτέλεσμα της αναγνώρισης δράσεων, όπου στα παράθυρα αποδοθεί ετικέτες από ένα σύνολο κατηγοριών δράσεων (*StandUp*, *Walk*, *Sit Down*, *Background class*) (Σχήμα 5.3β′), όπως αυτές έχουν προκύψει από τη σύγκριση των πιθανοτικών εξόδων των SVMs για κάθε κλάση (Σχήμα 5.3α′). Η είσοδος των SVM ταξινομητών ήταν η ℓ_1-κανονικοποιημένη BoVW αναπαράσταση κάθε video segment, βασισμένη

στις τιμές των Trajectory περιγραφητών. Στο Σχήμα 5.3γ΄ παρουσιάζονται οι πιθανότητες των SVMs μετά από φιλτράρισμα με median φίλτρο. Το φιλτράρισμα ομαλοποιεί τις πιθανότες έτσι ώστε να διορθώνονται λάθος ταξινομήσεις κάποιων παραθύρων βάσει των ταξινομήσεων των γειτονικών παραθύρων. Ωστόσο, ανάλογα με το μέγεθος του παραθύρου φιλτραρίσματος ελοχεύει ο κίνδυνος να φιλτραριστούν δράσεις με πολύ σύντομη διάρκεια (π.χ. 2-3 παραθύρων). Π.χ. στη συγκεκριμένη περίπτωση, το ενδιάμεσο τμήμα ανάμεσα στη δράση *Walk* και στη δράση *Sit Down* που δεν περιλαμβάνει καμία δράση χάνεται, αφού η απότομη κορυφή της καμπύλης πιθανοτήτων της κλάσης *Background class* στο παράθυρο 165 κόβεται από το median φιλτράρισμα. Εντούτοις, με το φιλτράρισμα αποθορυβοποιούνται οι πιθανότητες. Λαμβάνοντας αυτές τις πιθανότητες ως παρατηρήσεις ενός HMM, ο αλγόριθμος Viterbi βρίσκει την πιο πιθανή ακολουθία δράσεων (Σχήμα 5.3δ΄). Έτσι καταλήγουμε σε ένα πιο ακριβή χρονικό εντοπισμό των δράσεων που εκτελούνται στο βίντεο.

5.3 Πειραματικά αποτελέσματα

5.3.1 Η βάση δεδομένων ανθρώπινων δράσεων MOBOT

Τα πειράματα αναγνώρισης συνεχόμενων δράσεων με τη χρήση πυκνών τροχιών πραγματοποιήθηκαν σε ένα υποσύνολο δεδομένων της πολυτροπικής και πολυ-αισθητηριακής βάσης δράσεων MOBOT. Η βάση αυτή δημιουργήθηκε για τις ανάγκες του ερευνητικού προγράμματος MOBOT [1], το οποίο έχει ως σκοπό τη σχεδίαση και κατασκευή μιας ρομποτικής πλατφόρμας (robotic rollator) που θα προσφέρει υποστήριξη κατά τη βάδιση σε ηλικιωμένα άτομα με ήπιες κινητικές ή γνωστικές διαταραχές. Ο σχεδιασμός ενός ευφυούς ρομποτικού συστήματος θα οδηγήσει σε βελτίωση της ποιότητας της ζωής και της ανεξάρτητης διαβίωσης των ηλικιωμένων. Για να μπορεί το ρομπότ να παρέχει ενεργή υποστήριξη και βοήθεια κατά την μετακίνηση των ατόμων, προσαρμοσμένη στον εκάστοτε χρήστη και περιβάλλον, θα πρέπει να μπορεί να παρακολουθεί και να κατανοεί συγκεκριμένες μορφές ανθρώπινης δραστηριότητας, για να συνάγει ποιες είναι οι ανάγκες του ανθρώπου όσον αφορά την κινητικότητά του. Έτσι αναδεικνύεται άλλη μία σημαντική πρακτική εφαρμογή της αναγνώρισης ανθρώπινων δράσεων.

Για να είναι σε θέση το ρομποτικό σύστημα να καταγράφει τις δράσεις του χρήστη, αλλά και για να αλληλεπιδρά μαζί του, εκμεταλλεύεται διάφορες τροπικότητες από ένα σύνολο αισθητήρων [6], όπως οπτικά αποστασιόμετρα (laser range finders), αισθητήρες δύναμης-ροπής, RGB και RGB-D κάμερες,

[1] http://www.mobot-project.eu/

κωδικοποιητές και συστοιχία μικροφώνων MEMS. Οι αισθητήρες οπτικής πληροφορίας που χρησιμοποιούνται παρατίθενται παρακάτω:

- Άνωθεν Kinect: Αισθητήρας Kinect-For-Windows τοποθετημένος οριζόντια πάνω στη ρομποτική πλατφόρμα, με σκοπό την καταγραφή του κορμού, της μέσης, των ισχίων και του άνω μέρους των ποδιών του ασθενή από κοντινή απόσταση. Παρέχει δεδομένα RGB (έγχρωμες εικόνες) και βάθους.

- Κάτωθεν Kinect: Αισθητήρας Kinect-For-Windows τοποθετημένος σε συμπληρωματική κατεύθυνση σε σχέση με το άνωθεν Kinect, με σκοπό την καταγραφή των κάτω άκρων του ασθενή και τη διευκόλυνση του εντοπισμού των ποδιών και της ανάλυσης διαταραχών βάδισης.

- Κάμερα GoPro: Κάμερα ευρείας γωνίας λήψης τοποθετημένη πάνω στο ρομποτικό βοηθό, με σκοπό την καταγραφή του άνω μέρους του σώματος του χρήστη. Παρέχει δεδομένα RGB μεγάλης ευκρίνειας.

Στο χώρο των πειραμάτων υπήρχαν επίσης δύο RGB κάμερες υψηλής ευκρίνειας για την καταγραφή όλου του πεδίου δράσης, καθώς και σύστημα τύπου Qualisys Motion Capture System, που καθιστά εφικτή την καταγραφή του σκελετού του χρήστη μέσω ειδικών σημειωτών (markers) τοποθετημένων σε συγκεκριμένα σημεία στο σώμα του χρήστη.

Για τον πειραματισμό μας στην αναγνώριση συνεχόμενων δράσεων χρησιμοποιήσαμε δεδομένα που είχαν ληφθεί από πραγματικούς ασθενείς από το άνωθεν Kinect. Πιο συγκεκριμένα, χρησιμοποιήθηκαν RGBD δεδομένα 6 ασθενών (ασθενείς 1,3,4,5,6 και 18) από τον αισθητήρα Kinect που καταγράφει το άνω μέρος του σώματος του ασθενή. Κάθε ασθενής εκτελεί μια ακολουθία κινήσεων που περιλαμβάνουν δράσεις (actions) και χειρονομίες και φωνητικές εντολές (audio-gestural commands) για την αλληλεπίδραση με το ρομπότ. Το σενάριο που εκτελεί ο κάθε χρήστης, με ή χωρίς τη βοήθεια ενός φροντιστή (carer), είναι το εξής: αρχικά ο χρήστης είναι καθισμένος σε καρέκλα και ζητάει βοήθεια από τη ρομποτική εκτελώντας μια κατάλληλη χειρονομία και εκφέροντας τη φωνητική εντολή "Come here". Αφού πλησιάσει το ρομπότ, ο χρήστης εκτελεί την εντολή "I want to stand up" και σηκώνεται από την καρέκλα (δράση *Stand Up*). Για να δηλώσει την πρόθεση να περπατήσει, εκτελεί την εντολή "Let's go" και ξεκινάει να περπατάει ευθεία προς ένα προκαθορισμένο στόχο (μια άλλη καρέκλα για να καθίσει) (δράση *Walk*). Όταν συναντήσει ένα στατικό εμπόδιο, εκτελεί την εντολή "Turn left/right" και στρίβει. Στη συνέχεια, περπατάει ξανά ευθεία μέχρι να φτάσει το στόχο (δράση *Walk*). Τέλος, μόλις φτάσει στην καρέκλα, δηλώνει ότι επιθυμεί να καθίσει με κατάλληλη χειρονομία και φωνητική εντολή "I want to sit down" και κάθεται (δράση *Sit Down*).

Οι δράσεις προς αναγνώριση είναι τρεις: *Stand Up*, *Walk*, *Sit Down* και, όπως έγινε σαφές από την αναλυτική περιγραφή του σεναρίου, ανάμεσά τους εκτελούνται και άλλες δράσεις που δεν ανήκουν στο λεξιλόγιό μας (όπως *Turn Left/Right*) καθώς και χειρονομίες. Όλα αυτά τα αποσπάσματα του βίντεο ανατίθενται στην κατηγορία *Background class*. Κάποια ενδεικτικά frames των δράσεων αυτού του υποσυνόλου της βάσης MOBOT απεικονίζονται στο Σχήμα 5.4.

(α′) Δράση *Stand up* (Ασθενής 3).

(β′) Δράση *Sit down* (Ασθενής 1).

(γ′) Δράση *Walk* (Ασθενής 6).

Σχήμα 5.4: Δράσεις που εκτελούνται από ασθενείς κατά τη διάρκεια του σεναρίου 3 της βάσης MOBOT (παραλλαγή 3.b).

Παρά το μικρό αριθμό δράσεων προς αναγνώριση, το υποσύνολο αυτό της βάσης δεδομένων MOBOT παρουσιάζει πολλές δυσκολίες και προκλήσεις, οι

οποίες είναι υπαρκτές και σε άλλες βάσεις, αλλά είναι πιο έντονες στη συγκεκριμένη βάση, λόγω της φύσης της εφαρμογής. Ενδεικτικά, οι κυριότερες δυσκολίες που πρέπει να αντιμετωπίζει ένα σύστημα αναγνώρισης δράσεων στη βάση MOBOT είναι:

- Κίνηση της κάμερας: Η κίνηση της κάμερας, εκτός του ότι οδηγεί σε συχνές αλλαγές του οπτικού πεδίου (view-point changes), οι οποίες μπορεί να είναι τέτοιες που να οδηγούν σε παροδική απουσία του ασθενούς από το οπτικό πεδίο, επηρεάζει και την εξαγωγή περιγραφητών οπτικής κίνησης, καθώς εκλαμβάνεται από το σύστημα ως οπτική ροή. Στη βάση MOBOT ο αισθητήρας Kinect είναι τοποθετημένος πάνω στην κινούμενη ρομποτική πλατφόρμα, οπότε η κίνηση της κάμερας είναι διαρκής.

- Διακύμανση της κλίμακας/διάρκειας εκτέλεσης των δράσεων: Οι ασθενείς μπορεί να καταγράφονται σε διαφορετικές κλίμακες (scales) (ανάλογα με τη σχετική θέση του ρομποτικού βοηθού και του ασθενούς) και οι πράξεις που εκτελούν μπορεί να έχουν μεταβλητή διάρκεια. Π.χ. ανάλογα με το βαθμό των κινητικών προβλημάτων του κάθε ασθενή, κάποιοι μπορεί να σηκώνονται πιο γρήγορα από την καρέκλα σε σχέση με τους υπόλοιπους.

- Επικαλύψεις (occlusions): Σε κάποιες περιπτώσεις, ο αισθητήρας Kinect δεν μπορεί να καταγράψει την κίνηση του ασθενή μιας και ο φροντιστής επικαλύπτει μερικώς τον ασθενή, καθώς τον βοηθάει να κινηθεί.

- Οπτικός θόρυβος υποβάθρου (background visual noise): Σε πολλά frames των βίντεο είναι ορατά στο υπόβαθρο άτομα από το προσωπικό που κάθονται ή μετακινούνται στο χώρο.

- Διακύμανση της εκτέλεσης των δράσεων: Διαφορετικοί άνθρωποι εν γένει εκτελούν την ίδια δράση με ελαφρώς διαφορετικό τρόπο. Το φαινόμενο οξύνεται εν προκειμένω λόγω των κινητικών/γνωστικών διαταραχών των ατόμων που εκτελούν τις δράσεις.

Για να μετρήσουμε την επίδοση του συστήματος αναγνώρισης δράσεων βασιζόμαστε σε μια εκτίμηση της ακρίβειας αναγνώρισης χρησιμοποιώντας τη μέθοδο Leave-One-Out (LOO). Πιο συγκεκριμένα, εκπαιδεύουμε το σύστημα χρησιμοποιώντας δεδομένα πέντε ασθενών και το αξιολογούμε στα δεδομένα του έκτου ασθενή (unseen patient), υπολογίζοντας την ακρίβεια αναγνώρισης ως το ποσοστό των σωστά ταξινομημένων παραθύρων (video segments που προέκυψαν από το κυλιόμενο παράθυρο) ως προς το συνολικό αριθμό παραθύρων των δεδομένων αξιολόγησης. Αυτή η διαδικασία επαναλαμβάνεται για όλους τους ασθενείς, κάθε φορά εξαιρώντας το βίντεο ενός ασθενή από το

σύνολο δεδομένων εκπαίδευσης και χρησιμοποιώντας το για αξιολόγηση. Εφόσον οι ασθενείς που εμφανίζονται στην εκπαίδευση και την αξιολόγηση του συστήματος είναι κάθε φορά διαφορετικοί, δεν υπάρχει κάποια συσχέτιση μεταξύ των συνόλων εκπαίδευσης (training set) και αξιολόγησης (testing set). Παρόλο που η συγκεκριμένη μέθοδος είναι υπολογιστικά απαιτητική και τα αποτελέσματα αναγνώρισης για κάθε unseen patient μπορεί να έχουν μεγάλη διακύμανση και να οδηγήσουν σε μικρότερο τελικό αποτέλεσμα αναγνώρισης, μας δίνει μια αμερόληπτη εκτίμηση του error rate και για λίγα δεδομένα, όπως στην περίπτωσή μας, εκτιμά αποτελεσματικά την ικανότητα γενίκευσης του μοντέλου μας σε ένα ανεξάρτητο σύνολο δεδομένων. Ως τελική μετρική αξιολόγησης χρησιμοποιείται ο μέσος όρος των ακριβειών αναγνώρισης για κάθε unseen patient.

5.3.2 Πειραματικό πλαίσιο

Όπως αναφέραμε, τα πειράματα μας εκτελούνται στο υποσύνολο της βάσης MOBOT που περιέχει βίντεο συνεχόμενων δράσεων από 6 ασθενείς και μας ενδιαφέρει να αναγνωρίσουμε 3 κατηγορίες δράσεων: *Stand Up*, *Walk* και *Sit Down*.

Μέθοδοι και παράμετροι εξαγωγής χαρακτηριστικών

Ως μέθοδοι εξαγωγής χαρακτηριστικών χρησιμοποιούνται οι πυκνές τροχιές (Dense Trajectories - DTs) και οι βελτιωμένες πυκνές τροχιές (improved Dense Trajectories - iDTs), όπως αυτές είναι υλοποιημένες στον πηγαίο κώδικα που παραχωρούν οι συγγραφείς [2] [3] . Χρησιμοποιούμε τις τιμές των παραμέτρων που προτείνουν οι συγγραφείς [30], [31]. Πιο συγκεκριμένα, χρησιμοποιούμε τροχιές με μήκος $l = 15$ πλαίσια, ενώ οι περιγραφητές υπολογίζονται εντός ενός χωροχρονικού όγκου διάστασης $l \times N \times N$ ευθυγραμμισμένου με κάθε τροχιά, όπου το μέγεθος της γειτονιάς είναι $N = 32$ pixels. Για την ενσωμάτωση πληροφορίας δομής, αυτός ο χωροχρονικός όγκος υποδιαιρείται σε ένα χωροχρονικό πλέγμα (spatio-temporal grid) μεγέθους $n_\sigma \times n_\sigma \times n_\tau$, όπου $n_\sigma = 2$ είναι το πλήθος των χωρικών κελιών του πλέγματος και $n_\tau = 3$ είναι το πλήθος των χρονικών κελιών. Επίσης, η πυκνή δειγματοληψία γίνεται ανά $W = 5$ pixels το πολύ σε 8 χωρικές κλίμακες. Για την εξαγωγή των βελτιωμένων πυκνών τροχιών δε χρησιμοποιηθήκαν ορθογώνια πλαίσια που να περικλείουν τους ανθρώπους (human bounding box). Οι περιγραφητές που χρησιμοποιήθηκαν είναι: Trajectory, HOG, HOF, MBHx και MBHy και οι αρχικές διαστάσεις τους είναι: 30, 96, 108, 96 και 96 αντίστοιχα. Στις

[2] `http://lear.inrialpes.fr/people/wang/dense_trajectories`

[3] `http://lear.inrialpes.fr/people/wang/improved_trajectories`

περιπτώσεις που δηλώνεται ρητά ότι χρησιμοποιείται PCA με whitening, οι διαστάσεις των περιγραφητών είτε παραμένουν ως έχουν (*PCANoReduce*) είτε μειώνονται σε 15, 64, 54, 64 και 64 αντίστοιχα. Για το χρονικό εντοπισμό των δράσεων χρησιμοποιούμε ένα μη επικαλυπτόμενο κυλιόμενο παράθυρο μήκους $\Delta t = 15$ frames.

Μέθοδοι και παράμετροι υπολογισμού αναπαράστασης βίντεο

Για την αναπαράσταση των βίντεο χρησιμοποιούνται οι μέθοδοι BoVW με K_{BoVW} και VLAD με K_{VLAD} οπτικές λέξεις, οι οποίες έχουν υπολογιστεί με K-means ομαδοποίηση 100000 τυχαία επιλεγμένων χαρακτηριστικών εκπαίδευσης, ομοιόμορφα κατανεμημένων μεταξύ των κατηγοριών δράσεων.

Σε όλες τις μεθόδους που εξετάζουμε, το μέγεθος του λεξικού, δηλαδή ο αριθμός των clusters χαρακτηριστικών που παράγονται κατά τη διαδικασία ομαδοποίησης, αποτελεί μια βασική παράμετρο που καθορίζει το μέγεθος του διανύσματος αναπαράστασης που παράγεται. Η επιλογή του μεγέθους του λεξικού δεν επηρεάζει μόνο το υπολογιστικό κόστος, αλλά εισάγει και ένα trade-off ανάμεσα στην ικανότητα διαχωρισμού (discrimination) των διαθέσιμων video και στην ικανότητα γενίκευσης σε άγνωστα video. Όταν το λεξιλόγιο είναι πολύ μικρό, η αναπαράσταση αποτυγχάνει να διαχωρίσει επιτυχώς τα video που ανήκουν σε διαφορετικές κατηγορίες, καθώς τα χαρακτηριστικά που διαφέρουν αρκετά μπορεί να αντιστοιχιστούν (ανατεθούν) στην ίδια οπτική λέξη. Αντιθέτως, όταν το πλήθος των οπτικών λέξεων αυξάνει, τα video του συνόλου εκπαίδευσης διαχωρίζονται καλύτερα, αλλά μειώνεται η ικανότητα γενίκευσης και η ευρωστία στο θόρυβο, καθώς παρόμοια χαρακτηριστικά μπορεί να ανατεθούν σε διαφορετικές λέξεις. Στη διεθνή βιβλιογραφία έχει παρατηρηθεί ότι ένα λεξικό $K_{BoVW} = 4000$ λέξεων οδηγεί σε ικανοποιητικά αποτελέσματα με την αναπαράσταση BoVW, ενώ η μέθοδος VLAD, η οποία ενσωματώνει στατιστική πληροφορία πρώτης τάξεως για κάθε οπτική λέξη οδηγεί σε παρόμοια και ενίοτε καλύτερα αποτελέσματα με πολύ μικρότερο αριθμό οπτικών λέξεων. Γι'αυτό επιλέγουμε $K_{VLAD} = 256$.

Υπενθυμίζουμε ότι η ποιότητα της ομαδοποίησης των χαρακτηριστικών εξαρτάται από την αρχικοποίηση των κεντροειδών του αλγορίθμου K-means και μπορεί να επηρεάσει σημαντικά το τελικό αποτέλεσμα. Γι'αυτό πολύ συχνά ο αλγόριθμος εκτελείται πολλές φορές με τυχαία αρχικοποίηση των κεντροειδών και διατηρείται η ομαδοποίηση με το μικρότερο τελικό μέτρο παραμόρφωσης. Εν προκειμένω, η ομαδοποίηση εκτελείται μία φορά αλλά με ίδιο seed της γεννήτριας τυχαίων αριθμών για όλα τα πειράματα έτσι ώστε να μην επηρεάζει τη σύγκριση των μεθόδων. Χρησιμοποιήθηκε η υλοποίηση του αλγορίθμου

K-means από τη βιβλιοθήκη Yael[4] και η υλοποίηση του VLAD από τη βιβλιοθήκη VLfeat[5]. Τα BoVW ιστογράμματα είναι κανονικοποιημένα με την ℓ_1 νόρμα, ενώ για τα διανύσματα VLAD συγκρίνουμε τις κανονικοποιήσεις ℓ_2-normalization, power-normalization kai intra-normalization.

Μέθοδοι και παράμετροι για την ταξινόμηση

Για την ταξινόμηση των παραθύρων κάθε βίντεο με χρήση SVM ταξινομητών χρησιμοποιήθηκε η βιβλιοθήκη LIBSVM [6]. Τα SVMs που χρησιμοποιούνται για την ταξινόμηση των βίντεο που έχουν αναπαρασταθεί με BoVW ιστογράμματα είναι μη γραμμικά με πυρήνα χ^2. Αντιθέτως, χρησιμοποιούμε γραμμικές μηχανές διανυσματικής υποστήριξης για την περίπτωση των VLAD διανυσμάτων. Σε κάθε περίπτωση, η παράμετρος κόστους C των SVMs, η οποία ελέγχει το trade-off ανάμεσα στη λανθασμένη ταξινόμηση των στιγμιοτύπων εκπαίδευσης και στην πολυπλοκότητα του μοντέλου, τέθηκε εμπειρικά ίση με 2. Εξετάζεται επίσης η επίδοση του συνδυασμού των περιγραφητών μέσω πολυκαναλικής σύμμειξης (multi-channel fusion) στην περίπτωση του BoVW και μέσω της συνένωσης των περιγραφητών σε έναν ενιαίο περιγραφητή στην περίπτωση του VLAD.

Ο συνδυασμός των εξόδων των SVM ταξινομητών με one-against-all τεχνική, δηλαδή οι ετικέτες που ανατίθενται ανεξάρτητα σε κάθε παράθυρο των βίντεο αξιολόγησης, χρησιμοποιούνται για τον υπολογισμό της μέσης ακρίβειας αναγνώρισης στα baseline πειράματά μας. Για να βελτιώσουμε αυτά τα baseline αποτελέσματα, ομαλοποιούμε τις πιθανοτικές εξόδους των SVMs εφαρμόζοντας median φίλτρο με παράθυρο 19 στοιχείων και χρησιμοποιώντας αυτές τις ομαλοποιημένες πιθανότητες ως παρατηρήσεις ενός HMM μοντέλου, υπολογίζουμε τη βέλτιστη ακολουθία κρυφών καταστάσεων (δράσεων) χρησιμοποιώντας αποκωδικοποίηση Viterbi. Κάθε κρυφή κατάσταση του HMM έχει την ίδια αρχική πιθανότητα $\pi_i = 0.25$. Ο πίνακας μεταβάσεων είναι:

$$A = \begin{bmatrix} 0.8 & 0.1 & 0 & 0.1 \\ 0 & 0.8 & 0.1 & 0.1 \\ 0 & 0 & 1 & 0 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{bmatrix} \tag{5.7}$$

Όπως είναι προφανές, στον πίνακα μετάβασης έχουμε ενσωματώσει την πρότερη γνώση μας σχετικά με την διαδοχή των δράσεων που εκτελούν οι ασθενείς. Π.χ. από την κατάσταση *Stand Up* σε ένα παράθυρο έχουμε 0.8 πιθα-

[4] `http://yael.gforge.inria.fr/`

[5] `www.vlfeat.org`

[6] `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

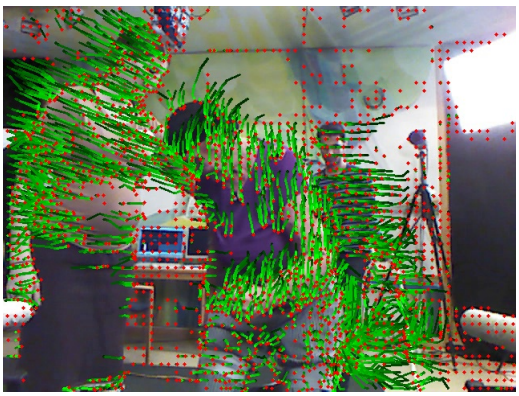
νότητα να παραμείνουμε στην ίδια κατάσταση, 0.1 πιθανότητα να μεταφερθούμε στην κατάσταση *Walk*, μηδενική πιθανότητα να μεταφερθούμε στην κατάσταση *Sit Down* και 0.1 πιθανότητα να μεταφερθούμε στην κατάσταση *Background class*. Επίσης γνωρίζοντας ότι η δράση *Sit Down* είναι η τελευταία που εκτελείται, όταν φτάσει εκεί το HMM, αναγκάζεται να παραμένει στην ίδια κρυφή κατάσταση μέχρι το τέλος. Όσον αφορά το πέναλτυ μετάβασης από κατάσταση σε κατάσταση που χρησιμοποιούμε στον αλγόριθμο Viterbi, επιλέγουμε για κάθε περίπτωση τη σταθερά από ένα εύρος $0 - 30$ με βήμα 0.5 η οποία μεγιστοποιεί την τελική ακρίβεια αναγνώρισης. Ωστόσο, όσο αυτή η σταθερά είναι μεγαλύτερη του μηδενός, οι διαφορετικές τιμές της έχουν πολύ περιορισμένη επίδραση στο τελικό αποτέλεσμα.

5.3.3 Σύγκριση μεθόδων εξαγωγής χαρακτηριστικών και περιγραφητών

Τα αποτελέσματα της αναγνώρισης δράσεων με διαφορετικούς ανιχνευτές χαρακτηριστικών και περιγραφητές παρατίθενται στον Πίνακα 5.1. Τα αποτελέσματα αφορούν την αναπαράσταση BoVW με ℓ_1 κανονικοποίηση.

Όσον αφορά τα αποτελέσματα του βασικού (baseline) συστήματος, παρατηρούμε ότι η ακρίβεια αναγνώρισης δράσεων με τις βελτιωμένες τροχιές δεν είναι σημαντικά καλύτερη σε σύγκριση με τις πυκνές τροχιές. Οι περιγραφητές Trajectory, HOG και ο συνδυασμός όλων των περιγραφητών είναι τα μόνο χαρακτηριστικά που παρουσιάζουν μια αισθητή βελτίωση (από $\sim 0.1\%$ μέχρι $\sim 3\%$). Οι άλλοι περιγραφητές δίνουν περίπου τα ίδια ή λίγο χαμηλότερα αποτελέσματα. Αυτό μπορεί να οφείλεται στη χρήση των βελτιωμένων πυκνών τροχιών χωρίς ανιχνευτή των ανθρώπων. Επειδή το άνωθεν Kinect που βρίσκεται πάνω στο ρομπότ βρίσκεται σε μικρή απόσταση από το χρήστη, σε αρκετά πλαίσια κυριαρχεί ο άνθρωπος και συνεπώς η κίνησή του. Έτσι, ο αλγόριθμος των βελτιωμένων πυκνών τροχιών αποτυγχάνει να εκτιμήσει σωστά την κίνηση της κάμερας, καθώς αυτή δε συνάδει με την κίνηση του ανθρώπου. Πιο συγκεκριμένα, όπως περιγράψαμε και στο Κεφάλαιο 2, ο αλγόριθμος των βελτιωμένων πυκνών τροχιών (iDTs) βρίσκει τις αντιστοιχίες χαρακτηριστικών SURF και δειγματοληπτημένων διανυσμάτων οπτικής ροής μεταξύ δύο frames και βάσει αυτών εκτιμά την ομογραφία χρησιμοποιώντας τον αλγόριθμο RANSAC. Όταν η πλειοψηφία αυτών των χαρακτηριστικών που έχουν αντιστοιχιστεί ανήκει στον κινούμενο άνθρωπο, τότε η εκτιμούμενη κίνηση του υποβάθρου είναι λανθασμένη. Γι'αυτό άλλωστε προτείνεται και από τους συγγραφείς η χρήση ενός human detector. Ειδικά στην περίπτωσή μας που η κίνηση της κάμερας είναι έντονη και η κίνηση των ανθρώπων κυριαρχεί στα frames, η χρήση ενός ανιχνευτή ανθρώπων για την απόρριψη των χαρα-

κτηριστικών που έχουν αντιστοιχιστεί εντός των ορθογώνιων πλαισίων που περικλείουν τους ανθρώπους θα οδηγούσε σε βελτίωση της ακρίβειας αναγνώρισης δράσεων. Εντούτοις, η ανίχνευση των ανθρώπων σε ένα βίντεο είναι ένα δύσκολο πρόβλημα, κυρίως λόγω των σημαντικών αλλαγών στην πόζα των ανθρώπων που διαδραματίζονται κατά τη διάρκεια του βίντεο και των επικαλύψεων. Επίσης, πολύ συχνά οι άνθρωποι βγαίνουν μερικώς ή ολικώς από το οπτικό πεδίο της κάμερας. Εκτός από την έλλειψη του ανιχνευτή ανθρώπων, παρόλο που η μέθοδος των iDTs βελτιώνει την επίδοση των περιγραφητών που βασίζονται στην οπτική ροή και παράγει λιγότερα χαρακτηριστικά επιταχύνοντας τη διαδικασία της αναγνώρισης των δράσεων, απορρίπτει πληροφορία που σχετίζεται με το υπόβαθρο ή την κίνηση της κάμερας που είναι συσχετισμένη με συγκεκριμένες δράσεις, όπως η αλλαγή της οπτικής γωνίας της κάμερας λόγω της ταυτόχρονης στροφής του ρομποτικού βοηθού και του ασθενή. Για παράδειγμα, αν ο ρομποτικός βοηθός, και κατ'επέκταση ο αισθητήρας Kinect, βρίσκεται κοντά στον ασθενή, η κίνηση του υποβάθρου που προκαλείται από την κίνηση της κάμερας (στροφή) είναι σε αρκετές περιπτώσεις ο μόνος τρόπος να καταλάβουμε ότι ο ασθενής πραγματοποιεί μια δράση *Turn Left/Right*. Τέλος, παρόλο που η επίδοση των DTs και των iDTs είναι παρόμοια, η υψηλότερη τελική μέση ακρίβεια αναγνώρισης επιτυγχάνεται με το συνδυασμό όλων των περιγραφητών που έχουν εξαχθεί από τις βελτιωμένες πυκνές τροχιές (84%).



Σχήμα 5.5: Ενδεικτικές πυκνές τροχιές ενός βίντεο της βάσης δεδομένων MOBOT. Παρατηρούμε ότι οι τροχιές καταγράφουν όχι μόνο την κίνηση της ασθενούς, αλλά και τις κινήσεις της βοηθού και του ανθρώπου στο υπόβαθρο.

Παρατηρώντας την επίδοση των 5 περιγραφητών και του συνδυασμού τους, μπορούμε να συνάγουμε τα εξής συμπεράσματα:

- Οι περιγραφητές Trajectory και HOF παρουσιάζουν τη χειρότερη επίδοση τόσο στην περίπτωση των Dense Trajectories όσο και των improved

Dense Trajectories. Αντιθέτως, ο άλλος περιγραφητής που βασίζεται στην οπτική ροή, ο MBHy, ο οποίος είναι πιο εύρωστος στην κίνηση της κάμερας, έχει την καλύτερη επίδοση στην περίπτωση των Dense Trajectories και τη δεύτερη καλύτερη επίδοση στην περίπτωση των improved Dense Trajectories. Επομένως, ενισχύεται η υπόθεσή μας ότι η μέθοδος των βελτιωμένων πυκνών τροχιών δε διαχειρίζεται αποδοτικά την κίνηση της κάμερας και αυτό επηρεάζει την ακρίβεια αναγνώρισης των δράσεων όταν χρησιμοποιούνται περιγραφητές που βασίζονται στην οπτική ροή.

- Ο περιγραφητής HOG οδηγεί σε υψηλή μέση ακρίβεια αναγνώρισης υποδηλώνοντας ότι η στατική εμφάνιση και η πόζα είναι ένα σημαντικό εργαλείο διαχωρισμού των δράσεών μας.

- Ο περιγραφητής MBHy εν γένει οδηγεί σε υψηλότερη ακρίβεια αναγνώρισης σε σύγκριση με τον MBHx. Αυτή η διαφορά μπορεί να αποδωθεί στο γεγονός ότι ο MBHy περιγράφει καλύτερα δράσεις που περιλαμβάνουν κινήσεις στον κατακόρυφο άξονα (όπως *Stand Up* και *Sit Down*).

- Η αναγνώριση δράσεων φαίνεται να ωφελείται από το συνδυασμό όλων των περιγραφητών, αφού είναι σε μεγάλο βαθμό συμπληρωματικοί μεταξύ τους. Αυτη η υπόθεση ενισχύεται από το γεγονός ότι ο συνδυασμός τους στην περίπτωση των improved Dense Trajectories οδηγεί σε καλύτερη επίδοση από αυτές που είχαν επιτευχθεί με τη χρήση του κάθε περιγραφητή ξεχωριστά.

Το φιλτράρισμα των πιθανοτικών εξόδων των πιθανοτικών εξόδων των SVMs και η χρήση του αλγορίθμου Viterbi βελτιώνουν την επίδοση του baseline συστήματος για όλους τους περιγραφητές, από $\sim 1\%$ έως και $\sim 15\%$ στην περίπτωση του χαρακτηριστικού DT Trajectory. Ειδικά, τα αποτελέσματα των περιγραφητών Trajectory και HOF βελτιώνονται σημαντικά. Αυτή η βελτίωση μπορεί να εξηγηθεί από το γεγονός ότι λαμβάνουμε υπόψη γειτονικά παράθυρα και ο αλγόριθμος Viterbi πραγματοποιεί έναν καλύτερο εντόπισμο των δράσεων στο χρόνο.

	Dense Trajectories					
	Trajectory	HOG	HOF	MBHx	MBHy	Combined
Baseline	65.62	79.23	71.57	75.21	**79.71**	78.98
SmoothProb	81.10	80.07	80.55	81.78	**82.86**	80.88
	Improved Trajectories					
	Trajectory	HOG	HOF	MBHx	MBHy	Combined
Baseline	68.43	**79.42**	69.75	75.18	78.19	79.41
SmoothProb	79.84	81.54	78.95	78.46	83.27	**84.00**

Πίνακας 5.1: Σύγκριση ανιχνευτών χαρακτηριστικών και περιγραφητών ως προς την μέση ακρίβεια αναγνώρισης δράσεων. "Combined": Συνδυασμός των BoVW ιστογραμμάτων όλων των περιγραφητών με πολυκαναλική σύμμειξη. "Baseline": Υπολογίζουμε τη μέση ακρίβεια αναγνώρισης δράσεων χρησιμοποιώντας τις ετικέτες που αναθέτουν οι SVM ταξινομητές σε κάθε τμήμα των βίντεο αξιολόγησης. "SmoothProb": Χρησιμοποιούμε ομαλοποιημένες πιθανοτικές εξόδους των SVM ταξινομητών ως παρατηρήσεις ενός Κρυφού Μαρκοβιανού Μοντέλου και βρίσκουμε την πιο πιθανή ακολουθία δράσεων με χρήση αποκωδικοποίησης Viterbi.

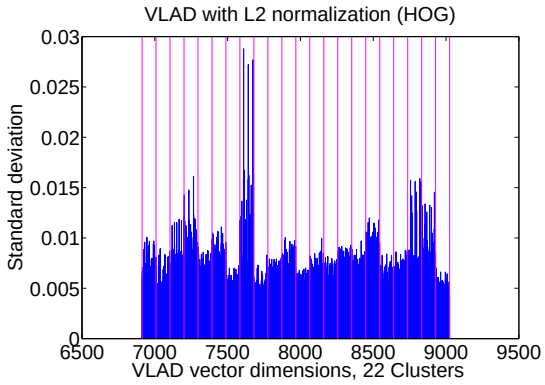
5.3.4 Σύγκριση μεθόδων αναπαράστασης βίντεο

Στα πειράματα που ακολουθούν αξιολογούμε την επίδραση της μεθόδου αναπαράστασης του βίντεο στην επίδοση του συστήματος αναγνώρισης συνεχόμενων δράσεων, συγκρίνοντας τα αποτελέσματα που προκύπτουν με χρήση των BoVW και VLAD μεθόδων. Επίσης, πειραματιζόμαστε με τρεις διαφορετικές μεθόδους κανονικοποίησης του διανύσματος VLAD (ℓ_2 νόρμας, power-normalization, intra-normalization) και με τη χρήση της μεθόδου PCA με και χωρίς μείωση της διάστασης των χαρακτηριστικών.

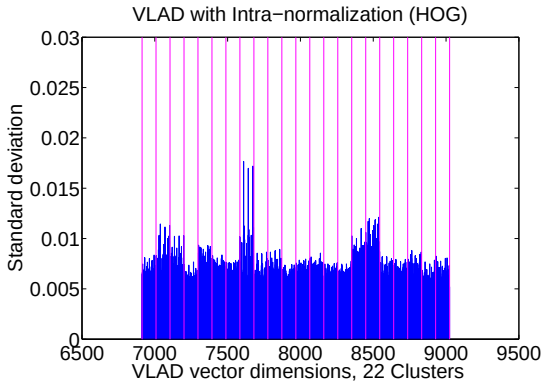
Τα αποτελέσματα του πειραματισμού μας στη βάση δεδομένων MOBOT παρατίθενται στον Πίνακα 5.2 και απεικονίζονται γραφικά στο Σχήμα 5.8. Βλέπουμε ότι παρόλο που για την αναπαράσταση VLAD χρησιμοποιεί κανείς πολύ μικρότερο αριθμό οπτικών λέξεων (256 οπτικές λέξεις) σε σχέση με την BoVW αναπαράσταση και αποδοτικά υπολογισμένους γραμμικούς SVM ταξινομητές, η ακρίβεια αναγνώρισης δράσεων για τον καλύτερο συνδυασμό κανονικοποίησης και προεπεξεργασίας σε κάθε περιγραφητή (εκτός από τον περιγραφητή Trajectory) είναι μεγαλύτερη από την ακρίβεια που δίνει το BoVW με 4000 οπτικές λέξεις και μη γραμμικούς SVM ταξινομητές.

Σχήμα 5.6: Διαισθητική απεικόνιση της αναπάρασης BoVW. Οι υποθετικές οπτικές λέξεις έχουν σημειωθεί με πράσινες ελλείψεις.

Όσον αφορά τη σύγκριση των μεθόδων κανονικοποίησης που εφαρμόστηκαν στην αναπαράσταση VLAD των χαρακτηριστικών, η μέθοδος intra-normalization ήταν η στρατηγική κανονικοποίησης που έδωσε τα καλύτερα αποτελέσματα για κάθε περιγραφητή. Όπως αναλύθηκε και στο Κεφάλαιο 3, η κανονικοποίηση intra-normalization εξαλείφει τις μεγάλες συνιστώσες του διανύσματος VLAD, που αντιστοιχούν σε clusters που έχουν πολλά χαρακτηριστικά ή χαρακτηριστικά με μεγάλη απόσταση από το κεντροειδές του cluster, όπως φαίνεται και στο Σχήμα 5.7. Διαφορετικά, αυτές οι συνιστώσες θα επηρέαζαν σημαντικά την ομοιότητα ανάμεσα σε δύο βίντεο που έχουν αναπαρασταθεί με το διάνυσμα VLAD και κατ'επέκταση την ικανότητα του εκπαιδευμένου SVM ταξινομητή να αποφασίσει αν ανήκουν στην ίδια κλάση ή όχι.



(α′) ℓ_2-κανονικοποίηση.



(β′) Intra-normalization.

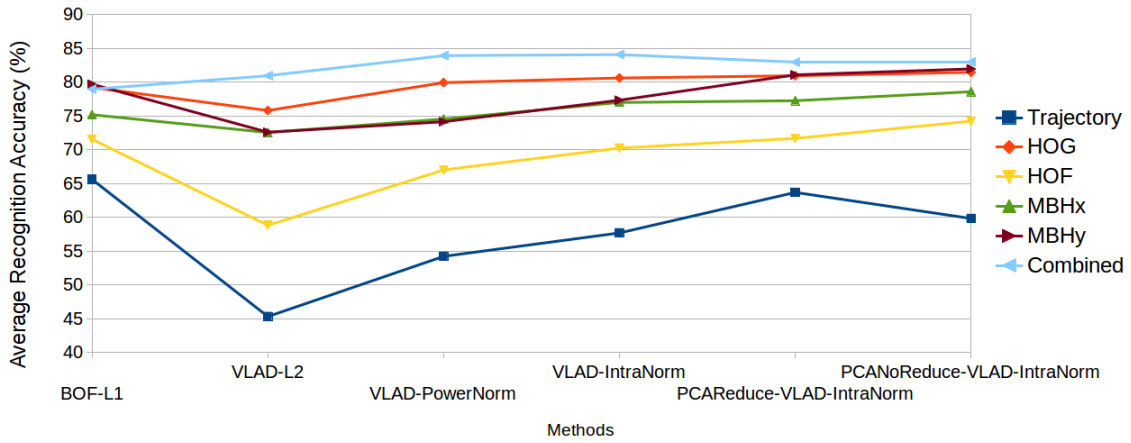
Σχήμα 5.7: Τυπική απόκλιση (ενέργεια) των τιμών κάθε στοιχείου της αναπαράστασης VLAD υπολογισμένη από 780 τμήματα βίντεο εκπαίδευσης της βάσης MOBOT για δύο διαφορετικές στρατηγικές κανονικοποίησης: ℓ_2-κανονικοποίηση και intra-normalization. Οι μωβ γραμμές διαχωρίζουν τα τμήματα του VLAD που σχετίζονται με την κάθε ομάδα (cluster). Όπως παρατηρούμε, η ενέργεια είναι συγκεντρωμένη σε λίγες συνιστώσες στην περίπτωση της ℓ_2 κανονικοποίησης, ενώ η μέθοδος intra-normalization εξομαλύνει αποτελεσματικά αυτές τις κορυφές. Για λόγους ευκρίνειας, η τυπική απόκλιση απεικονίζεται μόνο για ένα υποσύνολο 22 clusters από τα 256 συνολικά. Τα clusters προέκυψαν από K-means ομαδοποίηση των HOG περιγραφητών των video segments.

Από τα αποτελέσματα μπορεί να συμπεράνει κανείς ότι η χρήση της μεθόδου PCA, η οποία αποσυσχετίζει τα χαρακτηριστικά, σε συνδυασμό με τη μέθοδο Whitening, έτσι στε τα ασυσχέτιστα χαρακτηριστικά να έχουν και την ίδια διακύμανση, βελτιώνει την επίδοση του συστήματος αναγνώρισης δράσεων και μειώνει τον υπολογιστικό χρόνο επεξεργασίας όταν οι διαστάσεις

των χαρακτηριστικών μειώνονται. Η διατήρηση όλων των κύριων συνιστωσών οδηγεί σε περαιτέρω βελτίωση της επίδοσης σχεδόν σε όλες τις περιπτώσεις, αυξάνοντας ωστόσο τις απαιτήσεις σε υπολογιστική ισχύ και αποθηκευτικό χώρο. Μόνο στην περίπτωση του περιγραφητή Trajectory, η ακρίβεια αναγνώρισης με τη μείωση της διάστασής του από 30 σε 15 στοιχεία είναι υψηλότερη σε σχέση με τη διατήρηση όλων των συνιστωσών. Στον Πίνακα βλέπουμε το αποτέλεσμα της μετεπεξεργασίας των πιθανοτικών εξόδων των SVM ταξινομητών για την προαναφερθείσα αναπαράσταση των βίντεο.

	Dense Trajectories					
	Trajectory	HOG	HOF	MBHx	MBHy	Combined
BoVW-L1	**65.62**	79.23	71.57	75.21	79.71	78.98
VLAD-L2	45.34	75.82	58.83	72.57	72.61	80.97
VLAD-PowerNorm	54.25	79.95	67.04	74.57	74.16	83.94
VLAD-IntraNorm	57.71	80.63	70.27	77.02	77.35	**84.10**
VLAD-PCAReduce	63.70	80.97	71.70	77.28	81.10	82.97
VLAD-PCANoReduce	59.84	**81.48**	**74.25**	**78.61**	**81.98**	83.01

Πίνακας 5.2: Σύγκριση διάφορων μεθόδων αναπαράστασης και κανονικοποίησης ως προς της μέση ακρίβεια αναγνώρισης με τη χρήση χαρακτηριστικών Dense Trajectories. "Combined": συνδυασμός όλων των περιγραφητών με πολυκαναλική σύμμειξη στην περίπτωση της αναπαράστασης BoVW και με συνένωση των VLAD αναπαραστάσεων στην περίπτωση της αναπαράστασης VLAD. "PowerNorm": Κανονικοποίηση power-normalization, "IntraNorm": κανονικοποίηση intra-normalization. "PCAReduce": χρήση PCA και whitening για την αποσυσχέτιση των χαρακτηριστικών και τη μείωση των διαστάσεών τους (το διάνυσμα VLAD έχει κανονικοποιηθεί με intra-normalization), "PCANoReduce": χρήση PCA και whitening για την αποσυσχέτιση των χαρακτηριστικών με διατήρηση όλων των κύριων συνιστωσών, χωρίς μείωση της διάστασης των χαρακτηριστικών (το διάνυσμα VLAD έχει κανονικοποιηθεί με intra-normalization)



Σχήμα 5.8: Γραφική απεικόνιση των αποτελεσμάτων των πειραμάτων σύγκρισης περιγραφητών και μεθόδων κωδικοποίησης στη βάση MOBOT.

	Dense Trajectories					
	Trajectory	HOG	HOF	MBHx	MBHy	Combined
Baseline	59.84	81.48	74.25	78.61	81.98	83.01
SmoothProb	79.72	80.38	83.62	79.39	82.18	83.12

Πίνακας 5.3: Επίδραση της ομαλοποίησης των εκτιμήσεων των πιθανοτήτων των SVM ταξινομητών και της εφαρμογής του αλγορίθμου Viterbi για την περίπτωση αναπαράστασης των βίντεο με VLAD και PCA-Whitening χωρίς μείωση των διαστάσεων των περιγραφητών.

5.3.5 Αξιοποίηση της πληροφορίας βάθους

Ο αισθητήρας Kinect είναι ιδιαίτερα δημοφιλής γιατί εκτός από RGB εικόνες παρέχει και εικόνες βάθους (depth). Η πληροφορία του βάθους μπορεί να βοηθήσει σημαντικά στη βελτίωση των συστημάτων αναγνώρισης δράσεων, κυρίως για τρεις λόγους.

1. Η πρόσθετη αυτή οπτική πληροφορία μπορεί να βοηθήσει στην απλοποίηση των διαφορών μεταξύ δειγμάτων της ίδιας δράσης. Γενικά, η μεγάλη ποικιλία στις συνθήκες φωτισμού των βίντεο δεν περνάει στο κανάλι βάθους και οι διαφορές στην εμφάνιση των ανθρώπων, που οφείλονται π.χ. στο ρουχισμό, εξαλείφονται.

2. Κάποιες δράσεις μπορεί να έχουν παρόμοιες δισδιάστατες σιλουέττες όταν παρατηρούνται από μία συγκεκριμένη οπτική γωνία. Ειδικά κινήσεις

που για παράδειγμα διαφέρουν μόνο στο αν η κίνηση γίνεται προς τα μπροστά ή πίσω δεν μπορούν να διαχωριστούν μέσω της απλής RGB εικόνας.

3. Το κανάλι βάθους παρέχει χρήσιμη πληροφορία για την αφαίρεση του υποβάθρου (background) και την ανίχνευση επικαλύψεων.

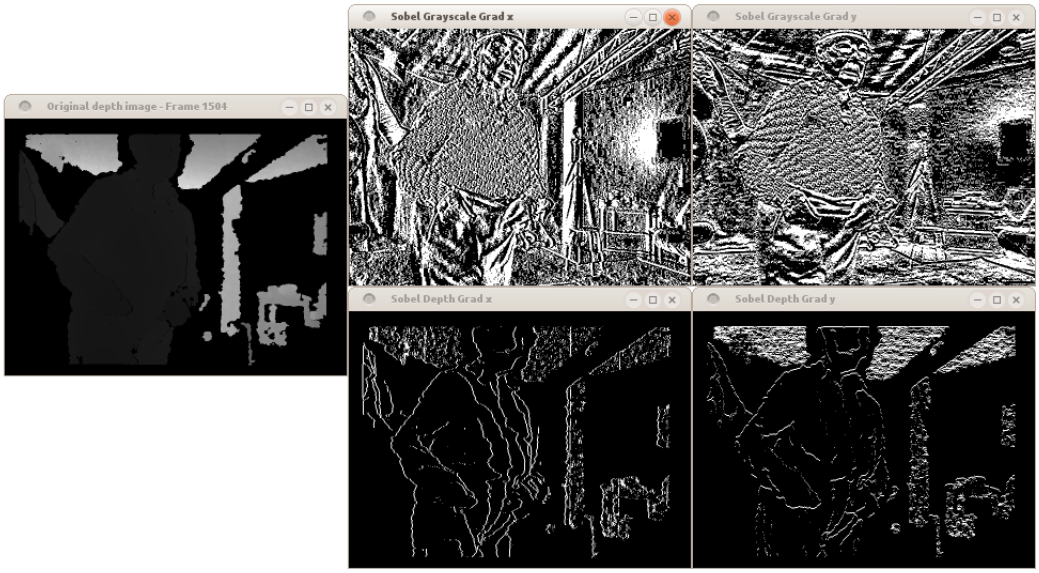
Λόγω αυτών των ιδιοτήτων του το κανάλι βάθους μπορεί να βοηθήσει στην εξαγωγή πιο αντιπροσωπευτικών και εύρωστων χαρακτηριστικών. Γι'αυτό έχουν αναπτυχθεί αρκετές μέθοδοι εξαγωγής χαρακτηριστικών αποκλειστικά από αυτό, πολλές εκ των οποίων αποτελούν επεκτάσεις μεθόδων εξαγωγής χαρακτηριστικών από RGB βίντεο. Ενδεικτικά μπορούμε να αναφέρουμε τον ανιχνευτή DSTIP, ο οποίος εισήχθη από τους Xia και Aggarwal [60] για να εντοπίζει σημεία ενδιαφέροντος που σχετίζονται με δράσεις σε βίντεο βάθους και βασίζεται στο φιλτράρισμα του βίντεο με ένα δισδιάστατο Γκαουσιανό φίλτρο στο πεδίο του χώρου και ένα μονοδιάστατο φίλτρο Gabor στο πεδίο του χρόνου. Οι Hadfield et al. [61] εξέλιξαν γνωστούς αλγορίθμους ανίχνευσης χαρακτηριστικών από RGB εικόνες (γωνίες Harris, σημεία Hessian και φίλτρα Gabor) σε 3.5D και 4D. Βασισμένοι σε μελέτες [62] που έδειξαν ότι για την αναγνώριση αντικειμένων, το σχήμα του αντικειμένου μπορεί να περιγραφεί καλύτερα χρησιμοποιώντας τα κάθετα διανύσματα (normal vectors) στις εικόνες βάθους σε σύγκριση με τις κλίσεις (gradients) που χρησιμοποιούνται στις έγχρωμες εικόνες, οι Oreifej και Liu πρότειναν τον περιγραφητή HON4D [63], ο οποίος υπολογίζει την κατεύθυνση του κάθετου διανύσματος επιφάνειας στον 4D χώρο (χρόνος, κανάλι βάθους και χωρικές συντεταγμένες) και οι Yang et al. πρότειναν το 2014 [64] τον περιγραφητή Super Normal Vector. Μια αναλυτική επισκόπηση των μεθόδων που έχουν αναπτυχθεί για την εξαγωγή χαρακτηριστικών από το κανάλι βάθους μπορεί να βρεθεί στο [65].

Πολύ πρόσφατα δοκιμάστηκε ο εμπλουτισμός των τροχιών με πληροφορία από το κανάλι βάθους. Οι Koperski et al. πρότειναν το 2014 [66] την επέκταση του περιγραφητή Trajectory, εμπλουτίζοντας κάθε σημείο $P(x, y)$ των τροχιών που έχουν εξαχθεί από τα RGB frames με τη συνιστώσα z, η οποία υπολογίζεται ως μέση τιμή φωτεινοτήτων σε μια γειτονιά γύρω από αυτό το σημείο στο κανάλι βάθους. Οι Xiao et al. [67] ασχολήθηκαν παράλληλα με αυτό το πρόβλημα, αντιστοιχίζοντας και αυτοί τις 2D θέσεις των σημείων των τροχιών που έχουν εξαχθεί από το RGB κανάλι, με τις αντίστοιχες θέσεις στο βίντεο βάθους, ανακτώντας έτσι την 3D τροχιά των σημείων ενδιαφέροντος, η οποία περιέχει σημαντική πληροφορία και στην κατεύθυνση του βάθους. Για να περιγράψουν τις 3D τροχιές, εφάρμοσαν το ιστόγραμμα ορίων κίνησης (MBH) στο κανάλι του βάθους και πρότειναν τρισδιάστατους περιγραφητές του σχήματος της τροχιάς.

Με παρόμοιο σκεπτικό, θα δοκιμάσουμε να εξάγουμε από τις εικόνες βάθους τον περιγραφητή HOG, που ως γνωστόν περιγράφει στατική εμφάνιση. Κίνητρό μας αποτελεί η ικανότητα του καναλιού βάθους να εκμηδενίζει διαφορές στην εμφάνιση των ανθρώπων που οφείλονται π.χ. στο ρουχισμό, διατηρώντας και ενισχύοντας βέβαια τις διαφορές στο σχήμα του σώματος, που ορίζει την πόζα.

Η μέθοδος που ακολουθήσαμε εξάγει πυκνές τροχιές από τις RGB εικόνες κατά τα γνωστά, αλλά αντί να εξάγει από τις ίδιες RGB εικόνες (frames) τον περιγραφητή HOG στο χωροχρονικό όγκο γύρω από κάθε τροχιά, απεικονίζει τα σημεία των τροχιών στις εικόνες βάθους και εξάγει από εκεί τον περιγραφητή HOG. Στο Σχήμα 5.9 βλέπουμε ένα frame ενός RGB βίντεο της βάσης MOBOT από το οποίο εξάγονται πυκνές τροχιές, όπου με κόκκινο χρώμα συμβολίζεται το τρέχον σημείο της τροχιάς και με πράσινο τα σημεία της τροχιάς που έχουν ανιχνευθεί σε προηγούμενα frames. Στην παραλλαγή της μεθόδου Dense Trajectories με περιγραφητή HOG έτσι ώστε να λαμβάνει πληροφορία βάθους, θα απεικονίσουμε αυτές τις τροχιές στα frames του καναλιού βάθους και θα υπολογίσουμε εκεί τον περιγραφητή HOG, ο οποίος τώρα πια αντί για Histogram of Oriented Gradients ονομάζεται Histogram of Oriented Depths (HOD). Ο περιγραφητής HOD προτάθηκε από τους Spinello et al. [68] για το πρόβλημα της ανίχνευσης ανθρώπων.

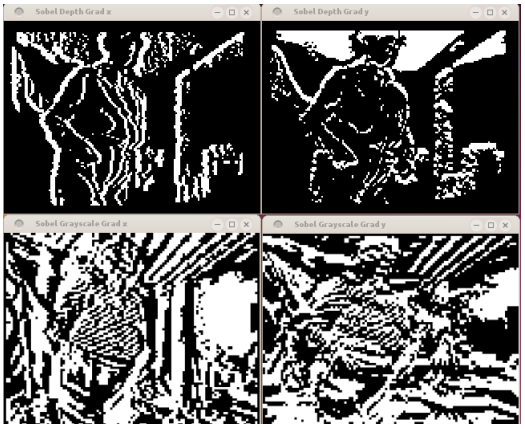
Σχήμα 5.9: Εξαγωγή πυκνών τροχιών από τα RGB frames ενός video της βάσης δεδομένων MOBOT.



Σχήμα 5.10: Εξαγωγή ακμών με χρήση του τελεστή Sobel από το RGB frame του σχήματος 5.9 (άνω σειρά εικόνων) και το αντίστοιχο depth frame ενός video της βάσης δεδομένων MOBOT (κάτω σειρά εικόνων).

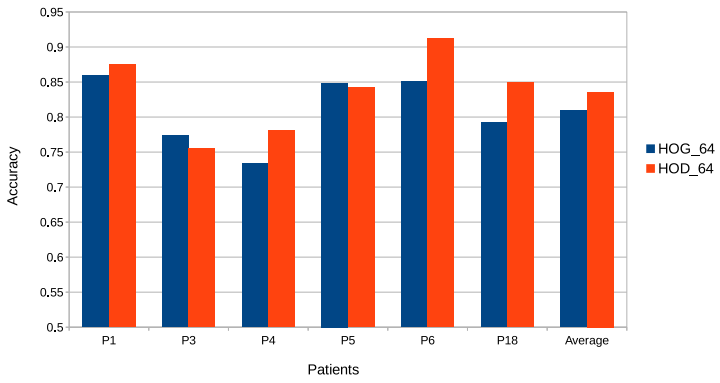
Στο Σχήμα 5.10 μπορούμε να δούμε την αντίστοιχη εικόνα βάθους της RGB εικόνας του Σχήματος 5.9. Παρατηρούμε ότι η φιγούρα του χρήστη διαχωρίζεται εύκολα από το υπόβαθρο (αν εξαιρέσουμε το χέρι του φροντιστή, το οποίο βρίσκεται στο ίδιο επίπεδο βάθους). Επίσης, παρατηρούμε ότι οι ακμές που έχουν εξαχθεί από την RGB εικόνα με χρήση του τελεστή Sobel αντικατοπτρίζουν διάφορες υφές της εικόνας, όπως τα ρούχα της ασθενούς, τα χαρακτηριστικά του προσώπου της και τα αντικείμενα του background. Αντιθέτως, οι ακμές που εξάγονται από τον ίδιο τελεστή στην εικόνα βάθους αντικατοπτρίζουν κυρίως τη στάση της ασθενούς. Έτσι ο περιγραφητής HOD θα ενσωματώσει την πληροφορία της πόζας, που είναι χαρακτηριστική της εκάστοτε δράσης (εν προκειμένω *Walk*) και εμφανίζεται παρόμοια σε κάθε στιγμιότυπο της δράσης. Από την άλλη, ο τελεστής HOG, που θα εξαχθει μέσω των ακμών των RGB εικόνων, θα ενσωματώσει στοιχεία θορυβώδους στατικής εμφάνισης που εμφανίζονται μόνο στο συγκεκριμένο στιγμιότυπο.

Το Σχήμα 5.11 αναδεικνύει άλλο ένα πλεονέκτημα του HOD περιγραφητή σε σχέση με το HOG. Όπως έχουμε προαναφέρει στο Κεφάλαιο 2, ο αλγόριθμος πυκνών τροχιών εκτελεί πυκνή δειγματοληψία και tracking των σημείων το πολύ σε 8 χωρικές κλίμακες. Σε μεγάλη κλίμακα οι ακμές από την εικόνα βάθους διατηρούν την πληροφορία για το περίγραμμα του σώματος, ενώ οι ακμές από την RGB εικόνα είναι αρκετά πιο θορυβώδεις.



Σχήμα 5.11: Εξαγωγή ακμών με χρήση του τελεστή Sobel από ένα depth frame (άνω σειρά εικόνων) και το αντίστοιχο RGB frame ενός video της βάσης δεδομένων MOBOT (κάτω σειρά εικόνων) σε μεγάλη χωρική κλίμακα.

Η ποιοτική υπεροχή του περιγραφητή HOD σε σύγκριση με τον HOG επιβεβαιώνεται και από τα πειραματικά αποτελέσματα, όπου παρατηρούμε ότι η ακρίβεια αναγνώρισης δράσεων πριν τη χρήση του αλγορίθμου Viterbi βελτιώνεται σχεδόν για όλους τους ασθενείς, με τη μέση ακρίβεια αναγνώρισης να αυξάνεται από 80.97% σε 83.59%.



Σχήμα 5.12: Σύγκριση περιγραφητών HOG και HOD εξαγμένων γύρω από κάθε τροχιά. Ακρίβειες αναγνώρισης δράσεων για κάθε ασθενή ξεχωριστά (unseen patient) και μέση ακρίβεια αναγνώρισης. Έχει γίνει χρήση VLAD αναπαράστασης με κανονικοποίηση Intra-Normalization και μείωση των διαστάσεων των περιγραφητών από 96 σε 64 στοιχεία με χρήση PCA.

Κεφάλαιο 6

Σχέσεις ομοιότητας ανάμεσα στις συστάδες χαρακτηριστικών

6.1 Εισαγωγή

Στα προηγούμενα κεφάλαια παρουσιάστηκαν μερικές από τις πιο διαδεδομένες αναπαραστάσεις βίντεο δράσεων της βιβλιογραφίας, όπως οι BoVW και VLAD αναπαραστάσεις. Ένα κοινό χαρακτηριστικό αυτών των μεθόδων είναι ότι συλλέγουν στατιστικά για την κατανομή των χαρακτηριστικών του βίντεο σε κάθε cluster ενός οπτικού λεξικού, είτε αυτά αφορούν τον αριθμό των χαρακτηριστικών που έχουν ανατεθεί σε κάθε συστάδα είτε τη σχετική θέση (ή και διακύμανσή) τους ως προς την οπτική λέξη/κεντροειδές της συστάδας, χωρίς να λαμβάνονται υπόψη σχέσεις μεταξύ των clusters.

Οι συστάδες αυτές ομαδοποιούν χαρακτηριστικά που μπορεί να έχουν παρόμοια κίνηση ή εμφάνιση, ανάλογα με τον περιγραφητή που χρησιμοποιείται. Επομένως, τα χαρακτηριστικά της κάθε συστάδας μπορεί να ανήκουν σε ένα μέρος (part) των κινούμενων αντικειμένων που κινείται με ένα συγκεκριμένο τρόπο (π.χ. σε σε ένα συγκεκριμένο μέλος του σώματος του ανθρώπου που εκτελεί μια δράση). Κατά την εκτέλεση των δράσεων υπάρχει αλληλεπίδραση μεταξύ των κινούμενων parts. Ο βαθμός και ο τρόπος που αλληλεπιδρούν μεταξύ τους τα διαφορετικά μέρη του σώματος διαφέρει από δράση σε δράση και έτσι μπορεί να χρησιμοποιηθεί για το διαχωρισμό τους. Κατά τη βάδιση υπάρχει συντονισμός των κινήσεων των άνω και κάτω άκρων και η κίνηση π.χ. του δεξιού χεριού ακολουθεί την κίνηση του αριστερού εισάγοντας μια σχέση εξάρτησης (αιτίου-αιτιατού) μεταξύ τους. Παρόμοια, κατά τη δράση *Jump* η κίνηση των γονάτων δεν παρουσιάζει σχέση αιτίας-αιτιατού, αλλά υπάρχει μια

σχέση ομοιότητας.

Σε αυτό το κεφάλαιο θα εξετάσουμε το κατά πόσο οι σχέσεις ομοιότητας/εξάρτησης μεταξύ των συστάδων (οπτικών λέξεων) μπορούν να οδηγήσουν σε εμπλουτισμένες αναπαραστάσεις, οι οποίες θα βοηθούν περαιτέρω στην επίλυση του προβλήματος της αναγνώρισης ανθρώπινων δράσεων. Αρχικά θα γίνει μια σύντομη επισκόπηση της μεθόδου αναπαράστασης βίντεο με περιγραφή της αιτιατότητας μεταξύ των τροχιών και στη συνέχεια θα παρουσιαστεί μια νέα μέθοδος αναπαράστασης βίντεο, η οποία βασίζεται στην κατευθυνόμενη ομοιότητα (directional similarity) μεταξύ των συστάδων τροχιών.

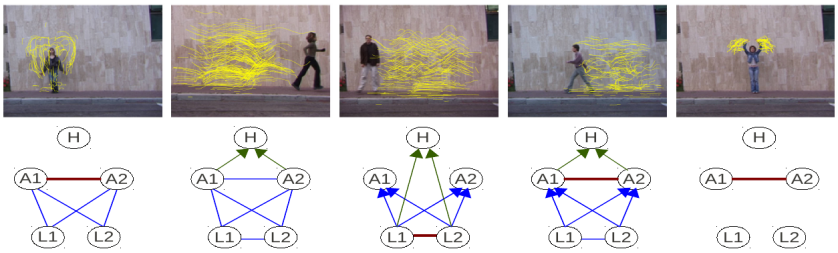
6.2 Σχέσεις αιτίου-αιτιατού μεταξύ των τροχιών

Οι Narayan et al. μελέτησαν το 2014 [69] την αλληλεπίδραση μεταξύ ζευγών από τροχιές και κατ'επέκταση μεταξύ ζευγών από συστάδες τροχιών και τη μοντελοποίησαν μετρώντας την αιτιατότητα κατά Granger (Granger causality measure). Στην εργασία τους υποθέτουν ότι οι τροχιές έχουν ισχυρή συσχέτιση μεταξύ τους και εμφανίζονται διαδοχικά, εξαρτώμενες από άλλες κινήσεις που είναι μέρος της δράσης. Κατά τη διάρκεια της βάδισης η αιώρηση του ενός χεριού ακολουθεί την αιώρηση του άλλου και επομένως οι τροχιές αυτών των κινήσεων έχουν σχέση αιτίου-αιτιατού (cause and effect). Η ισχύς της σχέσης αιτιατότητας μεταξύ δύο χρονοσειρών, εν προκειμένω τροχιών, μπορεί να ποσοτικοποιηθεί μέσω μιας μετρικής που βασίζεται στην αιτιατότητα κατά Granger.

Έστω δύο σήματα $\mathbf{p}_t$ και $\mathbf{q}_t$ (π.χ. δύο τροχιές). Αν το σφάλμα ϵ_1 της πρόβλεψης του $\mathbf{p}_t$, χρησιμοποιώντας μόνο τα προηγούμενα δείγματα του $\mathbf{p}_t$ είναι μεγαλύτερο από το σφάλμα ϵ_2 της πρόβλεψης χρησιμοποιώντας προηγούμενα δείγματα των $\mathbf{p}_t$ και $\mathbf{q}_t$, τότε λέμε ότι το σήμα $\mathbf{q}_t$ αιτιάζει κατά Granger (Granger-causes) το $\mathbf{p}_t$. Πιο συγκεκριμένα,

$$\begin{aligned} \mathbf{p}_t &= \mathbf{A}^T \mathbf{p}_{t-k}^{(m)} + \epsilon_1 \\ \mathbf{p}_t &= \mathbf{B}^T \mathbf{p}_{t-k}^{(m)} + \mathbf{C}^T \mathbf{q}_{t-k}^{(m)} + \epsilon_2 \end{aligned} \tag{6.1}$$

όπου $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{R}^{mn \times n}$ πίνακες με συντελεστές πρόβλεψης, ϵ_1, ϵ_2 σφάλματα πρόβλεψης, τα οποία μοντελοποιούνται ως γκαουσιανός θόρυβος μηδενικής μέσης τιμής και με πίνακες συνδιακύμανσης $\mathbf{\Sigma}_1, \mathbf{\Sigma}_2$ αντίστοιχα. Τα $\mathbf{p}_{t-k}^{(m)}, \mathbf{q}_{t-k}^{(m)}$



Σχήμα 6.1: Σχέσεις αιτιατότητας μεταξύ των κινήσεων των μελών του σώματος για διάφορες κατηγορίες δράσεων [69]. Τα μέλη του σώματος που εξετάζονται είναι: Κεφάλι (H), Βραχίονας 1 (A1), Βραχίονας 2 (A2), Πόδι 1 (L1), Πόδι 2 (L2). Η ισχύς της αιτιατότητας μεταξύ δύο κόμβων (μελών του σώματος) απεικονίζεται μέσω του πάχους και του χρώματος της αντίστοιχης ακμής του γράφου. Οι παχιές, κόκκινες γραμμές συμβολίζουν ισχυρές σχέσεις αιτιατότητας, οι μεσαίου πάχους, μπλε ακμές αντιστοιχούν σε σχέσεις αιτιατότητας μεσαίας ισχύος και οι λεπτές πράσινες σηματοδοτούν μικρή αιτιατότητα. Η έλλειψη ακμής ανάμεσα σε δύο κόμβους σηματοδοτεί την απουσία σχέσης αιτιατότητας. Οι κατευθυνόμενες ακμές υποδηλώνουν ο λόγος αιτιατότητας είναι μεγαλύτερος σε αυτή την κατεύθυνση σε σύγκριση με την αντίθετη. Αντίστοιχα, οι μη κατευθυνόμενες ακμές υποδηλώνουν ότι οι λόγοι αιτιατότητας είναι παρόμοι και προς τις δύο κατευθύνσεις.

είναι σήματα τάξης m με καθυστέρηση k:

$$\begin{aligned} \mathbf{p}_{t-k}^{(m)} &= (\mathbf{p}_{t-k}^T, \mathbf{p}_{t-k-1}^T, \ldots, \mathbf{p}_{t-k-m+1}^T)^T \\ \mathbf{q}_{t-k}^{(m)} &= (\mathbf{q}_{t-k}^T, \mathbf{q}_{t-k-1}^T, \ldots, \mathbf{q}_{t-k-m+1}^T)^T \end{aligned} \tag{6.2}$$

Η ισχύς της αιτιατότητας από το $\mathbf{q}$ στο $\mathbf{p}$ μετράται με το λόγο αιτιατότητας (causality ratio):

$$CR_{\mathbf{q}\to\mathbf{p}} = \frac{\text{trace}(\mathbf{\Sigma}_1)}{\text{trace}(\mathbf{\Sigma}_2)} \tag{6.3}$$

Στο Σχήμα 6.1 οπτικοποιούνται οι σχέσεις αιτιατότητας μεταξύ των μελών του σώματος για μερικές κατηγορίες δράσεων. Όπως παρατηρούμε, οι σχέσεις αυτές διαφέρουν μεταξύ των δράσεων, όπου για παράδειγμα οι τροχιές των δύο χεριών έχουν ισχυρή σχέση αιτιατότητας στην περίπτωση της δράσης *Wave* αλλά δε σχετίζονται κατά την εκτέλεση της δράσης *Side walk*.

Ωστόσο η αναπαράσταση ενός βίντεο με χρήση του causality ratio ανάμεσα σε κάθε ζεύγος από τροχιές είναι προβληματική, καθώς κάθε βίντεο περιέχει διαφορετικό πλήθος τροχιών και επομένως ο πίνακας που θα περιείχε τα causality ratios για κάθε ζεύγος τροχιών θα είχε διαφορετικό μέγεθος για κάθε βίντεο, δυσχεραίνοντας τη σύγκριση μεταξύ των βίντεο. Γι'αυτό οι Narayan

et al. προτείνουν την κατασκευή ενός οπτικού λεξικού με τον ίδιο τρόπο που κατασκευάζεται στη μέθοδο BoVW με χρήση K-means πάνω σε ένα τυχαία επιλεγμένων περιγραφητών, π.χ. Trajectory descriptors. Έτσι θα μπορούσε να υπολογιστεί για κάθε βίντεο ένας πίνακας αιτιατότητας $K \cdot K$ στοιχείων, του οποίου το (i, j)-οστό στοιχείο θα ήταν ο μέσος όρος των causality ratios μεταξύ των τροχιών που ανήκουν στη συστάδα (cluster) i και των τροχιών που ανήκουν στη συστάδα j. Αυτή η αναπαράσταση όμως έχει πολύ υψηλό υπολογιστικό κόστος, καθώς πρέπει να υπολογιστεί ο λόγος αιτιατότητας ανάμεσα σε κάθε ζεύγος τροχιών, οι οποίες, όπως προαναφέραμε, είναι πολυάριθμες. Για αυτό το λόγο, προτείνεται μια τελική αναπαράσταση κατά την οποία υπολογίζεται ένας ενδιάμεσος πίνακας αναφοράς R, του οποίου κάθε στοιχείο (i, j) είναι το causality ratio ανάμεσα στη μέση τροχιά του cluster i και τη μέση τροχιά του cluster j, και στη συνέχεια το κάθε στοιχείο του πίνακα αιτιατότητας (causality descriptor) $CD(i, j)$ υπολογίζεται ως:

$$CD(i, j) = N_i N_j R(i, j) \tag{6.4}$$

όπου N_i είναι ο αριθμός των τροχιών που έχουν ανατεθεί στην οπτική λέξη i και N_j είναι ο αριθμός των τροχιών που έχουν ανατεθεί στην οπτική λέξη j. Το σημαντικό σε αυτή την προσέγγιση είναι ότι ο πίνακας $R(i, j)$, που απαιτεί τον υπολογισμό της μετρικής αιτιότητας, μπορεί να κατασκευαστεί μία φορά για ένα ολόκληρο dataset (σύνολο βίντεο δράσεων) από τις τροχιές που χρησιμοποιούνται στο clustering και στη συνέχεια αρκεί να μετράμε σε κάθε βίντεο πόσες τροχιές ανατίθεται σε κάθε βίντεο, όπως κάναμε άλλωστε και με το γρήγορο και απλό αλγόριθμο BoVW. Η αναπαράσταση του βίντεο είναι τελικά το διάνυσμα που περιέχει τα στοιχεία του πίνακα CD.

Χρησιμοποιώντας τα χαρακτηριστικά πυκνών τροχιών (dense trajectories) (Κεφ. 2), την αναπαράσταση βίντεο που παρουσιάστηκε και μη γραμμικούς SVM ταξινομητές με πυρήνα τομής ιστογραμμάτων (Histogram Intersection Kernel) (Κεφ. 4), οι Narayan et al. οδηγήθηκαν σε state-of-the-art αποτελέσματα αναγνώρισης δράσεων.

6.3 Ποσοτικοποίηση της ομοιότητας μεταξύ συστάδων

Η μέθοδος που μόλις αναλύθηκε, λοιπόν, εκμεταλλεύεται την πληροφορία αιτίας-αιτιατού της μέσης τροχιάς ενός cluster σε σχέση με τη μέση τροχιά ενός άλλου cluster (predictive causality) με χρήση ενός γραμμικού μοντέλου, και έτσι εμπλουτίζει με αυτή τη συμπληρωματική πληροφορία τις υπάρχουσες αναπαραστάσεις τύπου BoVW. Στα πλαίσια αυτής της εργασίας, θα γενικεύ-

σουμε αυτή τη μέθοδο έτσι ώστε ο πίνακας αναφοράς R να ενσωματώσει πληροφορία σχετική με την ομοιότητα των ζευγών από clusters ως προς κάποιο χαρακτηριστικό.

Η βασική ιδέα είναι να υπολογίσουμε μια μετρική κατευθυνόμενης ομοιότητας ανάμεσα στους χώρους των οπτικών χαρακτηριστικών των συστάδων. Οι χώροι οπτικών χαρακτηριστικών κατασκευάζονται εφαρμόζοντας τη μέθοδο PCA πάνω στα χαρακτηριστικά (Trajectory, HOG, HOF, MBH κλπ.) που έχουν ανατεθεί σε κάθε συστάδα. Η ομοιότητα κάθε ζεύγους συστάδων υπολογίζεται αν προβάλλουμε τα χαρακτηριστικά της μιας συστάδας στο χώρο PCA της δεύτερης και αντίστροφα, δηλαδή υπάρχουν δύο κατευθύνσεις ομοιότητας.

Αυτή η μοντελοποίηση και ποσοτικοποίηση της ομοιότητας είναι εμπνευσμένη από τη δουλειά των Lee et. al [70], οι οποίοι είχαν ποσοτικοποιήσει με τον ίδιο τρόπο την κατευθυνόμενη ομοιότητα μεταξύ των χρονοσειρών των φωνητικών χαρακτηριστικών δύο ομιλητών, με απώτερο στόχο την εξαγωγή συμπερασμάτων για το συντονισμό (entrainment) των δύο ομιλητών. Στις αλληλεπιδράσεις μεταξύ των ανθρώπων, το entrainment είναι ένα φαινόμενο το οποίο συμβαίνει όταν τα άτομα που αλληλεπιδρούν προσαρμόζουν αμοιβαία τις συμπεριφορές τους κατά τη διάρκεια της αλληλεπίδρασης. Σύμφωνα με μελέτες της ψυχολογίας, αυτό το φαινόμενο συμβαίνει αυθόρμητα και υπηρετεί πολλαπλούς σκοπούς, όπως τη βελτίωση της αποδοτικότητας της επικοινωνίας, την αύξηση του ενδιαφέροντος και της συμμετοχής, ενώ ταυτόχρονα ευνοεί και την αμοιβαία κατανόηση [71]. Στα πλαίσια αυτή της προσαρμογής της συμπεριφοράς, συντονίζονται και τα χαρακτηριστικά της φωνής των ομιλητών, καθώς ο ένας ομιλητής παρασύρει τον άλλον. Μέσω της μετρικής της κατευθυνόμενης ομοιότητας μπορεί κανείς να καταλάβει πόσο τείνουν να προσαρμοστούν τα φωνητικά χαρακτηριστικά του ενός ομιλητή στα χαρακτηριστικά του συνομιλητή του και αντίστροφα. Κατά ανάλογο τρόπο, στην αναπαράσταση που προτείνουμε μας ενδιαφέρει να δούμε πόσο προσομοιάζουν τα χαρακτηριστικά ενός cluster στα χαρακτηριστικά ενός άλλου cluster, μεταφέροντας τεχνικές από τη μελέτη της αλληλεπίδρασης μεταξύ των ανθρώπων στη μελέτη της αλληλεπίδρασης μεταξύ των διαφορετικών ομάδων τροχιών που συνιστούν μια δράση.

Από την επιτυχία της μεθόδου αναπαράστασης Fisher μπορούμε να συμπεράνουμε πόσο σημαντική είναι η πληροφορία του σχήματος της κατανομής των χαρακτηριστικών που ανήκουν σε κάθε συστάδα (cluster). Γι'αυτό προτείνουμε να συγκρίνονται τα ζεύγη των clusters ως προς το σχήμα των κατανομών των χαρακτηριστικών. Αυτή η πρόταση γενικεύει τη δουλειά των Narayan, η οποία εστιάζει στο αν η πληροφορία της μέσης τροχιάς ενός cluster βοηθάει στην πρόβλεψη της μέσης τροχιάς ενός άλλου cluster (predictive causality). Ο υπολογισμός των ομοιοτήτων μεταξύ των σχημάτων των χώρων

χαρακτηριστικών των clusters μάς επιτρέπει να λάβουμε υπόψη οποιοδήποτε περιγραφητή θέλουμε, χωρίς να περιοριζόμαστε στην ποσοτικοποίηση της αιτιατότητας μεταξύ των χωροχρονικών σημάτων των τροχιών που ανήκουν σε κάθε cluster. Μπορούμε για παράδειγμα να εστιάσουμε στην ομοιότητα των clusters ως προς τα χαρακτηριστικά στατικής εμφάνισης HOG και έτσι να εκμεταλλευτούμε πλουσιότερη πληροφορία.

Για να ποσοτικοποιήσουμε την ομοιότητα του σχήματος της κατανομής των δεδομένων ενος cluster ως προς το σχήμα ενός άλλου cluster, δεν αρκεί να συγκρίνουμε απλώς τις κατευθύνσεις των κύριων συνιστωσών των χώρων χαρακτηριστικών. Παραδείγματος χάριν, στο δισδιάστατο χώρο, ένα cluster του οποίου τα στοιχεία έχουν μικρή συσχέτιση θα έχει σχεδόν κυκλικό σχήμα, ενώ αντιθέτως ένα cluster του οποίου τα στοιχεία έχουν μέγιστη συσχέτιση θα έχει όλα τα στοιχεία του κατά μήκος μιας γραμμής. Μπορεί στις δύο προηγούμενες περιπτώσεις, οι κύριες συνιστώσες των δύο συστάδων, όπως αυτές προκύπτουν μέσω της μεθόδου PCA, να είναι ευθυγραμμισμένες, αλλά το σχήμα των clusters, που αποτυπώνεται στη διακύμανσή των στοιχείων τους, είναι προφανώς διαφορετικό. Για να ποσοτικοποιήσουμε αυτές τις διαφορές των σχημάτων των κατανομών των χαρακτηριστικών δύο clusters, εξετάζουμε τη διαφορά στις κατανομές της διακύμανσης κατά μήκος των κύριων συνιστωσών [72], χρησιμοποιώντας την απόκλιση Kullback-Leibler [73].

Πιο συγκεκριμένα, στην παραλλαγή της μεθόδου των Narayan et al. που προτείνουμε, το στοιχείο (i, j) του πίνακα αναφοράς R υπολογίζεται από τα χαρακτηριστικά που έχουν χρησιμοποιηθεί για την κατασκευή του οπτικού λεξικού ως εξής:

Έστω $\mathbf{X}_i$ το σύνολο των χαρακτηριστικών, π.χ. τροχιών που περιγράφονται από τον περιγραφητή MBHy, που έχουν ανατεθεί στο cluster i, δηλαδή η οπτική λέξη i είναι ο κοντινότερος γείτονάς τους, και $\mathbf{X}_j$ το σύνολο των χαρακτηριστικών που αντιστοιχούν στο cluster j.

- Εκτελούμε PCA στο σύνολο $\mathbf{X}_i$ έτσι ώστε να προσδιορίσουμε τον πίνακα προβολής $\mathbf{A}_i$ (βλ. Κεφάλαιο 3), δηλαδή τα ιδιοδιανύσματα του πίνακα συνδιακύμανσης $\mathbf{X}_i^T\mathbf{X}_i$, και τις κύριες συνιστώσες $\mathbf{Y}_i$:

$$\mathbf{Y}_i = \mathbf{X}_i^T\mathbf{A}_i.$$

- Εκτελούμε PCA στο σύνολο $\mathbf{X}_j$ έτσι ώστε να προσδιορίσουμε τον πίνακα προβολής $\mathbf{A}_j$ και τις κύριες συνιστώσες $\mathbf{Y}_j$:

$$\mathbf{Y}_j = \mathbf{X}_j^T\mathbf{A}_j.$$

- Υπολογίζουμε τα εξής δύο διανύσματα κανονικοποιημένων διακυμάνσεων (normalized variances) $\{\boldsymbol{\lambda}_{ij}^n, \boldsymbol{\lambda}_{jj}^n\}$, που αντιστοιχούν στις διακυμάνσεις κατά μήκος των κύριων συνιστωσών αν: α) τα χαρακτηριστικά

του cluster i προβληθούν στις κύριες συνιστώσες του cluster j και β) τα χαρακτηριστικά του cluster j προβληθούν στις κύριες συνιστώσες του cluster j, αντίστοιχα.

1. Προβάλλουμε τα χαρακτηριστικά $\mathbf{X}_i$ χρησιμοποιώντας τον πίνακα $\mathbf{A}_j$: $\mathbf{Y}_{ij} = \mathbf{X}_i^T \mathbf{A}_j$.
2. Υπολογίζουμε το διάνυσμα διακυμάνσεων: $\boldsymbol{\lambda}_{ij} = \mathrm{var}(\mathbf{Y}_{ij})$.
3. Κανονικοποιούμε το διάνυσμα διακυμάνσεων, έτσι ώστε να έχει μοναδιαία L_1 νόρμα, δηλαδή τα στοιχεία του να αθροίζουν στη μονάδα: $\boldsymbol{\lambda}_{ij}^n = \frac{\boldsymbol{\lambda}_{ij}}{\sum_k \boldsymbol{\lambda}_{ij,k}}$.
4. Προβάλλουμε τα χαρακτηριστικά $\mathbf{X}_j$ χρησιμοποιώντας τον πίνακα $\mathbf{A}_j$: $\mathbf{Y}_{jj} = \mathbf{X}_j^T \mathbf{A}_j$.
5. Υπολογίζουμε το διάνυσμα διακυμάνσεων: $\boldsymbol{\lambda}_{jj} = \mathrm{var}(\mathbf{Y}_{jj})$.
6. Κανονικοποιούμε το διάνυσμα διακυμάνσεων: $\boldsymbol{\lambda}_{jj}^n = \frac{\boldsymbol{\lambda}_{jj}}{\sum_k \boldsymbol{\lambda}_{jj,k}}$.

Αυτά τα κανονικοποιημένα διανύσματα θα μας επιτρέψουν να υπολογίσουμε την ομοιότητα ανάμεσα στις κατανομές των χαρακτηριστικών στα δύο cluster i και j, όταν προβάλλουμε το $\mathbf{X}_i$ στο χώρο PCA του $\mathbf{X}_j$. Κάθε στοιχείο αυτών των διανυσμάτων υποδηλώνει το ποσοστό της διακύμανσης που εξηγείται καθώς προβάλλουμε τα χαρακτηριστικά σε καθεμία από τις κύριες συνιστώσες του cluster j. Αν διατηρηθούν όλες οι συνιστώσες, τότε τα στοιχεία του κάθε διανύσματος αθροίζουν στη μονάδα. Επομένως, μπορούμε να θεωρήσουμε ότι το κάθε διάνυσμα αντιπροσωπεύει μια κατανομή πιθανότητας της τυχαίας μεταβλητής V_j: $P_{ij} = P(V_j = k) = \boldsymbol{\lambda}_{ij,k}^n$ και $P_{jj} = P(V_j = k) = \boldsymbol{\lambda}_{jj,k}^n$. Αξίζει να σημειωθεί, ωστόσο, ότι για να έχει νόημα η χρήση της μεθόδου PCA για τον προσδιορισμό των κύριων κατευθύνσεων σε κάθε συστάδα, πρέπει ο αριθμός των χαρακτηριστικών (τροχιών) που έχουν ανατεθεί στη συστάδα να είναι μεγαλύτερος ή ίσος της διάστασης του περιγραφητή που χρησιμοποιείται. Για παράδειγμα, αν χρησιμοποιηθεί ο περιγραφητής Trajectory με διάσταση 30 στοιχεία, πρέπει να έχουμε τουλάχιστον 30 τροχιές αντιστοιχισμένες σε κάθε συστάδα έτσι ώστε να παραχθεί μια μοναδική αναπαράσταση PCA. Άρα πρέπει ο αριθμός των τυχαία επιλεγμένων χαρακτηριστικών που θα χρησιμοποιηθούν για την κατασκευή του οπτικού λεξικού να είναι πολύ μεγαλύτερος του αριθμού των συστάδων, έτσι ώστε να ανατεθεί μεγάλος αριθμός χαρακτηριστικών σε κάθε συστάδα, σημαντικά μεγαλύτερος της διάστασης του περιγραφητή.

- Ομοίως, υπολογίζουμε τα εξής δύο διανύσματα κανονικοποιημένων διακυμάνσεων (normalized variances) $\{\boldsymbol{\lambda}_{ji}^n, \boldsymbol{\lambda}_{ii}^n\}$, που αντιστοιχούν στις

διακυμάνσεις κατά μήκος των κύριων συνιστωσών αν: α) τα χαρακτηριστικά του cluster j προβληθούν στις κύριες συνιστώσες του cluster i και β) τα χαρακτηριστικά του cluster i προβληθούν στις κύριες συνιστώσες του cluster i, αντίστοιχα.

1. Προβάλλουμε τα χαρακτηριστικά $\mathbf{X}_j$ χρησιμοποιώντας τον πίνακα $\mathbf{A}_i$: $\mathbf{Y}_{ji} = \mathbf{X}_j^T \mathbf{A}_i$.
2. Υπολογίζουμε το διάνυσμα διακυμάνσεων: $\boldsymbol{\lambda}_{ji} = \text{var}(\mathbf{Y}_{ji})$.
3. Κανονικοποιούμε το διάνυσμα διακυμάνσεων: $\boldsymbol{\lambda}_{ji}^n = \frac{\boldsymbol{\lambda}_{ji}}{\sum_k \boldsymbol{\lambda}_{ji,k}}$.
4. Προβάλλουμε το $\mathbf{X}_i$ χρησιμοποιώντας τον πίνακα $\mathbf{A}_i$: $\mathbf{Y}_{ii} = \mathbf{X}_i^T \mathbf{A}_i$.
5. Υπολογίζουμε το διάνυσμα διακυμάνσεων: $\boldsymbol{\lambda}_{ii} = \text{var}(\mathbf{Y}_{ii})$.
6. Κανονικοποιούμε το διάνυσμα διακυμάνσεων: $\boldsymbol{\lambda}_{ii}^n = \frac{\boldsymbol{\lambda}_{ii}}{\sum_k \boldsymbol{\lambda}_{ii,k}}$.

Με αυτά τα διανύσματα κανονικοποιημένων διακυμάνσεων θα υπολογίσουμε την ομοιότητα μεταξύ των $\mathbf{X}_i$ και $\mathbf{X}_j$ όταν αναπαριστούμε το $\mathbf{X}_j$ στο χώρο PCA του $\mathbf{X}_i$. Όπως ήδη εξηγήσαμε, μπορούμε να θεωρήσουμε ότι το κάθε διάνυσμα αντιπροσωπεύει μια κατανομή πιθανότητας της τυχαίας μεταβλητής V_i: $P_{ji} = P(V_i = k) = \boldsymbol{\lambda}_{ji,k}^n$ και $P_{ii} = P(V_i = k) = \boldsymbol{\lambda}_{ii,k}^n$.

- Έχοντας εκφράσει τα διανύσματα κανονικοποιημένων διακυμάνσεων ως κατανομές πιθανότητας μπορούμε να εκφράσουμε την ομοιότητα μεταξύ των διανυσμάτων διακύμανσης σαν ομοιότητα ανάμεσα σε δύο κατανομές πιθανότητας. Μια πολύ διαδεδομένη μετρική της διαφοράς (και επομένως της ομοιότητας) μεταξύ κατανομών πιθανότητας μιας τυχαίας μεταβλητής $\mathcal{X}$ είναι η συμμετρική Kullback-Leibler απόκλιση (SKLD). Όσο πιο παρόμοιες είναι δύο συστάδες χαρακτηριστικών, τόσο μικρότερη είναι η τιμή της SKLD.

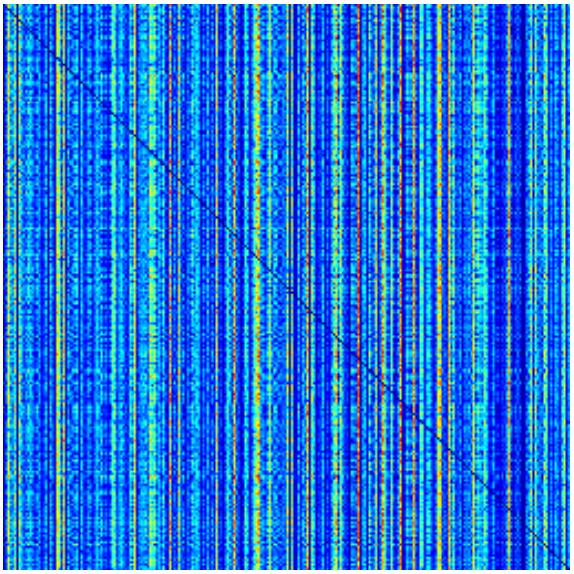
$$R(i,j) = SKLD(P_{jj}||P_{ij}) = \frac{1}{2}(D_{KL}(P_{jj}||P_{ij}) + D_{KL}(P_{ij}||P_{jj})) \quad (6.5)$$

Η απόκλιση Kullback-Leibler (ή σχετική εντροπία) D_{KL} δύο κατανομών πιθανότητας P και Q μιας τυχαίας μεταβλητής $\mathcal{X}$ συγκρίνει την εντροπία των δύο πιθανοτικών κατανομών. Η απόκλιση KL προέρχεται από το πεδίο της Θεωρίας Πληροφορίας και διαισθητικά είναι ο αριθμός των πρόσθετων bits που απαιτούνται αν κωδικοποιήσουμε μια τυχαία μεταβλητή που ακολουθεί μια κατανομή $P(i)$ χρησιμοποιώντας μια εναλλακτική κατανομή $Q(i)$.

$$D_{KL}(P||Q) = \sum_{i \in \mathcal{X}} P(i) \log_2 \frac{P(i)}{Q(i)} \quad (6.6)$$

Από τον ορισμό της Kullback-Leibler απόκλισης βλέπουμε ότι είναι μη αρνητική $D_K L(P||Q) \geq 0$, μη συμμετρική $D(P||Q) \neq D(Q||P)$ και είναι ίση με το μηδέν για ίσες κατανομές $D(P||Q) = 0 \quad iff\ P(i) = Q(i)\, \forall i \in \mathcal{X}$.

Όπως είναι προφανές, εν γένει ισχύει ότι $R(i,j) \geq 0$, $R(i,j) \neq R(j,i)$ και $R(i,i) = 0$. Επομένως κατασκευάστηκε μια μετρική κατευθυνόμενης ομοιότητας. Στο Σχήμα 6.2 απεικονίζονται οι τιμές της μετρικής κατευθυνόμενης ομοιότητας για κάθε ζεύγος από 256 clusters. Όσο χαμηλότερη είναι τιμή της μετρικής, τόσο πιο όμοια είναι τα clusters. Τα στοιχεία της διαγωνίου είναι μηδενικά. Αντιθέτως, υψηλές τιμές της μετρικής αποκαλύπτουν ζεύγη clusters με έντονες διαφορές στην κατανομή των χαρακτηριστικών τους. Οι τιμές που απεικονίζονται αποτελούν τα στοιχεία του πίνακα αναφοράς R που αντικαθιστά τον πίνακα R της μεθόδου των Narayan, ο οποίος περιείχε τιμές αιτιατότητας κατά Granger.



Σχήμα 6.2: Πίνακας αναφοράς με χρήση της μετρικής SKLD. Η μετρική έχει υπολογιστεί για όλα τα ζεύγη 256 clusters. Τα clusters έχουν υπολογιστεί με τη βοήθεια του αλγορίθμου K-means, εφαρμοσμένου σε MBHy χαρακτηριστικά τυχαία επιλεγμένα από τα χαρακτηριστικά ενός σύνολου βίντεο δράσεων. Όσο χαμηλότερη είναι τιμή της μετρικής (μπλε αποχρώσεις), τόσο πιο όμοια είναι τα clusters. Αντιθέτως, υψηλές τιμές της μετρικής αποκαλύπτουν ζεύγη clusters με έντονες διαφορές στην κατανομή των χαρακτηριστικών τους.

Για κάθε βίντεο που επιθυμούμε να αναπαραστήσουμε αρκεί να αναθέσουμε τα χαρακτηριστικά του (τροχιές) στις συστάδες/οπτικές λέξεις του λεξικού που έχει κατασκευαστεί, να μετρήσουμε τον αριθμό των τροχιών N_i που έχουν ανατεθεί με αυστηρή ανάθεση σε κάθε συστάδα i και να υπολογίσουμε τα KxK στοιχεία του διανύσματος SD (similarity descriptor) της αναπαράστασής μας ως εξής:

$$SD = N_i N_j R(i,j) \tag{6.7}$$

όπου R είναι ο πίνακας αναφοράς (6.5). Εισήχθηκε λοιπόν μια αναπαράσταση βίντεο, η οποία ενσωματώνει την πληροφορία της ομοιότητας μεταξύ των clusters καθώς και την πληροφορία του αριθμού των χαρακτηριστικών που ανατίθενται σε κάθε cluster.

6.4 Χρήση GMM clustering

Τόσο η μέθοδος των Narayan et al. όσο και η παραλλαγή της που προτάθηκε παραπάνω, βασίζονται στην κατασκευή ενός λεξικού με τη χρήση K-means και την καταμέτρηση των τροχιών που έχουν ανατεθεί σε κάθε cluster, δηλαδή βασίζονται έμμεσα στον υπολογισμό των BoVW. Όπως είχαμε αναλύσει ωστόσο στο Κεφ. 3 η αυστηρή ανάθεση των χαρακτηριστικών στις οπτικές λέξεις οδηγεί σε απώλεια πληροφορίας, η οποία μπορεί να περιοριστεί αν χρησιμοποιηθεί χαλαρή ανάθεση των χαρακτηριστικών μέσω π.χ. GMM clustering. Για να εξετάσουμε αν όντως η αντικατάσταση του K-means με GMM βοηθάει στην περίπτωσή μας, θα τροποποιήσουμε ελαφρώς τα βήματα υπολογισμού της αναπαράστασής μας.

Αρχικά προσδιορίζονται οι παράμετροι $\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}$ του Μοντέλου Μείγματος Γκαουσιανών σε κάποια τυχαία επιλεγμένα χαρακτηριστικά από βίντεο εκπαίδευσης, στα οποία έχει εφαρμοστεί PCA για να εξαλειφθούν οι συσχετίσεις μεταξύ τους, έτσι ώστε να ισχύει η υπόθεση των διαγώνιων πινάκων συνδιακύμανσης των Γκαουσιανών (βλ. ενότητα 3.2.2).

Στη συνέχεια, πρέπει υπολογίσουμε τις τιμές του πίνακα αναφοράς R για κάθε ζεύγος συστάδων (Γκαουσιανών). Για να υπολογιστούν αυτές οι μετρικές πρέπει να υπολογίσουμε τον πίνακα προβολής PCA για την κάθε συστάδα, βάσει των χαρακτηριστικών που ανήκουν σε αυτή. Εν προκειμένω, το n-οστό χαρακτηριστικό ανήκει στη συστάδα k με μία ύστερη πιθανότητα (responsibility) γ_{nk} (εξίσωση 3.8). Για λόγους απλότητας κατά τον υπολογισμό των πινάκων προβολής $\mathbf{A}_i$, $\mathbf{A}_j$ αναθέτουμε το κάθε χαρακτηριστικό στο cluster που είχε τη μεγαλύτερη ευθύνη για την παραγωγή του. Αφού μοιράσουμε τα χαρακτηριστικά στα clusters, υπολογίζουμε κατά τα γνωστά τον πίνακα R. Τέλος, αντικαθιστούμε το γινόμενο $N_i N_j$ που αντιστοιχούσε στο

γινόμενο του αριθμού των τροχιών που είχαν ανατεθεί βάσει της Ευκλείδειας απόστασης στις οπτικές λέξεις i και j (hard assignment), με το γινόμενο του ενεργού αριθμού των τροχιών που έχουν ανατεθεί στις Γκαουσιανές (soft assignment). Άρα η αναπαράσταση ενός βίντεο με χρήση πληροφορίας ομοιότητας μεταξύ των clusters και GMM clustering (SD-soft assignment) γίνεται:

$$SD(i,j) = \sum_{n=1}^{N} \gamma_{ni} \sum_{n=1}^{N} \gamma_{nj} R(i,j) \tag{6.8}$$

όπου N είναι ο αριθμός των τροχιών του βίντεο.

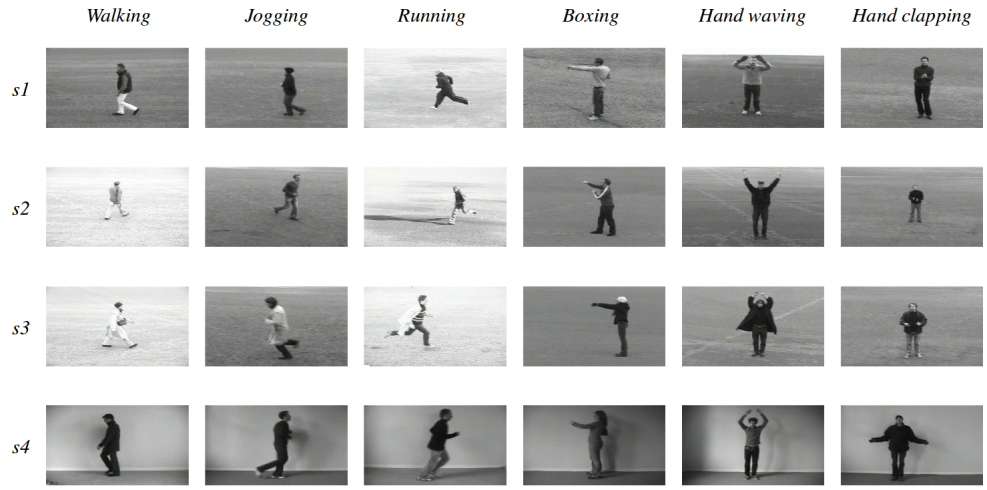
6.5 Πειράματα ταξινόμησης ανθρώπινων δράσεων

Σε αυτή την ενότητα αξιολογούμε τη μέθοδο αναπαράστασης αναγνώρισης δράσεων που προτείνουμε στη δημοφιλή βάση δεδομένων ανθρώπινων δράσεων KTH, η οποία είναι ιδιαίτερα δημοφιλής. Αρχικά, παρουσιάζουμε τη βάση KTH, τις λεπτομέρειες υλοποίησης της προσέγγισης μας και των συστημάτων σύγκρισης. Στη συνέχεια, παραθέτουμε τα αποτελέσματα των πειραμάτων στη βάση KTH.

6.5.1 Η βάση ανθρώπινων δράσεων KTH

Η βάση δεδομένων ανθρώπινων δράσεων KTH είναι από τις πιο γνωστές βάσεις δεδομένων και χρησιμοποιείται ευρέως στη διεθνή βιβλιογραφία. Καλύπτει ένα μικρό αριθμό κατηγοριών δράσεων, οι οποίες εκτελούνται σε ελεγχόμενο περιβάλλον. Πιο συγκεκριμένα, έχει 6 κατηγορίες δράσεων προς ταξινόμηση: *Walking*, *Jogging*, *Running*, *Boxing*, *Hand Waving* και *Hand Clapping*. Περιέχει συνολικά 2391 βίντεο από τις 6 κατηγορίες δράσεων, εκτελεσμένες σε ελεγχόμενο περιβάλλον πολλαπλές φορές από 25 άτομα υπό 4 διαφορετικές συνθήκες (s_1: σε εξωτερικό χώρο, s_2: σε εξωτερικό χώρο με μεταβλητή κλίμακα, s_3: σε εξωτερικό χώρο με διαφορετικά ρούχα και s_4: σε εσωτερικό χώρο) (Σχήμα 6.3). Τα βίντεο έχουν ληφθεί σε ομογενές υπόβαθρο με στατική κάμερα και συχνότητα δειγματοληψίας 25 frames ανά δευτερόλεπτο (fps). Επίσης, τα βίντεο έχουν υποστεί υποδειγματοληψία έτσι ώστε να έχουν μια ανάλυση 160×160 pixels και διαρκούν κατά μέσο όρο 4 δευτερόλεπτα. Χρησιμοποιούμε τον ίδιο διαχωρισμό (splits) των βίντεο σε σύνολο εκπαίδευσης και αξιολόγησης με αυτόν που χρησιμοποιήθηκε στο αρχικό στήσιμο της βάσης [14], με τα βίντεο από 9 άτομα (2,3,5,6,7,8,9,10 και 22) να ανήκουν στο

σύνολο αξιολόγησης και τα βίντεο από τα υπόλοιπα άτομα να χρησιμοποιούνται ως βίντεο εκπαίδευσης. Η ακρίβεια ταξινόμησης δράσεων υπολογίζεται ως το ποσοστό των σωστά ταξινομημένων βίντεο προς το σύνολο των βίντεο της βάσης.



Σχήμα 6.3: Ενδεικτικά frames από τη βάση KTH που περιέχει τις 6 κατηγορίες δράσεων: *Walking*, *Jogging*, *Running*, *Boxing*, *Hand Waving* και *Hand Clapping*.

6.5.2 Πειραματική διαδικασία

Στα πειράματα που ακολουθούν εξάγουμε πυκνές τροχιές (Dense Trajectories) χρησιμοποιώντας τον κώδικα που παρέχουν οι Wang et al. [30] με τις default παραμέτρους, όπως αυτές αναλύθηκαν στα προηγούμενα κεφάλαια. Υπολογίζονται οι περιγραφητές Trajectory, HOG, HOF, MBHx και MBHy. Το οπτικό λεξικό για κάθε περιγραφητή κατασκευάζεται κατά τα γνωστά με K-means ή GMM ομαδοποίηση 100000 τυχαία επιλεγμένων στιγμιοτύπων εκπαίδευσης, με την ίδια τυχαία αρχικοποίηση ($seed = 10$) κάθε φορά. Το μέγεθος του λεξικού επιλέχθηκε ίσο με $K = 256$ οπτικές λέξεις.

Εξετάζουμε τις δύο παραλλαγές της μεθόδου μας (εξισώσεις 6.5 και 6.8) και τις συγκρίνουμε με τρεις διαφορετικές μεθόδους αναπαράστασης (BoVW, VLAD και Fisher Vector). Τα ιστογράμματα BoVW κανονικοποιούνται με χρήση της ℓ_1 νόρμας, ενώ τα διανύσματα VLAD και Fisher κανονικοποιούνται με intra-normalization και power-normalization, αντίστοιχα. Για την αναπαράσταση Fisher, εφαρμόζουμε PCA-Whitening στα χαρακτηριστικά με διατήρηση όλων των μισών συνιστωσών. Χρησιμοποιήθηκε η υλοποίηση του

VLAD από τη βιβλιοθήκη VLfeat[1] και η υλοποίηση των GMMs, Fisher Vector της βιβλιοθήκης Yael[2].

Για να αξιολογήσουμε τη μετρική ομοιότητας, τη συγκρίνουμε με μία εναλλακτική μετρική, καθώς ο πίνακας αναφοράς R εν γένει μπορεί να περιέχει οποιαδήποτε μετρική. Εν προκειμένω χρησιμοποιούμε τους συντελεστές αυτοσυσχέτισης μεταξύ των τροχιών, οι οποίοι μπορούν να μετρήσουν την αλληλεπίδραση μεταξύ των τροχιών, χωρίς όμως να μας πληροφορούν σχετικά με την κατεύθυνση της αλληλεπίδρασης, καθώς πρόκειται για μια συμμετρική μετρική. Υπολογίζουμε ξεχωριστά τους συντελεστές αυτοσυσχέτισης για τις x και y συνιστώσες τροχιών, τους οποίους μετασχηματίζουμε έτσι ώστε να έχουν τιμές από 0 έως 1 (αντί για -1 έως 1). Για κάθε ζεύγος συστάδων υπολογίζονται δύο συντελεστές αυτοσυσχέτισης $R_x(i,j)$ και $R_y(i,j)$ και επομένως μπορούν να παραχθούν για κάθε βίντεο δύο αναπαραστάσεις: $CorD_x = N_i N_j R_x(i,j)$ και $CorD_y = N_i N_j R_y(i,j)$. Ωστόσο, λόγω της συμμετρίας της μετρικής αυτοσυσχέτισης, αρκούν $\binom{4000}{2}$ στοιχεία από κάθε αναπαράσταση. Τα στοιχεία αυτά συνενώνονται και προκύπτει η τελική αναπαράσταση του βίντεο με χρήση συντελεστών αυτοσυσχέτισης $CorD$.

Σε όλες τις αναπαραστάσεις χρησιμοποιήθηκαν $K = 256$ οπτικές λέξεις/συστάδες. Οι διαστάσεις των αναπαραστάσεων για αυτό τον αριθμό οπτικών λέξεων είναι: 256 στοιχεία για το BoVW, $256 \cdot D$ στοιχεία για την αναπαράσταση $VLAD$, όπου D είναι η διάσταση του περιγραφητή, $2 \cdot 256 \cdot D'$ στοιχεία για το διάνυσμα Fisher, όπου D' είναι η διάσταση του περιγραφή μετά τη μείωση διάστασης μέσω PCA και $256 \cdot 256$ στοιχεία για τις αναπαραστάσεις CorD, SD-hard assignment, SD-soft assignment. Οι διαστάσεις των περιγραφητών μετά την PCA μειώθηκαν στο μισό.

Για την ταξινόμηση των βίντεο στις κατηγορίες δράσεων, χρησιμοποιούμε SVM ταξινομητές (υλοποίηση LIBSVM [74]) με τη χρήση μη γραμμικού ταξινομητή στην περίπτωση των BoVW, CorD, SD-hard assignment και SD-soft assignment αναπαραστάσεων, και πιο συγκεκριμένα χρησιμοποιείται πυρήνας τομής ιστογραμμάτων (Histogram Intersection Kernel). Για τις αναπαραστάσεις VLAD και Fisher Vector χρησιμοποιούμε γραμμικούς ταξινομητές. Για την ταξινόμηση σε πολλαπλές κλάσεις ακολουθούμε την προσέγγιση "one-against-all" (Ένας-Εναντίον-Όλων), ταξινομώντας το βίντεο στην κατηγορία με το μεγαλύτερο score.

[1] `www.vlfeat.org`

[2] `http://yael.gforge.inria.fr/`

6.5.3 Πειραματικά αποτελέσματα

Στον πίνακα 6.1 παρατίθενται τα αποτελέσματα των πειραμάτων μας. Στις δύο πρώτες γραμμές βλέπουμε την ακρίβεια αναγνώρισης δράσεων με χρήση των δύο αναπαραστάσεων αναφοράς, δηλαδή την αναπαράσταση BoVW, η οποία δεν περιέχει πληροφορία σχετικά με την αλληλεπίδραση μεταξύ των clusters, και την αναπαράσταση CorD, η οποία χρησιμοποιεί την απλή, μη κατευθυνόμενη μετρική της αυτοσυσχέτισης μεταξύ των τροχιών. Στις γραμμές 3 και 4 παρατίθενται οι επιδόσεις των δύο παραλλαγών της μεθόδου μας. Όπως παρατηρούμε, και οι δύο παραλλαγές οδηγούν για όλους τους περιγραφητές σε αύξηση της ακρίβειας αναγνώρισης σε σχέση με το BoVW της τάξης του $\sim 2\%$ για την αναπαράσταση SD - hard assignment.

Η χρήση GMM clustering και η αντίστοιχη προσαρμογή της αναπαράστασης ομοιότητας οδηγούν σε περαιτέρω βελτίωση, σχεδόν για όλους τους περιγραφητές, εκτός των Trajectory και HOF, με τις μεγαλύτερες αυξήσεις να παρατηρούνται στους περιγραφητές HOG και MBHy (από 82.5% σε 91.08% και από 89.57% σε 94.09% αντίστοιχα). Επομένως μπορούμε να συνάγουμε το συμπέρασμα ότι η αξιοποίηση της πληροφορίας σχετικά με την ομοιότητα μεταξύ clusters βελτιώνει την ικανότητα διαχωρισμού των κλάσεων και ότι η χρήση GMM clustering οδηγεί σε μια καλύτερη μοντελοποίηση του χώρου χαρακτηριστικών, με μικρότερη απώλεια πληροφορίας, και συνεπώς σε αισθητά βελτιωμένες επιδόσεις. Επίσης, παρατηρούμε ότι η μέθοδος μας με χρήση της συμμετρικής Kullback-Leibler απόκλισης μεταξύ των κατανομών των διακυμάνσεων στις κύριες συνιστώσες κάθε cluster, οδηγεί σε μια μικρή βελτίωση σε σχέση με την αναπαράσταση CorD. Αυτό επιβεβαιώνει την υπόθεσή μας ότι η μετρική μας ποσοτικοποιεί την αλληλεπίδραση μεταξύ των clusters και ότι η πληροφορία της κατεύθυνσης της αλληλεπίδρασης οδηγεί σε υψηλότερες ακρίβειες αναγνώρισης, μιας και περιγράφει καλύτερα τις ανθρώπινες δράσεις.

Τέλος, συγκρίνουμε τη μέθοδό μας με τις state-of-the-art μεθόδους αναπαράστασης VLAD και Fisher Vector. Στον πίνακα έχουν σημειωθεί με έντονη γραμματοσειρά τα δύο καλύτερα αποτελέσματα αναγνώρισης δράσεων στη βάση KTH για κάθε περιγραφητή. Όπως φαίνεται, η μέθοδός μας (άλλοτε με τη χρήση αυστήρης ανάθεσης και άλλοτε με τη χρήση χαλαρής ανάθεσης) ανήκει στις δύο καλύτερες μεθόδους του συγκεκριμένου πειράματος για όλους τους περιγραφητές εκτός από τον HOF. Μάλιστα,με τη χρήση της αναπαράστασης SD - soft assignment οδηγούμαστε σε αποτελέσματα τα οποία είναι ανώτερα για κάποιους περιγραφητές σε σχέση με την Fisher Vector προσέγγιση. Λόγω της συμπληρωματικής φύσης της μεθόδου μας σε σχέση με τις μεθόδους BoVW, VLAD και Fisher vector, η σύμμειξη τους στο μέλλον αναμένεται να οδηγήσει σε ακόμα καλύτερα αποτελέσματα.

	Trajectory	HOG	HOF	MBHx	MBHy
BoVW	88.88	82.50	90.03	91.31	89.57
CorD	**90.27**	84.01	91.08	92.47	89.46
SD - hard assignment	**90.50**	84.47	91.89	92.70	90.03
SD - soft assignment	87.72	**91.08**	89.46	**93.86**	**94.09**
VLAD	88.06	87.83	**92.70**	93.40	**94.55**
Fisher Vector	88.53	**88.41**	**92.47**	**93.97**	93.63

Πίνακας 6.1: Ακρίβεια αναγνώρισης ανθρώπινων δράσεων διάφορων μεθόδων αναπαράστασης στη βάση δεδομένων KTH. Οι δύο παραλλαγές της μεθόδου μας (Similarity Descriptor (SD) - hard assignment και SD - soft assignment) ξεπερνούν τη μέθοδο BoVW, ενώ ταυτόχρονα οδηγούν σε επιδόσεις συγκρίσιμες με αυτές κάποιων από τις state-of-the-art μεθόδους της βιβλιογραφίας (VLAD, Fisher Vector).

Κεφάλαιο 7

Αναπαράσταση βίντεο με χρονική ακολουθία οπτικών λέξεων

Στα προηγούμενα κεφάλαια εξετάσαμε αναπαραστάσεις βίντεο που περιγράφουν τη στατιστική κατανομή των χαρακτηριστικών στις ομάδες (clusters) που έχουν υπολογιστεί με κάποιον αλγόριθμο συσταδοποίησης (K-means, GMM clustering), όπως το Bag-of-Visual Words, το VLAD vector καθώς και αναπαραστάσεις βίντεο που λαμβάνουν υπόψη τις σχέσεις συσχέτισης, αιτιατότητας ή ομοιότητας των χώρων χαρακτηριστικών μεταξύ των συστάδων. Ωστόσο, ας αναλογιστούμε το γενικό ορισμό της έννοιας *δράση* που αναφέραμε στο Κεφάλαιο 1: μια δράση είναι μια αλληλουχία βασικών κινήσεων (motion primitives) που εκτελείται από ένα άτομο. Από αυτόν τον ορισμό γίνεται σαφές ότι μία δράση είναι ένα δυναμικό φαινόμενο που εξελίσσεται στο χρόνο και ότι εκτός από το είδος των κινήσεων, η σειρά εκτέλεσής τους αποτελεί διαχωριστικό στοιχείο μεταξύ διαφορετικών δράσεων.

Η πλειονότητα των μεθόδων αναπάραστασης βίντεο δράσεων που χρησιμοποιούνται ως είσοδοι σε SVM ταξινομητές δεν κωδικοποιούν τη δυναμική εξέλιξη του φαινομένου στο χρόνο. Η χρονική πληροφορία εμπεριέχεται έμμεσα στα χαρακτηριστικά που έχουν εξαχθεί, π.χ. στους περιγραφητές των πυκνών τροχιών, αλλά χάνεται κατά την αναπαράσταση του βίντεο σε διανύσματα που δεν έχουν κάποια *διάταξη* (orderless representations). Παραδείγματος χάριν, το ιστόγραμμα BoVW αποθηκεύει την πληροφορία της συχνότητας εμφάνισης των οπτικών λέξεων στο βίντεο, χωρίς να διατηρεί κάποια πληροφορία σχετικά με το πότε εμφανίστηκαν. Φυσικά, η απώλεια της πληροφορίας της χρονικής (αλλά και χωρικής) διάταξης των χαρακτηριστικών/οπτικών λέξεων είναι ένα γνωστό πρόβλημα στο ερευνητικό πεδίο της Αναγνώρισης Δράσεων, για το οποίο έχουν προταθεί διάφορες μέθοδοι διόρθωσης μέσω νέων

αναπαραστάσεων ή με τη χρήση δυναμικών μοντέλων. Στην παρούσα εργασία προτείνουμε μια νέα μέθοδο αναπαράστασης βίντεο, η οποία αξιοποιεί τη χρονική αλληλουχία των οπτικών λέξεων. Κατασκευάζουμε υπο-ακολουθίες από συχνά εμφανιζόμενες οπτικές λέξεις, οι οποίες εν τέλει συνενώνονται σε μια τελική ακολουθία οπτικών λέξεων που αναπαριστά μια δράση (*Sequence of Dominant Visual Words*). Επίσης προτείνουμε και μια μετρική της ομοιότητας μεταξύ των διανυσμάτων που προκύπτουν με αυτή τη μέθοδο αναπαράστασης, βασισμένη σε αλγόριθμο τοπικής στοίχισης των ακολουθιών, έτσι ώστε να γίνει δυνατή η αξιοποίησή της από μηχανές διανυσματικής υποστήριξης. Τέλος, εξετάζουμε τη σύμμειξη της νέας μεθόδου αναπαράστασης με τη μέθοδο BoVW. Τα αποτελέσματα των πειραμάτων μας στην ευρέως χρησιμοποιούμενη βάση δεδομένων δράσεων KTH και την απαιτητική βάση μεγάλης κλίμακας HMDB51 [75] δείχνουν ότι η προτεινόμενη αναπαράσταση είναι συμπληρωματική της αναπαράστασης BoVW και ο συνδυασμός τους οδηγεί σε αποτελέσματα συγκρίσιμα με αυτά των καλύτερων μεθόδων της τρέχουσας διεθνούς βιβλιογραφίας, πολλές εκ των οποίων χρησιμοποιούν πιο πολύπλοκα μοντέλα.

Στις ενότητες που ακολουθούν θα επιχειρήσουμε να σκιαγραφήσουμε τις σημαντικότερες από τις μεθόδους αναπαράστασης βίντεο της διεθνούς βιβλιογραφίας που προσπαθούν να λάβουν υπόψη τη χρονική πληροφορία και στη συνέχεια θα παρουσιάσουμε ενδελεχώς την προτεινόμενη μέθοδο και τα αποτελέσματα της αξιολόγησής της σε δύο βάσεις δεδομένων ανθρώπινων δράσεων.

7.1 Σχετική Βιβλιογραφία

Τα τελευταία χρόνια έχουν γίνει αρκετές προσπάθειες για τον εμπλουτισμό του BoVW ιστογράμματος με χρονική πληροφορία και εν γένει την αναπαράσταση των βίντεο με τρόπο που κωδικοποιεί την πληροφορία της χρονικής διάταξης. Μια πλούσια επισκόπηση των μεθόδων αναπαράστασης βίντεο με ενσωμάτωση χρονικής πληροφορίας μπορεί να βρει κανείς στην εργασία των Ramana et al. [76]. Παρακάτω, θα αναφερθούμε ενδεικτικά σε μερικές από αυτές τις μεθόδους.

Ο πιο απλός τρόπος ενσωμάτωσης πληροφορίας σχετικά με τη χωρική και χρονική δομή των βίντεο είναι η χρήση χωροχρονικών πυραμίδων [22], οι οποίες αποσυνθέτουν τον όγκο του βίντεο σε χωροχρονικές υποδιαιρέσεις και υπολογίζουν ένα διαφορετικό διάνυσμα BoVW για κάθε υποδιαίρεση. Η τελική αναπαράσταση του βίντεο προκύπτει από τη συνένωση αυτών των διανυσμάτων. Οι Cheng et al. [77] μοντελοποιούν τις χρονικές σχέσεις μεταξύ συστάδων χαρακτηριστικών χρησιμοποιώντας ένα υποσύνολο των χρονικών σχέσεων του Allen. Οι Augusti et al. [78] ενσωματώνουν έμμεσα χρονικές

σχέσεις μέσα στο BoVW μοντέλο, κωδικοποιώντας τις συνεμφανίσεις οπτικών λέξεων κατά τη διάρκεια του βίντεο για διαφορετικά χρονικά διαστήματα. Οι Savarese et al. [79] χρησιμοποιούν correlograms για να μοντελοποιήσουν τις τοπικές χωροχρονικές σχέσεις ανάμεσα σε ζεύγη οπτικών λέξεων, υπολογίζοντας ιστογράμματα συνεμφανιζόμενων οπτικών λέξεων σε τοπικές γειτονιές. Άλλες μέθοδοι προσπαθούν να χρησιμοποιήσουν ακολουθιακά μοντέλα έτσι ώστε να αναπαραστήσουν το δυναμικό φαινόμενο της δράσης. Εμπνευσμένοι από τις εξελίξεις στην αναγνώριση ομιλίας, αρκετοί ερευνητές έχουν χρησιμοποιήσει κρυφά Μαρκοβιανά μοντέλα [80]–[82], αλλά προς το παρόν δεν έχουν επιτύχει state-of-the-art απόδοση σε απαιτητικές βάσεις.

Λόγω της μεγάλης μεταβλητότητας των μοτίβων κίνησης στις ανθρώπινες δράσεις, τα ακολουθιακά μοντέλα δεν μπορούν να τις μοντελοποιήσουν ικανοποιητικά και χρειάζονται μεγάλο πλήθος στιγμιοτύπων εκπαίδευσης. Γι'αυτό πολλές ερευνητικές ομάδες έχουν προσπαθήσει να μοντελοποιήσουν τις δράσεις με χρονικές ακολουθίες. Οι Hatun και Duygulu [83] κατασκευάζουν το οπικό λεξικό από χαρακτηριστικά πόζας και αναπαριστούν το βίντεο ως μια χρονικά διατεταγμένη ακολουθία οπτικών λέξεων πόζας. Για τον υπολογισμό της ομοιότητας μεταξύ των ακολουθιών χρησιμοποιούν έναν αλγόριθμο στοίχισης συμβολοσειρών. Οι Lan et al. το 2014 [84] πειραματίστηκαν με διάφορες παραλλαγές των πυραμίδων, έτσι ώστε να λάβουν τη χωρική και χρονική δομή κατά τη διάρκεια του βίντεο, και με μια παραλλαγή των περιγραφητών, η οποία λαμβάνει υπόψη τη θέση των χαρακτηριστικών στο βίντεο. Πιο συγκεκριμένα, η χρονική πληροφορία αξιοποιείται επεκτείνοντας τα διανύσματα των χαρακτηριστικών με ένα timestamp που αφορά τη σχετική χρονική θέση κάθε χαρακτηριστικού (π.χ. τροχιάς) στο βίντεο. Στη μέθοδο που προτείνουμε εμπλουτίζουμε με τον ίδιο τρόπο τους περιγραφητές, αλλά σε αντίθεση με τους Lan et al. δεν ομαδοποιούμε απλώς τα εμπλουτισμένα διανύσματα, αλλά χρησιμοποιούμε τα timestamps για την άμεση χρονική διάταξη των οπτικών λέξεων σε ακολουθίες.

Η αναπαράσταση βίντεο που παρουσιάζει τις περισσότερες ομοιότητες με τη δική μας είναι η αναπαράσταση που πρότειναν οι Glaser et al. [85]. Οι συγγραφείς χρησιμοποιούν τα ίδια χρονικά εμπλουτισμένα διανύσματα για να σχηματίσουν ακολουθίες οπτικών λέξεων, οι οποίες έχουν προέλθει από clustering των εμπλουτισμένων περιγραφητών. Στη συνέχεια, ομαδοποιούν γειτονικές εμφανίσεις μιας οπτικής λέξης σε ένα "action part". Η απόσταση μεταξύ των ακολουθιών υπολογίζεται ως η μέση απόσταση μεταξύ των ζευγών "action parts", η οποία μπορεί να αφορά είτε τη χ^2 απόσταση μεταξύ ιστογραμμάτων συχνότητας εμφάνισης οπτικών λέξεων σε μικρές χρονικές γειτονιές πριν και μετά το part ή την "Damerau - Levenshtein" απόσταση των ακολουθιών οπτικών λέξεων σε μικρή χρονική γειτονιά γύρω από το action part. Από αυτή τη συνοπτική περιγραφή της μεθόδου, προκύπτουν οι κύριες διαφορές

της με τη μέθοδο αναπαράστασης που προτείνουμε. Η πρώτη κύρια διαφορά είναι ότι οι Glaser et al. κατασκευάζουν την ακολουθία όλων των οπτικών λέξεων που εμφανίζονται σε ένα βίντεο, ενώ όπως θα αναλύσουμε παρακάτω, η πληροφορία αυτή είναι πολύ θορυβώδης, καθώς οι περισσότερες οπτικές λέξεις εμφανίζονται με αμελητέα συχνότητα και μπορεί να αντιστοιχούν σε θορυβώδη χαρακτηριστικά. Επίσης, ο τρόπος υπολογισμού αποστάσεων που χρησιμοποιούν λαμβάνει υπόψη τη μέση τιμή των αποστάσεων μεταξύ των action parts, ενώ εμείς προτείνουμε τη χρήση ενός αλγορίθμου τοπικής στοίχισης χρονικών ακολουθιών, ο οποίος έχει τη δυνατότητα να εντοπίζει την περιοχή μέγιστης ομοιότητας ανάμεσα σε ακολουθίες, συγκρίνοντας τμήματα όλων των πιθανών μηκών. Επιπρόσθετα, σε αντίθεση με τους Glaser et al. βασίζουμε την αναπαράστασή μας στα state-of-the-art χαρακτηριστικά των πυκνών τροχιών και όχι σε χωροχρονικά σημεία ενδιαφέροντος (STIP), ενώ ταυτόχρονα αξιολογούμε τη μέθοδο μας σε απαιτητικές δημοφιλείς βάσεις ανθρώπινων δράσεων.

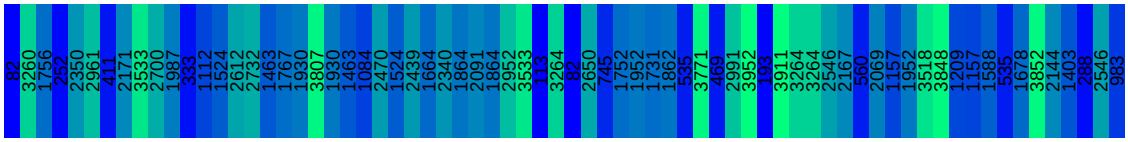
7.2 Επισκόπηση του συστήματος

Η μέθοδος που προτείνουμε αναπαριστά κάθε δράση με μία ακολουθία συχνά εμφανιζόμενων οπτικών λέξεων. Η πληροφορία αυτή είναι διαισθητικά συμπληρωματική ως προς την πληροφορία της αναπαράστασης BoVW, που καταγράφει απλώς τη συχνότητα εμφάνισης των λέξεων. Η αναπαράσταση BoVW εκμηδενίζει οποιαδήποτε πληροφορία σχετίζεται με τη σειρά εμφάνισης των οπτικών λέξεων και αυτό μειώνει την ικανότητα του συστήματος αναγνώρισης δράσεων να διαχωρίσει κατηγορίες δράσεων που μπορεί να παρουσιάζουν παραπλήσια στατιστική κατανομή των χαρακτηριστικών σε οπτικές λέξεις αλλά διαφορετική ακολουθία αυτών των χαρακτηριστικών. Για παράδειγμα, οι δράσεις *Stand* (σηκώνομαι) και *Sit* (κάθομαι) είναι σε μεγάλο βαθμό χρονικά συμμετρικές, δηλαδή τα άτομα εκτελούν περίπου τις ίδιες κινήσεις κατά τη διάρκεια αυτών των δράσεων αλλά με αντίστροφη χρονική σειρά. Έτσι περιμένουμε οι BoVW αναπαραστάσεις αυτών των δύο δράσεων, στην ιδανική περίπτωση που είναι τελείως συμμετρικές, να είναι παρόμοιες. Αντιθέτως, αν λάβουμε υπόψη τη χρονική αλληλουχία των οπτικών λέξεων τότε είμαστε σε θέση να διαχωρίσουμε αυτές τις δύο κατηγορίες δράσεων.

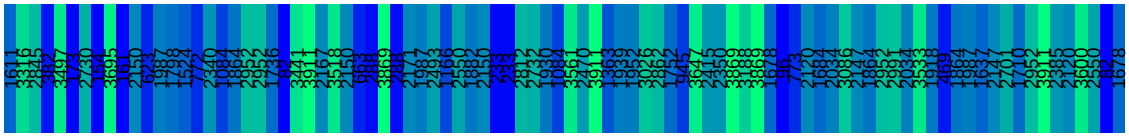
Πιο συγκεκριμένα, η αναπαράσταση SoDVW ξεκινάει με το διαχωρισμό του βίντεο που περιέχει μια δράση σε τμήματα με τη βοήθεια ενός παραθύρου. Σε κάθε τμήμα βρίσκουμε το σύνολο των οπτικών λέξεων που έχουν μη μηδενική συχνότητα εμφάνισης. Από αυτό το σύνολο διατηρούμε μόνο τις πιο συχνά εμφανιζόμενες λέξεις, τις οποίες διατάσσουμε στη συνέχεια χρονικά. Με αυτό τον τρόπο παράγουμε από όλα τα τμήματα του βίντεο χρονικές

(α′) Δράση *Stand*.

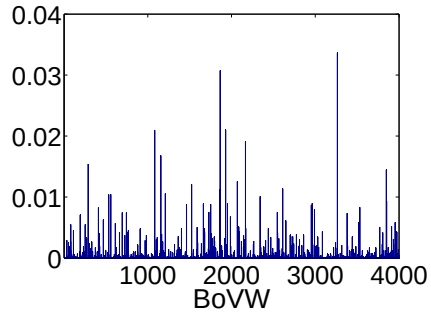
(β′) *Sit*.



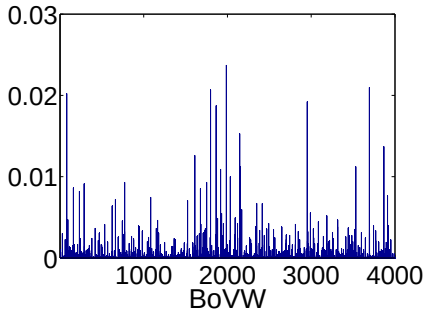
(γ′) Ακολουθία οπτικών λέξεων που αναπαριστά ένα βίντεο της δράσης *Stand*.



(δ′) Ακολουθία οπτικών λέξεων που αναπαριστά ένα βίντεο της δράσης *Sit*.



(ε′) Ιστόγραμμα BoVW που αναπαριστά ένα βίντεο της δράσης *Stand*.



(στ′) Ιστόγραμμα BoVW που αναπαριστά ένα βίντεο της δράσης *Sit*.

Σχήμα 7.1: Ενδεικτικά frames βίντεο από τη βάση HMDB51 [75] για τις δράσεις (α′) *Stand* και (β′) *Sit*. Το προτεινόμενο σύστημα αναγνώρισης συνδυάζει την πληροφορία της αλληλουχίας των κυρίαρχων οπτικών λέξεων (SoDVW) ((γ′),(δ′)) καθώς και την πληροφορία της συχνότητας εμφάνισης όλων των οπτικών λέξεων (BoVW) ((ε′),(στ′)) για να βελτιώσει την ακρίβεια αναγνώρισης των δράσεων. Η αναπαράσταση SoDVW ενσωματώνει πλούσια χρονική πληροφορία, σε αντίθεση με την αναπαράσταση BoVW που την αγνοεί.

υπο-ακολουθίες συχνά εμφανιζόμενων οπτικών λέξεων. Η συνένωση αυτών των υπο-ακολουθιών αποτελεί την τελική αναπαράσταση της δράσης. Αξίζει να τονιστεί σε αυτό το σημείο πώς η διάταξη των οπτικών λέξεων γίνεται σε δύο επίπεδα. Αρχικά, διατάσσονται χρονικά ανάλογα με τη εμφάνισή τους στο κάθε παράθυρο (τοπική διάταξη) και στη συνέχεια διατάσσονται μεταξύ τους βάσει της χρονικής αλληλουχίας των παραθύρων (ολική διάταξη). Μετά από αυτή τη διαδικασία κάθε βίντεο αναπαρίσταται από μια χρονική ακολουθία "συμβόλων" (οπτικών λέξεων). Για να αποτιμήσουμε την ομοιότητα μεταξύ δύο βίντεο δράσεων, χρησιμοποιούμε έναν αλγόριθμο τοπικής στοίχισης ακολουθιών κατά ζεύγη (pairwise local alignment) και ενσωματώνουμε σε αυτόν την πληροφορία σχετικά με τις σχέσεις ομοιότητας των οπτικών λέξεων. Τέλος, συνδυάζουμε τις δύο ροές πληροφορίας, BoVW και SoDVW, με γραμμικό συνδυασμό των συναρτήσεων πυρήνα στο επίπεδο του SVM ταξινομητή. Το συνολικό σύστημα αξιολογείται στις βάσεις δράσεων KTH [14] και HMDB51 [75], οδηγώντας σε συνεπείς βελτιώσεις της επίδοσης της αναπαράστασης BoVW έως και 5% .

7.3 Χρονικές ακολουθίες οπτικών λέξεων

Το σύστημα μας βασίστηκε στην αναπάραστση χαρακτηριστικών πυκνών τροχιών, χωρίς όμως να υπάρχει κάποιος περιορισμός στη φύση των χαρακτηριστικών που μπορούν να χρησιμοποιηθούν. Όπως απεικονίζεται στο Σχήμα 7.2 η αναπαράσταση δράσεων που προτείνουμε υπολογίζεται από τα εξαγμένα χαρακτηριστικά ακολουθώντας τα εξής βήματα:

1. Χωρίζουμε το βίντεο δράσης σε τμήματα με τη χρήση ενός χρονικού παραθύρου.

2. Για κάθε τμήμα του βίντεο δράσης αναθέτουμε τροχιές στην κοντινότερή τους οπτική λέξη/κεντροειδές βάσει των Ευκλείδειων αποστάσεων μεταξύ του εκάστοτε περιγραφητή της τροχιάς και των οπτικών λέξεων του λεξικού που έχει παραχθεί για αυτόν τον περιγραφητή.

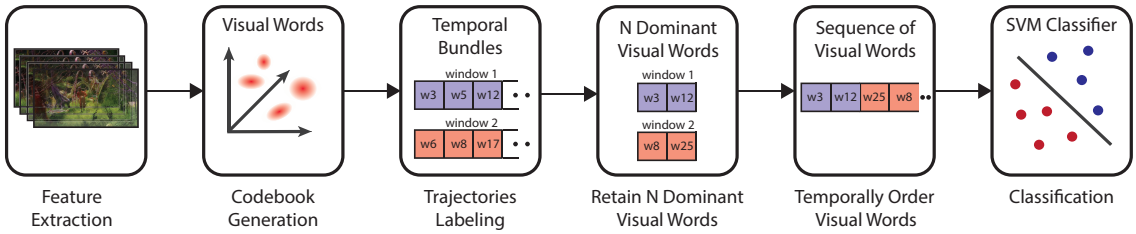
3. Συλλέγουμε τις οπτικές λέξεις που εμφανίζονται σε κάθε τμήμα του βίντεο σε σύνολα (*temporal bundles*).

4. Από κάθε σύνολο οπτικών λέξεων κρατάμε τις πιο κυρίαρχες λέξεις βασισμένοι στη συχνότητα εμφάνισής τους.

5. Έμμεσα εισάγουμε λεπτομερή χρονική πληροφορία, διατάσσοντας τις οπτικές λέξεις που κρατήσαμε σε κάθε σύνολο σε αύξουσα σειρά ως προς την μέση σχετική χρονική θέση των τροχιών που έχουν ανατεθεί

σε αυτές. Έτσι τα σύνολα μετατρέπονται σε χρονικές υπο-ακολουθίες (*temporal subsequences*).

6. Συνενώνουμε τις χρονικές υπο-ακολουθίες σε μία τελική χρονική ακολουθία κυρίαρχων οπτικών λέξεων (*Sequence of Dominant Visual Words - SoDVW*), η οποία αναπαριστά τη δράση.



Σχήμα 7.2: Μπλοκ διάγραμμα που απεικονίζει τα βήματα υπολογισμού της αναπαράστασης SoDVW.

Πιο συγκεκριμένα, έστω $\{\mathbf{x}_n\}$ το σύνολο των τροχιών που έχουν εξαχθεί από ένα βίντεο που περιέχει ένα στιγμιότυπο μιας δράσης και $\mathcal{D}$ το λεξικό K οπτικών λέξεων $w_1, \ldots, w_K$ που έχει κατασκευαστεί από τυχαία δειγματοληπτημένα χαρακτηριστικά εκπαίδευσης (π.χ. Dense Trajectories - HOF χαρακτηριστικά). Επεξεργαζόμαστε κάθε βίντεο χρησιμοποιώντας χρονικά παράθυρα και αναθέτουμε τροχιές σε παράθυρα (τμήματα του βίντεο). Για να αποφασίσουμε ποιες τροχιές ανήκουν σε κάθε παράθυρο, εξετάζουμε την επικάλυψή τους με αυτό. Τροχιές με επικάλυψη τουλάχιστον $\lceil \frac{L}{2} \rceil$ σημείων με το παράθυρο ανατίθενται σε αυτό, όπου L το μήκος της τροχιάς.

Οι τροχιές που ανήκουν σε κάθε παράθυρο αντιστοιχίζονται με την κοντινότερή τους οπτική λέξη. Βάσει αυτών των αντιστοιχίσεων, κάθε παράθυρο αναπαρίσταται αρχικά ως ένα σύνολο οπτικών λέξεων που εμφανίζονται εντός αυτού του χρονικού παραθύρου. Για κάθε οπτική λέξη που ανήκει σε ένα τέτοιο σύνολο, κρατάμε την πληροφορία της μέσης σχετικής χρονικής θέσης $tloc_i(w_j)$ των τροχιών που έχουν ανατεθεί σε αυτή, όπου η χρονική θέση κάθε τροχιάς $\mathbf{x}_n$ ορίζεται ως ο αύξοντας αριθμός του frame στο οποίο τελειώνει η τροχιά. Επίσης, λαμβάνουμε υπόψη τις τοπικές, βραχέος χρόνου, συχνότητες εμφάνισης των οπτικών λέξεων, υπολογίζοντας την BoVW αναπαράσταση εντός κάθε χρονικού παραθύρου. Μετά από αυτή τη διαδικασία, κάθε παράθυρο $i = 1 \ldots T$ αναπαρίσταται με ένα σύνολο οπτικών λέξεων, οδηγώντας σε μια αναπαράσταση βίντεο ως μιας συλλογής από *temporal bundles*:

$$VWSet_i = \{w_j | f_i(w_j) \neq 0\}, \; i = 1, \ldots, T, \; j = 1, \ldots, K \qquad (7.1)$$

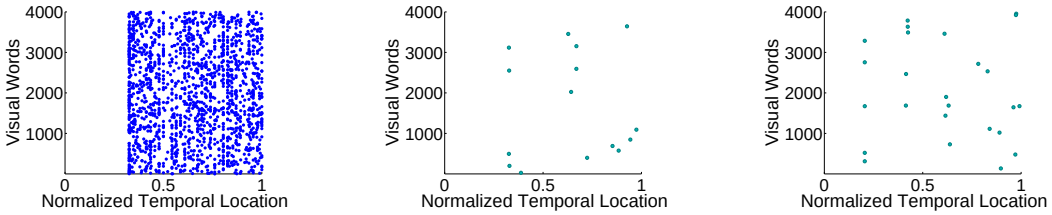
όπου $f_i(w_j)$ είναι το j-οστό στοιχείο του τοπικού ιστογράμματος BoVW, δηλαδή η συχνότητα εμφάνισης της j-οστής οπτικής λέξης εντός του παραθύρου:

$$BoVW_i = [f_i(w_1), f_i(w_2), \ldots, f_i(w_K)] \tag{7.2}$$

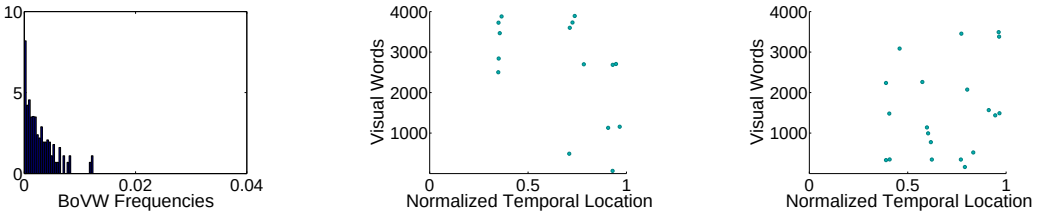
Το σχήμα 7.3α΄ απεικονίζει τις οπτικές λέξεις που εμφανίζονται σε ένα βίντεο της δράσης *Sit* κατά τη διάρκεια της εκτέλεσης της δράσης. Εύκολα παρατηρεί κανείς ότι οι τροχιές που ανήκουν σε ένα χρονικό τμήμα του βίντεο ανατίθενται σε ένα μεγάλο αριθμό οπτικών λέξεων. Αν χρησιμοποιήσουμε την BoVW αναπαράσταση ενός παραθύρου του παραπάνω βίντεο και καταγράψουμε την κατανομή των συχνοτήτων εμφάνισης των οπτικών λέξεων, βλέπουμε ότι οι περισσότερες οπτικές λέξεις έχουν αμελητέες συχνότητες εμφάνισης, με τη μεγαλύτερη συγκέντρωση να παρατηρείται σε συχνότητες ~ 0.0002 (Σχήμα 7.3δ΄). Μόνο λίγες οπτικές λέξεις κυριαρχούν στο παράθυρο, έχουν δηλαδή συχνότητες εμφάνισης μεγαλύτερες από π.χ. ~ 0.005. Οι Yang et al. [86] είχαν διαπιστώσει, στα πλαίσια του πειραματισμού τους με διαφορετικές μεθόδους μείωσης του μεγέθους του λεξικού των οπτικών λέξεων για την ταξινόμηση σκηνών (scene classification), ότι οι πιο συχνά εμφανιζόμενες οπτικές λέξεις εμπεριέχουν σημαντική πληροφορία για το διαχωρισμό των σκηνών σε αντίθεση με τις πιο σπάνιες οπτικές λέξεις. Οι οπτικές λέξεις με πολύ μικρές συχνότητες εμφάνισης μπορεί να οφείλονται στη θορυβώδη φύση των εικόνων ή στον αλγόριθμο συσταδοποίησης, ο οποίος ενδέχεται να παράγει μερικά πολύ μικρά clusters. Βασισμένοι στις παραπάνω παρατηρήσεις, διατηρούμε μόνο τις N_d πιο συχνά εμφανιζόμενες οπτικές λέξεις σε κάθε *temporal bundle*, έτσι ώστε να αυξήσουμε τη διαχωρισιμότητα ανάμεσα στις κλάσεις των δράσεων, να μειώσουμε το θόρυβο και να ενισχύσουμε την ευρωστία της αναπαράστασης. Η προκύπτουσα χρονική αλληλουχία των $N_d = 5$ κυρίαρχων οπτικών λέξεων του κάθε παραθύρου, η οποία απεικονίζεται στο σχήμα 7.3 για δύο βίντεο της κλάσης *Sit* και δύο βίντεο της κλάσης *Stand*, αποκαλύπτει καλώς σχηματισμένα μοτίβα που μπορούν να χρησιμοποιηθούν σαν μία πρόσθετη ροή πληροφορίας εμπλουτίζοντας τις τρέχουσες μεθόδους αναπαράστασης δράσεων.

Μετέπειτα, επεξεργαζόμαστε τα παραπάνω σύνολα ταξινομώντας τις οπτικές λέξεις που έχουν διατηρηθεί σε αύξουσα σειρά ως προς τη σχετική χρονική θέση τους $tloc_i(w_j)$. Με αυτό τον τρόπο, ενσωματώνουμε τη χρονική πληροφορία εντός των συνόλων, μετατρέποντάς τα σε χρονικές υπο-ακολουθίες οπτικών λέξεων:

$$\begin{aligned} &VWSeq_i = [w_{i1}, w_{i2}, \ldots, w_{iN_d}], \\ &tloc_i(w_{i1}) \leq tloc_i(w_{i2}) \leq \ldots \leq tloc_i(w_{iN_d}) \quad i = 1 \ldots T \end{aligned} \tag{7.3}$$



(α′) (β′) (γ′)



(δ′) (ε′) (στ′)

Σχήμα 7.3: (α′) Οπτικές λέξεις που εμφανίζονται κατά τη διάρκεια ενός στιγμιοτύπου της δράσης *Sit*. (δ′) Κατανομή συχνοτήτων εμφάνισης των οπτικών λέξεων που εμφανίζονται σε ένα χρονικό παράθυρο ενός στιγμιοτύπου της δράσης *Sit*. Έχει χρησιμοποιηθεί λογαριθμική κλίμακα στον κάθετο άξονα.(β′),(ε′) Κυρίαρχες (πιο συχνά εμφανιζόμενες) οπτικές λέξεις που εμφανίζονται κατά τη διάρκεια δύο στιγμιοτύπων της δράσης *Sit*. (γ′),(στ′) Κυρίαρχες οπτικές λέξεις που εμφανίζονται κατά τη διάρκεια δύο στιγμιοτύπων της δράσης *Stand*. Τα στιγμιότυπα των δράσεων έχουν ληφθεί από τη βάση HMDB51.

Με τη συνένωση αυτών των υπο-ακολουθιών οπτικών λέξεων καταλήγουμε σε μία χρονικά διατεταγμένη ακολουθία κυρίαρχων οπτικών λέξεων:

$$SoDVW = [VWSeq_1, VWSeq_2, \ldots, VWSeq_T] \tag{7.4}$$

7.3.1 Τοπική Στοίχιση Χρονικών Ακολουθιών Οπτικών Λέξεων

Για να είμαστε σε θέση να χρησιμοποιήσουμε την αναπαράσταση που ορίσαμε στην προηγούμενη ενότητα ως είσοδο ενός SVM ταξινομητή, πρέπει να οριστεί η απόσταση ανάμεσα σε δύο ακολουθίες κυρίαρχων οπτικών λέξεων. Για αυτό το σκοπό, προτείνουμε τον υπολογισμό της ομοιότητας δύο ακολουθιών οπτικών λέξεων με τη χρήση του αλγορίθμου τοπικής στοίχισης ακολουθιών (local sequence alignment) Smith-Waterman.

Η στοίχιση ακολουθιών μας επιτρέπει να συγκρίνουμε συμβολικές ακολουθίες και να εντοπίζουμε περιοχές ομοιότητας. Χρησιμοποιείται ευρέως στον κλάδο της Βιοπληροφορικής για τη σύγκριση ακολουθιών νουκλεοτιδιών ή αμινοξέων, με σκοπό την εξαγωγή χρήσιμης πληροφορίας για την λειτουργική, δομική ή/και εξελικτική σχέση των ακολουθιών των βιομακρομορίων των σύγχρονων οργανισμών. Κατά τη στοίχιση δύο ακολουθιών, κάθε σύμβολο της μίας ακολουθίας αντιστοιχίζεται είτε σε ένα σύμβολο της άλλης ακολουθίας είτε σε ένα κενό. Πιο συγκεκριμένα, θεωρούμε ότι οι δύο ακολουθίες έχουν κάποιο κοινό πρόγονο και προσπαθούμε να δούμε πώς αποκλίνουν από αυτόν μέσω αντικαταστάσεων (substitutions), ενθέσεων (insertions) και διαγραφών (deletions). Σκοπός των αλγορίθμων στοίχισης ακολουθιών κατά ζεύγη είναι η εύρεση της βέλτιστης στοίχισης των δύο ακολουθιών, η οποία διατηρεί τη σειρά των συμβόλων, εισάγει τα απαραίτητα κενά και μεγιστοποιεί ένα συνολικό score ομοιότητας.

Υπάρχουν δύο κατηγορίες αλγορίθμων στοίχισης ακολουθιών: οι αλγόριθμοι ολικής και οι αλγόριθμοι τοπικής στοίχισης. Οι αλγόριθμοι ολικής στοίχισης (global alignment) προσπαθούν να στοιχίσουν όσο το δυνατόν περισσότερα σύμβολα των δύο ακολουθίων σε όλο το μήκος τους μεγιστοποιώντας το συνολικό score, ακόμα και σε βάρος τμημάτων των ακολουθιών που έχουν προφανή ομοιότητα. Είναι λοιπόν κατάλληλοι για την εύρεση στοίχισης παραπλήσιων ακολουθιών παρόμοιου μήκους. Ο γνωστότερος αλγόριθμος ολικής στοίχισης είναι ο αλγόριθμος Needleman-Wunsch [87]. Αντιθέτως, οι αλγόριθμοι τοπικής στοίχισης προσπαθούν να εντοπίσουν περιοχές υψηλής ομοιότητας μεταξύ των δύο ακολουθιών μέσα σε μεγάλες ακολουθίες που μπορεί να διαφέρουν πολύ. Οι περιοχές που διαφέρουν σημαντικά μπορούν να αντικατασταθούν με κενά και να μην επηρεάσουν το score ομοιότητας των ακολουθιών. Έτσι μπορούν να ανακαλυφθούν περιοχές με όμοια ακολουθιακά

μοτίβα μέσα σε ακολουθίες διαφορετικού μήκους που μπορεί να αποκλίνουν σημαντικά μεταξύ τους. Ο πιο χαρακτηριστικός αλγόριθμος τοπικής στοίχισης είναι ο αλγόριθμος δυναμικού προγραμματισμού Smith-Waterman.

Ο αλγόριθμος Smith-Waterman [88] προτάθηκε από τους Temple F. Smith και Michael S. Waterman το 1981 για την εύρεση παρόμοιων μοριακών υποακολουθιών. Βρίσκει την περιοχή μέγιστης ομοιότητας ανάμεσα σε ακολουθίες, συγκρίνοντας τμήματα όλων των πιθανών μηκών και προσδιορίζει το βέλτιστο score ομοιότητας. Επομένως, είναι ο καταλληλότερος υπόψηφιος αλγόριθμος για την ανίχνευση ομοιοτήτων ανάμεσα στις ακολουθίες οπτικών λέξεων, οι οποίες πολύ συχνά αποκλίνουν σημαντικά μεταξύ τους λόγω των διακυμάνσεων στη διάρκεια και στον τρόπο εκτέλεσης των δράσεων, των αλλαγών της οπτικής γωνίας και των επικαλύψεων. Αυτό το μέτρο ομοιότητας ακολουθιών είναι επίσης εύρωστο στη μεταβλητότητα της θέσης της δράσης μέσα στο βίντεο. Παρόλο που ασχολούμαστε με το πρόβλημα της ταξινόμησης βίντεο που περιέχουν ένα στιγμιότυπο κάποιας δράσης, η έναρξη (το τέλος) της εκτέλεσης της δράσης συχνά δε συμπίπτει χρονικά με την έναρξη (το τέλος) του βίντεο και το χρονικό διάστημα που εκτελείται η κάθε πράξη διαφέρει από βίντεο σε βίντεο. Περιοχές των ακολουθιών με μεγάλη απόκλιση στην αρχή ή στο τέλος των ακολουθιών αγνοούνται και δεν επηρεάζουν το τελικό score ομοιότητας.

Δεδομένου ενός αλφαβήτου Σ, το οποίο στην περίπτωσή μας αποτελείται από τις K οπτικές λέξεις του λεξικού, και δύο ακολουθίες $A : w_{i1}w_{i2}\dots w_{im}$ και $B : w_{j1}w_{j2}\dots w_{jn}$, ο αλγόριθμος επιστρέφει το score ομοιότητας της τοπικής στοίχισής τους. Επίσης, επιστρέφει και τη βέλτιστη στοίχιση που οδηγεί σε αυτό το score ομοιότητας. Πρόσθετες παράμετροι του αλγορίθμου είναι η ποινή εισαγωγής κενών (gap penalty) p και ο πίνακας αντικαταστάσεων (substitution matrix) S, του οποίου το (i, j)-οστό στοιχείο υποδηλώνει την ομοιότητα ανάμεσα στα σύμβολα του αλφαβήτου w_i, w_j. Στα πλαίσια της μεθόδου μας για την αναπαράσταση βίντεο ανθρώπινων δράσεων, όπου το αλφάβητο αποτελείται από οπτικές λέξεις, προτείνουμε έναν πίνακα αντικαταστάσεων, ο οποίος βασίζεται στις συσχετίσεις ανάμεσα στις οπτικές λέξεις. Κατ'αναλογία με την ομοιότητα συμβολοσειρών, όπου η συμβολοσειρά "οπτική" μοιάζει περισσότερο με τη συμβολοσειρά "οπτυκή" αλλά λιγότερο με τη συμβολοσειρά "οπτφκή", επειδή το σύμβολο (χαρακτήρας) "ι" μοιάζει περισσότερο με το σύμβολο "υ" σε σχέση με το "φ", ορίζουμε τα στοιχεία του δικού μας πίνακα αντικαταστάσεων ως εξής:

$$S(w_i, w_j) = -2 \cdot \frac{d(w_i, w_j)}{\max\limits_{k,l=1\dots K} d(w_k, w_l)} + 1 \tag{7.5}$$

όπου $d(w_i, w_j)$ είναι η Ευκλείδεια απόσταση ανάμεσα σε δύο οπτικές λέξεις.

Όσο πιο όμοιες είναι δύο οπτικές λέξεις, τόσο μεγαλύτερο το κέρδος που έχουμε από την αντιστοίχισή τους, με την ομοιότητα μιας οπτικής λέξης με τον εαυτό της να ισούται με τη μονάδα. Αντιθέτως, η ομοιότητα των πιο "μακρινών" οπτικών λέξεων ισούται με -1.

Για την επίτευξη της στοίχισης των ακολουθιών και την ανάδειξη περιοχών υψηλής ομοιότητας συχνά χρειάζεται να εισαχθούν κενά (gaps). Τα κενά αυτά συνοδεύονται από μια ποινή κενού. Η ποινή κενού μπορεί να έχει σχέση με το πλήθος των συνεχόμενων κενών ή να διαφοροποιείται στην περίπτωση που εισάγεται το πρώτο κενό σε σχέση με τη επέκτασή του ή να είναι μια σταθερή τιμή. Όπως προαναφέραμε, στη συγκεκριμένη εφαρμογή χρησιμοποιούμε μια σταθερή τιμή p, η οποία οδηγεί σε πιο αποδοτικό υπολογισμό του score ομοιότητας. Προφανώς, η ποινή κενού επηρεάζει το τελικό score ομοιότητας και η τιμή της πρέπει να επιλεχθεί κατάλληλα, ανάλογα με τις τιμές του πίνακα αντικατάστασης. Πιο συγκεκριμένα, αν η ποινή κενού έχει πολύ υψηλή τιμή σε σχέση με τις τιμές του πίνακα αντικατάστασης, τότε μια στοίχιση δύο ακολουθιών χωρίς κενά (εισαγωγές/διαγραφές) συνεπάγεται μικρότερο κόστος σε σχέση με μια στοίχιση με κενά. Ωστόσο, όπως είπαμε, τα κενά ίσως συμβάλλουν στη στοίχιση πιο όμοιων περιοχών (υπο-δράσεων).

Βάσει αυτών των παραμέτρων και εισόδων, ο Smith-Waterman αλγόριθμος κατασκευάζει έναν πίνακα F, διαστάσεων $(m+1) \times (n+1)$, όπου $F(i,j)$ είναι το μεγαλύτερο score ομοιότητας δύο αποσπασμάτων των ακολουθιών που λήγουν στα σύμβολα $A(i)$ και $B(j)$, αντίστοιχα. Τα στοιχεία της πρώτης γραμμής και πρώτης στήλης του πίνακα είναι μηδενικά ($F(i,0) = F(0,j) = 0$). Τα υπόλοιπα στοιχεία του πίνακα F ορίζονται μέσω της ακόλουθης σχέσης:

$$F(i,j) = \max \begin{cases} 0 & \\ F(i-1,j-1) + S(A(i),B(j)), & (Association - Match/Mismatch) \\ F(i-1,j) - p, & (Deletion) \\ F(i,j-1) - p, & (Insertion) \end{cases} \tag{7.6}$$

Όπως παρατηρούμε, ο πίνακας F εξ'ορισμού δεν μπορεί να περιέχει αρνητικά στοιχεία. Η ομοιότητα ανάμεσα σε δύο ακολουθίες είναι το μέγιστο στοιχείο αυτού του πίνακα $f^\star$. Η τιμή αυτή κανονικοποιείται, έτσι ώστε το score ομοιότητας ανάμεσα σε δύο στιγμιότυπα της ίδιας ακολουθίας να είναι ίσο με 1.

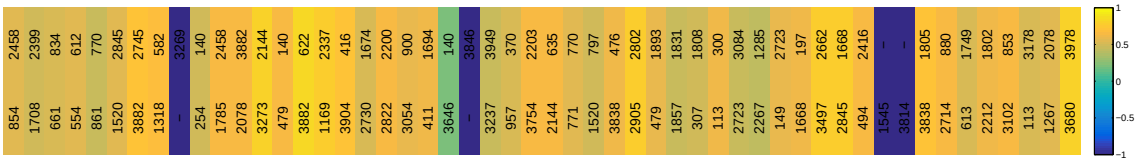
$$Similarity := \frac{f^\star}{\max(m,n)} \tag{7.7}$$

Άρα το score ομοιότητας ανάμεσα σε δύο χρονικές ακολουθίες οπτικών λέξεων είναι πάντα μια τιμή στο διάστημα $[0,1]$. Αυτό μας επιτρέπει να ορίσουμε

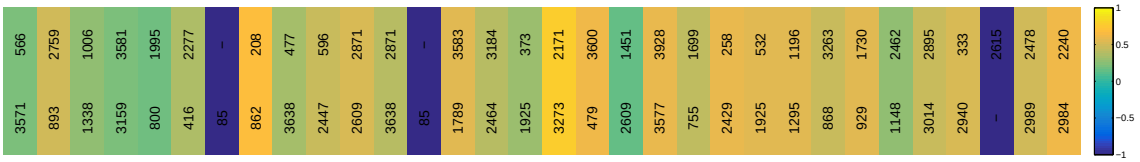
εύκολα την απόστασή τους:

$$D_{SW}(SoDVW_1, SoDVW_2) = 1 - Similarity(SoDVW_1, SoDVW_2) \quad (7.8)$$

Ένα μειονέκτημα του αλγορίθμου Smith-Waterman είναι η τετραγωνική πολυπλοκότητα του σε χρόνο και σε μνήμη. Εντούτοις, αυτός ο περιορισμός αναδεικνύει άλλο ένα όφελος της διατήρησης μόνο των N_d κυρίαρχων οπτικών λέξεων σε κάθε παράθυρο, εφόσον τόσο το N_d όσο και ο αριθμός των παραθύρων των βίντεο δράσεων είναι αρκετά μικρά.



(α′)



(β′)

Σχήμα 7.4

Σχήμα 7.5: Στοίχιση χρονικών ακολουθιών κυρίαρχων οπτικών λέξεων (α′) για ένα ζευγάρι ακολουθιών που ανήκουν στην ίδια κατηγορία δράσης (*Run*) και (β′) για ένα ζευγάρι ακολουθιών που ανήκουν σε διαφορετικές κατηγορίες δράσης (*Stand* και *Fall Floor*). Απεικονίζεται μόνο η περιοχή των ακολουθιών που έχει αντιστοιχιστεί με τον αλγόριθμο τοπικής στοίχισης Smith-Waterman, ενώ ενδεχόμενες ανόμοιες περιοχές στην αρχή και στο τέλος των ακολουθιών δεν απεικονίζονται και δεν επηρεάζουν το score ομοιότητας των ακολουθιών. Οι στήλες του πίνακα στοίχισης απεικονίζονται με χρώμα ανάλογο της ομοιότητας των οπτικών λέξεων που έχουν αντιστοιχιστεί, δηλαδή ανάλογο των τιμών του πίνακα αντικαταστάσεων. Όσο περισσότερο μοιάζουν δύο οπτικές λέξεις, τόσο πιο κοντά είναι η τιμή της ομοιότητάς τους στο 1. Το αντίστροφο συμβαίνει για ανόμοιες οπτικές λέξεις.

Ένα παράδειγμα στοιχισμένων ακολουθιών SoDVW απεικονίζεται στο Σχήμα 7.4. Αφού έχουμε υπολογίσει τον πίνακα F, η εύρεση μιας βέλτιστης στοίχισης γίνεται με απλή οπισθοδρόμηση, ξεκινώντας από το κελί του F με τη μεγαλύτερη τιμή και καταλήγοντας όταν συναντήσουμε ένα κελί με

τιμή 0, το οποίο αντιστοιχεί στην αρχή της στοίχισης. Η οπισθοδρόμηση γίνεται προς τα πάνω και αριστερά, ξεκινώντας από το κελί με τη μέγιστη τιμή και βλέποντας κάθε φορά από ποιο κελί προήλθε αυτή η τιμή. Αν προήλθε από το $F(i-1, j-1)$, τότε κινούμαστε διαγώνια και δεν εισάγουμε κενό (δηλαδή έχουμε την αντιστοίχηση δύο ίδιων (match) ή παραπλήσιων (mismatch) οπτικών λέξεων). Αν προήλθε από το $F(i-1, j)$ τότε προστίθεται ένα κενό στη δεύτερη ακολουθία (διαγραφή - deletion), ενώ αν προήλθε από το $F(i, j-1)$ τότε προστίθεται ένα κενό στην πρώτη ακολουθία (εισαγωγή - insertion). Η εύρεση μιας βέλτιστης στοίχισης μπορεί να γίνει για λόγους οπτικοποίησης, αλλά ο υπολογισμός της δεν είναι απαραίτητος στο πρόβλημά μας, εφόσον μας ενδιαφέρει μόνο ο υπολογισμός της απόστασης μεταξύ δύο ακολουθιών.

7.3.2 Σύμμειξη των αναπαραστάσεων SoDVW και BoVW

Για να μπορέσουμε να ταξινομήσουμε ένα βίντεο βασισμένοι στην αναπαράστασή του με μια ακολουθία οπτικών λέξεων SoDVW χρησιμοποιώντας μηχανές διανυσματικής υποστήριξης, ορίζουμε τον πυρήνα:

$$K(Seq_1, Seq_2) = e^{-\frac{D_{SW}(Seq_1, Seq_2)}{A}} \tag{7.9}$$

όπου A είναι η μέση απόσταση των αναπαραστάσεων των βίντεο του συνόλου εκπαίδευσης.

Επομένως, εφόσον οι αναπαραστάσεις BoVW και SoDVW είναι θεμελιωδώς διαφορετικές, με τη μεν να είναι ένα ιστόγραμμα και τη δε να είναι μια χρονική ακολουθία συμβόλων, έχουν διαφορετικές μετρικές ομοιότητας που αντιστοιχούν σε διαφορετικούς πυρήνες. Όπως αναφέραμε και στο Κεφάλαιο 4, ο συνδυασμός πολλαπλών ροών πληροφορίας που αντιστοιχούν σε διαφορετικούς πυρήνες μπορεί να επιτευχθεί μέσω του υπολογισμού ενός νέου πυρήνα ως το γραμμικό συνδυασμό των πυρήνων (LKC). Σε αυτή την εργασία, πειραματιζόμαστε με διαφορετικά διανύσματα βαρών $\boldsymbol{\theta} = [\theta_1, \theta_2]$ για το συνδυασμό του πυρήνα με την απόσταση Smith-Waterman (7.9)(K_1) και του πυρήνα χ^2 (K_2):

$$K = \theta_1 K_1 + \theta_2 K_2,\ \theta_1, \theta_2 \geq 0 \tag{7.10}$$

7.4 Πειράματα ταξινόμησης ανθρώπινων δράσεων

Σε αυτή την ενότητα αξιολογούμε τη μέθοδο αναγνώρισης δράσεων που προτείνουμε σε δύο μεγάλες βάσεις δεδομένων ανθρώπινων δράσεων: την

KTH, που χρησιμοποιήσαμε για τον πειραματισμό μας και στο Κεφάλαιο 6 και την HMDB. Αρχικά, παρουσιάζουμε τη βάση HMDB51 και τις λεπτομέρειες υλοποίησης της προσέγγισης μας και του βασικού συστήματος αναφοράς, το οποίο χρησιμοποιεί την αναπαράσταση BoVW. Ακολούθως, παρουσιάζουμε τα αποτελέσματα αναγνώρισης δράσεων σε αυτές τις δύο βάσεις, συγκρίνοντάς τα με το σύστημα αναφοράς. Επιπρόσθετα, ερευνούμε την επίδραση διαφορετικών παραμέτρων στην ακρίβεια αναγνώρισης, όπως ο αριθμός N_d των οπτικών λέξεων που διατηρούνται, η ποινή κενού του Smith-Waterman αλγορίθμου και το διάνυσμα βαρών για την σύμμειξη των πυρήνων με τη μέθοδο LKC.

7.4.1 Η Βάση Ανθρώπινων Δράσεων HMDB51

Η βάση δεδομένων ανθρώπινων δράσεων HMDB51 [75] είναι μια από τις πιο απαιτητικές σύγχρονες βάσεις δράσεων με 51 κατηγορίες δράσεων (Σχήμα 7.6). Κάθε κατηγορία έχει τουλάχιστον 101 στιγμιότυπα, ενώ η βάση περιέχει συνολικά 6766 ρεαλιστικά βίντεο τα οποία προέρχονται από ταινίες, δημόσιες βάσεις, όπως το αρχείο Prelinger, άλλα βίντεο διαθέσιμα στο internet καθώς και YouTube και Google videos. Φυσικά τα βίντεο από αυτές τις πηγές βίντεο ποικίλουν σε μέγεθος και frame rate. Για λόγους συνέπειας, όλα τα βίντεο κλιμακώθηκαν έτσι ώστε να έχουν ύψος 240 pixels και εύρος κατάλληλο έτσι ώστε να διατηρείται το aspect ratio του κάθε βίντεο. Ο ρυθμός δειγματοληψίας μετατράπηκε σε 30 frames ανά δευτερόλεπτο για όλα τα βίντεο.

Οι κατηγορίες των δράσεων της βάσης μπορούν να ομαδοποιηθούν σε 5 είδη δράσεων:

1. Γενικές δράσεις προσώπου (facial actions): *smile*, *laugh*, *chew*, *talk*

2. Δράσεις του προσώπου με χειρισμό αντικειμένου: *smoke*, *eat*, *drink*

3. Γενικές κινήσεις του σώματος: *cartwheel*, *clap hands*, *climb*, *climb stairs*, *dive*, *fall on the floor*, *backhand flip*, *handstand*, *jump*, *pull up*, *push up*, *run*, *sit down*, *sit up*, *somersault*, *stand up*, *turn*, *walk*, *wave*

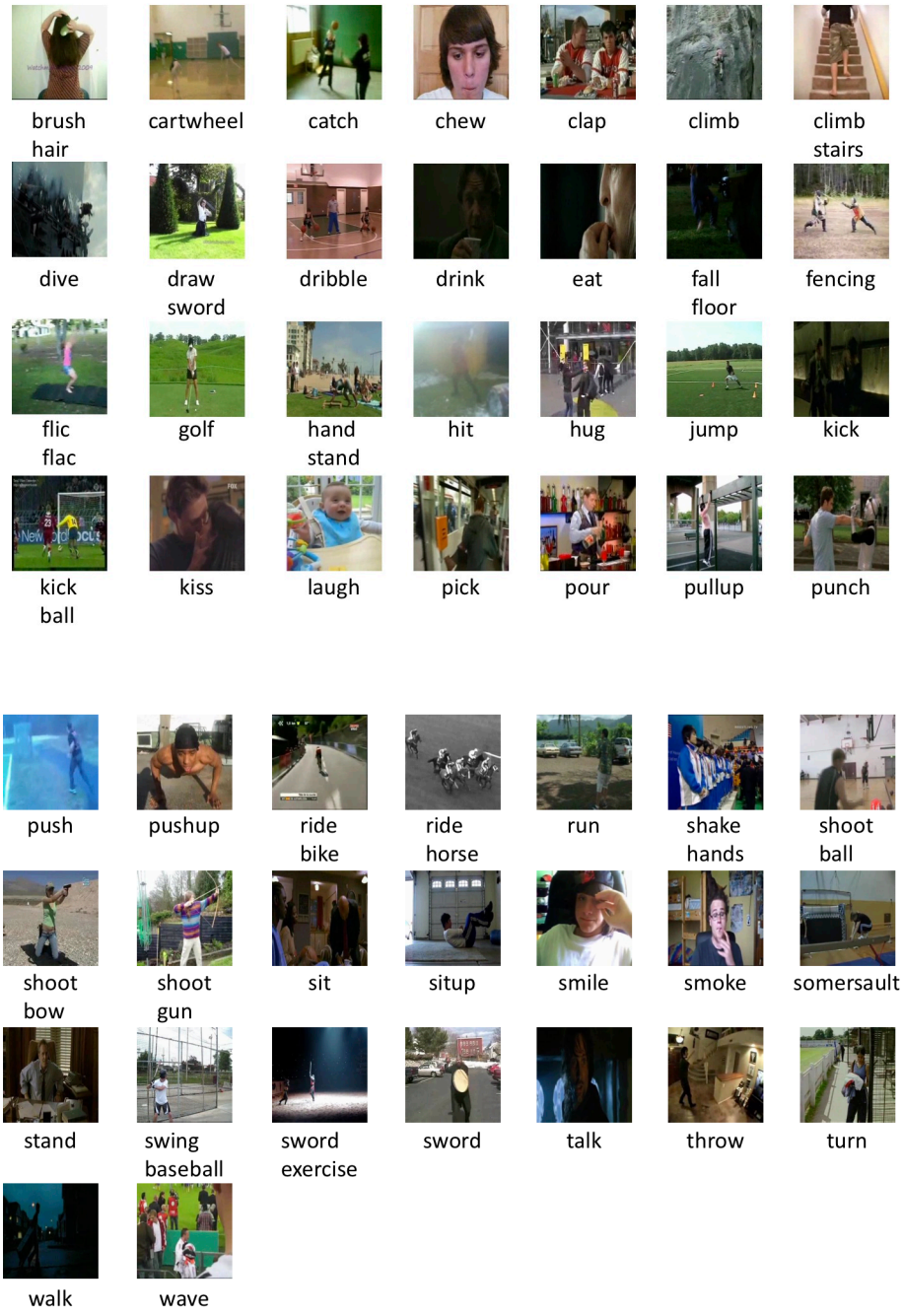
4. Κινήσεις του σώματος συνοδευόμενες από αλληλεπίδραση με αντικείμενο: *brush hair*,*catch*, *draw sword*, *dribble*, *golf*, *hit something*, *kick ball*, *pick*, *pour*, *push something*, *ride bike*, *ride horse*, *shoot ball*, *shoot bow*, *shoot gun*, *swing baseball bat*, *sword exercise*, *throw*

5. Κινήσεις του σώματος για την αλληλεπίδραση ατόμων: *fencing*, *hug*, *kick someone*, *kiss*, *punch*, *shake hands*, *sword fight*

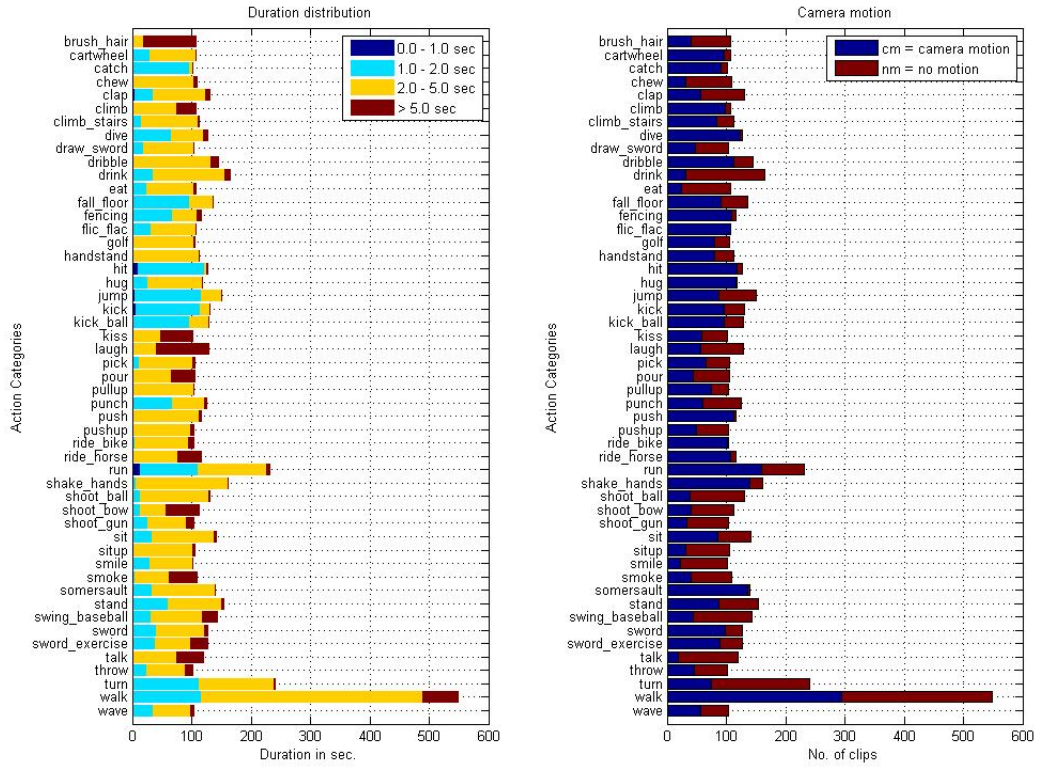
Εκτός από την ετικέτα της δράσης που περιέχεται σε κάθε βίντεο, η βάση HMDB δίνει πληροφορίες και για τα ορατά μέρη του σώματος/ επικαλύψεις,

για την κίνηση της κάμερας, για την οπτική γωνία της κάμερας και για τον αριθμό των ατόμων που συμμετέχουν σε μία δράση. Άρα εκτός του μεγάλου αριθμού κατηγοριών δράσεων, η βάση HMDB51 είναι ιδιαίτερα απαιτητική λόγω των ρεαλιστικών βίντεο, που έχουν ληφθεί σε μη ελεγχόμενο περιβάλλον και περιέχουν δράσεις πολλών διαφορετικών κατηγοριών, με πολλές διαφοροποιήσεις στον τρόπο εκτέλεσης της ίδιας δράσης, διάφορες οπτικές γωνίες και επικαλύψεις. Επίσης, είναι συχνή και η κίνηση της κάμερας κατά τη λήψη αυτών των ρεαλιστικών βίντεο. Η διάρκεια των περισσότερων βίντεο είναι μικρότερη των 5 δευτερολέπτων (Σχήμα 7.7α′).

Στα πειράματα που θα ακολουθήσουν χρησιμοποιούμε 3 splits, δηλαδή 3 διαφορετικά σύνολα εκπαίδευσης - αξιολόγησης, στα οποία περιέχονται 70 βίντεο εκπαίδευσης και 30 βίντεο αξιολόγησης από κάθε κατηγορία δράσης. Τα splits που χρησιμοποιούμε είναι τα ίδια με αυτά που χρησιμοποιήθηκαν στον αρχικό πειραματισμό πάνω στη βάση από τους Kuehne et al. [75] το 2011. Η τελική ακρίβεια ταξινόμησης δράσεων στη βάση HMDB προκύπτει από το μέσο όρο των ακριβειών ταξινόμησης στα 3 splits.



Σχήμα 7.6: Στιγμιότυπα των 51 κατηγοριών δράσεων της βάσης HMDB51 [75].



(α′) (β′)

Σχήμα 7.7: (α′)Διάρκεια των βίντεο των διαφορετικών κατηγοριών δράσεων και (β′) αναλογία βίντεο με και χωρίς κίνηση της κάμερας για τις διαφορετικές κατηγορίες δράσεων της βάσης HMDB51 [75].

7.4.2 Πειραματικό πλαίσιο

Στα πειράματα που ακολουθούν χρησιμοποιούμε τις ακόλουθες παραμέτρους για την αναπαράσταση SoDVW, εκτός και αν αναφέρονται ρητά άλλες τιμές για τις παραμέτρους: χρονικό παράθυρο 15 frames, $N_d = 10$ κυρίαρχες οπτικές λέξεις, $p = 0.1$ ποινή κενού του αλγορίθμου Smith-Waterman και $\boldsymbol{\theta} = [0.3, 0.7]$ διάνυσμα βαρών για το γραμμικό συνδυασμό των πυρήνων των αναπαραστάσεων BoVW και SoDVW.

Όσον αφορά τα χαρακτηριστικά που χρησιμοποιούμε, στη βάση KTH εξάγουμε πυκνές τροχιές (Dense Trajectories) χρησιμοποιώντας τον κώδικα που παρέχουν οι Wang et al. [38] με τις default παραμέτρους, όπως αυτές αναλύθηκαν στα προηγούμενα κεφάλαια. Για την απαιτητική βάση δε-

δομένων HMDB51, εξάγουμε βελτιωμένες πυκνές τροχιές (improved Dense Trajectories), χρησιμοποιώντας τον κώδικα και τα ορθογώνια πλαίσια, τα οποία περιβάλλουν τους ανθρώπους στα βίντεο, που παρέχουν οι συγγραφείς [1]. Ο λόγος που προτιμήθηκε αυτός ο ανιχνευτής χαρακτηριστικών είναι η κίνηση της κάμερας σε μεγάλο ποσοστό των βίντεο της βάσης (Σχήμα 7.7β′), καθώς οι βελτιωμένες πυκνές τροχιές προσπαθούν να αντισταθμίσουν την κίνηση της κάμερας βελτιώνοντας τους περιγραφητές που βασίζονται στην οπτική ροή και αγνοώντας τροχιές που οφείλονται στην κίνηση της κάμερας.

Σε κάθε περίπτωση, υπολογίζονται οι περιγραφητές Trajectory, HOG, HOF, MBHx, MBHy και MBH (συνένωση των MBHx και MBHy). Το οπτικό λεξικό για κάθε περιγραφητή κατασκευάζεται κατά τα γνωστά με K-means συσταδοποίηση 100000 τυχαία επιλεγμένων στιγμιοτύπων εκπαίδευσης, με την ίδια τυχαία αρχικοποίηση κάθε φορά. Το μέγεθος του λεξικού επιλέχθηκε ίσο με $K = 4000$ οπτικές λέξεις.

Για την ταξινόμηση των βίντεο στις κατηγορίες δράσεων, χρησιμοποιούμε SVM ταξινομητές (υλοποίηση LIBSVM [74]) με τις συναρτήσεις πυρήνα που παρουσιάστηκαν παραπάνω για τα διανύσματα SoDVW και BoVW. Για την ταξινόμηση σε πολλαπλές κλάσεις χρησιμοποιούμε την προσέγγιση "one-against-all" (Ένας-Εναντίον-Όλων), ταξινομώντας το βίντεο στην κατηγορία με το μεγαλύτερο score. Πολλαπλοί περιγραφητές συνδυάζονται μέσω του αθροίσματος των πυρήνων τους, ενώ η σύμμειξη των μεθόδων αναπαράστασης SoDVW και BoVW επιτυγχάνεται με έναν γραμμικό συνδυασμό των πυρήνων τους.

7.4.3 Πειραματικά αποτελέσματα και συγκρίσεις

Στο σημείο αυτό η μέθοδος αναπαράστασης δράσεων βάσει της χρονικής αλληλουχίας των οπτικών λέξεων, που περιγράφηκε παραπάνω, εφαρμόζεται στα βίντεο των βάσεων KTH και HMDB51. Αρχικά συγκρίνουμε την επίδοση της μεθόδου και της σύμμειξής της με τη μέθοδο BoVW με την επίδοση του συστήματος αναφοράς που χρησιμοποιεί μόνο τη μέθοδο BoVW. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 7.1. Η μέθοδος SoDVW οδηγεί σε ακρίβεια αναγνώρισης 87.83% στη βάση KTH και 38.39% στην απαιτητική βάση μεγάλης κλίμακας HMDB51, χρησιμοποιώντας μόνο χρονική πληροφορία με τη μορφή ακολουθιών συχνά εμφανιζόμενων οπτικών λέξεων, χωρίς να έχει καμία πληροφορία σχετικά με την στατιστική κατανομή των χαρακτηριστικών στο εκάστοτε cluster (π.χ. αριθμός χαρακτηριστικών που έχουν ανατεθεί στην κάθε οπτική λέξη, σχετική θέση με αυτή κλπ). Επίσης, πρέπει να τονιστεί ότι αυτές οι ακολουθίες έχουν πολύ μικρότερο μήκος από τα ιστογράμματα BoVW, τα οποία στα πειραματά μας έχουν μήκος 4000 στοι-

[1] `http://lear.inrialpes.fr/people/wang/improved_trajectories`

Descriptor	KTH			HMDB51		
	BoVW	SoDVW	BoVW+ SoDVW	BoVW	SoDVW	BoVW+ SoDVW
Trajectory	90.85	83.78	91.19	33.47	23.75	38.32
HOG	86.67	80.76	86.79	29.13	18.84	34.18
HOF	93.4	86.33	93.51	41.26	30.61	43.86
MBHx	93.86	87.49	94.9	35.08	19.43	38.17
MBHy	92.93	85.17	93.63	42.64	22.55	45.84
MBH	94.67	87.83	**95.13**	43.55	25.53	46.47
Combined	94.09	87.83	94.67	52.16	38.39	**54.05**

Πίνακας 7.1: Αποτελέσματα ακρίβειας ταξινόμησης δράσεων με χρήση των μεθόδων BoVW, SoDVW και του συνδυασμού τους στις βάσεις δεδομένων ανθρώπινων δράσεων KTH και HMDB51.

χείων. Ενδεικτικά αναφέρουμε ότι τα μέγιστα μήκη των SoDVW ακολουθιών που αναπαριστούν τα βίντεο εκπαίδευσης και αξιολόγησης στο πρώτο split της βάσης HMDB51 έχουν μέγιστη τιμή 474 και 350, αντίστοιχα και διάμεσο (median) 50 στοιχεία. Επομένως, τα αποτελέσματα στις βάσεις KTH και HMDB51 αποδεικνύουν την πλούσια πληροφορία που ενσωματώνεται σε αυτές τις μικρές ακολουθίες και την ικανότητά τους να διαχωρίζουν τις δράσεις. Επίσης, επιβεβαιώνουν την αρχική υπόθεσή μας περί της σημαντικότητας της πληροφορίας της χρονικής διάταξης για την αναγνώριση δράσεων.

Εντούτοις, η διαδεδομένη αναπαράσταση BoVW επιτυγχάνει υψηλότερη ακρίβεια αναγνώρισης, χωρίς να εκμεταλλεύεται όμως τη σημαντική πληροφορία της χρονικής διάταξης. Ο συνδυασμός τους μέσω του γραμμικού συνδυασμού των συναρτήσεων πυρήνα οδηγεί σε βελτιωμένα αποτελέσματα σε σύγκριση με το σύστημα αναφοράς και στις δύο βάσεις για όλους τους περιγραφητές και το συνδυασμό τους. Οι βελτιώσεις στη βάση KTH φτάνουν το 1%, με την καλύτερη επίδοση να επιτυγχάνεται με τον περιγραφητή MBH. Στη βάση HMDB51, οι βελτιώσεις σε σχέση με το βασικό σύστημα είναι μεγαλύτερες και κυμαίνονται από 2% έως 5%, οδηγώντας σε μια τελική ακρίβεια αναγνώρισης 54.05%. Αυτές οι βελτιώσεις αποδεικνύουν τη συμπληρωματικότητα της αναπαράστασής μας, η οποία επιτρέπει την εύκολη ενσωμάτωση χρονικής πληροφορίας στη διαδεδομένη μέθοδο BoVW.

Στη συνέχεια παρουσιάζουμε την επίδραση της ποινής κενού του αλγορίθμου Smith-Waterman στο τελικό αποτέλεσμα ακρίβειας αναγνώρισης δράσεων με την αναπαράσταση SoDVW. Διατηρώντας $N_d = 5$ κυρίαρχες οπτικές λέξεις σε κάθε παράθυρο, πειραματιζόμαστε με δύο διαφορετικές τιμές της ποινής $p = 0.1$ και $p = 2$ στη βάση δεδομένων KTH και παρουσιάζουμε

	Trajectory	HOG	HOF	MBHx	MBHy	MBH
SoDVW($N_d = 5, p = 2$)	82.97	76.94	82.97	84.01	79.37	85.05
SoDVW($N_d = 5, p = 0.1$)	84.24	79.84	86.67	86.33	83.08	89.34

Πίνακας 7.2: Αποτελέσματα αναγνώρισης δράσεων στη βάση KTH για μεταβαλλόμενη ποινή κενού (gap penalty).

	Trajectory	HOG	HOF	MBHx	MBHy
SoDVW($N_{d(fmin)} = 5$)	76.59	73.00	79.27	78.4	76.13
SoDVW($N_{d(fmax)} = 5$)	82.97	76.94	82.97	84.01	79.37

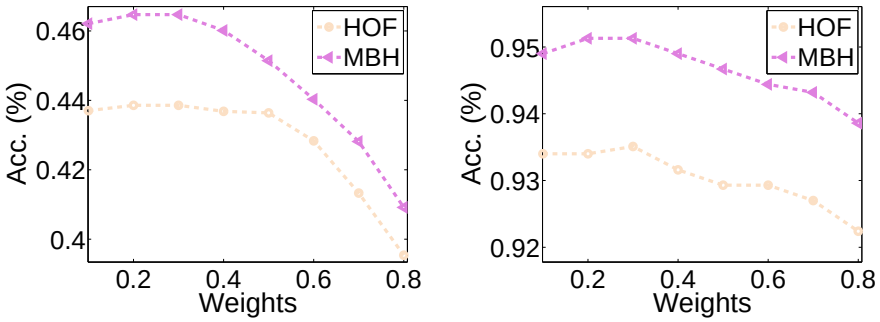
Πίνακας 7.3: Αποτελέσματα αναγνώρισης δράσεων στη βάση KTH για δύο διαφορετικούς τρόπους επιλογής οπτικών λέξεων.

τα αποτελέσματα στον Πίνακα 7.2. Όπως παρατηρούμε, το gap penalty με τιμή 0.1 οδηγεί σε καλύτερη επίδοση. Αυτό οφείλεται στην καλύτερη τοπική στοίχιση των ακολουθιών που επιτυγχάνεται με αυτή την τιμή της ποινής κενού. Όταν το gap penalty είναι ίσο με 2, τότε έχει πολύ μεγαλύτερη τιμή από τα στοιχεία του πίνακα αντικαταστάσεων και έτσι μια στοίχιση δύο ακολουθιών χωρίς τη χρήση κενών, δηλαδή ενθέσεων ή διαγραφών, συνεπάγεται μικρότερο κόστος σε σύγκριση με τις στοιχίσεις που χρησιμοποιούν κενά με μεγάλη ποινή. Ωστόσο, κατά τη στοίχιση ακολουθιών σχεδόν πάντα επιβάλλεται η προσθήκη κάποιων ενδιάμεσων κενών, έτσι ώστε να επιτευχθεί βέλτιστη στοίχιση και να ανακαλυφθούν περιοχές των ακολουθιών (sub-actions) με μεγάλη ομοιότητα, ιδίως όταν έχουμε μεγάλες μεταβολές στη διάρκεια και στον τρόπο εκτέλεσης των δράσεων.

Στον Πίνακα 7.3 συγκρίνουμε τα αποτελέσματα αναγνώρισης δράσεων στην περίπτωση που χρησιμοποιούνται οι $N_{d(fmax)} = 5$ πιο συχνά εμφανιζόμενες οπτικές λέξεις και στην περίπτωση που χρησιμοποιούνται οι $N_{d(fmin)} = 5$ πιο σπάνια εμφανιζόμενες οπτικές λέξεις. Η υπεροχή των πιο συχνά εμφανιζόμενων οπτικών λέξεων είναι προφανής στα αποτελέσματα και επιβεβαιώνει τόσο την υπόθεσή μας, όσο και τα πειραματικά αποτελέσματα των Yang et al. [86] στο πρόβλημα της μείωσης του μεγέθους του οπτικού λεξικού για την αναγνώριση σκηνών.

Επιπρόσθετα, οι δύο προηγούμενες παρατηρήσεις αναδεικνύουν την ευαισθησία της αναγνώρισης δράσεων στο περιέχομενο και τη στοίχιση των ακολουθιών, το οποίο είναι θεμιτό καθώς υποδεικνύει ότι οι προτεινόμενες ακολουθίες ενσωματώνουν ικανοποιητικά την πληροφορία της χρονικής διάταξης.

Τέλος, πειραματιζόμαστε με διαφορετικούς γραμμικούς συνδυασμούς των



Σχήμα 7.8: Ποσοστά μέσης ακρίβειας αναγνώρισης δράσεων στις βάσεις HMDB51 (αριστερά) και KTH (δεξιά) μεταβάλλοντας το βάρος θ_1 που ρυθμίζει τη συνεισφορά στο τελικό αποτέλεσμα της μεθόδου αναπαράστασης SoDVW κατά τη σύμμειξη με τη μέθοδο BoVW.

πυρήνων των SVMs που αντιστοιχούν στις αναπαραστάσεις SoDVW και BoVW, μεταβάλλοντας το διάνυσμα βαρών $\boldsymbol{\theta}$ και καταγράφουμε την ακρίβεια αναγνώρισης δράσεων μετά τη σύμμειξη της αναπαράστασής μας, που βασίζεται στη χρονική δομή των δράσεων, με την αναπαράσταση BoVW. Στο Σχήμα 7.8 απεικονίζεται η ακρίβεια αναγνώρισης δράσεων στις βάσεις KTH και HMDB51 για διαφορετικές τιμές της παραμέτρου θ_1 που σταθμίζει τον πυρήνα ομοιότητας των ακολουθιών κατά το συνδυασμό του με τον χ^2 πυρήνα. Αρχικά σχεδόν σε όλες τις περιπτώσεις το αποτέλεσμα της σύμμειξης των δύο μεθόδων είναι ανώτερο ή παραπλήσιο αυτού που επιτυγχάνεται μόνο με τη μέθοδο BoVW. Μεταβάλλοντας την παράμετρο θ_1 ρυθμίζουμε τη σχετική συνεισφορά της μεθόδου μας στο τελικό αποτέλεσμα. Σχεδόν σε όλες τις περιπτώσεις, η επίδοση του συστήματος μεγιστοποιείται όταν η τιμή της παραμέτρου θ_1 είναι περίπου ίση με 0.3. Επομένως, χρησιμοποιήσαμε αυτή την τιμή σε όλα τα αποτελέσματα που προηγήθηκαν, βασισμένοι στη συνεπή επίδοσή της σε όλες τις βάσεις δεδομένων και σε όλους τους περιγραφητές.

Τέλος, συγκρίνουμε τα αποτελέσματα της μεθόδου μας με τα αποτελέσματα άλλων μεθόδων, όπως αυτά παρατίθενται στις αντίστοιχες δημοσιεύσεις. Σε αυτές τις μεθόδους συγκαταλέγονται μέθοδοι, οι οποίες ενσωματώνουν κάποιο είδος χρονικής πληροφορίας, καθώς και πρόσφατες μέθοδοι της διεθνούς βιβλιογραφίας που έχουν οδηγήσει στα υψηλότερα αποτελέσματα στις βάσεις KTH και HMDB51.

Εύκολα μπορεί να συνάγει κανείς το συμπέρασμα ότι η προσέγγισή μας είναι καλύτερη σχεδόν σε όλες τις περιπτώσεις όταν συγκρίνεται με άλλες μεθόδους που αξιοποιούν χρονική πληροφορία. Επίσης βελτιώνουμε το αποτέλεσμα των μεθόδων που βασίζονται σε τροχιές συνδυασμένες με BoVW, όπως το σύστημα αναφοράς των improved/dense trajectories [30], [31], η μέ-

Method	Year	KTH	HMDB51
Augusti et al. [78]	2014	97.2	24.5
Cheng et al. [77]	2013	89.7	-
Lan et al. [84] (Fisher vector)	2014	-	65
Savarese et al. [79]	2008	86.83	-
Hatun et al. [83]	2008	92.0	-
Narayan & Ramakrishnan [69] (CD)	2014	96.5	53.4
Narayan & Ramakrishnan [69] (CD+Fisher Vector)	2014	-	58.7
Simonyan & Zisserman (SVM Fusion) [13]	2014	-	59.4
Oneata et al. (Fisher Vector) [55]	2013	-	54.8
Lan et al. [90]	2015	-	63.7
Dense Trajectories [30],[31]	2011	95.0	-
Improved Trajectories with BoW [31]	2013	-	52.1
Improved Trajectories with Fisher Vector [31]	2013	-	57.2
Jain et al. [32]	2013	-	52.3
Sadanand et al. [89]	2013	98.2	26.9
Our Method		95.1	54.1

Πίνακας 7.4: Σύγκριση της επίδοσης του συστήματος μας με άλλες προσεγγίσεις που αξιοποιούν χρονική πληροφορία (άνω μέρος) και πρόσφατες state-of-the-art μεθόδους, με τα υψηλότερα ποσοστά μέσης ακρίβειας αναγνώρισης δράσεων στις βάσεις KTH και HMDB51.

θοδος του περιγραφητή αιτιατότητας (causality descriptor) των Narayan et al. [69] , η μέθοδος action-bank [89] και η μέθοδος των Jain et al. [32] που διαχωρίζει τις τροχιές σε κυρίαρχες και δευτερεύουσες. Όπως ήταν αναμενόμενο, η επίδοσή μας είναι χαμηλότερη σε σύγκριση με τις μεθόδους που χρησιμοποιούν την ισχυρή μέθοδο αναπαράστασης Fisher Vector, η οποία εκμεταλλεύεται στατιστικά δεύτερης τάξης. Εντούτοις, το αποτέλεσμά μας είναι συγκρίσιμο με των Oneata et al. [55], παρόλο που χρησιμοποιούν Fisher Vector.

Εφόσον η μέθοδος αναπαράστασης που προτείνουμε προσφέρει συμπληρωματική πληροφορία σε σύγκριση με το Fisher Vector, περιμένουμε βελτιωμένη επίδοση από τη σύμμειξη αυτών των δύο μεθόδων στο μέλλον.

Κεφάλαιο 8

Συμπεράσματα

Στο κεφάλαιο αυτό ανακεφαλαιώνουμε τις βασικές συνεισφορές της διπλωματικής εργασίας και αναφερόμαστε συνοπτικά σε ορισμένες κατευθύνσεις όπου μπορεί να στραφεί η μελλοντική έρευνα, δεδομένων των συμπερασμάτων και των αποτελεσμάτων που προέκυψαν από την έρευνά μας.

8.1 Συμβολή της διπλωματικής εργασίας

Στην παρούσα διπλωματική εργασία ασχοληθήκαμε με το πρόβλημα της αναγνώρισης δράσεων υπό το πρίσμα των αναπαραστάσεων των βίντεο. Οι κύριες πτυχές της έρευνας αυτής της εργασίας συνοψίζονται ως εξής:

- Εισαγωγή στο πρόβλημα της αυτόματης αναγνώρισης ανθρώπινων δράσεων σε βίντεο, με έμφαση στις εφαρμογές, τις προκλήσεις και τις κυριότερες προσεγγίσεις επίλυσής του. Πιο συγκεκριμένα, παρουσιάστηκαν αναλυτικά ποικίλες σύγχρονες μέθοδοι εξαγωγής χαρακτηριστικών (όπως οι πυκνές τροχιές), αναπαράστασης και ταξινόμησης βίντεο, οι οποίες συνιστούν ένα σύστημα αναγνώρισης δράσεων.

- Εκτενής πειραματισμός με τις παραπάνω μεθόδους σε μία νέα βάση δεδομένων, την πολυτροπική και πολυαισθητηριακή βάση δεδομένων MOBOT, η οποία παρουσιάζει ιδιαίτερες προκλήσεις, καθώς τα βίντεο έχουν ληφθεί από οπτικό αισθητήρα τοποθετημένο πάνω σε κινούμενο ρομπότ και απεικονίζουν δράσεις που έχουν εκτελεσθεί από ηλικιωμένα άτομα με κινητικά, και πολλές φορές, και διανοητικά προβλήματα. Επίσης, επεκτείναμε κατάλληλα το σύστημα αναγνώρισης μεμονωμένων δράσεων, έτσι ώστε να μπορεί να αναγνωρίζει συνεχόμενες δράσεις. Μέσω των πειραματισμών μας, μπορέσαμε να συγκρίνουμε τους διάφορους ανιχνευτές χαρακτηριστικών, περιγραφητές και αναπαραστάσεις, να εντοπίσουμε

αδυναμίες και να μελετήσουμε την επίδραση των κανονικοποιήσεων των διανυσμάτων αναπαράστασης στο τελικό αποτέλεσμα αναγνώρισης.

- Αξιοποίηση του καναλιού βάθους (depth) στη βάση δεδομένων MOBOT για τη διευκόλυνση της αναγνώρισης δράσεων. Εξετάσαμε την εξαγωγή ενός περιγραφητή εμφάνισης από τα frames του βίντεο βάθους κατά μήκος των πυκνών τροχιών που εξάγονται από το κανάλι RGB. Αυτός ο συνδυασμός των δύο καναλιών πληροφορίας οδήγησε σε βελτίωση της ακρίβειας αναγνώρισης.

- Ανάπτυξη νέας μεθόδου αναπαράστασης βίντεο, η οποία λαμβάνει υπόψη την αλληλεπίδραση μεταξύ συστάδων οπτικών χαρακτηριστικών, χρησιμοποιώντας μια μετρική κατευθυνόμενης ομοιότητας μεταξύ τους. Η μέθοδος SD στηρίζεται στην εύρεση των κύριων συνιστωσών της κάθε συστάδας με χρήση PCA και στην προβολή των χαρακτηριστικών της μιας συστάδας στις κύριες κατευθύνσεις της άλλης, ενώ εξάγει τη μετρική χρησιμοποιώντας την απόκλιση Kullback-Leibler. Επιπρόσθετα, δοκιμάστηκε η χρήση GMM συσταδοποίησης και διαπιστώθηκε ότι οδηγεί σε βελτιωμένη απόδοση που εδράζεται στην καλύτερη μοντελοποίηση του χώρου χαρακτηριστικών και τη μικρότερη απώλεια πληροφορίας. Οι δύο παραλλαγές τις μεθόδου συγκρίθηκαν με άλλες γνωστές μεθόδους αναπαράστασης.

- Έμμεση μοντελοποίηση της χρονικής διάταξης των κινήσεων που αποτελούν μια δράση με την ανάπτυξη νέας μεθόδου αναπαράστασης, η οποία αναπαριστά κάθε βίντεο ως μία χρονική ακολουθία συχνά εμφανιζόμενων οπτικών λέξεων (SoDVW). Επίσης, προτάθηκε μετρική της απόστασης μεταξύ των διανυσμάτων αυτής της νέας αναπαράσταση με χρήση του αλγορίθμου τοπικής στοίχισης ακολουθιών Smith-Waterman. Δόθηκε ιδιαίτερη έμφαση στην ενσωμάτωση πληροφορίας σχετικά με την ομοιότητα των οπτικών λέξεων στον αλγόριθμο. Επίσης, διερευνήθηκε η σύμμειξη αυτής της μεθόδου, που λαμβάνει υπόψη αποκλειστικά τη χρονική πληροφορία με τη μέθοδο BoVW, η οποία κωδικοποιεί συχνότητες εμφάνισης οπτικών λέξεων. Η σύμμειξη των δύο μεθόδων με γραμμικό συνδυασμό των πυρήνων τους οδηγήσε σε βελτιωμένη ακρίβεια αναγνώρισης, η οποία είναι συγκρίσιμη και σε κάποιες περιπτώσεις ξεπερνά τις επιδόσεις state-of-the-art συστημάτων αναγνώρισης δράσεων.

- Εκτός από τη βάση MOBOT, έγιναν πειράματα και στις διεθνείς βάσεις δεδομένων KTH και HMDB. Πιο συγκεκριμένα, η μέθοδος αναπαράστασης SD αξιολογήθηκε στη δημοφιλή βάση δεδομένων KTH, ενώ για

τη μέθοδο SoDVW διεξήχθησαν μεγάλης κλίμακας πειράματα τόσο στη βάση KTH όσο και στην απαιτητική βάση HMDB51.

8.2 Κατευθύνσεις για μελλοντική έρευνα

Τα ενθαρρυντικά αποτελέσματα της παρούσας διπλωματικής εργασίας αποτελούν κίνητρο για περαιτέρω έρευνα, η οποία θα έχει ως σκοπό τη βελτίωση και την επέκταση των μεθόδων που προτάθηκαν. Ορισμένες κατευθύνσεις στις οποίες θα μπορούσε να στραφεί να επικεντρωθεί η μελλοντική έρευνα είναι:

- Μελέτη πρόσθετων μεθόδων σύμμειξης: οι αναπαραστάσεις που προτάθηκαν στην εργασία μας αξιοποιούν συμπληρωματική πληροφορία σε σχέση με state-of-the-art μεθόδους, όπως οι VLAD και Fisher Vector. Επομένως, η σύμμειξη τους είναι πολύ πιθανό να οδηγήσει σε ακόμα υψηλότερα αποτελέσματα. Μάλιστα, στην παρούσα εργασία χρησιμοποιήθηκε μόνο ο γραμμικός συνδυασμός με προκαθορισμένα βάρη των πυρήνων των διαφορετικών καναλιών πληροφορίας, π.χ. περιγραφητών, διανυσματικών αναπαραστάσεων. Η αυτόματη εκμάθηση των βαρών του γραμμικού συνδυασμού για κάθε δράση ξεχωριστά με χρήση αλγορίθμων Multiple Kernel Learning αποτελεί ένα ενδιαφέρον πεδίο έρευνας.

- Η χρήση του καναλιού βάθους για την εξαγωγή χαρακτηριστικών που επιτρέπουν τη μεγαλύτερη διαφοροποίηση μεταξύ των δράσεων είναι μια άλλη χρήσιμη ερευνητική κατεύθυνση, στην οποία ήδη έχουν αφιερωθεί πολλές διεθνείς εργασίες, αλλά έχει ακόμα πολλά περιθώρια βελτίωσης.

- Η επέκταση των πειραμάτων στη βάση MOBOT σε μεγαλύτερο αριθμό κατηγοριών δράσεων και η συσχέτιση των αναγνωρισμένων και χρονικά εντοπισμένων δράσεων με διάφορες παθολογίες παρουσιάζει ιδιαίτερο ενδιαφέρον. Για παράδειγμα, ο χρόνος που διαρκεί η δράση Sit-to-Stand θα μπορούσε να χρησιμοποιηθεί ως διαγνωστικός δείκτης.

- Η χρήση πληροφορίας από μεθόδους εκτίμησης πόζας (pose estimation) όχι μόνο θα βοηθήσει στο χωρικό εντοπισμό της δράσης, αλλά μπορεί να βελτιώσει τις υπάρχουσες αναπαραστάσεις ή να πυροδοτήσει την ανάπτυξη νέων. Πιο συγκεκριμένα, στη βάση MOBOT η έλλειψη της πληροφορίας της θέσης των ανθρώπων οδήγησε σε μη βέλτιστα αποτελέσματα με τη χρήση των improved trajectories. Επίσης, η μοντελοποίηση των αλληλεπιδράσεων μεταξύ των διάφορων μελών του ανθρώπινου σώματος, η οποία έγινε έμμεσα στην παρούσα εργασία μέσω της ανάλυσης των ομοιοτήτων μεταξύ συστάδων χαρακτηριστικών, είναι μια πολλά υποσχόμενη κατεύθυνση.

Bibliography

[1] L. Fogassi, P. F. Ferrari, B. Gesierich, S. Rozzi, F. Chersi, and G. Rizzolatti, "Parietal Lobe: From Action Organization to Intention Understanding," en, *Science*, vol. 308, no. 5722, pp. 662–667, Apr. 2005, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1106138. [Online]. Available: http://www.sciencemag.org/content/308/5722/662 (visited on 07/04/2015).

[2] S.-R. Ke, H. L. U. Thuc, Y.-J. Lee, J.-N. Hwang, J.-H. Yoo, and K.-H. Choi, "A review on video-based human activity recognition," *Computers*, vol. 2, no. 2, pp. 88–131, 2013. [Online]. Available: http://www.mdpi.com/2073-431X/2/2/88/htm (visited on 07/04/2015).

[3] M. Parajuli, D. Tran, W. Ma, and D. Sharma, "Senior health monitoring using Kinect," in *2012 Fourth International Conference on Communications and Electronics (ICCE)*, Aug. 2012, pp. 309–312. DOI: 10.1109/CCE.2012.6315918.

[4] G. Chalvatzaki, G. Pavlakos, K. Maninis, X. Papageorgiou, V. Pitsikalis, C. Tzafestas, and P. Maragos, "Towards an intelligent robotic walker for assisted living using multimodal sensorial data," in *2014 EAI 4th International Conference on Wireless Mobile Communication and Healthcare (Mobihealth)*, Nov. 2014, pp. 156–159. DOI: 10.1109/MOBIHEALTH.2014.7015934.

[5] C. Rougier, E. Auvinet, J. Rousseau, M. Mignotte, and J. Meunier, "Fall Detection from Depth Map Video Sequences," in *Proceedings of the 9th International Conference on Toward Useful Services for Elderly and People with Disabilities: Smart Homes and Health Telematics*, ser. ICOST'11, Berlin, Heidelberg: Springer-Verlag, 2011, pp. 121–128, ISBN: 978-3-642-21534-6. [Online]. Available: http://dl.acm.org/citation.cfm?id=2026187.2026206 (visited on 07/04/2015).

[6] X. S. Papageorgiou, C. S. Tzafestas, P. Maragos, G. Pavlakos, G. Chalvatzaki, G. Moustris, I. Kokkinos, A. Peer, B. Stanczyk, E.-S. Fotinea, and others, "Advances in intelligent mobility assistance

robot integrating multimodal sensory processing," in *Universal Access in Human-Computer Interaction. Aging and Assistive Environments*, Springer, 2014, pp. 692–703. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-319-07446-7_66 (visited on 07/04/2015).

[7] I. Pastor, H. Hayes, and S. Bamberg, "A feasibility study of an upper limb rehabilitation system using kinect and computer games," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Aug. 2012, pp. 1286–1289. DOI: 10.1109/EMBC.2012.6346173.

[8] Y.-J. Chang, S.-F. Chen, and J.-D. Huang, "A Kinect-based system for physical rehabilitation: a pilot study for young adults with motor disabilities," eng, *Research in Developmental Disabilities*, vol. 32, no. 6, pp. 2566–2570, Dec. 2011, ISSN: 1873-3379. DOI: 10.1016/j.ridd.2011.07.002.

[9] D. Leightley, J. Darby, B. Li, J. McPhee, and M. H. Yap, "Human Activity Recognition for Physical Rehabilitation," in *2013 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Oct. 2013, pp. 261–266. DOI: 10.1109/SMC.2013.51.

[10] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976–990, Jun. 2010, ISSN: 0262-8856. DOI: 10.1016/j.imavis.2009.11.014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0262885609002704 (visited on 07/04/2015).

[11] K. Soomro and A. R. Zamir, "Action Recognition in Realistic Sports Videos," in *Computer Vision in Sports*, Springer, 2014, pp. 181–208. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-319-09396-3_9 (visited on 07/06/2015).

[12] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," *arXiv preprint arXiv:1405.4506*, 2014. [Online]. Available: http://arxiv.org/abs/1405.4506 (visited on 07/10/2015).

[13] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems*, 2014, pp. 568–576. [Online]. Available: http://papers.nips.cc/paper/5353-two-stream-convolutional-networks-for-action-recognition-in-videos (visited on 07/10/2015).

[14] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing Human Actions: A Local SVM Approach," in *Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 3 - Volume 03*, ser. ICPR '04, Washington, DC, USA: IEEE Computer Society, 2004, pp. 32–36, ISBN: 978-0-7695-2128-2. DOI: 10.1109/ICPR.2004.747. [Online]. Available: http://dx.doi.org/10.1109/ICPR.2004.747 (visited on 07/10/2015).

[15] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[16] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005*, Oct. 2005, pp. 65–72. DOI: 10.1109/VSPETS.2005.1570899.

[17] G. Willems, T. Tuytelaars, and L. Gool, "An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector," in *Proceedings of the 10th European Conference on Computer Vision: Part II*, ser. ECCV '08, Berlin, Heidelberg: Springer-Verlag, 2008, pp. 650–663, ISBN: 978-3-540-88685-3. DOI: 10.1007/978-3-540-88688-4_48. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-88688-4_48 (visited on 07/10/2015).

[18] K. Maninis, P. Koutras, and P. Maragos, "Advances on action recognition in videos using an interest point detector based on multiband spatio-temporal energies," in *2014 IEEE International Conference on Image Processing (ICIP)*, Oct. 2014, pp. 1490–1494. DOI: 10.1109/ICIP.2014.7025298.

[19] A. Klaser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *BMVC 2008-19th British Machine Vision Conference*, British Machine Vision Association, 2008, pp. 275–1. [Online]. Available: https://hal.inria.fr/inria-00514853/ (visited on 07/10/2015).

[20] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-Up Robust Features (SURF)," *Comput. Vis. Image Underst.*, vol. 110, no. 3, pp. 346–359, Jun. 2008, ISSN: 1077-3142. DOI: 10.1016/j.cviu.2007.09.014. [Online]. Available: http://dx.doi.org/10.1016/j.cviu.2007.09.014 (visited on 07/09/2015).

[21] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005*, vol. 1, Jun. 2005, 886–893 vol. 1. DOI: 10.1109/CVPR.2005.177.

[22] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008*, Jun. 2008, pp. 1–8. DOI: 10.1109/CVPR.2008.4587756.

[23] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *BMVC 2009-British Machine Vision Conference*, BMVA Press, 2009, pp. 124–1. [Online]. Available: https://hal.inria.fr/inria-00439769/ (visited on 07/10/2015).

[24] P. Matikainen, M. Hebert, and R. Sukthankar, "Trajectons: Action recognition through the motion analysis of tracked features," in *2009 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, IEEE, 2009, pp. 514–521. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5457659 (visited on 07/09/2015).

[25] R. Messing, C. Pal, and H. Kautz, "Activity recognition using the velocity histories of tracked keypoints," in *2009 IEEE 12th International Conference on Computer Vision*, Sep. 2009, pp. 104–111. DOI: 10.1109/ICCV.2009.5459154.

[26] J. Shi and C. Tomasi, "Good features to track," in *, 1994 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94*, Jun. 1994, pp. 593–600. DOI: 10.1109/CVPR.1994.323794.

[27] J. Sun, X. Wu, S. Yan, L.-F. Cheong, T.-S. Chua, and J. Li, "Hierarchical spatio-temporal context modeling for action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009*, Jun. 2009, pp. 2004–2011. DOI: 10.1109/CVPR.2009.5206721.

[28] J. Sun, Y. Mu, S. Yan, and L.-F. Cheong, "Activity recognition using dense long-duration trajectories," in *2010 IEEE International Conference on Multimedia and Expo (ICME)*, Jul. 2010, pp. 322–327. DOI: 10.1109/ICME.2010.5583046.

[29] T. Brox and J. Malik, "Object Segmentation by Long Term Analysis of Point Trajectories," en, in *Computer Vision – ECCV 2010*, ser. Lecture Notes in Computer Science 6315, K. Daniilidis, P. Maragos, and N. Paragios, Eds., Springer Berlin Heidelberg, 2010, pp. 282–295, ISBN: 978-3-642-15554-3 978-3-642-15555-0. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-15555-0_21 (visited on 07/09/2015).

[30] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2011, pp. 3169–3176. DOI: 10.1109/CVPR.2011.5995407.

[31] H. Wang and C. Schmid, "Action Recognition with Improved Trajectories," in *2013 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2013, pp. 3551–3558. DOI: 10.1109/ICCV.2013.441.

[32] M. Jain, H. Jegou, and P. Bouthemy, "Better Exploiting Motion for Better Action Recognition," in *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2013, pp. 2555–2562. DOI: 10.1109/CVPR.2013.330.

[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-w (visited on 07/10/2015).

[34] J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," *arXiv preprint arXiv:1503.08909*, 2015. [Online]. Available: http://arxiv.org/abs/1503.08909 (visited on 07/10/2015).

[35] ☒. Μαραγκός, *Ανάλυση Εικόνων και Όραση Υπολογιστών*. Εκδόσεις ΕΜΠ, 2005.

[36] B. D. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision," in *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, ser. IJCAI'81, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1981, pp. 674–679. [Online]. Available: http://dl.acm.org/citation.cfm?id=1623264.1623280 (visited on 07/09/2015).

[37] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Image Analysis*, Springer, 2003, pp. 363–370.

[38] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense Trajectories and Motion Boundary Descriptors for Action Recognition," en, *International Journal of Computer Vision*, vol. 103, no. 1, pp. 60–79, Mar. 2013, ISSN: 0920-5691, 1573-1405. DOI: 10.1007/s11263-012-0594-8. [Online]. Available: http://link.springer.com/article/10.1007/s11263-012-0594-8 (visited on 07/17/2015).

[39] R. Arandjelovic and A. Zisserman, "Three things everyone should know to improve object retrieval," in *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2012, pp. 2911–2918. DOI: 10.1109/CVPR.2012.6248018.

[40] N. Dalal, B. Triggs, and C. Schmid, "Human Detection Using Oriented Histograms of Flow and Appearance," en, in *Computer Vision – ECCV 2006*, ser. Lecture Notes in Computer Science 3952, A. Leonardis, H. Bischof, and A. Pinz, Eds., Springer Berlin Heidelberg, 2006, pp. 428–441, ISBN: 978-3-540-33834-5 978-3-540-33835-2. [Online]. Available: http://link.springer.com/chapter/10.1007/11744047_33 (visited on 07/09/2015).

[41] T. Vincent and R. Laganiere, "Detecting planar homographies in an image pair," in *Proceedings of the 2nd International Symposium on Image and Signal Processing and Analysis, 2001. ISPA 2001*, 2001, pp. 182–187. DOI: 10.1109/ISPA.2001.938625.

[42] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981. [Online]. Available: http://dl.acm.org/citation.cfm?id=358692 (visited on 07/10/2015).

[43] H. Jegou, M. Douze, C. Schmid, and P. Perez, "Aggregating local descriptors into a compact image representation," in *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2010, pp. 3304–3311. DOI: 10.1109/CVPR.2010.5540039.

[44] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *Computer Vision–ECCV 2010*, Springer, 2010, pp. 143–156. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-15561-1_11 (visited on 07/10/2015).

[45] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006. [Online]. Available: http://cds.cern.ch/record/998831/files/9780387310732_TOC.pdf (visited on 07/06/2015).

[46] R. Arandjelovic and A. Zisserman, "All About VLAD," in *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2013, pp. 1578–1585. DOI: 10.1109/CVPR.2013.207.

[47] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid, "Aggregating Local Image Descriptors into Compact Codes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1704–1716, Sep. 2012, ISSN: 0162-8828. DOI: 10.1109/TPAMI.2011.235.

[48] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995. [Online]. Available: http://link.springer.com/article/10.1007/bf00994018 (visited on 07/10/2015).

[49] S. Theodoridis and K. Koutroumbas, *Pattern Recognition, Fourth Edition*, 4th. Academic Press, 2008, ISBN: 978-1-59749-272-0.

[50] T. Fletcher, "Support vector machines explained," 2009. [Online]. Available: http://sutikno.blog.undip.ac.id/files/2011/11/SVM-Explained.pdf (visited on 07/10/2015).

[51] J. C. Platt, "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods," in *ADVANCES IN LARGE MARGIN CLASSIFIERS*, Citeseer, 1999. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.41.1639&g (visited on 07/10/2015).

[52] J. Zhang, M. Marsza\lek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *International journal of computer vision*, vol. 73, no. 2, pp. 213–238, 2007. [Online]. Available: http://link.springer.com/article/10.1007/s11263-006-9794-4 (visited on 07/10/2015).

[53] M. J. Swain and D. H. Ballard, "Color indexing," en, *International Journal of Computer Vision*, vol. 7, no. 1, pp. 11–32, Nov. 1991, ISSN: 0920-5691, 1573-1405. DOI: 10.1007/BF00130487. [Online]. Available: http://link.springer.com/article/10.1007/BF00130487 (visited on 07/10/2015).

[54] Z. Shu, K. Yun, and D. Samaras, "Action Detection with Improved Dense Trajectories and Sliding Window," en, in *Computer Vision - ECCV 2014 Workshops*, ser. Lecture Notes in Computer Science 8925, L. Agapito, M. M. Bronstein, and C. Rother, Eds., Springer International Publishing, Sep. 2014, pp. 541–551, ISBN: 978-3-319-16177-8 978-3-319-16178-5. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-319-16178-5_38 (visited on 07/10/2015).

[55] D. Oneata, J. Verbeek, and C. Schmid, "Action and Event Recognition with Fisher Vectors on a Compact Feature Set," in *2013 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2013, pp. 1817–1824. DOI: 10.1109/ICCV.2013.228.

[56] A. Jain, S. V. N. Vishwanathan, and M. Varma, "SPF-GMKL: generalized multiple kernel learning with a million kernels," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2012, pp. 750–758. [Online]. Available: http://dl.acm.org/citation.cfm?id=2339648 (visited on 07/10/2015).

[57] W. Lin, M.-T. Sun, R. Poovandran, and Z. Zhang, "Human activity recognition for video surveillance," in *IEEE International Symposium on Circuits and Systems, 2008. ISCAS 2008*, May 2008, pp. 2737–2740. DOI: 10.1109/ISCAS.2008.4542023.

[58] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=18626 (visited on 07/10/2015).

[59] P. Giannoulis and G. Potamianos, "A hierarchical approach with feature selection for emotion recognition from speech.," in *LREC*, 2012, pp. 1203–1206. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2012/pdf/917_Paper.pdf (visited on 07/10/2015).

[60] L. Xia and J. Aggarwal, "Spatio-temporal Depth Cuboid Similarity Feature for Activity Recognition Using Depth Camera," in *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2013, pp. 2834–2841. DOI: 10.1109/CVPR.2013.365.

[61] S. Hadfield and R. Bowden, "Hollywood 3d: Recognizing Actions in 3d Natural Scenes," in *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2013, pp. 3398–3405. DOI: 10.1109/CVPR.2013.436.

[62] S. Tang, X. Wang, X. Lv, T. X. Han, J. Keller, Z. He, M. Skubic, and S. Lao, "Histogram of oriented normal vectors for object recognition with a depth sensor," in *Computer Vision–ACCV 2012*, Springer, 2013, pp. 525–538. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-37444-9_41 (visited on 07/10/2015).

[63] O. Oreifej and Z. Liu, "HON4d: Histogram of Oriented 4d Normals for Activity Recognition from Depth Sequences," in *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2013, pp. 716–723. DOI: 10.1109/CVPR.2013.98.

[64] X. Yang and Y. Tian, "Super Normal Vector for Activity Recognition Using Depth Sequences," in *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2014, pp. 804–811. DOI: 10.1109/CVPR.2014.108.

[65] M. Ye, Q. Zhang, L. Wang, J. Zhu, R. Yang, and J. Gall, "A survey on human motion analysis from depth data," in *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*, Springer, 2013, pp. 149–187. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-44964-2_8 (visited on 07/10/2015).

[66] M. Koperski, P. Bilinski, and F. Bremond, "3d trajectories for action recognition," in *2014 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2014, pp. 4176–4180. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7025848 (visited on 07/10/2015).

[67] Y. Xiao, G. Zhao, J. Yuan, and D. Thalmann, "Activity recognition in unconstrained RGB-D video using 3d trajectories," in *SIGGRAPH Asia 2014 Autonomous Virtual Humans and Social Robot for Telepresence*, ACM, 2014, p. 4. [Online]. Available: http://dl.acm.org/citation.cfm?id=2668961 (visited on 07/10/2015).

[68] L. Spinello and K. O. Arras, "People detection in RGB-D data," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2011, pp. 3838–3843. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6095074 (visited on 07/10/2015).

[69] S. Narayan and K. Ramakrishnan, "A Cause and Effect Analysis of Motion Trajectories for Modeling Actions," in *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2014, pp. 2633–2640. DOI: 10.1109/CVPR.2014.337.

[70] C.-C. Lee, A. Katsamanis, M. P. Black, B. R. Baucom, A. Christensen, P. G. Georgiou, and S. S. Narayanan, "Computing vocal entrainment: A signal-derived PCA-based quantification scheme with application to affect analysis in married couple interactions," *Computer Speech & Language*, vol. 28, no. 2, pp. 518–539, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0885230812000472 (visited on 07/10/2015).

[71] T. L. Chartrand and J. A. Bargh, "The chameleon effect: the perception–behavior link and social interaction.," *Journal of personality and social psychology*, vol. 76, no. 6, p. 893, 1999. [Online]. Available: http://psycnet.apa.org/journals/psp/76/6/893/ (visited on 07/10/2015).

[72] M. E. O. S. Parthasarathy, "A dissimilarity measure for comparing subsets of data: Application to multivariate time series," *Temporal data mining: algorithms, theory and applications (TDM 2005)*, p. 101, 2005. [Online]. Available: http://web.cse.ohio-state.edu/dmrl/papers/tdm-otey.pdf (visited on 07/17/2015).

[73] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, pp. 79–86, 1951. [Online]. Available: http://www.jstor.org/stable/2236703 (visited on 07/10/2015).

[74] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011. [Online]. Available: http://dl.acm.org/citation.cfm?id=1961199 (visited on 07/10/2015).

[75] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *2011 IEEE International Conference on Computer Vision (ICCV)*, Nov. 2011, pp. 2556–2563. DOI: 10.1109/ICCV.2011.6126543.

[76] O. Ramana Murthy and R. Goecke, "The Influence of Temporal Information on Human Action Recognition with Large Number of Classes," in *2014 International Conference on Digital Image Computing: Techniques and Applications (DlCTA)*, Nov. 2014, pp. 1–8. DOI: 10.1109/DICTA.2014.7008131.

[77] G. Cheng, Y. Wan, W. Santiteerakul, S. Tang, and B. Buckles, "Action Recognition with Temporal Relationships," in *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2013, pp. 671–675. DOI: 10.1109/CVPRW.2013.101.

[78] P. Agustí, V. J. Traver, and F. Pla, "Bag-of-words with aggregated temporal pair-wise word co-occurrence for human action recognition," *Pattern Recognition Letters*, vol. 49, pp. 224–230, Nov. 2014, ISSN: 0167-8655. DOI: 10.1016/j.patrec.2014.07.014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865514002311 (visited on 03/31/2015).

[79] S. Savarese, A. DelPozo, J. Niebles, and L. Fei-Fei, "Spatial-Temporal correlatons for unsupervised action classification," in *IEEE Workshop on Motion and video Computing, 2008. WMVC 2008*, Jan. 2008, pp. 1–8. DOI: 10.1109/WMVC.2008.4544068.

[80] R. Hamid, S. Maddi, A. Johnson, A. Bobick, I. Essa, and C. Isbell, "A novel sequence representation for unsupervised analysis of human activities," *Artificial Intelligence*, vol. 173, no. 14, pp. 1221–1244, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0004370209000629 (visited on 07/11/2015).

[81] C.-C. Chen and J. Aggarwal, "Modeling human activities as speech," in *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2011, pp. 3425–3432. DOI: 10.1109/CVPR.2011.5995555.

[82] H. Kuehne, A. Arslan, and T. Serre, "The Language of Actions: Recovering the Syntax and Semantics of Goal-Directed Human Activities," in *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2014, pp. 780–787. DOI: 10.1109/CVPR.2014.105.

[83] K. Hatun and P. Duygulu, "Pose sentences: A new representation for action recognition using sequence of pose words," in *19th International Conference on Pattern Recognition, 2008. ICPR 2008*, Dec. 2008, pp. 1–4. DOI: 10.1109/ICPR.2008.4761702.

[84] Z. Lan, X. Li, and A. G. Hauptmann, "Temporal Extension of Scale Pyramid and Spatial Pyramid Matching for Action Recognition," *arXiv:1408.7071 [cs]*, Aug. 2014, arXiv: 1408.7071. [Online]. Available: http://arxiv.org/abs/1408.7071 (visited on 04/21/2015).

[85] T. Glaser and L. Zelnik-Manor, "Incorporating temporal context in Bag-of-Words models," in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, Nov. 2011, pp. 1562–1569. DOI: 10.1109/ICCVW.2011.6130436.

[86] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo, "Evaluating bag-of-visual-words representations in scene classification," in *Proceedings of the international workshop on Workshop on multimedia information retrieval*, ACM, 2007, pp. 197–206. [Online]. Available: http://dl.acm.org/citation.cfm?id=1290111 (visited on 07/10/2015).

[87] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of molecular biology*, vol. 48, no. 3, pp. 443–453, 1970. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0022283670900574 (visited on 07/10/2015).

[88] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *Journal of molecular biology*, vol. 147, no. 1, pp. 195–197, 1981. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0022283681900875 (visited on 07/10/2015).

[89] S. Sadanand and J. Corso, "Action bank: A high-level representation of activity in video," in *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2012, pp. 1234–1241. DOI: 10.1109/CVPR.2012.6247806.

[90] Z. Lan, X. Li, M. Lin, and A. G. Hauptmann, "Long-short Term Motion Feature for Action Classification and Retrieval," *arXiv preprint arXiv:1502.04132*, 2015. [Online]. Available: http://arxiv.org/abs/1502.04132 (visited on 07/11/2015).