Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Τομέας Σημάτων, Ελέγχου και Ρομποτικής

Εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων

Αναγνώριση και Ταξινόμηση Ανθρώπινων Δράσεων σε Video

Διπλωματική Εργασία

του

Κέβη-Κοκίτσι Ι. Μανίνη

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2014

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

Αναγνώριση και Ταξινόμηση Ανθρώπινων Δράσεων σε Video

Διπλωματική Εργασία

του

Κέβη-Κοκίτσι Ι. Μανίνη

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 29η Ιουλίου 2014.

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

Αθήνα, Ιούλιος 2014

(Υπογραφή)

.....
Κέβης-Κοκίτσι Ι. Μανίνης
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Σκοπός της παρούσας διπλωματικής εργασίας είναι η αντιμετώπιση του προβλήματος της αναγνώρισης και ταξινόμησης ανθρώπινων δράσεων στην Όραση υπολογιστών. Το πρόβλημα αυτό θεωρείται θεμελιώδες στην ανάλυση και ερμηνεία video και για αυτό ερευνάται και εφαρμόζεται ευρέως σε διάφορους τομείς όπως η ανάκτηση δεδομένων από video, η οπτική επιτήρηση και παρακολούθηση, η ρομποτική και η αλληλεπίδραση ανθρώπου-υπολογιστή.Η προσέγγιση που χρησιμοποιήσαμε εκμεταλλεύεται μεθόδους αναπαράστασης των video μέσω τοπικών χαρακτηριστικών. Σκοπεύουμε στην εύρεση εύρωστων χωροχρονικών σημείων ενδιαφέροντος που να αποτελούν πηγή μιας συμπαγούς αναπαράστασης των video, στα πλαίσια της οποίας αναπτύσσουμε και παρουσιάζουμε δύο νέους αλγόριθμους ανίχνευσης. Για την εξαγωγή χαρακτηριστικών χρησιμοποιούμε δημοφιλείς περιγραφητές, όπως οι Histograms of Oriented Gradients/Histograms of Optical Flow (HOG/HOF) και οι Histograms of Oriented 3D Gradients (HOG3D). Γίνεται προσπάθεια μοντελοποίησης και αναγνώρισης των ανθρώπινων δράσεων με χρήση ισχυρών εργαλείων όπως οι Μηχανές Διανυσμάτων Υποστήριξης, τα Κρυφά Μαρκοβιανά Μοντέλα και οι ταξινομητές k-Nearest Neighbour, σε συνδυασμό με γνωστές τεχνικές όπως Bag-of-Features και Γραμμική Πρόβλεψη. Οι αλγόριθμοί μας αξιολογούνται πειραματικά σε δύο πολύ γνωστές βάσεις δεδομένων ανθρώπινων δράσεων, όπου και ξεπερνούν τις επιδόσεις που επιτεύχθηκαν από δημοφιλείς αλγόριθμους της βιβλιογραφίας. Ο πειραματισμός μας επεκτάθηκε σε μια νέα πολυαισθητηριακή βάση δεδομένων, όπου και εφαρμόσαμε νέες τεχνικές αναγνώρισης συνεχόμενων δράσεων.

Λέξεις Κλειδιά

Όραση Υπολογιστών, αναγνώριση ανθρώπινων δράσεων, τοπικά χωροχρονικά χαρακτηριστικά, ανιχνευτές χωροχρονικών σημείων ενδιαφέροντος, οπτικοί περιγραφητές, Μηχανές Διανυσμάτων Υποστήριξης, Κρυφά Μαρκοβιανά Μοντέλα, φιλτράρισμα σε πολλαπλές ζώνες συχνοτήτων, φίλτρα Gabor, ανάλυση κυρίαρχης ενέργειας, ανίχνευση ενέργειας σε video

Abstract

The aim of this thesis is to deal with the task of human action classification and recognition in videos. This task is considered fundamental in video analysis and video understanding and because of that it is widely researched and applied in several domains such as video retrieval, video surveillance, robotics and human-computer interaction. Our approach takes advantage of video representation with local features. Our aim is to find robust spatio-temporal interest points that lead to compact representation of videos. We develop and propose two new algorithms that search for local spatio-temporal interest points. For feature extraction we use popular descriptors as Histograms of Oriented Gradients/Histograms of Optical Flow (HOG/HOF) and Histograms of Oriented 3D Gradients (HOG3D). We try to model and recognize human actions using powerful machine learning tools such as Support Vector Machines, Hidden Markov Models and k-Nearest Neighbour classifiers combined with known techniques such as Bag-of-Features and Linear Predictive Coding. Our algorithms are experimentally evaluated in two popular action databases, in which they ourperform state-of-the-art detectors. The experimental evaluation was extended to a new multi-sensor database, where we applied new methods for continuous action recognition.

Keywords

computer vision, human action recognition, local spatio-temporal features, spatio-temporal interest point detectors, visual descriptors, Support Vector Machines, Hidden Markov Models, multi-band filtering, Gabor filters, dominant energy analysis, energy tracking in videos

Ευχαριστίες

Θα ήθελα κατάρχήν να ευχαριστήσω τον καθηγητή Πέτρο Μαραγκό για την ανάθεση της παρούσας διπλωματικής εργασίας. Η παρακολούθηση των μαθημάτων Ψηφιακή Επεξεργασία Σήματος, Όραση Υπολογιστών και Αναγνώριση Προτύπων που διδάσκονται από τον κ. Μαραγκό υπήρξε καθοριστική για τη διαμόρφωση των ακαδημαϊκών μου ενδιαφερόντων καθώς και την έναρξη της συνεργασίας μας στα πλαίσια μιας διπλωματικής εργασίας. Για το ζήλο που δείχνει στο μάθημα, το εύρος γνώσεών του και τον κινητοποιητικό τρόπο διδασκαλίας του τον θεωρώ υπόδειγμα καθηγητή, ενώ για την εμπιστοσύνη που μου έδειξε, το χρόνο που μου αφιέρωσε, το ενδιαφέρον που έδειξε για την επίβλεψη της παρούσας διπλωματικής εργασίας και τις ευχάριστες συζητήσεις κατά την εκπόνησή της τον θεωρώ υπόδειγμα επαγγελματία και ανθρώπου.

Θα ήθελα επίσης να ευχαριστήσω όλα τα μέλη του εργαστηρίου, και ιδιαίτερα τον Πέτρο Κούτρα για τη άψογη συνεργασία που είχαμε, την επίβλεψή του και τις συμβουλές του. Χωρίς τον Πέτρο Κούτρα αυτή η διπλωματική σίγουρα θα είχε διαφορετική μορφή. Ευχαριστώ επίσης το συμφοιτητή και φίλο μου Γιώργο Παυλάκο για τη συνεργασία που είχαμε, τις συζητήσεις που είχαμε και που όποτε χρειάστηκα τη βοήθειά του ήταν εκεί. Ευχαριστώ το Χρήστο Γεωργάκη για την παραχώρηση του κώδικα με τον οποίο ξεκίνησα να σχολούμαι με το πρόβλημα της αναγνώρισης δράσεων. Ευχαριστώ τη Νάνσυ για τις πολύτιμες συμβουλές της, το Νάσσο για τις "τρελές" επιστημονικές του ιδέες, το Βασίλη που με έμαθε να είμαι οργανωμένος, τη Ξανθή και τη Γεωργία για τη συνεργασία μας, το Σταύρο, τον Παναγιώτη, την Αντιγόνη και τον Ισίδωρο για το ευχάριστο κλίμα που δημιουργούν στο εργαστήριο.

Ευχαριστώ από τα βάθη της καρδιάς μου τους γονείς μου, Γιάννη και Kazuyo για την υποστήριξή της καθ' όλη τη διάρκεια των σπουδών μου. Ευχαριστώ όλους τους φίλους μου και τους άνθρωπους που συνάντησα όλα αυτά τα χρόνια και αποτελούν πηγή έμπνευσης και θαυμασμού για εμένα. Τέλος θέλω να ευχαριστήσω εσένα Ευαγγελία για την υποστήριξη, την αγάπη και την υπομονή σου.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Γενικά για την Όραση Υπολογιστών

Η Όραση Υπολογιστών είναι η επιστήμη και η τεχνολογία που κάνει τις μηχανές να "βλέπουν", με την έννοια της εξαγωγής συμβολικής πληροφορίας από μια εικόνα ή ακοκολουθία εικόνων, η οποία είναι χρήσιμη για την επίλυση διάφορων προβλημάτων [38]. Στόχος είναι να μπορούμε να κατασκευάσουμε συστήματα που να μπορούν να αντιλαμβάνονται τον ορατό κόσμο με ανθρώπινη ευφυία. Η Όραση Υπολογιστών έχει τις ρίζες της στην τεχνητή νοημοσύνη και από τότε έχει αναπτυχθεί με μεγάλους ρυθμούς συνδυάζοντας μαζί πολλά επιστημονικά πεδία, όπως η επεξεργασία σημάτων, η αναγνώριση προτύπων, τα μαθηματικά, η φυσική, η νευροβιολογία, ο αυτόματος έλεγχος και ρομποτική.

Η εξαγωγή συμβολικών αναπαραστάσεων από εικόνες και η μηχανική οπτική αντίληψη του φυσικού κόσμου μέσω υπολογιστικών μεθόδων, όπως αυτές που μετέρχεται η Όραση Υπολογιστών παραμένει ακόμα και σήμερα ένα ιδιαίτερα πολύπλοκο και απαιτητικό πρόβλημα. Τα κυριότερα προβλήματα που προσπαθεί να επιλύσει αποτελούν πρόκληση για περαιτέρω έρευνα και ανάπτυξη καινοτόμων προσεγγίσεων αποτελεσματικής αντιμετώπισής τους.

Η Όραση Υπολογίστων βρίσκει πολλές εφαρμογές σε διάφορους τομείς όπως η επεξεργασία εικόνων, η ρομποτική, η βιοιατρική τεχνολογία, η επικοινωνία ανθρώπου-υπολογιστή, η τηλεπισκόπηση, τα ευφυή συστήματα, η οργάνωση πληροφορίας, ο κινηματογράφος και οι άλλες ψηφιακές τέχνες. Τυπικά προβλήματα της όρασης υπολογιστών είναι:

- η Αναγνώριση Αντικειμένων,
- η Εκτίμηση Κίνησης,
- η τρισδιάστατη Ανακατασκευή Σκηνής από διάφορες εικόνες της

- η Αποκατάσταση Εικόνας.

1.2 Το πρόβλημα της Αναγνώρισης Δράσεων σε Video

Το πρόβλημα της αναγνώρισης ενός αντικειμένου σε μία εικόνα ή video (ακολουθία εικόνων) αποτελεί ένα από τα βασικότερα προβλήματα της Όρασης Υπολογιστών που βρίσκει τεράστιες εφαρμογές στην επικοινωνία ανθρώπου υπολογιστή. Ο βασικός στόχος είναι να εξάγουμε συμβολικές πληροφορίες από τα δεδομένα, δηλαδή τα video, οι οποίες θα μας οδηγήσουν να αναγνωρίσουμε το ζητούμενο αντικείμενο και την κατάσταση στην οποία αυτό εμφανίζεται μέσα του. Στη συνέχεια, σε αρκετές περιπτώσεις, το αρχικό πρόβλημα ανάγεται σε ένα απλό πρόβλημα της Αναγνώρισης Προτύπων. Στις περισσότερες περιπτώσεις είναι απαραίτητο να διαθέτουμε εξ' αρχής ένα αρκετά σημαντικό σύνολο δεδομένων, το οποίο αποτελεί τα δεδομένα εκπαίδευσης. Τα βασικά στάδια ενός προβλήματος αναγνώρισης είναι:

- Δημιουργία ενός μοντέλου που να αναπαριστά το ζητούμενο αντικείμενο με βάση το σύνολο των δεδομένων εκπαίδευσης.
- Επεξεργασία των νέων δεδομένων και προσαρμογή του μοντέλου σε αυτά.
- Εξαγώγη της συμβολικής πληροφορίας είτε απευθείας από τις παραμέτρους του μοντέλου, είτε από κάποια επεξεργασία τους.
- Αναγνώριση του αντικειμένου και της κατάστασης του με βάση τη συμβολική πληροφορία ή εφαρμογή κάποιας τεχνικής από την περιοχή της αναγνώρισης προτύπων με σκοπό την τελική αναγνώριση του αντικειμένου.

Παρόμοια λογική ακολουθείται και για το πρόβλημα της αναγνώρισης δράσεων σε video, το οποίο αποτελεί ένα ιδιαίτερα ενεργό πεδίο της Όρασης Υπολογιστών και έχει κερδίσει έντονο ερευνητικό ενδιαφέρον τελευταία, χάρη στις πολλαπλές θεμελιώδεις εφαρμογές του σε τομείς όπως η ρομποτική, η οπτική επιτήρηση, η αλληλεπίδραση ανθρώπου-υπολογιστή και η ανάκτηση πολυμέσων. Ως αναγνώριση δράσεων ορίζεται η διαδικασία της ονοματοδοσίας των ανθρώπινων δράσεων με τη χρήση αισθητηριακών παρατηρήσεων και συνιστά εννοιολογικά ένα πρόβλημα ταξινόμησης. Στα πλαίσια της παρούσας διπλωματικής εργασίας θα ασχοηθούμε με την αυτόματη αναγνώριση δράσεων καθώς και την αναγνώριση συνεχόμενων δράσεων σε video, που προκύπτει αποκλειστικά από την πληροφορία του οπτικού καναλιού, δηλαδή των οπτικών παρατηρήσεων σε ακολουθίες εικόνων.

Μια ανθρώπινη δράση μπορεί να ιδωθεί ως μια αλληλουχία κινήσεων που εκτελεί ο δρων-υποκείμενο κατά την εκτέλεση μιας εργασίας. Υπό αυτόν τον ορισμό

η ανθρώπινη δράση συντίθεται από επιμέρους ανθρώπινες κινήσεις και ως έννοια περιλαμβάνει τη συνολική κίνηση ολόκληρου του σώματος. Το πρόβλημα της αναγνώρισης δράσεων αναφέρεται στην απόδοση στη δράση μιας ετικέτας-ονόματος που δύναται, ακόμα και στη μορφή ενός απλού ρήματος, να την περιγράψει καλύτερα, ανεξάρτητα από τις μεταβολές στην εμφάνιση των δρώντων, το ρυθμό εξέλιξής της, το περιβάλλον στο οποίο εκτυλίσσεται, την οπτική γωνία λήψης ή τις συνθήκες εγγραφής. Αποτελεί αφενός ένα πρόβλημα διαχωρισμού, αφού χρειάζεται να διαχωρίσουμε τις δράσεις ανάλογα με τη φύση τους, και αφετέρου ένα πρόβλημα γενίκευσης, που αφορά εκτελέσεις της ίδιας δράσης με μεγάλη μεταβλητότητα.

Σύμφωνα με τον Poppe [48], υπάρχουν δύο κατηγορίες αναπαράστασης μιας εικόνας ή μιας ακολουθίας εικόνων (video): οι τοπικές (local representations) και οι ολικές αναπαραστάσεις (global representations). Οι ολικές αναπαραστάσεις βασίζονται σε πρώτο στάδιο σε τεχνικές αφαίρεσης background ή στον εντοπισμό ανθρώπινης μορφής σε εικόνες, ορίζοντας έτσι την περιοχή ενδιαφέροντος. Οι κωδικοποιημένες αναπαραστάσεις εικόνων υπολογίζονται επί του συνόλου της περιοχής ενδιαφέροντος και προκύπτουν από χαρακτηριστικά σχήματος (ανθρώπινες σιλουέτες, ακμές) ή μετρήσεις οπτικής ροής. Κάποιες προσεγγίσεις επιλέγουν μια χωρική δομή πλέγματος εισάγοντας έτσι ένα τοπικό επίπεδο κωδικοποίησης της παρατήρησης σε επιμέρους τμήματα-κελιά των εικόνων με στόχο να μετριάσουν την ευαισθησία της ολικής αναπαράστασης σε παράγοντες όπως ο θόρυβος, η γωνία λήψης και τα οπτικά εμπόδια. Οι ολικές αναπαραστάσεις συνιστούν μια πλούσια κωδικοποίηση της περιοχής ενδιαφέροντος με χαρακτηριστικά περιγραμμάτων των ανθρώπων ή οπτικής ροής. Ωστόσο η εφαρμογή τους απαιτεί τον ακριβή εντοπισμό των ανθρώπων ή την αφαίρεση του παρασκηνίου και κατά συνέπεια καθίστανται ευάλωτες σε παράγοντες όπως η μερική απόκρυψη ανθρώπινης φιγούρας, η οπτική γωνία και ο θόρυβος.

Οι τοπικές αναπαραστάσεις έχουν κερδίσει έντονα το ερευνητικό ενδιαφέρον τα τελευταία χρόνια χάρη και στην πρόσφατη επιτυχία τους σε άλλα προβλήματα Όρασης Υπολογιστών, όπως η αναγνώριση αντικειμένων, υφής και ανθρώπων σε ακίνητες εικόνες. Η λογική τους βασίζεται στην περιγραφή των αρχικών ακολουθιών εικόνων μέσω μιας συλλογής ανεξάρτητων τοπικών τεμαχίων περιγραφής. Η εξαγωγή χαρακτηριστικών συντίθεται στην ανίχνευση οπτικά σημαντικών χωροχρονικών σημείων ενδιαφέροντος στο δείγμα video και στην μετέπειτα περιγραφή της χωροχρονικής γειτονιάς τους με χαρακτηριστικά εμφάνισης και κίνησης. Η τελική αναπαράσταση του αρχικού τρισδιάστατου όγκου video προκύπτει συνήθως από την εμπειρική στατιστική κατανομή σε αυτό προτύπων τοπικών τεμαχίων που έχουν εξαχθεί από ένα σετ εκπαίδευσης, χωρίς να διατηρείται η χωροχρονική πληροφορία των συντεταγμένων των χαρακτηριστικών. Για το λόγο αυτό δεν απαιτούνται τεχνικές αφαίρεσης του

background και απομόνωσης της περιοχής ενδιαφέροντος, όπως συμβαίνει στις ολικές αναπαραστάσεις, ενώ οι μέθοδοι αυτές έχουν επιδείξει μεγάλη ανοχή στον θόρυβο. Η παρούσα διπλωματική εργασία μελετά το πρόβλημα της αναγνώρισης ανθρώπινων δράσεων σε βίντεο υπό το πρίσμα τέτοιων τοπικών αναπαραστάσεων, οι οποίες θα αναλυθούν εκτενέστερα στην Ενότητα 2.1.

1.3 Οργάνωση του περιεχομένου της διπλωματικής

Το περιεχόμενο της διπλωματικής είναι οργανωμένο σε 7 κεφάλαια ως εξής:

- Στο **Κεφάλαιο 2** περιγράφονται συνοπτικά αρκετές μέθοδοι της βιβλιογραφίας σχετικά με τα θέματα αναγνώρισης και ταξινόμησης ανθρώπινων δράσεων σε video. Σκοπός του κεφαλαίου δεν είναι να αναπτυχθούν διεξοδικά οι λεπτομέρειες της κάθε τεχνικής αλλά να επισημανθούν τα κύρια στάδια των μεθόδων, καθώς και οι εφαρμογές τους. Ιδιαίτερη έμφαση δίνεται σε τεχνικές που βασίζονται στη χρήση των τοπικών χαρακτηριστικών και σε ανιχνευτές τοπικών χωροχρονικών σημείων ενδιαφέροντος.

- Στο **Κεφάλαιο 3** γίνεται αναλυτική παρουσίαση δύο νέων μεθόδων ανίχνευσης τοπικών χωροχρονικών σημείων ενδιαφέροντος που στηρίζονται σε Gabor φίλτρα και ανάλυση κυρίαρχης Teager-Kaiser ενέργειας. Αρχικά παρέχουμε συνοπτικά το απαραίτητο μαθηματικό υπόβαθρο σχετικά με τον τελεστή ενέργειας Teager-Kaiser και τις επεκτάσεις του και στη συνέχεια αναλύουμε δύο ανιχνευτές, τον ανιχνευτή TDE που στηρίζεται σε μονοδιάστατα χρονικά φίλτρα Gabor και στον ανιχνευτή Gabor3D που χρησιμοποιεί φιλτράρισμα στις 3 διαστάσεις, στις 2 χωρικές και στη χρονική. Αναλύονται διεξοδικά οι επιλογές παραμέτρων που κάναμε κατά την ανάπτυξη του αλγόριθμου Gabor3D και οι τεχνικές που εφαρμόσαμε για να μειώσουμε την πολυπλοκότητά του.

- Στο **Κεφάλαιο 4** αναλύονται οι τεχνικές που χρησιμοποιήθηκαν για να φτάσουμε από τα χωροχρονικά σημεία ενδιαφέροντος στην αναγνώριση και ταξινόμηση ανθρώπινων δράσεων. Συγκεκριμένα αναλύθηκαν οι περιγραφητές που χρησιμοποιήθηκαν για την αναπαράσταση των σημείων ενδιαφέροντος, δηλαδή οι HOG/HOF και HOG3D καθώς και η διαδικασία δημιουργίας των Bag-of-Features ιστογραμμάτων συνδυάζοντας την πληροφορία από όλους τους περιγραφητές. Γίνεται μια σύντομη εισαγωγή σε διάφορες τεχνικές ταξινόμησης των εν λόγω ιστογραμμάτων χρησιμοποιώντας διάφορους ταξινομητές που στηρίζονται σε διαφορετικές φιλοσοφίες. Συγκεκριμένα αναλύεται ο τρόπος λειτουργίας των Μηχανών

Διανυσμάτων Υποστήριξης και του ταξινομητή k-NN, η δύναμη των Κρυφών Μαρκοβιανών Μοντέλων και τα προβλήματα που επιλύονται με τη χρήση των παραπάνω τεχνικών.

- Στο **Κεφάλαιο 5** παρατίθενται τα πειράματα ταξινόμησης ανθρώπινων δράσεων που εκπονήθηκαν κατά την παρούσα διπλωματική εργασία. Γίνεται αρχικά ανάλυση του τρόπου με τον οποίο επιτυγχάνεται συνδυασμός των εργαλείων που παρουσιάστηκαν στο Κεφάλαιο 4 για να επιτευχθεί η διαδικασία ταξινόμησης, με τη χρήση των διάφορων ταξινομητών. Παρουσιάζουμε τα αποτελέσματά μας σε δύο πολύ δημοφιλείς βάσεις δεδομένων ανθρώπινων δράσεων, την KTH Action Dataset και την Hollywood2 Action Dataset. Παραθέτουμε τα αποτελέσματα διάφορων μεμονομένων πειραμάτων που έγιναν με σκοπό να καθοριστούν διάφορες παράμετροι των αλγορίθμων μας, καθώς και μεγαλύτερης κλίμακας πειράματα με σκοπό να συγκρίνουμε τις μεθόδους μας με τις μέθοδους της βιβλιογραφίας (οι οποίες παρουσιάζονται στο Κεφάλαιο 2). Τα αποτελέσματα των πειραμάτων δείχνουν στις περισσότερες περιπτώσεις ότι οι νέοι ανιχνευτές που αναπτύχθηκαν ξεπερνούν σε ακρίβεια ταξινόμησης τους ήδη υπάρχοντες ανιχνευτές της βιβλιογραφίας. Σε κάθε πείραμα παραθέτουμε το framework, τα αποτελέσματα καθώς και τα συμπεράσματα στα οποία καταλήξαμε.

- Στο **Κεφάλαιο 6** παρατίθενται τα πειράματα που εκπονήθηκαν σε μια καινούρια πολυτροπική και πολυαισθητηριακή βάση δεδομένων, τη βάση MOBOT. Αναλύονται οι διαφορές μεταξύ της απλής ταξινόμησης δράσεων και της αναγνώρισης συνεχόμενων ανθρώπινων δράσεων σε video και γίνεται εκτενής περιγραφή των μεθόδων που χρησιμοποιήθηκαν στα πειράματα αναγνώρισης. Ακόμα, εκτός από την πληροφορία που προέρχεται από το RGB κανάλι της κάμερας, εκμεταλλευόμαστε και το κανάλι βάθους (depth) για να διευκολύνουμε την αναγνώριση κάνοντας χρήση πολύ πιο εστιασμένων σημείων ενδιαφέροντος για το σύστημά μας. Εισάγουμε το Background μοντέλο για τις χρονικές στιγμές ενός βίντεο όπου δεν εκτελείται κάποια ανθρώπινη δράση. Τέλος, παραθέτουμε τα αποτελέσματα των πειραμάτων ταξινόμησης και αναγνώρισης καθώς και τα συμπεράσματά μας.

- Στο **κεφάλαιο 7** παρουσιάζονται τα συμπεράσματα που προκύπτουν από το σύνολο της διπλωματικής και συνοψίζονται οι επιστημονικές συνεισφορές της. Επίσης, αναφέρονται και κάποιες μελλοντικές κατευθύνσεις καθώς και προεκτάσεις της.

Κεφάλαιο 2

Σχετική Έρευνα για το Πρόβλημα της Αναγνώρισης και Ταξινόμησης Δράσεων σε Video

2.1 Τοπικά Χωροχρονικά Χαρακτηριστικά

Τα τοπικά χαρακτηριστικά (local features) έχουν δείξει τεράστια επιτυχία σε διάφορα προβλήματα αναγνώρισης της Όρασης Υπολογιστών, όπως η αναγνώριση αντικειμένων ή σκηνών [33], και τα τελευταία χρόνια στο πρόβλημα της Αναγνώρισης Ανθρώπινων Δράσεων σε Video [60]. Οι τοπικές αναπαραστάσεις περιγράφουν την παρατήρηση με μια σειρά από τοπικούς περιγραφητές ή τεμάχια που υπολογίζονται στη γειτονιά ανιχνευθέντων σημείων ενδιαφέροντος. Τα τοπικά αυτά τεμάχια συνήθως δε διατηρούν τοπική ή χρονική πληροφορία με άμεσο τρόπο, αλλά αναπαριστούν μια εικόνα (ή ένα video) ως μια συλλογή από διάφορα ήδη χαρακτηριστικών τα οποία θεωρούνται σημαντικά μέσω κάποιου συγκεκριμένου κανόνα και τη χαρακτηρίζουν με ικανοποιητικό τρόπο. Τελικά, η συλλογή των τοπικών χαρακτηριστικών ενσωματώνεται σε μια τελική αναπαράσταση ικανή να συλλάβει τη στατιστική κατανομή τους και να προχωρήσει στα επόμενα στάδια της αναγνώρισης.

Η αναπαράσταση μέσω τοπικών χαρακτηριστικών έχει επικρατήσει και στην αναγνώριση ανθρώπινων δράσεων, όπου γίνεται μια επιλογή από δεδομένα που αφ' ενός μειώνουν κατά πολύ τη διάστασή των video και αφ' ετέρου τα μετασχηματίζουν σε μια αναπαράσταση που τα κάνει κατηγοριοποιήσιμα. Τα τοπικά τεμάχια είναι δισδιάστατα ή τρισδιάστατα και εξάγονται απευθείας από το video εισόδου, συνήθως μεγιστοποιώντας έναν δείκτη σημαντικότητας (saliency function). Οι λόγοι που απλές ιδέες όπως τα τοπικά χωροχρονικά χαρακτηριστικά βιώνουν μια άνθηση στο χώρο της αναγνώρισης ανθρώπινων δράσεων αλλά και γενικότερα της αναγνώρισης

προτύπων είναι ότι εξασφαλίζουν αρκετά πλεονεκτήματα. Κατ' αρχήν ανιχνεύουν επιτυχώς χαρακτηριστικά της κίνησης και της εμφάνισης των εικόνων ή video στα οποία εφαρμόζονται. Σε αντίθεση με τις ολικές προσεγγίσεις, οι τοπικές αναπαραστάσεις δεν απαιτούν σαφή ανίχνευση και εντοπισμό του ανθρώπου, διαχωρισμό προσκηνίου και παρασκηνίου, ούτε καμία άλλη top-down πληροφορία και μπορούν να εφαρμοστούν χωρίς κανέναν περιορισμό. Ακόμα, είναι ανθεκτικά στο θόρυβο και παρέχουν μια σχετικά ανεξάρτητη αναπαράσταση σε σχέση με χωροχρονικές μετατοπίσεις και κλίμακες.

Τα τοπικά χωροχρονικά χαρακτηριστικά αποτελούν το εργαλείο μέσω του οποίου μια εικόνα ή μια ακολουθία εικόνων αναπαρίσταται και περιγράφεται αποτελεσματικά. Ωστόσο, στο πρόβλημα της αναγνώρισης ανθρώπινων δράσεων έχει δημιουργηθεί η ανάγκη να βρεθούν τα σημεία από τα οποία θα εξαχθούν αυτά τα χωροχρονικά χαρακτηριστικά. Είναι κατανοητό ότι σε ένα μεγάλο όγκο δεδομένων όπως είναι ένα video, το να εξάγει κανείς•παντού τοπικά χαρακτηριστικά δεν είναι ούτε υπολογιστικά ούτε και πρακτικά αποτελεσματικό. Πρέπει λοιπόν να πραγματοποιηθεί ανίχνευση ή πυκνή δειγματοληψία χωροχρονικών σημείων τα οποία έχουν κάποια κοινή ιδιαιτερότητα, και ονομάζονται στη βιβλιογραφία τοπικά χωροχρονικά σημεία ενδιαφέροντος (local spatio-temporal interest points). Η ανίχνευση σημείων ενδιαφέροντος επιτυγχάνεται με τη χρήση κατάλληλων ανιχνευτών (detectors), τα κριτήρια ανίχνευσης των οποίων θα συζητηθούν παρακάτω. Από ένα video τελικά, γίνεται μέσω ενός ανιχνευτή η επιλογή των σημείων που υποθετικά το χαρακτηρίζουν αποτελεσματικά. Οι τοπικοί περιγραφητές υπολογίζουν το σχήμα και την κίνηση στη γειτονιά των επιλεγμένων σημείων, δηλαδή εφαρμόζονται σε μια περιοχή (επιφάνεια ή όγκος) η οποία εμπεριέχει το προς μελέτη σημείο. Αυτοί οι υπολογισμοί συχνά αφορούν τα 2D ή 3D gradients ή/και την οπτική ροή των συνεχόμενων frames. Η ανίχνευση και η περιγραφή των σημείων ενδιαφέροντος αποτελεί εννοιολογικά τα τοπικά χαρακτηριστικά.

2.2 Ανιχνευτές Τοπικών Χωροχρονικών Χαρακτηριστικών

Οι ανιχνευτές τοπικών χαρακτηριστικών αναζητούν χωροχρονικά σημεία και κλίμακες ενδιαφέροντος που αντιστοιχούν σε περιοχές που χαρακτηρίζονται από σύνθετη κίνηση ή απότομες μεταβολές στην εμφάνιση του video εισόδου μεγιστοποιώντας μια συνάρτηση οπτικής σημαντικότητας. Πολλοί ανιχνευτές έχουν επινοηθεί τα τελευταία χρόνια αντλώντας αραιά αλλά εύρωστα σημεία. Θεμελιώνονται συνήθως με βάση μια μαθηματική συνάρτηση απόκρισης της οποίας τα τοπικά μέγιστα αντιστοιχούν σε εξέχουσες περιοχές ενδιαφέροντος στο

χώρο και στο χρόνο. Οι διαφορές μεταξύ των ανιχνευτών έγκειται κυρίως στη μορφή της συνάρτησης σημαντικότητας που χρησιμοποιούν, τις δομές των σημείων που αναζητούν, την ποιότητα των χωρικών και χρονικών κλιμάκων που εξάγουν καθώς και την υπολογιστική τους πολυπλοκότητα. Παρακάτω θα παρουσιάσουμε τους κύριους ανιχνευτές χωροχρονικών σημείων ενδιαφέροντος που υπάρχουν στη βιβλιογραφία, με τη χρήση των οποίων έγιναν πειράματα και συγκρίθηκαν οι μέθοδοι της παρούσας διπλωματικής εργασίας. Οι περισσότερες από τις μεθόδους ανίχνευσης που θα αναλυθούν στη συνέχεια υπάρχουν υλοποιημένες στο διαδίκτυο και είναι εύκολα προσβάσιμες για απεύθειας σύγκριση των μεθοδολογιών μας. Για τους διάφορους περιγραφητές που εξάγονται από τα ανιχνευμένα σημεία ενδιαφέροντος θα αναφερθούμε εκτενέστερα στην Ενότητα 4.1.

2.2.1 Ο Ανιχνευτής Harris3D

Ο ανιχνευτής χωροχρονικών σημείων ενδιαφέροντος Harris3D είναι ίσως ο πιο δημοφιλής στη βιβλιογραφία [32]. Εισήχθη από τους Laptev και Lindeberg και αποτελεί επέκταση του ανιχνευτή γωνιών Harris [19] στις τρεις διαστάσεις. Ο ανιχνευτής Harris είναι πολύ γνωστός για την αποτελεσματική ανίχνευση γωνιών σε εικόνες. Η τρισδιάστατη έκδοσή του έχει τη δυνατότητα να ανιχνεύει, ακολουθώντας την ίδια λογική με την οποία ο δισδιάστατος Harris ανιχνεύει σημεία που επιδεικνύουν υψηλή μεταβολή των τιμών της εικόνας στο χώρο, σημεία που επιδεικνύουν και μη σταθερή κίνηση στο χρόνο. Για την ακρίβεια ο Harris3D πυροδοτεί ανιχνεύσεις σε περιοχές που πληρούν και τις δύο παραπάνω προϋποθέσεις, δηλαδή παρουσιάζουν διακριτική εμφάνιση στο χώρο και διέπονται από μη σταθερή κίνηση στο χρόνο. Οι συγγραφείς ισχυρίζονται ότι τέτοια χαρακτηριστικά αντιστοιχούν σε γεγονότα με έντονο πληροφοριακό περιεχόμενο στο video εισόδου.

Ο Harris3D ψάχνει στο video εισόδου για σημεία που μεγιστοποιούν μια συνάρτηση χωροχρονικής γωνιότητας. Έστω $f(x, y, t)$ η τιμές της εικόνας στις δύο χωρικές και στη χρονική διάσταση. Αν θεωρήσουμε ως τοπική χωρική κλίμακα της παρατήρησης την σ_l^2 και ως χωρική κλίμακα ολοκλήρωσης την $\sigma_i^2 = s \cdot \sigma_l^2$, όπου s μια σταθερά. Αντίστοιχα η τοπική κλίμακα και η κλίμακα ολοκλήρωσης που αφορούν το χρόνο τ_l^2 και $\tau_i^2 = s \cdot \tau_l^2$. Η ακολουθία εικόνων $f : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$ μοντελοποιείται με τη γραμμική αναπαράσταση κλίμακας - χώρου $L : \mathbb{R}^2 \times \mathbb{R} \times \mathbb{R}_+^2 \to \mathbb{R}$ συνελίσσοντας την $f(x, y, t)$ με το γκαουσιανό πυρήνα με χωρική και χρονική μεταβλητότητα αντίστοιχα σ_l^2 και τ_l^2

$$L\left(\cdot; \sigma_l^2, \tau_l^2\right) = L\left(x, y, t; \sigma_l^2, \tau_l^2\right) = g\left(\cdot, \sigma_l^2, \tau_l^2\right) * f \tag{2.1}$$

όπου $f = f(x, y, t)$ και για το χωροχρονικά διαχωρίσιμο γκαουσιανό πυρήνα

ισχύει:

$$g\left(\cdot,\sigma_l^2,\tau_l^2\right)=g\left(x,y,t,\sigma_l^2,\tau_l^2\right)=\frac{1}{\sqrt{(2\pi)^3\sigma_l^4\tau_l^2}}\exp\left(-\frac{(x^2+y^2)}{2\sigma_l^2}-\frac{t^2}{2\tau_l^2}\right) \quad (2.2)$$

Κατασκευάζουμε την παραθυροποιημένη μήτρα δεύτερων στιγμών

$$M=g\left(\cdot,\sigma_l^2,\tau_l^2\right)*\begin{pmatrix}L_x^2 & L_xL_y & L_xL_t\\ L_xL_y & L_y^2 & L_yL_t\\ L_xL_t & L_yL_t & L_t^2\end{pmatrix} \quad (2.3)$$

Για την αποφυγή σύγχυσης μεταξύ των κλιμάκων σ_l^2,τ_l^2 και σ_i^2,σ_i^2 σημειώνουμε ότι οι πρώτες είναι οι τοπικές κλίμακες, με τις οποίες υπολογίζονται οι ομαλές γκαουσιανές παράγωγοι L_x, L_y και L_t, ενώ οι τελευταίες είναι κλίμακες ολοκλήρωσης του γκαουσιανού πυρήνα $g\left(\cdot,\sigma_l^2,\tau_l^2\right)$ που επιτελεί averaging της μήτρας M. Αν συμβολίσουμε με λ_1,λ_2 και λ_3 τις ιδιοτιμές της μήτρας M, με $\lambda_1\leq\lambda_2\leq\lambda_3$, το κριτήριο γωνιότητας που ορίζουν οι Laptev και Lindeberg στον 3D χώρο ακολουθεί τη λογική του δισδιάστατου κριτηρίου:

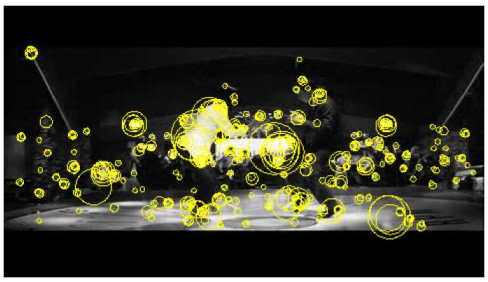
$$H=det(M)-k\cdot trace^3(M)=\lambda_1\cdot\lambda_2\cdot\lambda_3-k\cdot(\lambda_1+\lambda_2+\lambda_3)^3 \quad (2.4)$$

όπου $det(M)$ η ορίζουσα του πίνακα M και $trace(M)$ το ίχνος του. Για να δείξουν ότι μεγάλες τιμές των $\lambda_1,\lambda_2,\lambda_3$ μπορούν να αναζητηθούν στα τοπικά μέγιστα της συνάρτησης H αναδιατυπώνουν τη σχέση 2.4 ως εξής:

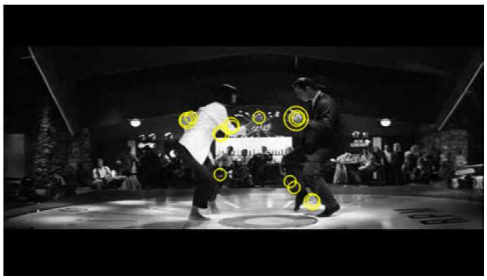
$$H=\lambda_1^3\left(\alpha\cdot\beta-k\cdot(1+\alpha+\beta)^3\right) \quad (2.5)$$

όπου $\alpha=\lambda_2/\lambda_1$ και $\beta=\lambda_3/\lambda_1$ και από την απαίτηση μη αρνητικότητας της 2.5 καταλήγουν πως $k\leq\alpha\cdot\beta/(1+\alpha+\beta)^3$. Όταν το k πλησιάζει στη μέγιστη τιμή του $k=1/27$ οι λόγοι α και β τείνουν στη μονάδα και επομένως τα τοπικά μέγιστα της H αντιστοιχούν σε περιοχές με μεγάλες μεταβολές τόσο στο πεδίο του χώρου όσο και στο πεδίο του χρόνου. Η συνάρτηση H αποτελεί το μετρικό σημαντικότητας του ανιχνευτή Harris3D με την έννοια ότι τα τοπικά χωροχρονικά μέγιστα αυτής αποκαλύπτουν χωροχρονικά σημεία της ακολουθίας εικόνων f με σημαντικό περιεχόμενο. Στο [31] οι ίδιοι συγγραφείς εισάγουν και μια μέθοδο ανάκτησης του χωροχρονικού περιεχομένου της κάθε ανίχνευσης, δηλαδή την αυτόματη επιλογή της χωρικής και χρονικής κλίμακας μέσω της εύρεσης των τοπικών μεγίστων του κανονικοποιημένου χωροχρονικού λαπλασιανού τελεστή και την προσαρμογή των θέσεων των ανιχνεύσεων βάσει των καινούριων εκτιμημένων τιμών των κλιμάκων. Ωστόσο, λόγω του γεγονότος ότι ο επαναληπτικός αλγόριθμος αποκλίνει σε ορισμένες περιπτώσεις, οι συγγραφείς εγκατέλειψαν την ιδέα της αυτόματης ανίχνευσης και υιοθέτησαν την ανίχνευση σε πολλαπλά επίπεδα προεπιλεγμένων χωρικών και χρονικών κλιμάκων [34]. Στο Σχήμα. 2.1

απεικονίζονται τα ανιχνευμένα σημεία που πληρούν τα κριτήρια γωνιότητας, στις 2 και στις 3 διαστάσεις αντίστοιχα. Παρατηρείται ότι τα σημεία που ανιχνεύονται από τον αλγόριθμο Harris3D να είναι περισσότερο εστιασμένα στις περιοχές της κάθε εικόνας που συντελείται η δράση. Για περισσότερες πληροφορίες για τις τεχνικές που χρησιμοποιούνται καλούμε τον αναγνώστη να ανατρέξει στα [31, 34], καθώς η αναλυτική παρουσίασή τους ξεφεύγει από τους σκοπούς της παρούσας διπλωματικής εργασίας.



(α΄) Ανιχνευμένα σημεία ενδιαφέροντος (χωρικές γωνίες) από τον ανιχνευτή Harris



(β΄) Ανιχνευμένα σημεία ενδιαφέροντος (χωροχρονικές γωνίες) από τον ανιχνευτή Harris3D

Σχήμα 2.1: Τα σημεία ενδιαφέροντος για τον ανιχνευτή Harris και την επέκτασή του στις τρεις διαστάσεις, Harris3D. Φαίνεται ο Harris3D να δίνει ως αποτέλεσμα πιο εστιασμένα σημεία στις δράσεις, και για αυτό το λόγο να αποτελεί καλύτερη επιλογή για το πρόβλημα της αναγνώρισης ανθρώπινων δράσεων σε video (επανεκτύπωση από το [41]).

Ο ανιχνευτής αυτός χρησιμοποιήθηκε και στην παρούσα διπλωματική εργασία ως μέτρο σύγκρισης με τους αλγόριθμους που αναπτύχθηκαν από εμάς. Για την εκπόνηση των πειραμάτων χρησιμοποιήθηκε η online διαθέσιμη έκδοση[1] που παρέχεται από την ιστοσελίδα του πρώτου συγγραφέα. Για τις παραμέτρους του Harris3D που χρησιμοποιήθηκαν στα πειράματά μας ακολουθήθηκαν οι προτεινόμενες τιμές των συγγραφέων. Για τις τιμές των τοπικών χωρικών και χρονικών κλιμάκων (σ_l^2, τ_l^2) ανίχνευσης επιλέχτηκαν οι τιμές $\sigma_{l,i} = 2^{(1+i)/2}, i = 1, ..., 6$ και $\tau_{l,i} = 2^{i/2}, i = 1, 2$ καταλήγοντας έτσι σε δεκαέξι συνδυασμούς χωρικών και χρονικών κλιμάκων στις οποίες αναζητήθηκαν οι ανιχνεύσεις, σύμφωνα με το [34]. Η παράμετρος k της συνάρτησης H της σχέσης (2.5) τέθηκε στην τιμή $k = 5 \cdot 10^{-4}$ ενώ το κατώφλι για την απόρριψη αδύναμων ανιχνεύσεων στην τιμή 10^{-9}. Τέλος για την αποφυγή επίπλαστων ανιχνεύσεων στα όρια των frames απορρίφθηκαν ανιχνεύσεις που βρίσκονται σε απόσταση μέχρι και πέντε pixels από τις συντεταγμένες ορίων κάθε frame.

[1] `http://www.di.ens.fr/~laptev/download.html`

2.2.2 Ο Ανιχνευτής Cuboid

Αντίθετα με τους Laptev et al., οι Dollár et al. ισχυρίζονται ότι απλές τρισδιάστατες επεκτάσεις των κοινών δισδιάστατων ανιχνευτών σημείων ενδιαφέροντος είναι ακατάλληλες για την εξαγωγή ποιοτικών τοπικών χωροχρονικών χαρακτηριστικών [12]. Εισάγουν τον ανιχνευτή Cuboid, που βασίζεται σε χρονικά Gabor φίλτρα. Τα τοπικά μέγιστα της συνάρτησης απόκρισης στην οποία θεμελιώνουν τον ανιχνευτή, αντιστοιχούν σε περιοχές με διακριτικά χαρακτηριστικά στο χώρο, τα οποία περιέχουν περιοδικούς συντελεστές συχνότητας ή γενικότερα υφίστανται μια σύνθετη κίνηση. Ο Cuboid δε συλλαμβάνει περιοχές που διέπονται από απλή μεταγραφική κίνηση ούτε περιοχές που δεν εμπεριέχουν διακριτές μεταβολές στην ένταση της εικόνας και απαιτεί την επιλογή χωρικής και χρονικής κλίμακας από το χρήστη. Αποτελεί έναν από τους δημοφιλέστερους ανιχνευτές αραιών χωροχρονικών σημείων ενδιαφέροντος σε ακολουθίες εικόνων και συχνά χρησιμοποιείται ως μηχανή παραγωγής σημαντικών σημείων για την αξιολόγηση τοπικών περιγραφητών λόγω της ευρωστίας του.

Ένα στοιχείο για τον ανιχνευτή Cuboid είναι ότι δεν αποτελεί επέκταση κάποιου χωρικού δισδιάστατου ανιχνευτή, καθώς οι συγγραφείς του ισχυρίζονται πως μια τέτοια προσέγγιση είναι εσφαλμένη. Οι ίδιοι παρατηρούν πως οι ανιχνεύσεις του Harris3D είναι αποτελεσματικές μόνο σε περιπτώσεις ανθρώπινων δράσεων που χαρακτηρίζονται ικανοποιητικά από την αντιστροφή της κατεύθυνσης κίνησης των χεριών και των ποδιών. Τέτοιες είναι οι κατηγορίες δράσεων της βάσης δεδομένων KTH Action Dataset, όπου ο Harris3D ξεπερνά τον Cuboid σε ποσοστά ακρίβειας αναγνώρισης [60]. Ο Cuboid στηρίζεται πάνω σε χρονικό φιλτράρισμα του video εισόδου με ένα ζευγάρι τετραγωνισμού (quadrature pair) δύο Gabor φίλτρων, αφού το ίδιο έχει πρώτα υποστεί εξομάλυνση στις χωρικές διαστάσεις μέσω ενός γκαουσιανού πυρήνα. Η συνάρτηση απόκρισης του Cuboid είναι:

$$R = (I * g * h_{ev})^2 + (I * g * h_{odd})^2 \tag{2.6}$$

όπου $g = g(x, y; \sigma)$ είναι ο δισδιάστατος γκαουσιανός πυρήνας εξομάλυνσης που εφαρμόζεται στις δύο διαστάσεις που αφορούν το χώρο, $I = I(x, y, t)$ είναι η ακολουθία εικόνων και h_{ev} και h_{odd} είναι το ζευγάρι χρονικών Gabor φίλτρων σε quadrature που εφαρμόζονται στο χρόνο, και δίνονται από τις σχέσεις:

$$h_{ev} = -\cos(2\pi t\omega)\, e^{-t^2/\tau^2} \tag{2.7}$$

$$h_{odd} = -\sin(2\pi t\omega)\, e^{-t^2/\tau^2} \tag{2.8}$$

$$\tag{2.9}$$

Για την κυκλική συχνότητα ω των φίλτρων Gabor οι συγγραφείς προτείνουν τη σχέση $\omega = 4/\tau$ αφήνοντας για τη συνάρτηση R δύο παραμέτρους που απαιτούν ορισμό, τις σ και τ, που προσεγγιστικά συσχετίζονται με τη χωρική και χρονική

κλίμακα ανίχνευσης της μεθόδου τους. Γενικότερα, οι περιοχές που πυδοροτούν ανιχνεύσεις του Cuboid ικανοποιούν δύο κριτήρια: διακριτικά χαρακτηριστικά στο χώρο και σύνθετη κίνηση στο χρόνο. Διαθέσιμη έκδοση του ανιχνευτή υπάρχει online[2] και προσφέρεται για περεταίρω πειραματισμό.

2.2.3 Ο Ανιχνευτής Hessian

Ο ανιχνευτής Hessian προτάθηκε από τους Willems et al.[63] και αποτελεί και αυτός με τη σειρά του τρισδιάστατη επέκταση του μετρικού σημαντικότητας Hessian που χρησιμοποιήθηκε για την ανίχνευση δομών "σταγόνας" (blobs) σε εικόνες. Οι ανιχνεύσεις αντιστοιχούν στα τοπικά μέγιστα της ορίζουσας της τρισδιάστατης μήτρας Hessian ενώ οι χωρικές και η χρονική κλίμακα επιλέγονται αυτόματα χωρίς τη χρήση επαναληπτικού σχήματος. Για την επιτάχυνση της υλοποίησης γίνεται χρήση των ολοκληρωτικών video (integral videos) και η ορίζουσα της 3D Hessian υπολογίζεται σε διάφορες οκτάβες καθεμία από τις οποίες αποτελείται από πέντε διακριτές χωρικές ή χρονικές κλίμακες. Μετά τον υπολογισμό όλων των κυβοειδών ενεργοποιείται ένας αλγόριθμος καταστολής των μη μεγίστων για την ανάκτηση των μεγίστων στον χώρο των πέντε διαστάσεων που αποτελείται από τις συντεταγμένες (x, y, t) και τις κλίμακες (σ^2, τ^2). Για περισσότερες λεπτομέρειες ο αναγνώστης παραπέμπεται στο [63], ενώ μια έκδοση του ανιχνευτή υπάρχει και online[3].

2.2.4 Ο Ανιχνευτής DCA3D

Έχοντας μελετήσει και χρησιμοποιήσει μεταξύ άλλων και τους ανιχνευτές που παρατέθηκαν προηγουμένως σε αυτήν την Ενότητα, οι Georgakis et al. προχώρησαν στη δημιουργία ενός ανιχνευτή χωροχρονικών σημείων ενδιαφέροντος, που συνδυάζει ιδέες από φιλτράρισμα Gabor σε πολλαπλές κλίμακες με Ανάλυση Κυρίαρχης Συνιστώσας (Dominant Component Analysis) [17]. Ο αλγόριθμός τους, ο DCA3D, ο οποίος αποτέλεσε και πηγή άντλησης ιδεών των νέων αλγορίθμων που αναπτύχθηκαν στην παρούσα διπλωματική εργασία, στηρίχτηκε σε δύο διακριτά στάδια επεξεργασίας. Στο πρώτο στάδιο το video εισόδου υφίσταται χωρικό φιλτράρισμα από μια συστοιχία 40 φίλτρων Gabor, πανομοιότυπη με εκείνη των Havlicek et al. [22]. Στις φιλτραρισμένες εξόδους εφαρμόζεται Ανάλυση Κυρίαρχης Συνιστώσας [22] σύμφωνα με την οποία η κυρίαρχη χωρική συνιστώσα για κάθε pixel αποτελείται από τη μέγιστη τιμή ανάμεσα στις 40 φιλτραρισμένες με τα χωρικά φίλτρα εξόδους. Ο προκύπτων όγκος κυρίαρχης χωρικής συνιστώσας θα υποστεί σε ένα δεύτερο στάδιο φιλτράρισμα με

[2] http://vision.ucsd.edu/~pdollar/research.html
[3] http://homes.psat.kuleuven.be/~gwillems/research/Hes-STIP/

μια συστοιχία μονοδιάστατων χρονικών Gabor φίλτρων, για τη μορφή της οποίας οι συγγραφείς εμπνεύστηκαν από το [11].

Στις χωροχρονικά πλέον φιλτραρισμένες εξόδους από όλα τα χρονικά φίλτρα του video εισόδου υπολογίζεται ως μετρικό σημαντικότητας του ανιχνευτή DCA3D η χρονική Teager - Kaiser ενέργεια. Ο τελικός χάρτης ενέργειας συνίσταται από τη μέγιστη τιμή των ενεργειών που προήλθαν από το χρονικό φιλτράρισμα (με τα 5 χρονικά φίλτρα), για κάθε pixel. Επάνω σε αυτόν το χάρτη ενέργειας οι Georgakis et al. έψαξαν τα χωροχρονικά σημεία ενδιαφέροντος, τα οποία προήλθαν από την ανίχνευση των τοπικών μεγίστων στις 3 διαστάσεις του. Χρησιμοποιήθηκε η μέθοδος της καταστολής των μη μεγίστων σε ένα ποσοστό της μέγιστης ανιχνευμένης ενέργειας ώστε να απορριφθούν οι ασθενείς εσφαλμένες ανιχνεύσεις. Οι κλίμακες ανίχνευσης τέθηκαν ίσες με τις κλίμακες των Gabor φίλτρων (χωρικών και χρονικών) για τις οποίες έχει επέλθει η συγκεκριμένη τιμή ενέργειας σε κάθε σημείο ενδιαφέροντος. Οι συγγραφείς επεκτάθηκαν σε πειράματα ταξινόμησης ανθρώπινων δράσεων χρησιμοποιώντας περιγραφητές HOG/HOF, αναπαράσταση Bag-Of-Features και ταξινομητές SVM με Γκαουσιανό πυρήνα, τα αποτελέσματα των οποίων δημοσιεύονται στο [17]. Ο ανιχνευτής DCA3D χρησιμοποιήθηκε ως μέτρο σύγκρισης και στα πειράματά μας, λόγω του ότι αποτελεί αποτέλεσμα επεξεργασίας καινοτόμων ιδεών για το πρόβλημα της αναγνώρισης δράσεων. Για την Teager - Kaiser ενέργεια, που χρησιμοποιήθηκε διεξοδικά και για τις ανάγκες της παρούσας διπλωματικής εργασίας θα γίνει ανάλυση στην υποενότητα 3.1.

2.3 Άλλες προσεγγίσεις για τη πρόβλημα της Αναγνώρισης Δράσεων

Το πρόβλημα της αναγνώρισης ανθρώπινων δράσεων σε video έχει αντιμετωπιστεί και με προσεγγίσεις που ξεφεύγουν από τις ιδέες ανίχνευσης τοπικών χαρακτηριστικών σε video (Ενότητα 2.2). Στις μεθόδους που θα αναφερθούν στη συνέχεια δεν θα επεκταθούμε τόσο όσο στους ανιχνευτές τοπικών χαρακτηριστικών, καθώς αφ' ενός δεν χρησιμοποιήθηκαν στα πειράματα της παρούσας διπλωματικής εργασίας και αφ' ετέρου δεν υπάρχει μέτρο σύγκρισης καθώς συνήθως οι συγγραφείς χρησιμοποιούν διαφορετικές βάσεις δεδομένων για να εκτελέσουν τα πειράματά τους, ενώ σε όσες μέθοδους υπάρχει πειραματισμός σε κοινή βάση δεδομένων, η πειραματική διαδικασία (experimental framework) είναι διαφορετική από τη δική μας.

Η πυκνή δειγματοληψία (Dense Sampling) είναι μέθοδος η οποία παρόλο που ανήκει σε αυτήν την παράγραφο σχετίζεται αρκετά με τις μεθόδους ανίχνευσης τοικών χαρακτηριστικών. Οι Wang et al. [60] το 2009 παρουσίασαν για πρώτη φορά μια εναλλακτική προσέγγιση που συνίσταται στην πυκνή δειγματοληψία

σε τακτές θέσεις και κλίμακες στο χώρο και στο χρόνο. Η ιδέα τους αντλεί έμπνευση από την πρόσφατη επιτυχία τεχνικών πυκνής δειγματοληψίας σε ακίνητες εικόνες για την αναγνώριση αντικειμένων. Σύμφωνα με τα πειράματα των συγγραφέων, αυτή η μέθοδος πυκνής εξαγωγής σημείων παράγει 15-20 φορές περισσότερα τοπικά χαρακτηριστικά ανά frame κατά μέσο όρο σε σχέση με τους γνωστούς ανιχνευτές αραιών χωροχρονικών σημείων. Οι συγγραφείς προτείνουν δειγματοληψία σε προκαθορισμένες κανονικές θέσεις και κλίμακες, συνολικά δηλαδή στο χώρο πέντε διαστάσεων (x, y, y, σ, τ) όπου (x, y, t) οι συντεταγμένες θέσης και (σ, τ) η χωρική και χρονική κλίμακα αντίστοιχα. Η επιτυχία της πυκνής δειγματοληψίας στο πρόβλημα της αναγνώρισης ανθρώπινων δράσεων σε video ενθαρρύνει την πρόοδο της έρευνας που αποσκοπεί την αύξηση της αποτελεσματικότητας μεθόδων που αποτελούν συμβιβασμό μεταξύ πυκνής δειγματοληψίας και ανίχνευσης σημείων ενδιαφέροντος. Πρόσφατα ερευνήθηκε για πρώτη φορά [57] μια υβριδική προσέγγιση συνδυασμού των πλεονεκτημάτων των σημείων ενδιαφέροντος και της πυκνής δειγματοληψίας σε ένα επιτυχημένο σχήμα που ονομάζεται Dense Interest Points.

Οι Sadanand και Corso στο [52] προτείνουν τη μέθοδο της Τράπεζας Δράσεων (Action Bank), μέσω της οποίας ισχυρίζονται πως ένα μεγάλο σετ από ανιχνευτές δράσεων μπορούν να αποτελέσουν βάση μιας αναπαράστασης πλούσιας σε semantic πληροφορία, δημιουργώντας template δράσεις. Ως ανιχνευτές δράσεων θεωρούν αναπαραστάσεις ενέργειας μέσω κατευθυντικών φίλτρων Γκαουσιανών παραγώγων. Κατασκευάζουν έναν "χώρο δράσεων", τον οποίο εκμεταλλεύονται για να βρουν συσχετίσεις μεταξύ των ανιχνευτών δράσεων σε διάφορες κλίμακες. Παρόλη τη μεγάλη επιτυχία της μεθόδου τους για το πρόβλημα της αναγνώρισης ανθρώπινων δράσεων σε video, η οποία είναι εντυπωσιακή, τα μειονεκτήματα της μεθόδου είναι πως το training set για τον πειραματισμό τους αποφασίζεται με χειροκίνητο τρόπο καθώς και ότι το σύστημα αναγνώρισης, λόγω της φιλοσοφίας του, που βασίζεται στην κατασκευή των template δράσεων και την εύρεση συσχετίσεων, υποφέρει από μεγάλη πολυπλοκότητα που οδηγεί σε μεγάλους χρόνους εκτέλεσης.

Οι Wang et al. [58] προσπαθούν να επεκτείνουν την ιδέα της πυκνής δειγματοληψίας κάνοντας παρακολούθηση στο χρόνο των σημείων ενδιαφέροντος (πυκνές τροχιές - dense trajectories) μέσω πυκνής οπτικής ροής. Εισάγουν επίσης τα Motion Boundary Histograms, περιγραφητές που στηρίζονται στη διαφορική οπτική ροή, που δείχνουν να συνεργάζονται πολύ καλά με τις πυκνές τροχιές που προτείνουν. Ο Wang και Schmid επεκτείνονται περεταίρω [59] παρουσιάζοντας έναν τρόπο να ισοσταθμίσουν μεταβολές που προκύπτουν από την κίνηση της κάμερας χρησιμοποιώντας χαρακτηριστικά SURF. Οι πολύ ενδιαφέρουσες και πολλά υποσχόμενες ιδέες τους απέδωσαν και πρακτικά, κερδίζοντας στο διεθνή διαγωνισμό THUMOS για ταξινόμηση πολλαπλών κατηγοριών δράσεων που έγινε

σε συνεργασία με το International Conference on Computer Vision - ICCV του 2013.

Κεφάλαιο 3

Ανιχνευτές Σημείων Ενδιαφέροντος Βασισμένοι σε Φίλτρα Gabor

Στο Κεφάλαιο 2 περιγράψαμε δημοφιλείς ανιχνευτές τοπικών χωροχρονικών σημείων ενδιαφέροντος, όπως ο Harris3D, ο Cuboid και ο Hessian. Στο παρόν κεφάλαιο θα περιγράψουμε δύο νέους ανιχνευτές οι οποίοι μελετώνται για πρώτη φορά για το πρόβλημα της αναγνώρισης ανθρώπινων δράσεων σε video και χρησιμοποιούν τεχνικές φιλτραρίσματος με πολλαπλά Gabor φίλτρα καθώς και ανάλυση κυρίαρχης Teager-Kaiser ενέργειας. Οι ανιχνεύσεις που παράγονται από τους αλγόριθμους που περιγράφονται, σε συνδυασμό με τις τεχνικές που αναλύονται στο Κεφάλαιο 4, θα αξιολογηθούν με πειραματισμό στο Κεφάλαιο 5 όπου θα γίνει σύγκριση των δύο νέων ανιχνευτών με τις υπάρχουσες μεθόδους της βιβλιογραφίας.

Θεωρούμε ότι στο σημείο αυτό, πριν συνεχίσουμε με την παρουσίαση των νέων ανιχνευτών θα ήταν δόκιμο να γίνει μια μελέτη των ιδεών του τελεστή ενέργειας Teager-Kaiser, πάνω στις οποίες βασίστηκε αρκετά η σχεδίασή τους.

3.1 Ο τελεστής Ενέργειας Teager-Kaiser

Ο μη γραμμικός ενεργειακός τελεστής Teager-Kaiser (TKEO) αναπτύχθηκε αρχικά από τους H. Teager και S. Teager [55] για το σκοπό της μη γραμμικής επεξεργασίας σημάτων φωνής. Ο TKEO βρήκε τη συστηματική καθιέρωση της ονομασίας και της χρήσης του από τον J.F Kaiser [25], στην εργασία του οποίου αποτέλεσε εργαλείο για την εξαγωγή της ενέργειας σημάτων απλών αρμονικών ταλαντωτών. Χρησιμοποιείται ευρέως για την αναπαράσταση της ενέργειας σημάτων AM-FM, διαμορφωμένου πλάτους $\alpha(t)$ και συχνότητας $\cos(\phi(t))$, που μοντελοποιούνται ως:

$$s(t) \equiv \alpha(t) \cdot \cos(\phi(t)) \tag{3.1}$$

Η εφαρμογή του ΤΚΕΟ πάνω στο σήμα της παραπάνω μορφής δίνεται από τον τύπο:

$$\Psi\left[s(t)\right] \equiv \left[s'(t)\right]^2 - s(t)s''(t) \tag{3.2}$$

Το ανάλογο του τελεστή για διακριτά σήματα κατά τους συγγραφείς ορίζεται διακριτοποιώντας τις παραγώγους ως:

$$\Psi_d\left[s[n]\right] \equiv s^2[n] - s[n-1]s[n+1] \tag{3.3}$$

Ο ΤΚΕΟ είναι μη γραμμικός, παραμένει αναλλοίωτος σε απλές μετατοπίσεις στο χρόνο και προϋποθέτει το γεγονός ότι το σήμα στο οποίο εφαρμόζεται έχει στενό συχνοτικό φάσμα [40], πράγμα το οποίο στους ανιχνευτές μας υλοποιείται μέσω ζωνοπερατών φίλτρων Gabor, όπως θα δούμε στη συνέχεια. Ο ενεργειακός τελεστής Ψ εφαρμοζόμενος στο σήμα $f(t)$ μπορεί να συλλάβει την ενέργεια της πηγής που παρήγαγε το σήμα ταλάντωσης αφού

$$\begin{aligned}\Psi\left[f(t)\right] &= \Psi\left[\alpha(t)\cos\left(\phi(t)\right)\right] \\ &= \Psi\left[\alpha(t)\cos\left(\int_0^t \omega_i(\tau)d\tau + \theta\right)\right] \\ &\approx \left[\alpha(t)\omega_i(t)\right]^2\end{aligned} \tag{3.4}$$

Οι Maragos και Bovik [39] επέκτειναν τον τελεστή σε μεγαλύτερης διάστασης σήματα. Μιας και στον αλγόριθμο Gabor3D που θα αναπτύξουμε χρησιμοποιούμε την 3D έκδοση του ΤΚΕΟ, παρκάτω παραθέτουμε τη μορφή της $\Phi(\cdot)$ η οποία ορίζεται ως:

$$\begin{aligned}\Phi(f) &\equiv \|\nabla f\|^2 - f\cdot\nabla^2 f \\ &= f_x^2 + f_y^2 + f_t^2 - f\cdot(f_{xx} + f_{yy} + f_{tt})\end{aligned} \tag{3.5}$$

όπου $f = f(x, y, t)$. Η διακριτή μορφή του ΤΚΕΟ, την οποία χρησιμοποιήσαμε στον πειραματισμό μας είναι το αποτέλεσμα της διακριτοποίησης των χωρικών και χρονικών παραγώγων και ορίζεται ως:

$$\begin{aligned}\Phi_d\left[f[x,y,t]\right] \equiv\ & 3f^2[x,y,t] \\ & - f[x-1,y,t]\cdot f[x+1,y,t] \\ & - f[x,y-1,t]\cdot f[x,y+1,t] \\ & - f[x,y,t-1]\cdot f[x,y,t+1]\end{aligned} \tag{3.6}$$

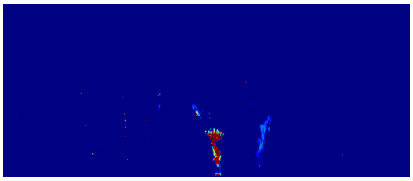
Το κίνητρο που έιχαμε για τη χρήση του ΤΚΕΟ είναι η ικανότητά του να εντοπίζει χωροχρονικές ταλαντώσεις της ενέργειας των σημάτων εισόδου και να τις διαχωρίζει σε συνιστώσες πλάτους και συχνότητας, με εξέχουσα χωροχρονική ανάλυση και πολύ μικρή πολυπλοκότητα. Περισσότερες πληροφορίες για τον τελεστή Teager-Kaiser και τη χρήση του μπορούν να αντληθούν από τα [55, 39, 40, 38], καθώς η λεπτομερής ανάλυσή του δεν είναι σκοπός της παρούσας διπλωματικής εργασίας.

3.2 Ανιχνευτής Βασισμένος σε 1D Φίλτρα Gabor - Ο Αλγόριθμος TDE

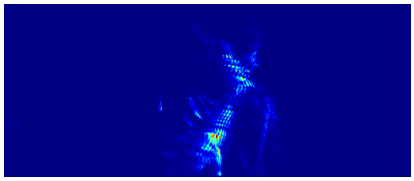
Ο ανιχνευτής TDE αναπτύχθηκε με βάση την πρόσφατη εργασία των Georgakis et al. [17] οι οποίοι εφήρμοσαν για πρώτη φορά τις ιδέες της Ανάλυσης Κυρίαρχων Συνιστωσών (Dominant Component Analysis) για το πρόβλημα της αναγνώρισης ανθρώπινων δράσεων. Ο αλγόριθμός τους, που περιγράφεται στην υποενότητα 2.2.4, στηρίχτηκε σε δύο διακριτά στάδια φιλτραρίσματος με Gabor φίλτρα. Σε πρώτο στάδιο εφαρμόζεται φιλτράρισμα σε κάθε frame του video εισόδου σε πολλαπλές ζώνες συχνοτήτων στο χώρο βάσει μιας συστοιχίας δισδιάστατων μιγαδικών φίλτρων Gabor. Αφού υπολογιστεί η χωρική Teager - Kaiser ενέργεια σε κάθε pixel, υπολογίζεται για αυτά η Κυρίαρχη Χωρική Συνιστώσα, δηλαδή η τιμή της φιλτραρισμένης εξόδου από το κανάλι που απέδωσε τη μέγιστη τιμή ενέργειας. Στη συνέχεια, στο δεύτερο στάδιο, ο όγκος κυρίαρχης (χωρικής) συνιστώσας φιλτράρεται από μια συστοιχία μονοδιάστατων χρονικών φίλτρων Gabor. Με τον ίδιο τρόπο υπολογίζεται η χρονική Teager - Kaiser ενέργεια για κάθε pixel. Η αναπαράσταση αυτή αποτελεί τον τελικό ενεργειακό χάρτη του video σύμφωνα με τον αλγόριθμο DCA3D, πάνω στον οποίο γίνεται η ανίχνευση των χωροχρονικών σημείων ενδιαφέροντος.

Ο ανιχνευτής TDE αποτελεί μια απλοποιημένη και πιο αποδοτική μορφή του DCA3D, όπου ουσιαστικά το πρώτο στάδιο του χωρικού φιλτραρίσματος και της εξαγωγής της Κυρίαρχης Χωρικής Συνιστώσας έχει παραληφθεί. Αντί για αυτό, χρησιμοποιήθηκαν μόνο μονοδιάστατα χρονικά φίλτρα για την επεξεργασία των video εισόδου. Ο λόγος που έγινε αυτή η απλούστευση είναι ότι η κυρίαρχη χωρική συνιστώσα του πρώτου σταδίου του ανιχνευτή DCA3D επιφέρει λόγω της φύσης της σε κάθε frame του video θόρυβο, ο οποίος ενισχύεται κατά το χρονικό φιλτράρισμα, δίνοντας ως αποτέλεσμα έναν θορυβώδη χάρτη ενέργειας ο οποίος με τη σειρά του οδηγεί σε λανθασμένες ανιχνεύσεις (false alarms). Ο αλγόριθμος TDE αντιθέτως, οδηγεί σε έναν χάρτη ενέργειας που έχει υψηλές τιμές μόνο στις περιοχές που χαρακτηρίζονται από έντονη κίνηση, όπως φάινεται στο Σχήμα 3.1β'. Στο Σχήμα 3.1 φαίνονται οι χάρτες ενέργειας και τα σημεία ενδιαφέροντος που προκύπτουν από τους δύο αλγόριθμους. Μπορούμε να παρατηρήσουμε πως τόσο η ενέργεια όσο και τα σημεία που ανιχνεύθηκαν με τον TDE αλγόριθμο είναι περισσότερο εστιασμένα στον κινούμενο άνθρωπο, και όχι στο background όπως στην περίπτωση του DCA3D.

Τα φίλτρα που χρησιμοποιήθηκαν διατάχτηκαν έτσι ώστε να καλύπτουν ολόκληρο το φάσμα των διακριτών συχνοτήτων όπως φαίνεται στο Σχήμα 3.2. Η συστοιχία φίλτρων είναι αυτή που χρησιμοποιήθηκε στο [17], με ιδέες που βασίζονται στην έρευνα των Dimitriadis και Maragos [11].



(α') Ενέργεια που ανιχνεύτηκε από τον αλγόριθμο DCA3D.



(β') Ενέργεια που ανιχνεύτηκε από τον αλγόριθμο TDE.

(γ') Χωροχρονικά σημεία ενδιαφέροντος ποτ αντιστοιχούν στην ενέργεια του αλγόριθμου DCA3D.

(δ') Χωροχρονικά σημεία ενδιαφέροντος ποτ αντιστοιχούν στην ενέργεια του αλγόριθμου TDE.

Σχήμα 3.1: Ενεργειακοί χάρτες για το ίδιο frame ενός video της Hollywood2 Action Database καθώς και οι ανιχνεύσεις χωροχρονικών σημείων ενδιαφέροντος για τους αλγόριθμους DCA3D και TDE. Οι λανθασμένες ανιχνεύσεις του DCA3D που οφείλονται στον όγκο κυρίαρχης χωρικής συνιστώσας δεν υπάρχουν στις ανιχνεύσεις του TDE.

Το μονοδιάστατο μη κανονικοποιημένο πραγματικό φίλτρο Gabor $g(t)$ έχει κρουστική απόκριση

$$g(t) = \exp(-\beta^2 t^2) cos(\omega_c t) \tag{3.7}$$

όπου β είναι η παράμετρος εύρους ζώνης και $\omega_c = 2\pi f_c$ η γωνιακή κεντρική συχνότητα του φίλτρου. Η πρώτη παράγωγος υπολογίζεται από τη σχέση:

$$\frac{dg(t)}{dt} = \left(-2\beta^2 t \cos(\omega_c t) - \omega_c \sin(\omega_c t)\right) \exp(-\beta^2 t^2) \tag{3.8}$$

ενώ η δεύτερη παράγωγος από τη σχέση:

$$\frac{d^2 g(t)}{dt^2} = \left(4\beta^2 \omega_c t sin(\omega_c t) - (4\beta^4 t^2 - 2\beta^2 - \omega_c^2) cos(\omega_c t)\right) \exp(-\beta^2 t^2) \tag{3.9}$$

Λόγω της αντιμεταθετικής ιδιότητας των τελεστών συνέλιξης και παραγώγισης [38], η έξοδος της φιλτραρισμένης συνιστώσας $s(t) = x(t) * g(t)$ από το μονοδιάστατο ενεργειακό τελεστή Teager-Kaiser δίνεται από τον τύπο:

$$\Psi\left[s(t)\right] = \left(x(t) * \frac{dg(t)}{dt}\right)^2 - \left(x(t) * g(t)\right)\left(x(t) * \frac{d^2 g(t)}{dt^2}\right) \tag{3.10}$$

Στην περίπτωση διακριτών σημάτων χρησιμοποιείται η διακριτή μορφή των σημάτων Gabor που δίνονται από τη σχέση:

$$g[n] = g(t)|_{t=nT} \tag{3.11}$$

όπου T είναι η περίοδος δειγματοληψίας, και $f_s = \frac{1}{T}$ η συχνότητα δειγματοληψίας που ισούται στην περίπτωσή μας με το frame rate του αρχικού video εισόδου. Οι Georgakis et al. αποφάσισαν στη χρήση 5 φίλτρων που οι κεντρικές συχνότητές τους εμπίπτουν στο πεδίο συχνοτήτων $[0, f_s/2]$. Τα φίλτρα που χρησιμοποιήθηκαν είναι κανονικοποιημένα μονοδιάστατα χρονικά φίλτρα Gabor που έχουν σταθερό εύρος ζώνης οκτάβας $B = 0.75$ οκτάβες και κεντρικές συχνότητες $f_{c,l}$ που ξεκινούν από τα $2Hz$ και έχουν σταθερό βήμα $1.5Hz$ (Σχήμα 3.2) για ένα video 25 frames ανά δευτερόλεπτο, το οποίο είναι κατά κανόνα το frame rate των δειγμάτων video των βάσεων δεδομένων πάνω στις οποίες έγινε πειραματισμός.

Οι παράμετροι β των φίλτρων προκύπτουν από τη σχέση:

$$\beta = K\frac{\pi f_{c,l}}{\sqrt{\ln 2}} \tag{3.12}$$

όπου $f_{c,l}$ η κεντρικές συχνότητες των 5 φίλτρων (για $l = 1, 2, ..., 5$) και K μια σταθερά που εξαρτάται από το εύρος ζώνης οκτάβας B ως εξής:

$$K = \frac{2^B - 1}{2^B + 1} \tag{3.13}$$

Αφού το video εισόδου υποστεί φιλτράρισμα με τα $L = 5$ χρονικά φίλτρα που προαναφέραμε, το επόμενο βήμα είναι να υπολογιστεί ο 3D όγκος της Χρονικά Κυρίαρχης Συνιστώσας E_{TDE}, σύμφωνα με τον τύπο:

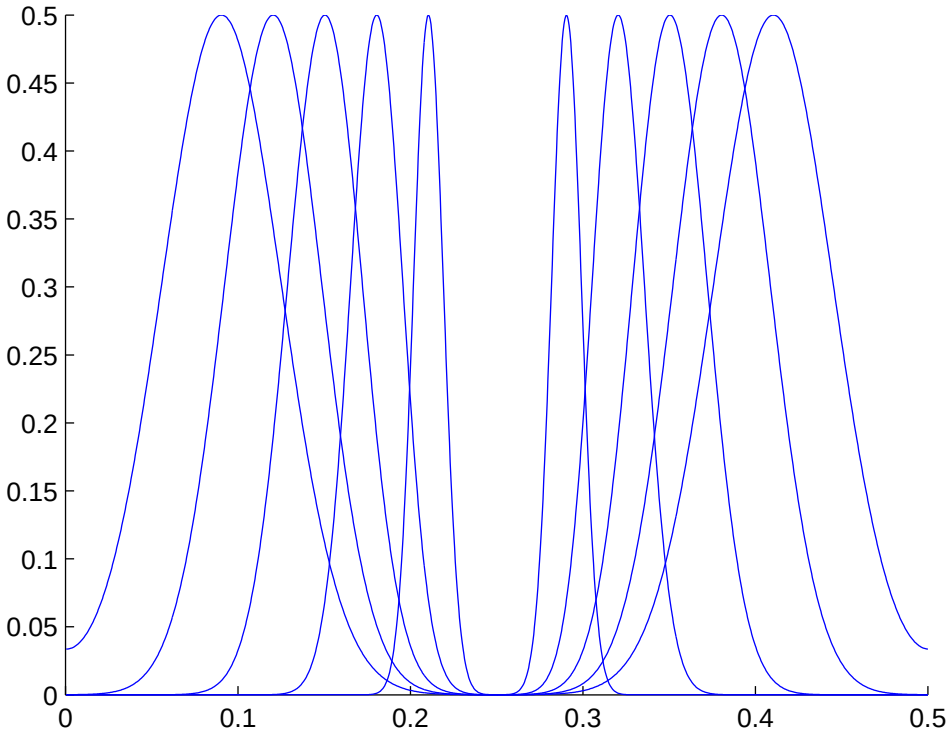
$$E_{TDE}(x, y, t) = \max_{0 \le l \le L} \Psi\left[V(x, y, t) * g_l(t)\right] \tag{3.14}$$

όπου l ο δείκτης του φίλτρου, $g_l(t)$ το l-οστό Gabor φίλτρο και $V(x, y, t)$ το αρχικό video εισόδου. Αντίστοιχα οι δείκτες που θα μας δώσουν το κυρίαρχο κανάλι της συστοιχίας των φίλτρων για κάθε pixel (x,y,t) δίνονται από τη σχέση:

$$i_{TDE}(x, y, t) = arg \max_{0 \le l \le L} \Psi\left[V(x, y, t) * g_l(t)\right] \tag{3.15}$$

Οι δείκτες που ανιχνεύτηκαν εν συνεχεία μας βοηθούν στον υπολογισμό της χρονικής κλίμακας ανίχνευσης (temporal scale) που δίνεται από την τυπική απόκλιση της Γκαουσιανής περιβάλλουσας του κυρίαρχου φίλτρου $i_{TDE}(x, y, t)$, πολλαπλασιασμένης επί τη συχνότητα δειγματοληψίας f_s, δηλαδή:

$$\tau_{TDE}(x, y, t) = round\left[\frac{1}{\sqrt{2}\beta_{i(x,y,t)}} \cdot f_s\right] \tag{3.16}$$



Σχήμα 3.2: Απόκριση συχνότητας της συστοιχίας μονοδιάστατων φίλτρων Gabor που χρησιμοποιήθηκε για τον ανιχνευτή TDE, για μοναδιαία τιμή της συχνότητας δειγματοληψίας.

Ο πολλαπλασιασμός με τη συχνότητα δειγματοληψίας (frame rate) γίνεται ώστε να λάβουμε την προσαρμοσμένη τιμή χρονικής κλίμακας σε frames ενώ η στρογγυλοποίηση γίνεται για να λάβουμε ακέραιες τιμές χρονικής κλίμακας, διαφορετικές για κάθε φίλτρο της συστοιχίας.

Τα χωροχρονικά σημεία ενδιαφέροντος που ανιχνεύονται στη συνέχεια για το πρόβλημα της αναγνώρισης ανθρώπινων δράσεων αναζητούνται στον τρισδιάστατο χάρτη ενέργειας που προκύπτει από την σχέση (3.14). Επιλέγουμε τα σημεία ενδιαφέροντος ως τα χωροχρονικά τοπικά μέγιστα σε μια γειτονιά $3 \times 3 \times 3$ του 3D όγκου της ενέργειας. Για να αποφύγουμε εσφαλμένες ανιχνεύσεις οι τιμές ενέργειας των ανιχνεύσεων κατωφλιώνονται σε ένα ποσοστό της μέγιστης τιμής της ενέργειας, σύμφωνα με την τεχνική καταστολής των μη μεγίστων (non maxima suppression). Στο σχήμα 3.3 φαίνεται η σημασία της κατωφλιοποίησης για την επιλογή χωροχρονικών σημείων ενδιαφέροντος που χαρακτηρίζουν την ανθρώπινη δράση.

(α′) (β′) (γ′)

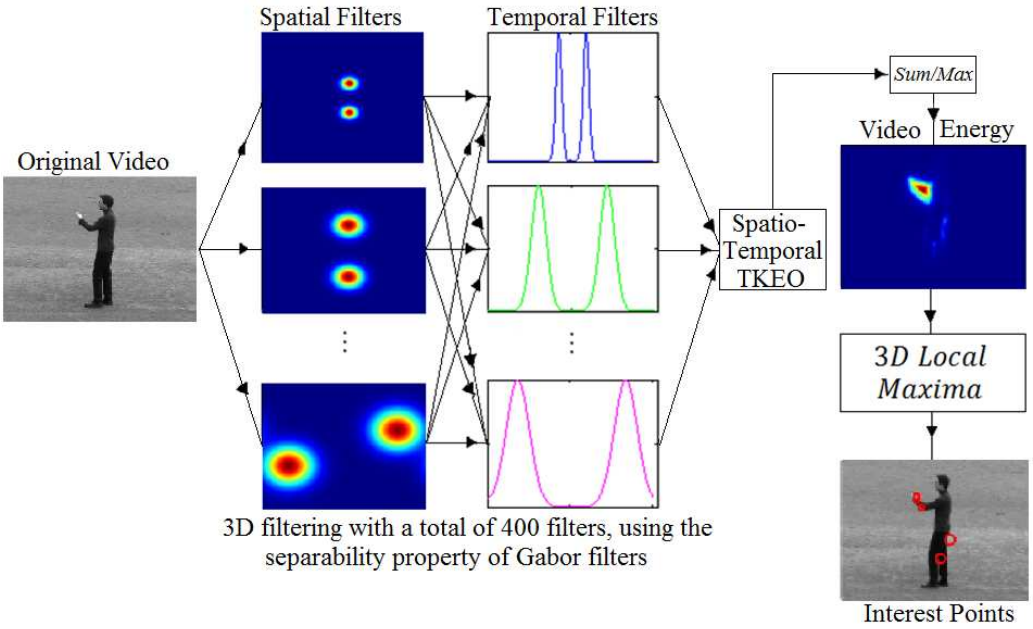
Σχήμα 3.3: Η σημασία της κατωφλιοποίησης για ένα video της KTH Action Dataset. Από αριστερά προς τα δεξιά: Ο χάρτης ενέργειας που προέκυψε από τον Ανιχνευτή TDE για ένα frame. Οι ανιχνεύσεις χωροχρονικών σημείων ενδιαφέροντος ως τα 3D τοπικά μέγιστα του χάρτη. Οι ανιχνεύσεις κατωφλιωμένες στο 7% της μέγιστης τιμής ενέργειας του video.

3.3 Ανιχνευτής Βασισμένος σε 3D Φίλτρα Gabor - Ο Αλγόριθμος Gabor3D

Ο ανιχνευτής Gabor3D αναπτύχθηκε ως μια γενικότερη περίπτωση των αλγορίθμων DCA3D (υποενότητα 2.2.4) και TDE. Η ιδέα στηρίχτηκε στην κατασκευή ενός κοινού frontend, τόσο για την προσέγγιση του προβλήματος της αναγνώρισης ανθρώπινων δράσεων σε video, όσο και της ανίχνευσης οπτικής προσοχής (Visual Saliency) [29]. Επίσης, το γεγονός ότι οι προαναφερθέντες αλγόριθμοι δεν αποδίδουν σε όλες τις περιπτώσεις καλύτερα από τους ήδη υπάρχοντες αλγόριθμους της βιβλιογραφίας - όπως θα δούμε στο Κεφάλαιο 5 - μας οδήγησε στην κατασκευή ενός αλγόριθμου που εφαρμόζει εξονυχιστικά διαδικασίες φιλτραρίσματος σε όλο το χωροχρονικό φάσμα συχνοτήτων. Οι ιδέες που θα παρουσιαστούν παρακάτω προήλθαν από διεξοδική έρευνα της βιβλιογραφίας και υλοποιήθηκαν με τη βοήθεια του ερευνητή Πέτρου Κούτρα, η συμβολή του οποίου ήταν το λιγότερο πολύτιμη.

Ο αλγόριθμος ανίχνευσης στηρίζεται σε φιλτράρισμα με Gabor φίλτρα, όπως και στην περίπτωση των DCA3D και TDE, τα οποία υποστηρίζονται από γνωσιακά πειράματα. Οι αναπαραστάσεις συχνότητας και κατεύθυνσης των εν λόγω φίλτρων λειτουργούν παρόμοια με το ανθρώπινο σύστημα όρασης, ενώ χρησιμοποιούνται ευρέως για αναπαραστάσεις και κατηγοριοποιήσεις υφής.Η διαφορά έγκειται στο ότι τα φίλτρα που χρησιμοποιούνται είναι τρισδιάστατα και στο ότι η διαδικασία του φιλτραρίσματος εφαρμόζεται ταυτόχρονα στις 2 χωρικές και τη μία χρονική διάσταση, ενώ η διάταξη των φίλτρων είναι τέτοια ώστε να καλύπτει όλο το χώρο συχνοτήτων. Αντίθετα με τον αλγόριθμο DCA3D όπου η διαδικασία του φιλτραρίσματος και της εξαγωγής της κυρίαρχης συνιστώσας χωρίζεται σε 2

διακριτά στάδια, το χωρικό και το χρονικό, στο νέο αλγόριθμό μας χρησιμοποιούμε τρισδιάστατο φιλτράρισμα με πολλαπλά φίλτρα σε ένα και μόνο στάδιο, και επεξεργασία των εξόδων ώστε να ληφθεί ο χάρτης ενέργειας του video εισόδου. Η διαδικασία φαίνεται στο Σχήμα 3.4. Αρχικά, τα αρχικά έγχρωμα frames του video μετατρέπονται σε γκρίζες εικόνες. Εν συνεχεία ακολουθεί η κύρια διαδικασία της επεξεργασίας, η Κυρίαρχη Χωρο-Χρονική Ανάλυση, η οποία εφαρμόζεται πάνω στον γκρίζο όγκο του video για να προκύψει ο τρισδιάστατος χάρτης ενέργειας, ο οποίος αποτελεί το μετρικό σημαντικότητας της ανίχνευσής μας (Saliency measure). Το τελευταίο στάδιο είναι το στάδιο της ανίχνευσης και εξαγωγής των σημείων ενδιαφέροντος, ως τα κατωφλιωμένα τοπικά μέγιστα του χάρτη ενέργειας, ακριβώς όπως περιγράφηκε στην Ενότητα 3.2.



Σχήμα 3.4: Διαδικασία εξαγωγής των χωροχρονικών σημείων ενδιαφέροντος με τον ανιχνευτή Gabor3D. Το video εισόδου φιλτράρεται από την συστοιχία των Gabor φίλτρων, ξεχωριστά για κάθε διάσταση, λόγω της ιδιότητας της διαχωρησιμότητας, και υπολογίζεται η ενέργειά του. Τα σημεία ενδιαφέροντος επιλέγονται ως τα τοπικά μέγιστα στις τρεις διαστάσεις, από τον όγκο ενέργειας που προκύπτει.

3.3.1 Λεπτομέρειες Υλοποίησης των τρισδιάστατων χωροχρονικών φίλτρων

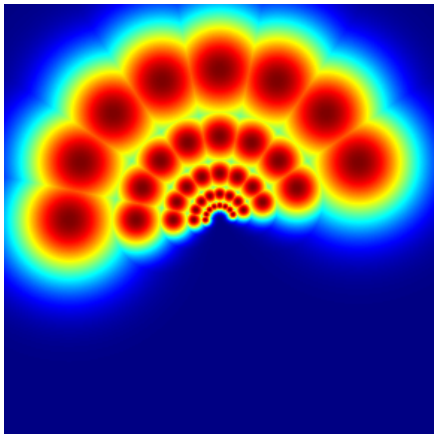
Το πρώτο στάδιο της Κυρίαρχης Χωρο-Χρονικής Ανάλυσης είναι το χωροχρονικό φιλτράρισμα του video εισόδου. Από τις διάφορες προσεγγίσεις φιλτραρίσματος

που έχουν προταθεί στη βιβλιογραφία και βασίζονται σε ψυχοφυσικά πειράματα, αυτές που έχουν κυριαρχήσει είναι δύο: τα φίλτρα Gabor και οι παράγωγοι Γκαουσιανών (Gaussian Derivatives). Επιλέξαμε να χρησιμοποιήσουμε μια τρισδιάστατη έκδοση προσανατολισμένων φίλτρων Gabor, τόσο λόγω της ομοιότητάς τους με το ανθρώπινο σύστημα όρασης, όσο και για την βελτιστιστη τιμή της αβεβαιότητας που παρουσιάζουν στο χώρο (ή χρόνο) και στη συχνότητα [15, 9]. Ακόμα, για μεγάλης τάξης παραγώγους, οι παράγωγοι Γκαουσιανών είναι προσεγγίσεις φίλτρων Gabor [28].

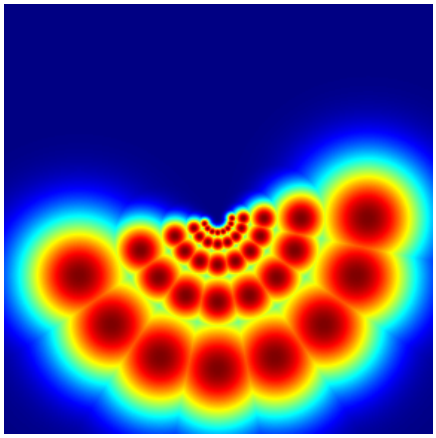
Τα φίλτρα Γκαουσιανών παραγώγων συνδυασμένα με το μετασχηματισμό Hilbert τους αποτελούν ένα ζευγάρι φίλτρων σε καθετότητα ή αλλιώς σε τετραγωνισμό, δηλαδή ένα ζευγάρι άρτιου και περιττού φίλτρου (quadrature pair). Τα ζεύγη φίλτρων σε τετραγωνισμό χρησιμοποιούνται ευρέως σε προβλήματα επεξεργασίας εικόνων και video [42, 10], κυρίως επειδή μπορούν να υλοποιηθούν με αποδοτικό τρόπο αφού είναι steerable[14]. Αυτό σημαίνει ότι είναι δυνατόν να υλοποιήσουμε μια κατευθυνόμενη έκδοση των φίλτρων Γκαουσιανών παραγώγων, απλά πολλαπλασιάζοντας την αρχική έκδοση του φίλτρου με ένα μητρώο, χωρίς να χρειάζονται επιμέρους συνελίξεις που είναι χρονοβόρες. Τα φίλτρα Gabor από την άλλη, δεν έχουν αυτή τη μαθηματική ιδιότητα, αλλά όπως έδειξε ο Heeger [23, 24] μπορούν να γίνουν διαχωρίσιμα, που σημαίνει ότι ένα φίλτρο μεγαλύτερης διάστασης μπορεί να χτιστεί από μονοδιάστατες αποκρίσεις της ίδιας εκδοχής φίλτρων.

Εφαρμόζουμε φιλτράρισμα με ζεύγη φίλτρων Gabor σε τετραγωνισμό, με ίδια κεντρική συχνότητα και εύρος ζώνης. Τα ζεύγη αυτά μπορούν να δημιουργηθούν από μονοδιάστατα (1D) φίλτρα Gabor με τον τρόπο που πρότεινε ο Daugman για δισδιάστατες (2D) προσανατολισμένες αποκρίσεις των ίδιων φίλτρων [8]. Το μονοδιάστατο μιγαδικό φίλτρο Gabor αποτελείται από ένα Γκαουσιανό παράθυρο διαμορφωμένο από ένα μιγαδικό ημίτονο. Η κανονικοποιημένη κρουστική του απόκριση έχει τη μορφή:

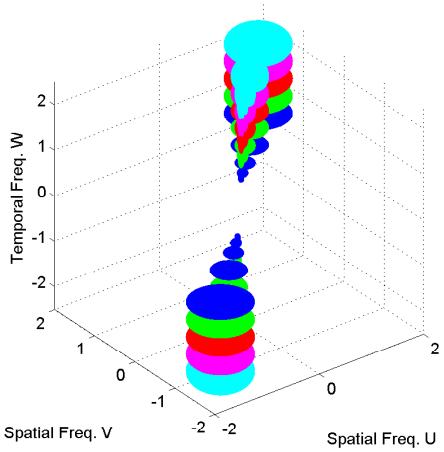
$$g(t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{t^2}{2\sigma^2}\right) \exp(j\omega_{t_0} t) = g_c(t) + j g_s(t) \tag{3.17}$$



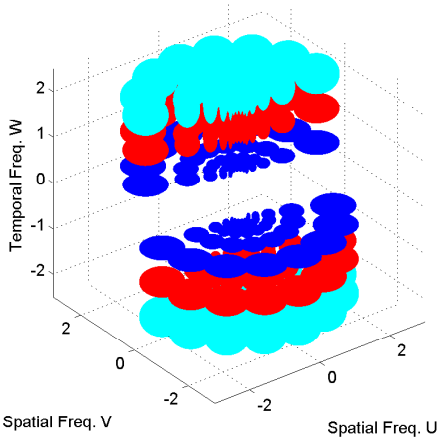
(α') Χωρική τράπεζα φίλτρων για χρονική συχνότητα ω_{t_0} (κάτοψη).



(β') Χωρική τράπεζα φίλτρων για χρονική συχνότητα $-\omega_{t_0}$ (κάτοψη).



(γ') Χωροχρονική τράπεζα φίλτρων σε 5 διαφορετικές χωρικές κλίμακες, 1 από τις 8 χωρικές κατευθύνσεις και 5 χρονικές συχνότητες.



(δ') Χωροχρονική τράπεζα φίλτρων σε 5 διαφορετικές χωρικές κλίμακες, 8 χωρικές κατευθύνσεις και 3 από τις 5 χρονικές συχνότητες.

Σχήμα 3.5: Ισοϋψείς καμπύλες της 3D χωροχρονικής τράπεζας φίλτρων, και μια κάτοψη μιας "φέτας" της τράπεζας σχεδιασμένη σε μια χρονική συχνότητα ω_{t_0}. Οι ισοϋψείς αντιστοιχούν στο 70% του πλάτους του εύρους ζώνης, ενώ διαφορετικά χρώματα χρησιμοποιούνται για διαφορετικές χρονικές συχνότητες. Μπορεί να παρατηρηθεί ότι ο συμμετρικός λωβός κάθε φίλτρου εμφανίζεται στο επίπεδο που ορίζεται από τη χρονική συχνότητα $-\omega_{t_0}$, σε αντίθεση με την περίπτωση της δισδιάστατης συστοιχίας φίλτρων. Παρατηρείται επίσης ότι το εύρος ζώνης κάθε φίλτρου αλλάζει ανάλογα με τη χρονική κλίμακα και τη χρονική συχνότητα.

Το παραπάνω μιγαδικό φίλτρο μπορεί να διαχωριστεί σε ένα περιττό (ημιτονικό $g_s(t)$) και σε ένα άρτιο (συνημιτονικό $g_c(t)$) φίλτρο, τα οποία αποτελούν ένα ζεύγος φίλτρων σε καθετότητα. Σχεδόν όλα τα φίλτρα Gabor είναι ζωνοπερατά με την κεντρική τους συχνότητα να ταυτίζεται με τη συχνότητα διαμόρφωσης ω_{t_0}. Μοναδική εξαίρεση αποτελεί η περίπτωση που γίνονται Γκαουσιανά βαθυπερατά φίλτρα, για συχνότητα $\omega_{t_0} = 0$. Έτσι, μπορούμε να καλύψουμε ολόκληρο το τρισδιάστατο χωροχρονικό πεδίο συχνοτήτων με Gabor φίλτρα των οποίων οι αποκρίσεις συχνότητας είναι κεντραρισμένες γύρω από συγκεκριμένες συχνότητες.

Η επέκταση των Gabor φίλτρων στις 3 διαστάσεις δίνεται για ένα συνημιτονοειδές (άρτιο) φίλτρο από την παρακάτω σχέση:

$$g_c(x,y,t) = \frac{1}{(2\pi)^{3/2}\sigma_x\sigma_y\sigma_t}\exp\left[-\left(\frac{x^2}{2\sigma_x^2}+\frac{y^2}{2\sigma_y^2}+\frac{t^2}{2\sigma_t^2}\right)\right] \cdot\cos(\omega_{x_0}x+\omega_{y_0}y+\omega_{t_0}t) \tag{3.18}$$

όπου τα $\omega_{x_0},\omega_{y_0},\omega_{t_0}$ είναι οι 2 χωρικές και η μια χρονική κεντρικές γωνιακές συχνότητες και τα $\sigma_x,\sigma_y,\sigma_t$ είναι οι τυπικές αποκλίσεις της 3D Γκαουσιανής περιβάλλουσας. Με τον ίδιο τρόπο προκύπτει η κρουστική απόκριση του ημιτονοειδούς (περιττού) φίλτρου, το οποίο συμβολίζουμε με $g_s(x,y,t)$.

Η απόκριση συχνότητας ενός 3D άρτιου συνημιτονικού φίλτρου έχει τη μορφή:

$$\begin{aligned} G_c(\omega_x,\omega_y,\omega_t) &= \frac{1}{2}\exp[-(\sigma_x^2(\omega_x-\omega_{x_0})^2/2 \\ &+ \sigma_y^2(\omega_y-\omega_{y_0})^2/2+\sigma_t^2(\omega_t-\omega_{t_0})^2/2)] \\ &+ \frac{1}{2}\exp[-(\sigma_x^2(\omega_x+\omega_{x_0})^2/2 \\ &+ \sigma_y^2(\omega_y+\omega_{y_0})^2/2+\sigma_t^2(\omega_t+\omega_{t_0})^2/2)] \end{aligned} \tag{3.19}$$

Συνεπώς η απόκριση συχνότητας ενός άρτιου φίλτρου Gabor αποτελείται από δύο Γκαουσιανά ελλειψοειδή τοποθετημένα συμμετρικά γύρω από τις συχνότητες $(\omega_{x_0},\omega_{y_0},\omega_{t_0})$ και $(-\omega_{x_0},-\omega_{y_0},-\omega_{t_0})$. Στο Σχήμα 3.5 φαίνονται οι ισοϋψείς της 3D συστοιχίας φίλτρων καθώς και μια κάτοψη μιας χωρικής τράπεζας φίλτρων της συστοιχίας σχεδιασμένη σε συγκεκριμένη χρονική συχνότητα ω_{t_0}. Αξίζει να παρατηρηθεί ότι οι συμμετρικοί λωβοί κάθε φίλτρου εμφανίζονται στο επίπεδο που ορίζεται από τη χρονική συχνότητα $-\omega_{t_0}$ σε αντίθεση με την περίπτωση των 2 διαστάσεων, όπου οι συμμετρικοί λωβοί απλά απεικονίζονται στα υπόλοιπα δύο τεταρτημόρια, λόγω απουσίας συχνότητας που να αφορά το χρόνο. Οπότε, εαν θέλουμε να καλύψουμε το χωρικό πεδίο συχνοτήτων για κάθε χρονική συχνότητα θα πρέπει να χρησιμοποιήσουμε και θετικές και αρνητικές χρονικές

συχνότητες. Έτσι, εισάγοντας αρνητικές χρονικές συχνότητες μπορούμε να καλύψουμε ολόκληρο το πεδίο συχνοτήτων που φαίνεται να μην καλύπτεται στο Σχήμα 3.5δ΄. Ακόμα, το εύρος ζώνης κάθε φίλτρου μεταβάλλεται καθώς μεταβάλλεται η χωρική κλίμακα και η συχνότητα που αφορά το χρόνο.

Για τη χωροχρονική τράπεζα φίλτρων χρησιμοποιήσαμε $N = 400$ Gabor φίλτρα (ισοτροπικά ως προς τη χωρική συνιστώσα, για μια τετράγωνη εικόνα) τα οποία τοποθετήθηκαν σε 5 χωρικές κλίμακες, 8 χωρικές κατευθύνσεις και 10 χρονικές κλίμακες (συχνότητες). Οι χωρικές κλίμακες και συχνότητες σχεδιάστηκαν έτσι ώστε να καλύπτουν ένα δισδιάστατο τετράγωνο πεδίο συχνοτήτων με τρόπο παρόμοιο με αυτόν που προτείνουν οι Havlicek et al. [22]. Εν συνεχεία οι κεντρικές συχνότητες και το εύρος ζώνης των Γκαουσιανών διαιρούνται με τη συχνότητα δειγματοληψίας στο χώρο για τη δημιουργία διακριτών φίλτρων με κανονικοποιημένες παραμέτρους συχνότητας που μπορούν να εφαρμοστούν σε εικόνες οποιασδήποτε διάστασης. Θα πρέπει να σημειωθεί ότι η διαδικασία μπορεί να οδηγήσει σε ανισοτροπικά χωρικά φίλτρα Gabor στο χώρο για μη τετράγωνες εικόνες, παρόλο που ο αρχικός σχεδιασμός περιλαμβάνει ισοτροπικά φίλτρα.

Χρησιμοποιούμε 10 χρονικά φίλτρα Gabor, 5 για θετικές και 5 για αρνητικές κεντρικές συχνότητες λόγω της ανάγκης να καλύψουμε ολόκληρο το τρισδιάστατο πεδίο συχνοτήτων που εξηγήσαμε στην προηγούμενη παράγραφο. Για τις 5 θετικές συχνότητες, η συστοιχία φίλτρων είναι πανομοιότυπη με αυτή που απεικονίζεται στο Σχήμα 3.2. Τονίζουμε ότι το να συμπεριληφθούν φίλτρα με θετικές και αρνητικές συχνότητες δεν έχει κάποια επίδραση στην πολυπλοκότητα του αλγορίθμου, διότι λόγω της ιδιότητας της διαχωρισιμότητας των φίλτρων Gabor, αρκεί να αλλαχτούν τα πρόσημα των εξισώσεων (3.29)-(3.30) για να λάβουμε τις φιλτραρισμένες εξόδους, χωρίς επιπλέον συνελίξεις που μειώνουν την απόδοση του συστήματος. Αυτά τα φίλτρα είναι τοποθετημένα γραμμικά ώστε να καλύπτουν ολόκληρο τον άξονα των κανονικοποιημένων συχνοτήτων, και το 3dB-εύρος ζώνης οκτάβας του καθενός είναι 0.75 οκτάβες. Στο Σχήμα 3.5 φαίνεται ο χωροχρονικός σχεδιασμός της 3D τράπεζας φίλτρων που χρησιμοποιήθηκε.

3.3.2 Παράμετροι του Αλγορίθμου Gabor3D

Κατά το σχεδιασμό του ανιχνευτή Gabor3D ήταν αναγκαίο να επιλέξουμε μεταξύ διαφορετικών εναλλακτικών ορισμένων παραμέτρων, όπως το είδος των φίλτρων, το είδος της ενέργειας που θα χρησιμοποιηθεί ως δείκτης σημαντικότητας για την αναπαράσταση των φιλτραρισμένων εξόδων, καθώς και τον τρόπο διαχείρισης των εξόδων των πολλαπλών φίλτρων.

3.3.2.1 Το είδος της ενέργειας

Πρώτα απ'ολα έπρεπε να επιλεγεί το είδος της ενέργειας. Είχαμε να επιλέξουμε ανάμεσα στην απλή τετραγωνική ενέργεια (square energy) και στην Teager-Kaiser ενέργεια.

Η τετραγωνική ενέργεια $E_S(\cdot)$ ορίζεται ως:

$$E_S(f(x,y,t)) = f^2(x,y,t) \tag{3.20}$$

όπου με $f(x,y,t)$ συμβολίζεται η φιλτραρισμένη έξοδος του video στο pixel (x,y,t). Αντίστοιχα, όπως προαναφέρθηκε στην Ενότητα 3.1, ο ενεργειακός τελεστής Teager-Kaiser στις τρεις διαστάσεις ορίζεται ως:

$$\begin{aligned}\Phi(f) &\equiv \|\nabla f\|^2 - f \cdot \nabla^2 f \\ &= f_x^2 + f_y^2 + f_t^2 - f \cdot (f_{xx} + f_{yy} + f_{tt})\end{aligned} \tag{3.21}$$

όπου $f = f(x,y,t)$. Η διακριτή μορφή του ΤΚΕΟ, την οποία χρησιμοποιήσαμε στον πειραματισμό μας είναι το αποτέλεσμα της διακριτοποίησης των χωρικών και χρονικών παραγώγων και ορίζεται ως:

$$\begin{aligned}\Phi_d[f[x,y,t]] \equiv\ & 3f^2[x,y,t] \\ & - f[x-1,y,t] \cdot f[x+1,y,t] \\ & - f[x,y-1,t] \cdot f[x,y+1,t] \\ & - f[x,y,t-1] \cdot f[x,y,t+1]\end{aligned} \tag{3.22}$$

Η ενέργεια Teager-Kaiser έχει χρησιμοποιηθεί με επιτυχία σε προβλήματα αναγνώρισης υφής. Το κίνητρο που έιχαμε για τη χρήση του ΤΚΕΟ είναι η ικανότητά του να εντοπίζει χωροχρονικές ταλαντώσεις της ενέργειας των σημάτων εισόδου και να τις διαχωρίζει σε συνιστώσες πλάτους και συχνότητας, με εξέχουσα χωροχρονική ανάλυση και πολύ μικρή πολυπλοκότητα. Για τη σύγκρισή του με τον τελεστή τετραγωνικής ενέργειας προχωρήσαμε σε πειράματα ταξινόμησης στη βάση KTH Action Dataset όπως θα δείξουμε στο Κεφάλαιο 5.

3.3.2.2 Το είδος των φίλτρων

Η επόμενη από τις επιλογές που έπρεπε να κάνουμε αφορά το είδος των φίλτρων Gabor. Συγκεκριμένα έπρεπε να επιλέξουμε ανάμεσα σε απλά συνημιτονικά φίλτρα και σε ζεύγη φίλτρων σε καθετότητα (quadrature), τα οποία υποστηρίζονται από τη βιβλιογραφία για εφαρμογές ανιχνεύσεων ακμών και κίνησης [27] και [1]. Η ιδέα πίσω από τα φίλτρα σε καθετότητα (ή φίλτρα με διαφορά φάσης $90°$) είναι ότι το καθένα θα εντοπίσει διαφορετικού τύπου ακμές από το κάθετό του. Στην περίπτωση των συναρτήσεων Gabor, η καθετότητα είναι εύκολο να

υλοποιηθεί χρησιμοποιώντας την ημιτονική και συνημιτονική έκδοση του ίδιου σήματος. Αφού φιλτράρουμε το video εισόδου με ένα τέτοιο ζεύγος φίλτρων, χρησιμοποιούμε ένα κλασσικό τέχνασμα, να υψώσουμε στο τετράγωνο και να αθροίσουμε τις φιλτραρισμένες με το ίδιο ημιτονικό και συνημιτονικό φίλτρο εξόδους, όπως περιγράφεται στην εξίσωση 3.24.

Συμβολίζουμε με $f_c(x, y, t)$ και $f_s(x, y, t)$ το αποτέλεσμα της συνέλιξης του video εισόδου με τη συνημιτονική και την ημιτονική έκδοση αντίστοιχα του ίδιου φίλτρου Gabor. Οι ενέργειες που προκύπτουν από το απλό συνημιτονικό φίλτρο αλλά και από το ζεύγος κάθετων φίλτρων δίνονται αντίστοιχα από τις σχέσεις:

$$E_{cos}(x, y, t) = E(f_c(x, y, t)) \tag{3.23}$$

$$E_{quad}(x, y, t) = E(f_c(x, y, t)) + E(f_s(x, y, t)) \tag{3.24}$$

όπου ορίζουμε ως $E(\cdot)$ το γενικό τελεστή ενέργειας. Στην περίπτωση που εξετάζουμε, οι εξισώσεις (3.23) και (3.24) μπορούν να οριστούν και για την απλή τετραγωνική αλλά και για την Teager-Kaiser ενέργεια, αντικαθιστώντας το $E(\cdot)$ με $E_S(\cdot)$ και $\Phi(\cdot)$, αντίστοιχα.

3.3.2.3 Ο τρόπος διαχείρισης των εξόδου των φίλτρων

Όπως φαίνεται και στο Σχήμα 3.4 με κάποιο τρόπο πρέπει να διαχειριστούμε την έξοδο των ενεργειών από τα $N = 400$ φίλτρα, για να λάβουμε τον τελικό χάρτη ενέργειας. Χρησιμοποιήσαμε ιδέες από Ανάλυση Κυρίαρχης Ενέργειας, παρόμοιες με αυτές που χρησιμοποιήθηκαν στο [17], όπου ως ενέργεια του τελικού χάρτη χρησιμοποιείται η ενέργεια του επικρατέστερου καναλιού, σύμφωνα με τον τύπο:

$$E_{max}(x, y, t) = \max_{1 \le k \le N} E_k(x, y, t) \tag{3.25}$$

όπου ως $E_k(x, y, t)$ συμβολίζουμε την έξοδο του k-οστού φίλτρου (ή ζεύγους φίλτρων στην περίπτωση των φίλτρων σε καθετότητα), αφού έχει επιδράσει πάνω της ο εκάστοτε τελεστής ενέργειας. Με N συμβολίζουμε τον συνολικό αριθμό των φίλτρων. Συγκρίναμε τη μέθοδο αυτή με μια διαφορετική προσέγγιση, όπου υπολογίσαμε το μέσο όρο των ενεργειών όλων των καναλιών ως την τελική αναπαράσταση ενέργειας, που είναι ισοδήναμο με το να υπολογίσουμε την υπέρθεση των ενεργειών όλων των καναλιών και να διαιρέσουμε με τον αριθμό τους (N).

$$E_{ave}(x, y, t) = \frac{\sum_{k=1}^{N} E_k(x, y, t)}{N} \tag{3.26}$$

Οι παράμετροι που περιγράφηκαν στην υποενότητα 3.3.2 αξιολογήθηκαν έτσι ώστε να προκύψει ο συνδυασμός που οδηγεί στην καλύτερη απόδοση για το πρόβλημα της αναγνώρισης ανθρώπινων δράσεων σε video. Συγκεκριμένα η αξιολόγηση έγινε στη βάση δεδομένων KTH Action Dataset [53] και ο συνδυασμός που οδήγησε στη μεγαλύτερη ακρίβεια ταξινόμησης (accuracy) επιλέχτηκε για την παραμετροποίηση του ανιχνευτή Gabor3D. Τα αποτελέσματα του πειραματισμού παρουσιάζονται στο Κεφάλαιο 5. Ενδεικτικά να προαναφέρουμε ότι η βέλτιστη συμπεριφορά του ανιχνευτή παρουσιάστηκε για το συνδυασμό Teager-Kaiser ενέργεια, ζεύγη φίλτρων Gabor σε quadrature και Ανάλυση Κυρίαρχης Ενέργειας.

3.3.2.4 Μείωση της πολυπλοκότητας του αλγορίθμου

Το φιλτράρισμα με 3D φίλτρα είναι μια χρονοβόρα διαδικασία λόγω της πολυπλοκότητάς της. Αν λάβει κανείς υπόψη την ύπαρξη μιας τράπεζας φίλτρων από πολλαπλά φίλτρα αυτού του είδους για τα οποία η διαδικασία του φιλτραρίσματος είναι πολύπλοκη, μπορεί να αντιληφθεί ότι ο αλγόριθμος θα είναι αρκετά αργός. Για να λύσουμε τέτοιου είδους προβλήματα χρησιμοποιούμε την ιδιότητα της διαχωρησιμότητας των φίλτρων Gabor [23], σύμφωνα με την οποία το φιλτράρισμα στις 3 διαστάσεις μπορεί να υλοποιηθεί σε κάθε διάσταση ξεχωριστά χρησιμοποιώντας μια κρουστική απόκριση που έχει τη μορφή 3.17. Με αυτόν τον τρόπο, εφαρμόζονται μόνο μονοδιάστατες συνελίξεις αντί για τρισδιάστατες, πράγμα που αυξάνει την απόδοση των υπολογισμών. Το αποτέλεσμα της διαδικασίας μπορεί να ληφθεί με τρεις 1D συνελίξεις χρησιμοποιώντας απλές τριγωνομετρικές ιδιότητες σε δύο βήματα, πρώτα για τον υπολογισμό του αποτελέσματος του 2D και εν συνεχεία του 3D φιλτραρίσματος. Αρχικά υπολογίζεται η έξοδος του 2D φιλτραρίσματος από τις κρουστικές αποκρίσεις $g_c(x), g_s(x), g_c(y), g_s(y)$ για το άρτιο και το περιττό φίλτρο ως εξής:

$$\begin{aligned} y_c^{2D}(x,y,t) &= (V(x,y,t) * g_c(x)) * g_c(y) \\ &- (V(x,y,t) * g_s(x)) * g_s(y) \qquad (3.27) \\ y_s^{2D}(x,y,t) &= (V(x,y,t) * g_s(x)) * g_c(y) \\ &+ (V(x,y,t) * g_c(x)) * g_s(y) \qquad (3.28) \end{aligned}$$

όπου $V(x,y,t)$ είναι το αρχικό video εισόδου μετασχηματισμένο στην κλίμακα του γκρί (grayscale). Ύστερα, το τελικό αποτέλεσμα του 3D χωροχρονικού φιλτραρίσματος μπορεί να υπολογιστεί συνελίσσοντας το αποτέλεσμα που λαμβάνεται από τις παραπάνω εξισώσεις με τις 1D χρονικές κρουστικές αποκρίσεις:

$$y_c^{3D}(x,y,t) = y_c^{2D}(x,y,t) * g_c(t) - y_s^{2D}(x,y,t) * g_s(t) \qquad (3.29)$$

$$y_s^{3D}(x,y,t) = y_c^{2D}(x,y,t) * g_s(t) + y_s^{2D}(x,y,t) * g_c(t) \qquad (3.30)$$

Για έναν 3D όγκο video διάστασης $n \times n \times n$ και ένα 3D φίλτρο μεγέθους $m \times m \times m$ η πολυπλοκότητα μειώνεται από $\mathcal{O}(n^3 \cdot m^3)$ που απαιτείται για τις 3D συνελίξεις σε $\mathcal{O}(3n^3 \cdot m)$ που χρειάζονται για τρεις 1D συνελίξεις, λόγω της ιδιότητας της διαχωρησιμότητας. Το χρώμα είναι επίσης ένας παράγοντας που μπορεί να αγνοηθεί για την προσέγγιση του προβλήματος της αναγνώρισης ανθρώπινων δράσεων σε video [60]. Ο μετασχηματισμός ενός έγχρωμου video στην κλίμακα του γκρι είναι μια διαδικασία που μειώνει τη διάστασή του στο $1/3$ της αρχικής, οπότε και οι διαδικασίες φιλτραρίσματος που απαιτούνται εφαρμόζονται σε έναν κατά πολύ μικρότερο όγκο πληροφορίας.

3.3.3 Χρήση μειωμένου αριθμού φίλτρων για τον αλγόριθμο Gabor3D

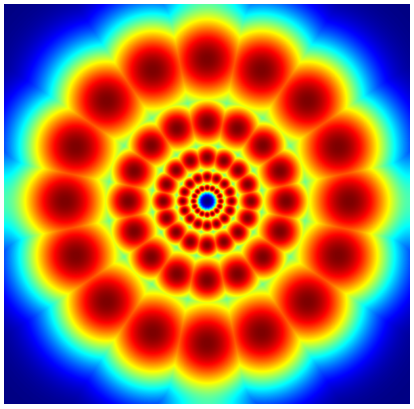
Η μείωση της πολυπλοκότητας του αλγορίθμου που περιγράφηκε στην προηγούμενη υποενότητα έγινε έτσι ώστε να καταστεί δυνατή η χρήση του σε συστήματα πραγματικού χρόνου. Παρόλαυτά, λόγω του μεγάλου αριθμού των φίλτρων που χρησιμοποιούνται, η διαδικασία της εξαγωγής των χωροχρονικών σημείων ενδιαφέροντος φαίνεται να διαρκεί σε χρόνο πάνω από 90% της συνολικής διαδικασίας αναγνώρισης ανθρώπινων δράσεων σε video. Για το λόγο αυτό έγινε μια προσπάθεια να μειωθεί ο αριθμός των φίλτρων που χρησιμοποιεί ο αλγόριθμος και ταυτόχρονα να μην πέσει η ακρίβεια αναγνώρισης του αλγορίθμου. Μετά από αρκετό πειραματισμό καταλήξαμε στη μειωμένη συστοιχία φίλτρων που φαίνεται Σχήμα 3.6β′ η οποία αν χρησιμοποιηθεί αντί της συστοιχίας που προτείνεται από τους Havlicek et al. δίνει προσεγγιστικά το ίδιο πειραματικό αποτέλεσμα (αναλυτικότερα στο Κεφάλαιο 5).

Οι παράμετροι που χρησιμοποιήθηκαν για την κατασκευή του νέου αυτού φίλτρου είναι οι παρακάτω. Το n-peak ακτινικό εύρος ζώνης οκτάβας (n-peak radial octave bandwidth που ορίζεται ως:

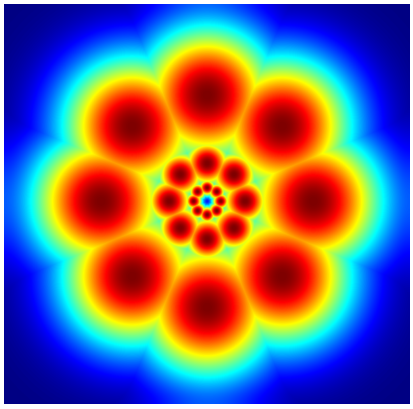
$$B = \log_2 \left[\frac{r_k + \frac{\sqrt{-\ln(n)}}{2\pi\sigma_k}}{r_k - \frac{\sqrt{-\ln(n)}}{2\pi\sigma_k}} \right] \tag{3.31}$$

ορίστηκε στην τιμή $B = 1.8$ οκτάβες. Η τιμή του n στις παραπάνω εξισώσεις ορίστηκε σε $n = \frac{1}{2}$, τιμή για την οποία το ακτινικό εύρος ζώνης οκτάβας ισοδυναμεί με το εύρος ζώνης $3 - dB$ φίλτρου. Ο λόγος μεταβολής της γεωμετρικής προόδου με αρχή το r_0 για τις κεντρικές συχνότητες των φίλτρων ορίστηκε στα $R = 2.8$, με $r_0 = 9.6$ κύκλοι ανά εικόνα. Περισσότερα για τις παραμέτρους των φίλτρων μπορούν να βρεθούν στα [21, 22].

Ενδεικτικά, παραθέτουμε το χρόνο εκτέλεσης του αλγορίθμου Gabor3D για τα video της KTH Action Dataset που έχουν διαστάσεις 120×160 pixels. Με



(α') Η πλήρης χωρική συστοιχία φίλτρων του ανιχνευτή Gabor3D.



(β') Η μειωμένη χωρική συστοιχία φίλτρων του ανιχνευτή Gabor3D.

Σχήμα 3.6: Η πλήρης και η μειωμένη συστοιχία χωρικών φίλτρων του Gabor3D. Η μειωμένη συστοιχία οδηγεί σε 120 χωροχρονικά φίλτρα έναντι 400 της πλήρους, και δίνει στον αλγόριθμο επιτάχυνση κατά έναν παράγοντα μεγαλύτερο του 3.

την αρχική έκδοση της τράπεζας φίλτρων με τα 400 φίλτρα, ο χρόνος εκτέλεσης είναι 1.6 δευτερόλεπτα ανά frame. Με τη μειωμένη έκδοση της συστοιχίας, ο χρόνος εκτέλεσης πέφτει στα 0.52 δευτερόλεπτα ανά frame. Η επίδοση του αλγορίθμου δοκιμάστηκε σε ένα κοινό laptop με επεξεργαστή Intel Core i5-430M, στο λογισμικό MATLAB.

Κεφάλαιο 4

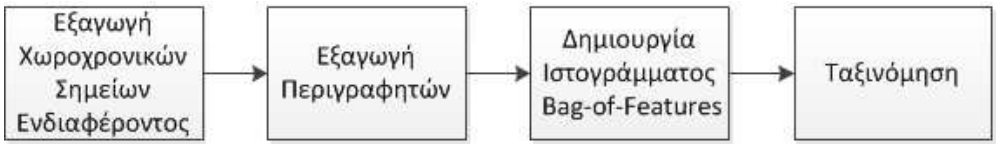
Μαθηματικά Εργαλεία - Τεχνικές

Η γενικότερη διαδικασία ταξινόμησης ανθρώπινων δράσεων σε video που χρησιμοποιήσαμε αποτελείται από 4 στάδια. Στο πρώτο στάδιο γίνεται η ανίχνευση χωροχρονικών σημείων ενδιαφέροντος από τον ανιχνευτή. Σε αυτό το στάδιο μπορεί να χρησιμοποιηθεί οποιοσδήποτε από τους ανιχνευτές που παρουσιάστηκαν (Harris3D, Cuboids, DCA3D, TDE, Gabor3D) ή οποιοσδήποτε άλλος ανιχνευτής. Τα χωροχρονικά αυτά σημεία ενδιαφέροντος αποτελούν voxels του video, γύρω από τα οποία γίνονται μετρήσεις και δημιουργείται ένας περιγραφητής. Οι περιγραφητές που δημιουργούνται (ένας για κάθε σημείο ενδιαφέροντος) ομαδοποιούνται ώστε τελικά να δημιουργηθεί ένα ιστόγραμμα από περιγραφητές, το οποίο ονομάζεται Bag-of-Features (BoF). Τα ιστογράμματα BoF, τα οποία αποτελούν την τελική αναπαράσταση των video, ταξινομούνται από κάποιον ταξινομητή, για να ληφθεί τελικά η απόφαση για το είδος της δράσης που ανιχνεύτηκε σε αυτό. Το block διάγραμμα της διαδικασίας που περιγράφηκε φαίνεται στο Σχήμα 4.1.

Στο παρόν Κεφάλαιο θα αναλύσουμε τη διαδικασία εξαγωγής ιστογραφικών περιγραφητών, θα παρουσιάσουμε περιγραφητές που χρησιμοποιούν μετρήσεις παραγώγων και οπτικής ροής, οι οποίοι χρησιμοποιήθηκαν στον πειραματισμό μας. Στη συνέχεια θα περιγραφεί λεπτομερώς η διαδικασία δημιουργίας ιστογραμμάτων Bag-of-Features και θα παρουσιαστούν οι ταξινομητές που χρησιμοποιήσαμε.

4.1 Τοπικοί Περιγραφητές Χωροχρονικών Σημείων Ενδιαφέροντος

Στα Κεφάλαια 2,3 μελετήσαμε ήδη υπάρχοντες αλλά και νέους ανιχνευτές χωροχρονικών σημείων ενδιαφέροντος που έχουν σχεδιαστεί για το πρόβλημα της αναγνώρισης ανθρώπινων δράσεων σε video. Στο δεύτερο στάδιο της αναγνώρισης,



Σχήμα 4.1: Block διάγραμμα της διαδικασίας ταξινόμησης ανθρώπινων δράσεων σε video. Αρχικά γίνεται η εξαγωγή χωροχρονικών σημείων ενδιαφέροντος με κάποιον ανιχνευτή, για καθένα από τα οποία δημιουργείται ένας περιγραφητής. Στη συνέχεια από τους περιγραφείς φτιάχνεται το ιστόγραμμα Bag-of-Features (BoF) και στη συνέχεια γίνεται ταξινόμηση των ιστογραμμάτων, με κάποιον ταξινομητή.

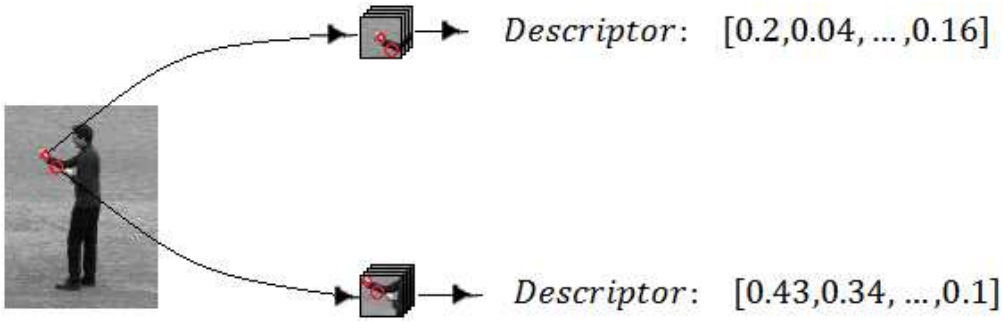
το οποίο είναι η εξαγωγή τοπικών χωροχρονικών χαρακτηριστικών από σημεία ενδιαφέροντος, αξιοποιείται η πληροφορία που προέρχεται από μια γειτονιά γύρω από κάθε σημείο για να γίνουν μετρήσεις οι οποίες θα είναι ικανές να αναπαραστήσουν την εμφάνιση ή/και την κίνηση των περιοχών ενδιαφέροντος. Ένας περιγραφητής μπορεί να υπολογιστεί σε μια δισδιάστατη γειτονιά γύρω από ένα σημείο έαν πρόκειται για εικόνα, σε μια τρισδιάστατη γειτονιά εαν πρόκειται για video, ενώ υπάρχουν και περιγραφείς που υπολογίζονται πάνω σε όλο τον όγκο πληροφορίας.

Ένας ιδανικός τοπικός περιγραφητής παρέχει μια αναπαράσταση ενός τοπικού τεμαχίου εικόνας ή video που παραμένει αναλλοίωτη στο θόρυβο παρασκηνίου, την εμφάνιση και τα οπτικά εμπόδια, και ενδεχομένως στην περιστροφή ή τις μεταβολές της κλίμακας. Είναι κοινή πρακτική ο καθορισμός της χρονικής και των χωρικών διαστάσεων του τρισδιάστατου όγκου υπολογισμού του περιγραφητή από τη χρονική και τις χωρικές κλίμακες αντίστοιχα, στις οποίες ανιχνεύθηκε το σημείο ενδιαφέροντος. Έτσι, τοπικοί περιγραφητές σημείων που ανιχνεύτηκαν σε πιο μεγάλες κλίμακες υπολογίζονται σε μεγαλύτερου μεγέθους χωροχρονικές γειτονιές, ενώ σημεία που πυροδότησαν τοπικούς ανιχνευτές σε πιο λεπτές, μικρές κλίμακες οδηγούν στον υπολογισμό ενός περιγραφητή σε μια μικρότερη γειτονιά. Η πλειονότητα των τρισδιάστατων τοπικών περιγραφητών αντλεί έμπνευση από περιγραφητές δύο διαστάσεων που χρησιμοποιήθηκαν επιτυχώς σε εικόνες, όπως είναι ο *SIFT* [35] και ο *HOG* [6].

Οι Dollár et al. [12] πειραματίζονται με διάφορους περιγραφητές που υπολογίζονται σε "κυβοειδείς" περιοχές γύρω από τα σημεία ενδιαφέροντος, με διαστάσεις ανάλογες με την αντίστοιχη κλίμακα ανίχνευσης που χρησιμοποιήθηκε, και στη συνέχεια τοποθετούνται στη σειρά για να σχηματίσουν ένα διάνυσμα

περιγραφής. Μελετούν τοπικές μετρήσεις όπως είναι οι κανονικοποιημένες τιμές εικόνας, τα gradients και η οπτική ροή. Παρόμοια λογική ακολουθούν και οι Niebles et al. [44] που χρησιμοποποιούν μετρήσεις παραγώγων για την περιγραφή των σημείων ενδιαφέροντος που εξάγουν με τον ανιχνευτή Cuboids.

Μια άλλη μορφή περιγραφητών περιλαμβάνει αυτούς που βασίζουν τους υπολογισμούς τους σε μια τρισδιάστατη δομή πλέγματος του τοπικού τεμαχίου. Ο τελικός περιγραφητής διαμορφώνεται από τη σύνοψη των παρατηρήσεων των επιμέρους κελιών από τα οποία απαρτίζεται το τοπικό τεμάχιο. Οι Scovanner et al.[54] προτείνουν μια επέκταση του δισδιάστατου περιγραφητή SIFT στις 3 διαστάσεις. Κάθε pixel περιγράφεται από το μέτρο και την κατεύθυνση του δισδιάστατου gradient αλλά και από μια επιπρόσθετη γωνία που αντιπροσωπεύει την απόκλιση από τη διεύθυνση του δισδιάστατου gradient. Σε κάθε υποπεριοχή της τρισδιάστατης γειτονιάς του σημείου ενδιαφέροντος υπολογίζεται ένα δισδιάστατο υπό-ιστόγραμμα των κατευθύνσεων των gradients. Το τελικό ιστόγραμμα του τοπικού τεμαχίου κατασκευάζεται από τη συσσώρευση των υπό-ιστογραμμάτων σε ένα διάνυσμα.



Σχήμα 4.2: Εξαγωγή τοπικού περιγραφητή από ένα χωροχρονικό σημείο ενδιαφέροντος. Ο υπολογισμός του περιγραφητή γίνεται γύρω από μια τρισδιάστατη γειτονιά του ανιχνευμένου σημείου, πρακτική που χρησιμοποιούν διάφοροι 3D περιγραφητές όπως ο HOG/HOF και ο HOG3D.

Παρακάτω παρουσιάζονται οι τοπικοί περιγραφείς (local descriptors) που χρησιμοποιήθηκαν στον πειραματισμό της παρούσας διπλωματικής εργασίας και έχουν δείξει πολύ ικανοποιητικά αποτελέσματα σε προβλήματα αναγνώρισης και ταξινόμησης, οι περιγραφητές HOG/HOF και HOG3D.

4.1.1 Ο περιγραφητής HOG/HOF

Ο τρισδιάστατος περιγραφητής HOG/HOF [34] αποτελεί επέκταση των δισδιάστατων περιγραφητών Histogram of Oriented Gradients (HOG) [6] και Histogram of oriented Optical Flow (HOF) [7], οι οποίοι χρησιμοποιήθηκαν εκτενώς σε προβλήματα ανίχνευσης ανθρώπων (human detection) σε εικόνες και σε video αντίστοιχα. Οι Laptev et al. εισήγαγαν το συνδυασμένο σχήμα HOG/HOF το οποίο εκμεταλλεύεται την πληροφορία των κατευθυνόμενων gradients αλλά και της οπτικής ροής, σε μια δομή πλέγματος στην οποία διαμερίζεται το κάθε τοπικό τεμάχιο. Τα εν λόγω ιστογράμματα στοχεύουν στο να συλλάβουν την τοπική εμφάνιση και κίνηση αντίστοιχα, στις γειτονιές των σημείων ενδιαφέροντος.

Ο περιγραφητής HOG/HOF συνίσταται στον υπολογισμό ιστογραμμάτων χωρικών gradients και οπτικής ροής στη χωροχρονική γειτονιά των σημείων ενδιαφέροντος που έχουν προηγουμένως εξαχθεί από κάποιον ανιχνευτή στο video εισόδου. Θεωρώντας χωρική και χρονική κλίμακα σ και τ αντίστοιχα στην οποία ανιχνεύθηκαν τα σημεία, οι διαστάσεις του τρισδιάστατου τοπικού τεμαχίου περιγραφής $(\Delta_x, \Delta_y, \Delta_t)$ ορίζονται ως εξής:

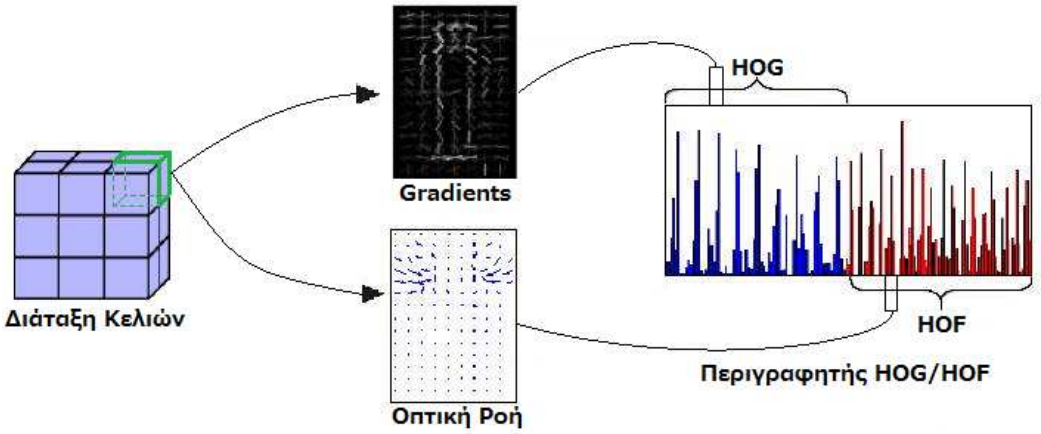
$$\Delta_x, \Delta_y = 2k_s\sigma \tag{4.1}$$

και

$$\Delta_t = 2k_t\tau \tag{4.2}$$

όπου k_s και k_t ακέραιοι πολλαπλασιαστές των τιμών της κλίμακας που ορίζονται από το χρήστη για το εκάστοτε πρόβλημα περιγραφής. Κάθε τοπικό τεμάχιο διαμερίζεται σε ένα τρισδιάστατο πλέγμα $n_x \times n_y \times n_t$ κελιών (blocks) για καθένα από τα οποία υπολογίζονται τα ιστογράμματα προσανατολισμού των gradients και της οπτικής ροής. Για τον εν λόγω περιγραφητή, αν και αντλούνται ιδέες από τους περιγραφητές HOG και HOF, ακολουθείται διαφορετική διαδικασία υπολογισμού των επιμέρους ιστογραμμάτων. Για τα ιστογράμματα HOG του HOG/HOF, οι προσανατολισμοί των χωρικών παραγώγων υπολογίζονται για όλα τα pixels του κελιού μέσω διάφορων μορφών παραγώγισης (πχ. Gaussian μάσκα παραγώγισης, τελεστής Sobel, κεντραρισμένες και μη κεντραρισμένες μάσκες παραγώγων κτλ). Στο [34] γίνεται εξαντλητικός πειραματισμός για το είδος των παραγώγων που λειτουργεί καλύτερα, και καταλήγουν ότι η απλή κεντραρισμένη μάσκα παραγώγων με διαφορές ($[-1, 0, 1]$ και $[-1, 0, 1]^T$) είναι αυτή που υπερτερεί όλων των άλλων σε ποσοστά σωστής αναγνώρισης. Οι τιμές των παραγώγων διακριτοποιούνται σε 4 κατευθύνσεις, παράγοντας ένα ιστόγραμμα 4 ράβδων, και η κάθε παράγωγος ψηφίζει ανάλογα με το βάρος της στη ράβδο στην οποία ανήκει. Για τα ιστογράμματα οπτικής ροής HOF αντίστοιχα, η κβαντοποίηση γίνεται σε 5 ράβδους, οι 4 από τις οποίες αντιστοιχούν σε διακριτές κατευθύνσεις κίνησης

και η πέμπτη σε απουσία κίνησης. Τα τοπικά ιστογράμματα όλων των "κελιών", αφού πρώτα κανονικοποιηθούν ξεχωριστά, συνενώνονται σε ένα κοινό διάνυσμα χαρακτηριστικών (features vector) που συνιστά την τελική αναπαράσταση του περιγραφητή, όπως φαίνεται στο Σχήμα 4.3.



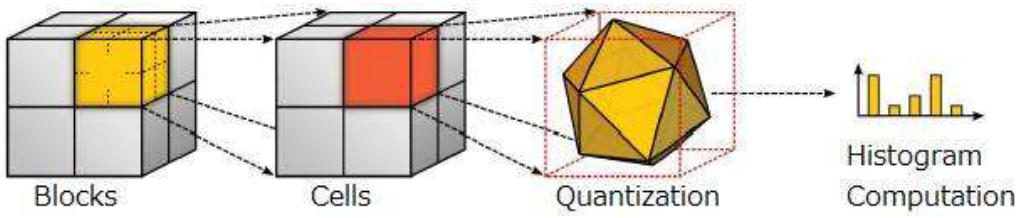
Σχήμα 4.3: Σχηματικό διάγραμμα κατασκευής του περιγραφητή HOG/HOF. Για κάθε κελί υπολογίζεται ένα υπο-ιστόγραμμα που αφορά την εμφάνιση (gradients) και ένα που αφορά την κίνηση (οπτική ροή), τα οποία κανονικοποιούνται ξεχωριστά. Ο τελικός περιγραφητής HOG/HOF αποτελεί συνένωση των επιμέρους υπό-ιστογραμμάτων.

Για παράδειγμα, παραθέτουμε τη διάταξη κελιών του περιγραφητή HOG/HOF που χρησιμοποιήθηκαν στα πειράματά μας, όπου $n_x = 3, n_y = 3, n_t = 2$. Σε κάθε κελί υπολογίζεται ένα υπο-ιστόγραμμα τεσσάρων ράβδων για τα χωρικά gradients και ένα πέντε ράβδων για την οπτική ροή. Οπότε ο τελικός περιγραφητής όλων των κελιών ($3 \times 3 \times 2 = 18$ κελιά) αποτελείται από $18 \times 4 = 72$ ράβδους που αφορούν την εμφάνιση (HOG) και $18 \times 5 = 90$ ράβδους που αφορούν την κίνηση (HOF). Συνενώνοντας τα υπο-ιστογράμματα παράγουμε ένα ιστόγραμμα $90 + 72 = 162$ ράβδων που αποτελεί τον συνδυασμένο περιγραφητή HOG/HOF.

4.1.2 Ο περιγραφητής HOG3D

Οι Kläser et al. προτείνουν μια διαφορετική επέκταση του HOG περιγραφητή στις 3 διαστάσεις, τον περιγραφητή HOG3D [26]. Στηρίζονται στον υπολογισμό κατευθύνσεων χωροχρονικών 3D gradients, οι οποίες κβαντοποιούνται βάσει

κανονικών πολυέδρων. Τα 3D gradients υπολογίζονται αποτελεσματικά με τη χρήση ολοκληρωτικών video, ενώ υιοθετείται και εδώ η λογική του υπολογισμού ιστογραμμάτων σε κελιά στα οποία χωρίζεται το χωροχρονικό τεμάχιο γύρω από το σημείο ενδιαφέροντος και της συνένωσής τους σε ένα τελικό διάνυσμα χρακτηριστικών που κανονικοποιείται με την $\mathcal{L}_2$ νόρμα. Και πάλι όπως και στην περίπτωση του HOG/HOF, οι διαστάσεις των τοπικών χωροχρονικών τεμαχίων και συνεπώς και των κελιών καθορίζονται από τη χωρική και χρονική κλίμακα ανίχνευσης των σημείων ενδιαφέροντος. Τα τοπικά χωροχρονικά τεμάχια χωρίζονται σε blocks, τα οποία με τη σειρά τους χωρίζονται σε cells, για καθένα από τα οποία γίνεται ξεχωριστά ο υπολογισμός των κβαντοποιημένων τιμών των 3D gradients. Για κάθε block το ιστόγραμμα προκύπτει από το άθροισμα των ιστογραμμάτων των επιμέρους Cells. Το τελικό ιστόγραμμα δημιουργείται από τη συνένωση των ιστογραμμάτων του κάθε block, τα οποία κανονικοποιούνται με την $\mathcal{L}_2$ νόρμα, όπως προτείνεται στο [36]. Στο Σχήμα. 4.4 φαίνεται περιγραφικά η διαδικασία για την εξαγωγή του περιγραφητή HOG3D.



Σχήμα 4.4

Θα πρέπει να σημειώσουμε ότι ο περιγραφητής HOG3D παρουσιάζει καλύτερα αποτελέσματα από τον περιγραφητή HOG/HOF στα πειράματά μας που αφορούν το πρόβλημα της αναγνώρισης ανθρώπινων δράσεων σε video, όπως θα δούμε στην Ενότητα 5.2.

4.2 Δημιουργία Ιστογράμματος Bag-of-Features (BoF)

Η πρόσφατη επιτυχία των τοπικών χαρακτηριστικών σε προβλήματα της Όρασης Υπολογιστών όπως η αναγνώριση αντικειμένων και η ταξινόμηση υφής κατέστησε αναγκαία την εύρεση μιας απλής, χαμηλής υπολογιστικής πολυπλοκότητας και αποτελεσματικής αναπαράστασης των εικόνων μέσω της κατηγοριοποίησης του

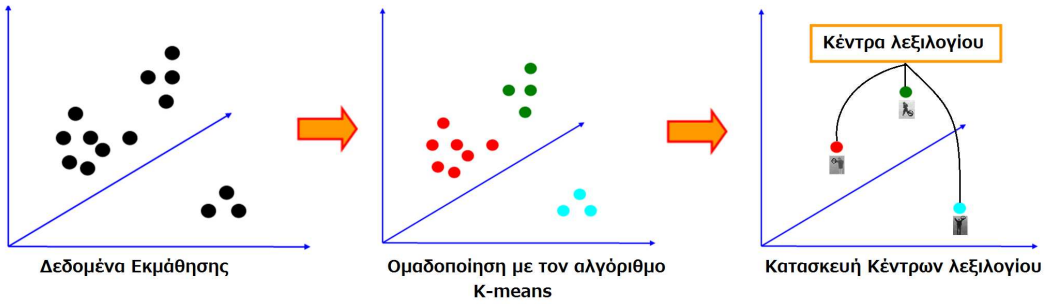
πλήθους των χαρακτηριστικών, ώστε η εκμάθηση των χαρακτηριστικών να είναι εύρωστη σε προβλήματα όπως μεγάλη μεταβλητότητα στην εμφάνιση, τις συνθήκες φωτισμού κτλ. Οι Willamowski et al. [62] προσέγγισαν το πρόβλημα της οπτικής κατηγοριοποίησης με την τεχνική Bag-of-Keypoints βάσει της οποίας μια εικόνα αναπαρίσταται από το ιστόγραμμα των συχνοτήτων εμφάνισης οπτικών προτύπων σε αυτήν. Πρόκειται λοιπόν για μια αναπαράσταση που ομαδοποιεί τους περιγραφητές τοπικών τεμαχίων που έχουν εξαχθεί για την εικόνα σε ένα διακριτό προκαθορισμένο σύνολο "οπτικών λέξεων" (visual words) χωρίς να διατηρεί καμία πληροφορία για τη γεωμετρική δομή των θέσεων των χαρακτηριστικών. Πρόκειται για μια προσέγγιση που οι ιδέες της πηγάζουν από την τεχνική Bag-of-Words (BoW) που είναι γνωστή από την επεξεργασία κειμένου [20].

Η συγκεκριμένη τεχνική υιοθετήθηκε και στο πρόβλημα αναγνώρισης ανθρώπινων δράσεων σε video αφού πρώτα μετονομάστηκε σε *Bag-of-Features (BoF)*. Αυτή η - απαλλαγμένη από τη γεωμετρία αλλά και από την πληροφορία του χρόνου - αναπαράσταση των video χρησιμοποιήθηκε πρόσφατα και συνεχίζει να χρησιμοποιείται για την προσέγγιση του προβλήματος [12, 63, 26, 34, 56, 44, 53]. Η δύναμη της συγκεκριμένης αναπαράστασης αντλείται από την αραιή αναπαράσταση του video, με οπτικά χαρακτηριστικά που είναι καλά καθορισμένα, και έχουν δείξει πειραματικά πως συνεργάζονται άψογα με Μηχανές Διανυσμάτων Υποστήριξης (SVMs) στο να διαχωριστούν επιτυχώς οι διαφορετικές κατηγορίες. Θα μπορούσαμε να πούμε πως παρόλο που η συγκεκριμένη αναπαράσταση δε χρησιμοποιεί κάποια άμεση πληροφορία για τη χωρική ή τη χρονική θέση των σημείων χωροχρονικών ενδιαφέροντος από τα οποία έχει εξαχθεί ο descriptor, έμεσσα το στοιχείο αυτό εμπεριέχεται τόσο στη δομή των blocks του τελευταίου, είτε από την οπτική ροή που αναπαρίσταται σε αυτόν, στην περίπτωση του HOG/HOF για παράδειγμα.

Για την κατασκευή του οπτικού λεξιλογίου που χρησιμοποιεί το BoF χρειάζονται δεδομένα εκμάθησης. Μια συνήθης πρακτική που χρησιμοποιείται είναι να κατασκευάζεται ένα λεξιλόγιο κάνοντας μια δειγματοληψία των δεδομένων εκπαίδευσης. Οι περιγραφητές εκπαίδευσης ομαδοποιούνται μέσω κάποιου γνωστού αλγόριθμου συσταδοποίησης σε συσκεκριμένο αριθμό ομάδων (clusters). Η επιλογή που προτείνεται είναι ο γνωστός αλγόριθμος K-means [13] που ωστόσο απαιτεί από πριν γνώση για τον αριθμό των clusters. Το λεξιλόγιο κατασκευάζεται ως τα κέντρα κάθε συστάδας που έχει προέλθει με τον αλγόριθμο K-means. Συνήθως ο K-means εκτελείται με αρκετές αρχικοποιήσεις, και χρησιμοποιείται αυτή που δίνει το μικρότερο σφάλμα.

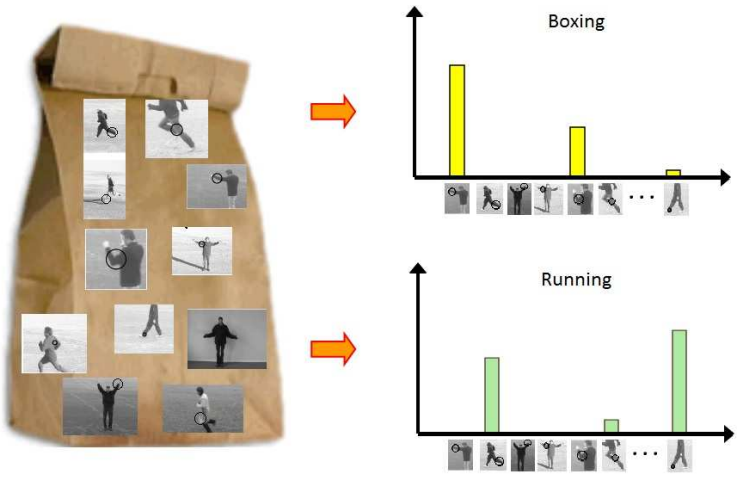
Μετά την κατασκευή του λεξιλογίου, κάθε περιγραφητής (τόσο των δεδομένων εκπαίδευσης όσο και των δεδομένων αξιολόγησης) αποδίδεται στην πλησιέστερη κλάση, με μετρικό την Ευκλείδια απόσταση από το κέντρο της καθεμίας. Η διαδικασία κατασκευής του οπτικού "χωροχρονικού" λεξιλογίου φαίνεται στο

Σχήμα. 4.5.



Σχήμα 4.5: Διαδικασία κατασκευής του οπτικού λεξιλογίου με τον αλγόριθμο K-means. Τα δεδομένα ομαδοποιούνται με τον αλγόριθμο K-means σε συγκεκριμένο αριθμό ομάδων. Το κέντρο κάθε ομάδας αποτελεί λέξη του λεξιλογίου του τελικού ιστογράμμαος.

Τα χαρακτηριστικά που ομαδοποιούνται στην προσέγγισή μας στην αναγνώριση ανθρώπινων δράσεων είναι περιγραφητές (descriptors) όπως αυτοί που περιγράφηκαν στην Ενότητα 4.1. Η κάθε λέξη του λεξιλογίου έχει προφανώς την ίδια διάσταση με το πλήθος στοιχείων του διανύσματος χαρακτηριστικών του τοπικού περιγραφητή που χρησιμοποιείται. Η τελική αναπαράσταση του video είναι το ιστόγραμμα της συχνότητας εμφάνισης των οπτικών λέξεων του λεξιλογίου, το οποίο κανονικοποιείται με βάση την $\mathcal{L}_2$ νόρμα (Σχήμα 4.6). Καταλήγουμε δηλαδή σε ένα "ιστόγραμμα ιστογραμμάτων" για την αναπαράσταση του εκάστοτε video.



Σχήμα 4.6: Το οπτικό λεξιλόγιο που κατασκευάζεται αποτελεί τις ράβδους (bins) του ιστογράμματος BoF. Οι ράβδοι στοιχίζονται σε τυχαία σειρά. Κάθε περιγραφητής ψηφίζει σε μία μόνο ράβδο του ιστογράμματος με βάση τη μικρότερη Ευκλείδια απόσταση από αυτή.

Το BoF είναι μια αναπαράσταση που είναι εύκολη στη σύλληψη, εύκολη στην υλοποίηση αλλά και πολύ αποδοτική στο διαχωρισμό κλάσεων σε συνεργασία με τα SVMs. Καταλήγει σε αναπαραστάσεις που είναι μικρής σχετικά διαστασιμότητας και ίσου μεγέθους για κάθε video, με αραιή (sparse) μορφή πληροφορίας. Ήδη έχουν συζητηθεί οι επεκτάσεις των ιδεών που παρατέθηκαν παραπάνω. Οι Bettadapura et al. [3] προτείνουν την επέκταση BoF σε μια μορφή που περιέχει πληροφορία της εξέλιξης των δράσεων με τη χρήση ιστογραμμάτων από n-grams από υπό-δράσεις, σχήμα το οποίο υποστηρίζουν ότι δουλεύει αποδοτικότερα. Η επέκταση των ιδεών του BoF έτσι ώστε να περιέχουν πιο σαφή πληροφορία πιστεύουμε πως θα απασχολήσει τους ερευνητές της περιοχής του Machine Learning στο κοντινό μέλλον.

4.3 Ταξινόμηση με διάφορους Classifiers

Στην Ενότητα αυτή θα μελετήσουμε διάφορων τύπων ταξινομητές (classifiers) τους οποίους χρησιμοποιήσαμε για να ταξινομήσουμε τα ιστογράμματα BoF ή τις παραλλαγές τους στις διάφορες κλάσεις που ορίσαμε. Έγινε χρήση στατικών ταξινομητών όπως είναι οι Μηχανές Διανυσμάτων Υποστήριξης - ή Support Vector Machines (SVMs) - και ο kNN, καθώς και αναγεννητικών μοντέλων

που παρακολουθούν τη χρονική εξέλιξη των γεγονότων όπως είναι τα Κρυφά Μαρκοβιανά Μοντέλα - ή Hidden Markov Models (HMMs). Στις υποενότητες που ακολουθούν δεν θα κάνουμε πλήρη παρουσίαση των ταξινομητών, καθώς η εκτενής τους μελέτη ξεφεύγει από τους ακοπούς της παρούσας διπλωματικής εργασίας, αλλά θα παραπέμψουμε τον αναγνώστη σε κλασσικές πηγές κάνοντας μια πιο high level περιγραφή τους.

4.3.1 Μηχανές Διανυσμάτων Υποστήριξης (SVMs)

Οι Μηχανές Διανυσμάτων Υποστήριξης (SVMs) [5] είναι ταξινομητές μηχανικής μάθησης που επιτυγχάνουν μεγάλα περιθώρια διαχωρισμού και έχουν πρόσφατη επιτυχία σε προβλήματα αναγνώρισης οπτικών προτύπων. Λόγω του ότι συνδυάζονται άψογα με τα ιστογράμματα BoF, αποτελούν τη συνήθη τεχνική ταξινόμησης ανθρώπινων δράσεων σε video. Θα δώσουμε μια γενική ιδέα της λειτουργίας τους.

Θεωρούμε το πρόβλημα του διαχωρισμού ενός συνόλου δεδομένων εκπαίδευσης $(\mathbf{x_1}, d_1), (\mathbf{x_2}, d_2), ..., (\mathbf{x_m}, d_m)$ όπου $\mathbf{x_i} \in \mathbb{R}^N$ είναι ένα διάνυσμα χαρακτηριστικών και $d_i \in \{-1, +1\}$ η ετικέτα της κλάσης στην οποία ανήκει. Θεωρώντας ότι τα δεδομένα - για τα οποία στη γενική περίπτωση δεν έχουμε γνώση της κατανομής τους - μπορούν να διαχωριστούν από ένα υπερεπίπεδο με επιφάνεια απόφασης $\mathbf{w} \cdot \mathbf{x} + b = 0$, τότε το βέλτιστο υπερεπίπεδο είναι αυτό που μεγιστοποιεί το περιθώριο διαχωρισμού μεταξύ των δεδομένων με θετική και αρνητική ετικέτα. Αποδεικνύεται ότι η εύρεση των βέλτιστων $\mathbf{w}$ και b είναι λύση του προβλήματος ελαχιστοποίησης με περιορισμούς:

$$\begin{aligned} &\text{minimize } \frac{1}{2}\|\mathbf{b}\|^2 \\ &\text{υπό τους περιορισμούς } d_i \cdot (\mathbf{w} \cdot \mathbf{x_i} + b) \geq 1, \forall i = 1, ..., m \end{aligned} \tag{4.3}$$

Το συγκεκριμένο πρόβλημα επιλύεται με τη χρήση πολλαπλασιαστών Lagrange και έτσι προκύπτει η ακόλουθη στνάρτηση απόφασης:

$$f(x) = sgn\left(\sum_{i=1}^{m} \alpha_i d_i \mathbf{w} \cdot \mathbf{x} + b\right) \tag{4.4}$$

όπου $\alpha_i, i = 1, ..., m$ οι πολλαπλασιαστές Lagrange. Τα διανύσματα χαρακτηριστικών $\mathbf{x_i}$ που αντιστοιχούν σε μη μηδενικούς πολλαπλασιαστές Lagrange ονομάζονται διανύσματα υποστήριξης και εννοιολογικά είναι αυτά που βρίσκονται πλησιέστερα στην επιφάνεια διαχωρισμού και επομένως διαδραματίζουν σπουδαιότερο ρόλο στον καθορισμό της.

Ένα SVM μπορεί να κατασκευαστεί με μια αντιστοίχηση του διανύσματος εισόδου σε ένα χώρο χαρακτηριστικών μεγαλύτερης διάστασης $\mathbf{x} \rightarrow \Phi(\mathbf{x})$ που δεν είναι ορατός από την είσοδο και την έξοδο, στον οποίο τα δεδομένα είναι γραμμικά διαχωρίσιμα. Με την εύρεσης μιας συνάρτησης πυρήνα K για την οποία ισχύει $K(\mathbf{x}, \mathbf{x_i}) = \Phi(\mathbf{x})\Phi(\mathbf{x_i})$, η αντικατάσταση του εσωτερικού γινομένου της σχέσης (4.4) από την τιμή της συνάρτησης πυρήνα δίνει τη συνάρτηση απόφασης ενός μη γραμμικού SVM:

$$f(x) = sng\left(\sum_{i=1}^{m} \alpha_i d_i K(\mathbf{x_i}, \mathbf{x}) + b\right) \tag{4.5}$$

Η αντικατάσταση με τη συνάρτηση πυρήνα δίνει εκτός από το πλεονέκτημα των "κρυφών" διαστάσεων στις οποίες τα δεδομένα μπορεί να είναι γραμμικά διαχωρήσιμα, και την απαλλαγή από προβλήματα που προκύπτουν από την ανάγκη βελτιστοποίησης πολύ μεγάλου αριθμού παραμέτρων. Υπάρχουν πολλά είδη συναρτήσεων πυρήνα όπως η γραμμική, η πολυωνυμική, η Radial Basis Function (RBF) και άλλες. Η επιλογή του κατάλληλου πυρήνα εξαρτάται από τα δεδομένα στα οποία χρειάζεται να εφαρμόσουμε ταξινόμηση με SVMs. Για το πρόβλημα αναγνώρισης ανθρώπινων δράσεων σε video, αυτός που φαίνεται να υπερτερεί με βάση τη βιβλιογραφία είναι ο χ^2-πυρήνας, ο οποίος χρησιμοποιεί την χ^2 απόσταση όπως φαίνεται στη σχέση (4.6).

Έχοντας $H_i = \{h_{i1}, h_{i2}, ..., h_{iV}\}$ και $H_j = \{h_{j1}, h_{j2}, ..., h_{jV}\}$ τα ιστογράμματα συχνότητας εμφάνισης οπτικών λέξεων και V το μέγεθος του οπτικού λεξιλογίου, ο χ^2 πυρήνας ορίζεται ως:

$$K(H_i, H_j) = \exp\left(-\frac{1}{2A}\sum_{n=1}^{V}\frac{(h_{in} - h_{jn})^2}{h_{in} + h_{jn}}\right) \tag{4.6}$$

Οι Zhang et al. [65] έδειξαν στα πειράματά τους ότι η τιμή της παραμέτρου A μπορεί να τεθεί ως η μέση χ^2 απόσταση μεταξύ όλων των δειγμάτων εκπαίδευσης.

Η πολύ μικρή τους πολυπλοκότητα, η επίδοσή τους και η άψογη συνεργασία τους με τα BoF ιστογράμματα, είναι μερικοί απ'τους λόγους που τα SVMs υπερτερούν στην έρευνα του πεδίου της αναγνώρισης δράσεων. Το μεγαλύτερο πειραματικό μέρος της παρούσας διπλωματικής, επίσης, στηρίζεται στη χρήση των SVMs.

4.3.2 Ταξινομητές k-Nearest Neighbours (κ-NN)

Ο k-NN [2] είναι ακόμα ένας ισχυρός ταξινομητής της αναγνώρισης προτύπων, τον οποίο χρησιμοποιήσαμε για τους πειραματισμούς μας, σε πολύ μικρότερη κλίμακα ωστόσο από τα SVMs. Η φιλοσοφία του k-NN βασίζεται στην ταξινόμηση

με βάση τα πλέον γειτονικά παραδείγματα εκμάθησης. Συνεπώς, ένα αντικείμενο θα αντιστοιχηθεί σε κάποια κλάση, βασιζόμενο στην πλειοψηφική ψήφο των k κοντινότερων γειτόνων του στο χώρο χαρακτηριστικών. Αν η τιμή του k γίνει ίση με 1, τότε η ανάθεση γίνεται απλά με βάση τον κοντινότερο γείτονα. Το μετρικό απόστασης που χρησιμοποιείται είναι ελεύθερο, τα πειράματά μας έγιναν ωστόσο χρησιμοποιώντας την Ευκλείδια απόσταση.

4.3.3 Κρυφά Μαρκοβιανά Μοντέλα (HMMs)

Η μεγάλη δύναμη των Κρυφών Μαρκοβιανών Μοντέλων HMMs είναι η δυνατότητα περιγραφής ενός δυναμικά εξελισσόμενου φαινομένου, και η στατιστική του μοντελοποίηση. Σε αντίθεση με άλλα εργαλεία της αναγνώρισης προτύπων και της μηχανικής μάθησης (όπως για παράδειγμα τα SVM ή τα kNN), τα HMMs μοντελοποιούν την εξέλιξη ενός φαινομένου σε σχέση με το χρόνο. Αυτή ήταν και το γεγονός που μας παρακίνησε να τα χρησιμοποιήσουμε αρχικά για το πρόβλημα της αναγνώρισης ανθρώπινων δράσεων σε video. Λόγω της φύσης του προβλήματος η χρονική εξέλιξη μιας δράσης δίνει επιμέρους πληροφορία στον ταξινομητή, σε αντίθεση με στατικούς ταξινομητές όπως είναι οι Μηχανές Διανυσμάτων Υποστήριξης. Στις υποενότητες που ακολουθούν, δεν επιχερούμε μια πλήρη παρουσίαση των HMMs (για αυτό παραπέμπουμε σε κάποιες κλασικές πηγές [49, 51]). Αντ' αυτού, προχωράμε αρχικά σε μία αναφορά της βασικής ορολογίας των HMMs, για λόγους πληρότητας και ευκολίας παρακολούθησης του υπολοίπου της παρούσας διπλωματικής εργασίας, και στη συνέχεια παρουσιάζουμε πιο εξειδικευμένα τον τρόπο αξιοποίησης των HMMs με στόχο την αναγνώριση ανθρώπινων δράσεων σε video.

Ένα Κρυφό Μαρκοβιανό Μοντέλο αποτελεί μια μηχανή πεπερασμένων καταστάσεων, οι οποίες δεν είναι άμεσα παρατηρήσιμες παρά μόνο μέσω της ακολουθίας των παραγόμενων συμβόλων του μοντέλου. Για την περιγραφή ενός HMM απαιτείται ο ορισμός ενός συνόλου παραμέτρων:

- Ένα HMM αποτελείται από ένα σύνολο N διαφορετικών καταστάσεων, και σε κάθε χρονική στιγμή $t = 1, 2, ...T$ το μοντέλο βρίσκεται σε μία από αυτές $q_t = \{q_1, q_2, ..., q_N\}$.
- Για την μετάβαση από μία κατάσταση i σε μία κατάσταση j ορίζεται η πιθανότητα μετάβασης: $\mathbf{A} = \{a_{ij}\} : a_{ij} = Pr(q_{t+1} = j | q_t = i), \quad 1 \leq i, j \leq N$. Οι πιθανότητες a_{ij} θα πρέπει να ικανοποιούν τις σχέσεις:

$$a_{ij} \geq 0 \tag{4.7}$$

$$\sum_{j=1}^{N} a_{ij} = 1, \forall i \tag{4.8}$$

- Σε κάθε κατάσταση εξάγεται μια παρατήρηση $\mathbf{o}_t$ από το σύνολο συμβόλων V των δυνατών παρατηρήσεων. Σε κάθε κατάσταση j αντιστοιχεί μια διαφορετική συνάρτηση κατανομής για την πιθανότητα εξαγωγής του κάθε συμβόλου: $\mathbf{B} = \{b_j\} : b_j(\mathbf{o}_t) = Pr(\mathbf{o}_t|q_t = j)$. Στην περίπτωση HMM με συνεχή συνάρτηση πυκνότητας πιθανότητας αυτή θα ορίζεται σαν το συνδυασμό M γκαουσιανών κατανομών:

$$b_j(\mathbf{o}_t) = \sum_{k=1}^{M} c_{jk}\mathcal{N}(\mathbf{o}_t, \boldsymbol{\mu}_{jk}, \mathbf{U}_{jk}), 1 \leq j \leq N \tag{4.9}$$

  Οι παράγοντες $c_{jk}, \boldsymbol{\mu}_{jk}, \mathbf{U}_{jk}$ αντιπροσωπεύουν τους συντελεστές των γκαουσιανών, το διάνυσμα των μέσων όρων και τον πίνακα συμμεταβλητότητας αντίστοιχα για την k κατανομή και την j κατάσταση.

  Επίσης, οι συντελεστές των γκαουσιανών θα πρέπει να ικανοποιούν τους παρακάτω περιορισμούς:

$$\sum_{k=1}^{M} c_{jk} = 1, \quad 1 \leq j \leq N \tag{4.10}$$

$$c_{jk} \leq 0, \quad 1 \leq j \leq N, \quad 1 \leq k \leq M \tag{4.11}$$

- Τέλος, ορίζεται και μια κατανομή πιθανότητας π_i η οποία εκφράζει την πιθανότητα το HMM να ξεκινά από την κατάσταση i:
  $\boldsymbol{\pi} = \{\pi_i\} : \pi_i = Pr(q_1 = i), 1 \leq i \leq N.$

Έτσι, με βάση τα παραπάνω, ένα HMM μπορεί να περιγραφεί πλήρως από το σύνολο παραμέτρων $\boldsymbol{\lambda} = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$.

Η χρήση των HMMs στο πεδίο της αναγνώρισης προτύπων Ιδιαίτερο ενδιαφέρον έχουν και κάποιοι βασικοί αλγόριθμοι που έχουν αναπτυχθεί και επιτρέπουν την εκπαίδευση αλλά και την αναγνώριση στα HMMs. Μιας και πρόκειται για πολύ γνωστούς αλγόριθμους, αναφερόμαστε μόνο επιγραμματικά σε αυτό το σημείο.

- Οι αλγόριθμοι forward και backward είναι δύο ιδιαίτερα αποδοτικοί αλγόριθμοι, οι οποίοι δεδομένης μιας ακολουθίας παρατηρήσεων $\mathbf{O} = (\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_T)$ και ενός HMM με παραμέτρους $\boldsymbol{\lambda} = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$ επιτρέπουν τον υπολογισμό της πιθανοφάνειας $Pr(\mathbf{O}|\boldsymbol{\lambda})$.

- Ο αλγόριθμος Viterbi για μια δεδομένη ακολουθία καταστάσεων $\mathbf{O} = (\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_T)$ και ενός μοντέλου $\boldsymbol{\lambda}$ μας επιτρέπει να βρούμε την ακολουθία καταστάσεων $Q = q_1, q_2, ..., q_T$ που μεγιστοποιεί την πιθανότητα $Pr(\mathbf{O}, \mathbf{q}|\boldsymbol{\lambda})$.

- Ο αλγόριθμος Baum-Welch, που πρόκειται για ειδική περίπτωση του EM (Expectation-Maximization), χρησιμοποιείται για την προσεγγιστική εύρεση των βέλτιστων παραμέτρων που μεγιστοποιούν την πιθανοφάνεια του μοντέλου $Pr(\mathbf{O}|\boldsymbol{\lambda})$ για ένα σύνολο εκπαίδευσης. Το πρόβλημα αυτό είναι και το πιο δύσκολο (δεν υπάρχει αναλυτική λύση), και αντιστοιχεί στην εκπαίδευση των HMMs, δεδομένου ενός συνόλου εκπαίδευσης.

Κεφάλαιο 5

Πειράματα Ταξινόμησης Ανθρώπινων Δράσεων

Στο παρόν κεφάλαιο θα γίνει αναλυτική παρουσίαση των μεθόδων που χρησιμοποιήσαμε για να αντιμετωπίσουμε το πρόβλημα της ταξινόμησης ανθρώπινων δράσεων σε video. Θα περιγράψουμε τη διαδικασία που χρησιμοποιήθηκε τόσο με τη χρήση στατικών ταξινομητών, όπως είναι τα SVMs και ο kNN όσο και με τη χρήση ταξινομητών που λαμβάνουν υπ' όψη τους τη χρονική εξέλιξη των γεγονότων, όπως είναι τα HMMs. Στη συνέχεια θα γίνει σύγκριση των αλγορίθμων μας με κάποιους δημοφιλείς ανιχνευτές χωροχρονικών σημείων ενδιαφέροντος, που παρουσιάστηκαν και στο Κεφάλαιο 2, σε δύο πολύ γνωστές βάσεις δεδομένων, την KTH Action Dataset και την Hollywood2 Action Dataset.

5.1 Περιγραφή της διαδικασίας

Πριν προχωρήσουμε στην παρουσίαση των πειραματικών μας αποτελεσμάτων θα περιγράψουμε τη διαδικασία που ακολουθήσαμε στις δύο βάσεις δεδομένων που προαναφέραμε για την ταξινόμηση δράσεων. Στόχος μας είναι να αποσαφηνίσουμε λεπτομέρειες υλοποίησης των ερευνητικών μας προσπαθειών, τόσο για το framework με τη χρήση SVMs, όσο και με τη χρήση HMMs. Για τις προσπάθειές μας με τη χρήση kNN ταξινομητών θα αναφερθούμε αρκετά λιγότερο αφ' ενός διότι η διαδικασία είναι αρκετά απλή και διαφέρει ελάχιστα από τη διαδικασία με χρήση SVMs και αφ' ετέρου διότι η χρήση τους δεν έγινε σε τόσο μεγάλη κλίμακα. Για κάποιες παραπάνω λεπτομέρειες που αφορούν την καθεμία από τις δύο βάσεις που χρησιμοποιήσαμε, θα περιγράψουμε το διαφορικό στην αντίστοιχη ενότητα.

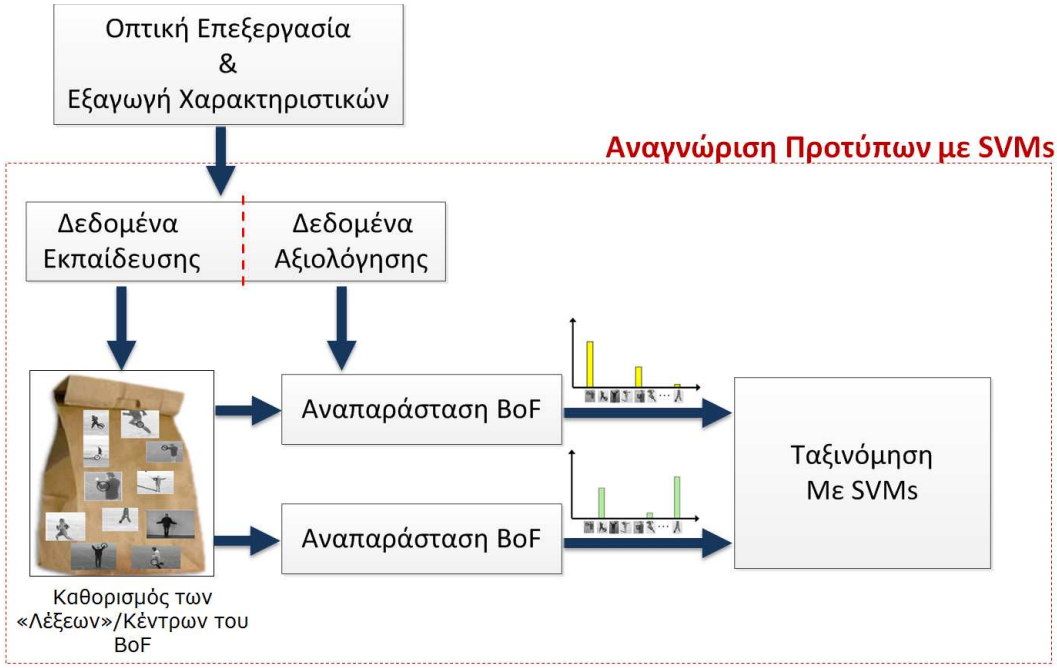
5.1.1 Ταξινόμηση Ανθρώπινων Δράσεων με SVM Classifiers

Η ταξινόμηση ανθρώπινων δράσεων σε video με χρήση SVM ταξινομητών είναι η πιο διαδεδομένη τεχνική για το συγκεκριμένο πρόβλημα. Όπως αναφέραμε και στην Ενότητα 4.2 τα SVMs δείχνουν πολύ καλή συμπεριφορά σε συνδυασμό με την τεχνική Bag-of-Features (BoF) και γενικότερα με το διαχωρισμό ιστογραφικών χαρακτηριστικών, και για αυτό το λόγο χρησιμοποιούνται πολύ συχνά στη βιβλιογραφία.

Η προσέγγιση που χρησιμοποιήθηκε στην παρούσα διπλωματική για την ταξινόμηση ανθρώπινων δράσεων με SVMs συνοψίζεται στο Σχήμα. 5.1. Αρχικά εκτελείται το στάδιο της Οπτικής Επεξεργασίας, δηλαδή η εξαγωγή των χωροχρονικών σημείων ενδιαφέροντος με τις κλίμακές τους για τα video κάθε βάσης, από έναν ανιχνευτή όπως αυτούς που αναλύσαμε στα προηγούμενα Κεφάλαια.

Στη συνέχεια εξάγεται ένας περιγραφητής σε μια χωροχρονική γειτονιά γύρω από το κάθε σημείο ενδιαφέροντος. Οι περιγραφητές που χρησιμοποιήθηκαν στα πειράματά μας ήταν οι HOG/HOF και HOG3D. Για τον περιγραφητή HOG/HOF χρησιμοποιήθηκε η υλοποίηση των Laptev et al. [34] με διάταξη κελιών $n_x = 3, n_y = 3, n_t = 2$. Για το εύρος γειτονιάς εφαρμογής του περιγραφητή χρησιμοποιήθηκαν οι τιμές παραμέτρων $k_s = 9$ και $k_t = 4$ για τις εξισώσεις (4.1)-(4.2) ενώ απορρίφθηκαν τα σημεία ενδιαφέροντος που ανιχνεύτηκαν στα σύνορα των εικόνων, και συγκεκριμένα στα 5 συνοριακά pixels, για να αποφύγουμε σφάλματα που προέρχονται από παραγωγίσεις στα συγκεκριμένα σημεία (τονίζουμε ότι στους ανιχνευτές που αναπτύξαμε έχουμε μεριμνήσει ώστε στα σύνορα των εικόνων η επεξεργασία να είναι όσο το δυνατόν πιο ομαλή). Η ανάλυση αυτή τελικά καταλήγει σε ένα διάνυσμα διάστασης 162 για κάθε σημείο ενδιαφέροντος. Για τον περιγραφητή HOG3D [26] χρησιμοποιήθηκαν διαφορετικές παράμετροι για καθεμία από τις 2 βάσεις, αφού ο δημιουργός του τις έχει προσαρμόσει και βελτιστοποιήσει για καθεμία από αυτές. Για την KTH Action Dataset χρησιμοποιήθηκαν οι παράμετροι: $k_s = 8$ και $k_t = 2$ για το μέγεθος των κελιών, $n_x = 5, n_y = 5, n_t = 4$ για τη διάταξή τους, ενώ για την κβαντοποίηση των κατευθύνσεων χρησιμοποιήθηκε το κανονικό εικοσάεδρο με "half orientation", δηλαδή τα gradients που έχουν την ίδια διεύθυνση αλλά διαφορετική φορά κβαντοποιούνται στην ίδια ράβδο (bin), τεχνική που χρησιμοποιείται και στο [6]. Οπότε παράγεται για κάθε σημείο ένας περιγραφητής μεγέθους $5 \times 5 \times 4 \times 20/2 = 1000$ ράβδων. Για τη Hollywood2 Action Dataset χρησιμοποιήθηκαν αντίστοιχα οι παράμετροι $k_s = 12, k_t = 6, n_x = 2, n_y = 2, n_t = 5$ ενώ για την κβαντοποίηση των κατευθύνσεων χρησιμοποιήθηκαν πολικές συντεταγμένες με συνολικό αριθμό 5 bins για τη χωρική διάσταση και 3 για τη χρονική. Οπότε αντίστοιχα καταλήγουμε σε περιγραφητή μεγέθους $2 \times 2 \times 5 \times 5 \times 3 = 300$ ράβδων. Αφού ολοκληρώνεται και η εξαγωγή των χαρακτηριστικών απομένει το κομμάτι της αναγνώρισης προτύπων

(ανθρώπινων δράσεων).



Σχήμα 5.1: Διαδικασία Ταξινόμησης με τη χρήση SVMs. Μετά την οπτική επεξεργασία και την εξαγωγή των περιγραφητών, τα δεδομένα χωρίζονται σε δεδομένα εκπαίδευσης και ταξινόμησης. Από τα δεδομένα εκπαίδευσης κατασκευάζεται το οπτικό λεξιλόγιο μέσω του οποίου για κάθε video φτιάχνεται η αναπαράσταση BoF. Τέλος, ακολουθεί η διαδικασία ταξινόμησης με γραμμικά και μη γραμμικά SVMs.

Κατόπιν, τα δεδομένα χωρίζονται σε δεδομένα εκπαίδευσης (Training Data) και δεδομένα αξιολόγησης (Test Data). Ένα υποσύνολο των δεδομένων (περιγραφητές) αξιολόγησης τροφοδοτείται στον αλγόριθμο συσταδοποίησης K-means για να εξαχθούν τα κέντρα των ομάδων, ώστε να ακολουθήσει η κατασκευή των Bag-of-Features ιστογραμμάτων. Οι Wang et al. [60] προτείνουν την τιμή $V = 4000$ για το πλήθος των ομάδων, αριθμός που καταλήξαμε ότι είναι βέλτιστος πειραματικά και στα πειράματα της παρούσας διπλωματικής εργασίας. Για το πλήθος των δεδομένων με τα οποία εξάγονται τα κέντρα από τον K-means, χρησιμοποιήσαμε ένα υποσύνολο 100000 τυχαία επιλεγμένων χαρακτηριστικών από τα δεδομένα εκπαίδευσης, όπως στο [60]. Για την εφαρμογή του αλγόριθμου K-means χρησιμοποιήθηκε η mex υλοποίηση του INRIA[1],με μέγιστο αριθμό επαναλήψεων

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

για σύγκλιση ίσο με 100.

Αφού εξαχθούν τα κέντρα, κατασκευάζεται για το κάθε video εισόδου ένα ιστόγραμμα Bag-of-Features από τα χαρακτηριστικά του. Κάθε περιγραφητής συμβάλλει στη διαμόρφωση του BoF ιστογράμματός του "ψηφίζοντας" στη ράβδο του κέντρου από το οποίο έχει τη μικρότερη Ευκλείδια απόσταση. Θα πρέπει να ξεκαθαρίσουμε ότι η παραπάνω διαδικασία αφορά και τα video που ανήκουν στο training set αλλά και αυτά που ανήκουν στο test set, ενώ τα κέντρα έχουν εξαχθεί χρησιμοποιώντας χαρακτηριστικά μόνο από το training set.

Το τελικό βήμα της διαδικασίας είναι η ταξινόμηση με τη χρήση SVMs. Για την υλοποίηση των SVMs χρησιμοποιήσαμε τη βιβλιοθήκη LIBSVM [4], που παρέχει ένα ολοκληρωμένο περιβάλλον επιλογών για τη χρήση τους. Σε ένα πρόβλημα όπως η αναγνώριση δράσεων σε video χρειάζεται ταξινόμηση των δεδομένων σε πολλές κλάσεις. Επειδή τα SVMs είναι discriminative μοντέλα, δηλαδή διαχωρίζουν τα δεδομένα με βέλτιστο τρόπο, έχει επικρατήσει (για το συγκεκριμένο πρόβλημα) η τεχνική να μη χρησιμοποιείται ένα SVM για όλες τις κλάσεις, αλλά η προσέγγιση *'Ενας-εναντίων-όλων* (*One-against-all*). Δηλαδή, αν έχουμε N διαφορετικές κλάσεις, εκπαιδεύουμε N διαφορετικά SVMs και κάθε δείγμα εξέτασης αποδίδεται στην κλάση της οποίας η συνάρτηση απόφασης έδωσε τη μεγαλύτερη τιμή, η οποία μεταφράζεται ως μια πιθανότητα το δείγμα να ανήκει στην συγκεκριμένη κλάση, σύμφωνα με το [47]. Με τη συγκεκριμένη τεχνική, γίνεται εκπαίδευση για κάθε κλάση ξεχωριστά, χρησιμοποιώντας ως αρνητικά παραδείγματα αυτά όλων των υπολοίπων κλάσεων.

Η χρήση μη γραμμικών SVMs στη βιβλιογραφία, συνήθως οδηγεί σε βελτιωμένες επιδόσεις σε σχέση με τα γραμμικά SVMs. Ειδικά για ιστογραφικά χαρακτηριστικά, όπως αυτά που προκύπτουν από την τεχνική Bag-of-Features, επιλέγεται συνήθως η χ^2-απόσταση, η οποία φαίνεται να αποδίδει καλύτερα σε σχέση με άλλες επιλογές. Συνεπώς και εμείς επεκτείναμε τον πειραματισμό μας χρησιμοποιώντα εκτός από γραμμικούς πυρήνες και σε μη γραμμικούς, με χ^2-απόσταση (η οποία δεν περιέχεται στην παρούσα έκδοση του LIBSVM). Για την βελτιστοποίηση της ταξινόμησης με SVMs, υπάρχει μια παράμετρος C η οποία αφορά την τιμωρία που επιβάλλει ο ταξινομητής σε περίπτωση λάθος ταξινόμησης των δεδομένων του Training Set, μετακινώντας το υπερεπίπεδο κατά μια απόσταση ανάλογη του C. Δηλαδή, για μια πολύ μικρή τιμή του C θα ψάξει για ένα διαχωριστικό υπερεπίπεδο με μεγάλα όρια, ακόμα και αν αυτό οδηγεί σε περισσότερες λάθος ταξινομήσεις. Αντίθετα για μεγαλύτερες τιμές του C ο ταξινομητής θα επιλέξει ένα διαχωριστικό υπερεπίπεδο με μικρότερα όρια, αν αυτός οδηγήσει σε περισσότερες σωστές ταξινομήσεις των δεδομένων του Training Set.

Η ταξινόμηση με kNN Classifiers ακολουθεί την ίδια ακριβώς διαδικασία μέχρι το κομμάτι της κατασκευής των BoF ιστογραμμάτων. Στη συνέχεια ορίζεται ο αριθμός των γειτόνων με βάση τους οποίους θα γίνει η ταξινόμηση, καθώς και το

μετρικό απόστασης που θα χρησιμοποιηθεί, και προχωράμε στην ταξινόμηση.

Το μετρικό επίδοσης που χρησιμοποιήθηκε σε όλα τα πειράματά ταξινόμησης είναι η Ακρίβεια Αναγνώρισης (Accuracy). Έαν H είναι ο αριθμός των σωστά αναγνωρισμένων δράσεων από τον αλγόριθμό μας και TN ο συνολικός αριθμός των εκτελέσεων δράσεων προς αναγνώριση, τότε το accuracy ορίζεται ως:

$$Accuracy = \frac{H}{TN}. \tag{5.1}$$

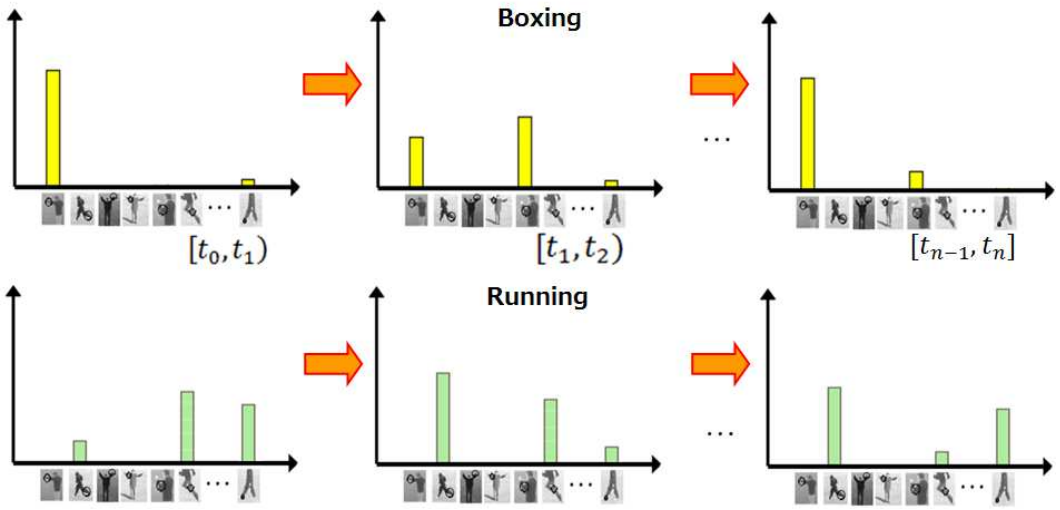
5.1.2 Ταξινόμηση Ανθρώπινων Δράσεων με HMMs

Για την υλοποίηση των HMMs χρησιμοποιήσαμε το γνωστό toolbox HTK [64], το οποίο παρέχει υλοποιήσεις των βασικών αλγορίθμων για HMMs και αποτελεί ένα ολοκληρωμένο περιβάλλον χρήσης HMMs. Όπως παρουσιάστηκε στην υποενότητα 4.3.3 τα HMMs έχουν το πλεονέκτημα πως μπορούν να περιγράψουν τη χρονική εξέλιξη ενός φαινομένου. Αυτό φαντάζει αρκετά αισιόδοξο για ένα πρόβλημα όπως η αναγνώριση δράσεων, το οποίο έχει την ιδιότητα να εξελίσσεται δυναμικά. Το χαρακτηριστικό που επιλέχτηκε να παρακολουθηθεί στο χρόνο στην προσέγγισή μας είναι το ιστόγραμμα Bag-of-Features. Αντίθετα με την περίπτωση της ταξινόμησης με χρήση SVMs, τα ιστογράμματα BoF που κατασκευάζουμε έχουν μια χρονική εξέλιξη, την οποία επιτυγχάνουμε ως εξής. Μέσω ενός κινούμενου στο χρόνο παραθύρου χρησιμοποιούμε χαρακτηριστικά που ανιχνεύθηκαν εντός του παραθύρου για την κατασκευή ενός ιστογράμματος. Με αυτόν τον τρόπο, κάθε χρονική στιγμή έχουμε ένα BoF το οποίο εξελίσσεται, και αυτήν την εξέλιξη προσπαθήσαμε να μοντελοποιήσουμε με τη χρήση των HMMs. Η ιδέα φαίνεται στο Σχήμα 5.2.

Για την εκπαίδευση των μοντέλων μας χρησιμοποιήθηκαν τα εργαλεία HERest του HTK, τα οποία αποτελούν υλοποιήσεις του αλγορίθμου Baum-Welch. Σε κάθε περίπτωση, με δεδομένο το σύνολο εκπαίδευσης, όλα τα παραδείγματα που εντάσσονταν σε αυτό, χρησιμοποιήθηκαν για την εκπαίδευση του συστήματος. Η διαδικασία της εκπαίδευσης περιελάμβανε την εκπαίδευση ενός μοντέλου για κάθε δράση του λεξιλογίου μας.

Ο αριθμός καταστάσεων αλλά και το πλήθος των Γκαουσιανών ανά κατάσταση δεν τέθηκαν σταθερές σε όλες τις περιπτώσεις, αλλά αποτέλεσαν αντικείμενο πειραματισμού. Χρησιμοποιήσαμε 3-5 καταστάσεις με 1-2 Γκαουσιανές ανά κατάσταση.

Τέλος, όσον αφορά την τοπολογία των HMMs, επιλέξαμε να χρησιμοποιήσουμε αποκλειστικά left-to-right μοντέλα (και πιο συγκεκριμένα μοντέλα που επιτρέπουν παραμονή στην ίδια κατάσταση, ή μετάβαση μόνο στην επόμενη κατάσταση, αντί για οποιαδήποτε επόμενη). Αυτή η επιλογή έγινε, γιατί θεωρούμε ότι τα left-to-right μοντέλα αποτελούν τη "φυσική" επιλογή για ένα φαινόμενο που εξελίσσεται στο χρόνο, χωρίς να έχουμε επιστροφή σε προηγούμενες καταστάσεις.

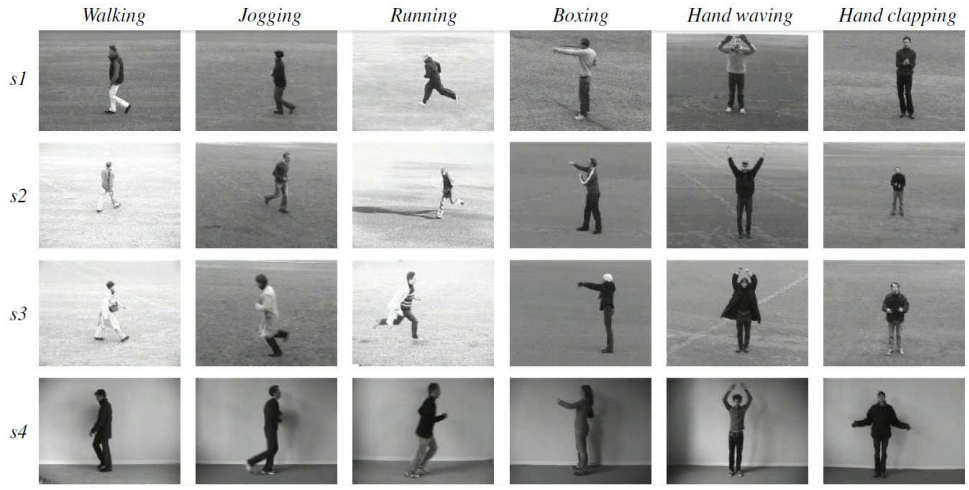


Σχήμα 5.2: Η ιδέα πίσω από τη χρήση των HMMs. Κάθε χρονική στιγμή έχουμε ένα BoF που έχει κατασκευαστεί από χαρακτηριστικά ανιχνευμένα εντός του χρονικού παραθύρου, δίνοντας μια χρονική εξέλιξη του ιστογράμματος.

5.2 Πειράματα Ταξινόμησης Ανθρώπινων Δράσεων στη Βάση Δεδομένων *KTH*

Η βάση δεδομένων *KTH Action Datset* [53] είναι ίσως η πιο γνωστή βάση δεδομένων ανθρώπινων δράσεων στη βιβλιογραφία. Αποτελείται από 6 ανθρώπινες δράσεις:Walking, Jogging, Running, Boxing, HandWaving και HandClapping. Υπάρχουν 25 υποκείμενα που εκτελούν τις εν λόγω δράσεις, σε 4 διαφορετικά σενάρια: σε εξωτερικό χώρο ($s1$), σε εξωτερικό χώρο με μεταβολές της κλίμακας ($s2$), σε εξωτερικό χώρο με διαφορετικά ρούχα ($s3$) και σε εσωτερικό χώρο ($s4$), όπως φαίνεται και στο Σχήμα. 5.3. Το κάθε υποκείμενο εκτελεί κατά κανόνα 4 επαναλήψεις σε κάθε σενάριο κάθε δράσης. Η βάση δεδομένων περιέχει 2391 εκτελέσεις δράσεων προς ταξινόμηση. Όλες οι λήψεις έγιναν σε ομογενές background με μια στατική κάμερα με συχνότητα δειγματοληψίας 25 frames ανά δευτερόλεπτο. Τα video της KTH Action Dataset έχουν ανάλυση 160×120 pixels και διαρκούν κατά μέσο όρο 4 δευτερόλεπτα. Η συνήθης τακτική που ακολουθείται [34, 60, 17, 66, 44, 12, 61] στη συγκεκριμένη βάση είναι να γίνονται πειράματα άγνωστου χρήστη (Unseen subject experiments), με 16 από τους 25 χρήστες να αποτελούν τα δεδομένα εκπαίδευσης και οι υπόλοιποι 9 να χρησιμοποιούνται για την αξιολόγηση του συστήματος. Η τακτική αυτή ακολουθήθηκε και στα

πειράματά μας ώστε να μπορούμε να συγκρίνουμε τα αποτελέσματά μας με τις ήδη υπάρχουσες μέθοδους της βιβλιογραφίας.



Σχήμα 5.3: Μερικά στιγμιότυπα της βάσης KTH Action Dataset [53] με λήψεις από διαφορετικούς χρήστες να εκτελούν διαφορετικές δράσεις σε διαφορετικά σενάρια.

Η συγκεκριμένη βάση δεδομένων[2], λόγω της απλότητάς της και του σχετικά μικρού όγκου δεδομένων που περιέχει ήταν και αυτή στην οποία πειραματιστήκαμε περισσότερο. Έγιναν πειράματα με πολλούς Classifiers (SVMS, HMMs, kNN), έγινε χρήση πολλών ανιχνευτών, ενώ ήταν και η βάση στην οποία καθορίστηκαν οι παράμετροι που χρησιμοποιήσαμε για την ανάπτυξη του αλγόριθμου Gabor3D και αναλύσαμε στην Ενότητα 3.3.2. Παρακάτω παραθέτουμε τα πειραματικά αποτελέσματά μας.

5.2.1 Πειράματα Ταξινόμησης με τη χρήση SVMs

Η διαδικασία με την οποία έγιναν τα πειράματα με SVMs στην KTH Action Dataset περιγράφηκε παραπάνω. Ακολουθούν τα αποτελέσματα που εξάγαμε αλλάζοντας το κομμάτι της εξαγωγής χωροχρονικών σημείων ενδιαφέροντος. Πειραματιστήκαμε με 4 ανιχνευτές (Harris3D, TDE, Gabor3D (Full Version), Gabor3D (Reduced Version)) και συγκρίναμε με επιπλέον 2 ανιχνευτές της βιβλιογραφίας (DCA3D και Cuboids). Επίσης πειραματιστήκαμε με δύο περιγραφητές, τον HOG/HOF και τον HOG3D. Σε κάθε περίπτωση ακολουθήθηκε πιστά η πειραματική διαδικασία που περιγράφεται στα [53, 60].

[2] `http://www.nada.kth.se/cvap/actions`

	Ακρίβεια Αναγνώρισης Μεθόδου					
Περιγραφητής	DCA3D [17]	Cuboids [60]	Harris3D [34]	TDE	Gabor3D (Full)[37]	Gabor3D (Reduced)[37]
HOG/HOF	78.8%	88.7 %	91.8%	91.2%	91.2%	-
HOG3D	-	90.0 %	89.0%	**93.75%**	**93.5%**	**93.4%**

Πίνακας 5.1: Ακρίβεια αναγνώρισης διάφορων μεθόδων για τη Βάση Δεδομένων KTH Action Dataset. Οι αλγόριθμοί μας φαίνεται να ξεπερνούν τις μεθόδους της βιβλιογραφίας.

Παρατηρούμε ότι στη γενικότερη περίπτωση ο περιγραφητής HOG3D δίνει καλύτερα αποτελέσματα από τον HOG/HOF. Αυτό δε συμβαίνει για την περίπτωση του ανιχνευτή Harris3D, ο οποίος παρουσιάζει καλύτερα αποτελέσματα με τον HOG/HOF. Με το συγκεκριμένο ανιχνευτή, εκτός από τα αποτελέσματα της βιβλιογραφίας που παραθέσαμε, έγινε μια προσπάθεια να αναπαραχθούν τα αποτελέσματα ακολουθώντας την ίδια πειραματική διαδικασία. Στα δικά μας πειράματα ο Harris3D απέδωσε ακρίβεια 90.32% με τον HOG/HOF και 90.03% με τον HOG3D, πράγμα που σημαίνει ότι η συνεργασία του με τον HOG3D περιγραφητή ήταν χειρότερη ξανά.

Ενδιαφέρον παρουσιάζει ότι ο αλγόριθμος TDE ο οποίος αποτελεί απλοποιημένη έκδοση του DCA3D παρουσιάζει τη βέλτιστη συμπεριφορά, με μικρή διαφορά από τις παραλλαγές του Gabor3D. Στη συγκεκριμένη βάση λόγω του ότι οι λήψεις γίνονται από σταθερή κάμερα, στις περισσότερες περιπτώσεις (εκτός του σεναρίου s2) όλες οι ανιχνεύσεις προέρχονται από κίνηση που εκτελεί το υποκείμενο. Ο TDE, λόγω της φύσης του (5 χρονικά Gabor φίλτρα) είναι και ο πιο ευαίσθητος αλγόριθμος στις μεταβολές της κίνησης, και παρουσιάζει τις περισσότερες ανιχνεύσεις, σωστά εστιασμένες πάνω στο υποκείμενο, φτιάχνοντας όπως φαίνεται και από το αποτέλεσμα ένα τελικό ιστόγραμμα που έχει τα πιο ευδιάκριτα χαρακτηριστικά για καθεμία από τις επιμέρους δράσεις. Η συμπεριφορά του αλγορίθμου αλλάζει, ωστόσο, όταν στα video προς ταξινόμηση προστεθεί κίνηση από την κάμερα, όπως θα δείξουμε στα πειράματά μας στη βάση Hollywood2 Action Dataset.

Άλλο ένα ενδιαφέρον στοιχείο των αποτελεσμάτων είναι ότι αν χρησιμοποιηθεί η απλοποιημένη μορφή του αλγόριθμου Gabor3D, δηλαδή μειώνοντας τον αριθμό των φίλτρων από 400 σε 120 όπως περιγράψαμε στην υποενότητα 3.3.3, επιτυγχάνεται η σχεδόν ίδια ακρίβεια αναγνώρισης στο $1/3$ του χρόνου που απαιτείται για τον αλγόριθμο στην πλήρη μορφή του. Το πολύ αισιόδοξο αποτέλεσμα δείχνει ότι ο αλγόριθμος στην πλήρη μορφή του ίσως χρησιμοποιεί υπερβολικό για το δεδομένο πρόβλημα αριθμό φίλτρων, πράγμα που αν αναλογιστεί κανείς και το κόστος σε χρόνο και πολυπλοκότητα, αξίζει σίγουρα

να μειωθεί.

Παραθέτουμε τους πίνακες σύγχυσης (Confusion Matrices) για τους ανιχνευτές TDE (Πίνακας 5.2) και Gabor3D (Πίνακας 5.3). Παρατηρούμε και στις 2 περιπτώσεις ότι η μεγαλύτερη σύγχυση προκαλείται μεταξύ των δράσεων Jogging και Running, που είναι λογικό λόγω της φύσης των δύο αυτών δράσεων. Επίσης παρατηρούμε πως δεν υπάρχει (σχεδόν) καμία σύγχυση μεταξύ των δράσεων που εκτελούνται κυρίως με τα χέρια (Boxing, Waving, Clapping) και των δράσεων που εκτελούνται με τα πόδια (Walking, Jogging, Running).

Δράση	Walking	Jogging	Running	Boxing	Waving	Clapping
Walking	1.00	0	0	0	0	0
Jogging	0.01	0.88	0.11	0	0	0
Running	0	0.17	0.83	0	0	0
Boxing	0.02	0	0	0.98	0	0
Waving	0	0	0	0	0.95	0.05
Clapping	0	0	0	0.01	0	0.99

Πίνακας 5.2: Ο πίνακας σύγχυσης (Confusion Matrix) του πειράματος ταξινόμησης στη βάση KTH με τον αλγόριθμο TDE. Η ακρίβεια αναγνώρισης είναι 93.75 %.

Δράση	Walking	Jogging	Running	Boxing	Waving	Clapping
Walking	1.00	0	0	0	0	0
Jogging	0.03	0.88	0.09	0	0	0
Running	0	0.20	0.80	0	0	0
Boxing	0	0	0	1.00	0	0
Waving	0	0	0	0	0.95	0.05
Clapping	0	0	0	0.02	0	0.98

Πίνακας 5.3: Ο πίνακας σύγχυσης (Confusion Matrix) του πειράματος ταξινόμησης στη βάση KTH με τον αλγόριθμο Gabor3D. Η ακρίβεια αναγνώρισης είναι 93.5 %.

Στην KTH Action Dataset καθορίστηκαν πειραματικά οι παράμετροι του αλγόριθμου Gabor3D. Χρησιμοποιήθηκαν δύο είδη ενεργειών (Τετραγωνική και Teager-Kaiser, δύο τρόπους διαχείρισης των εξόδων των φίλτρων (Ανάλυση Κυρίαρχης Ενέργειας-Max και Υπέρθεση Ενεργειών-Mean), και δύο πυρήνες

SVMs (ο γραμμικός και ο χ^2 πυρήνας). Έγιναν πειράματα με όλους τους δυνατούς συνδυασμούς, τα αποτελέσματα των οποίων φαίνονται στον Πίνακα. 5.4.

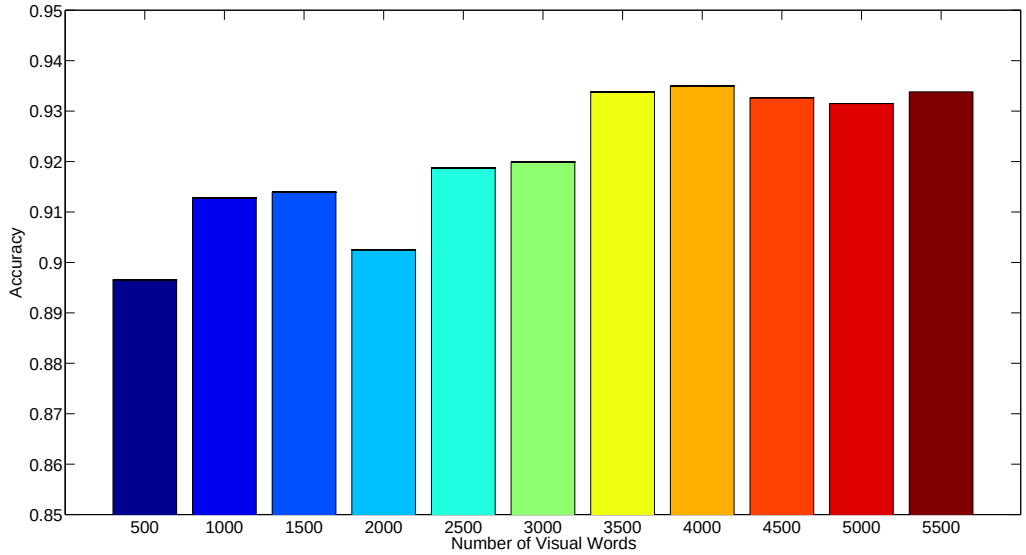
	Γραμμικά SVMs		χ^2 SVMs	
	Max	Sum	Max	Sum
Τετραγωνική Ενέργεια	90.85%	89.34%	92.35%	90.61%
Teager-Kaiser Ενέργεια	91.31%	90.73%	**93.50**%	92.47%

Πίνακας 5.4: Ακρίβεια ταξινόμησης για τη βάση KTH χρησιμοποιώντας τον αλγόριθμο Gabor3D, αλλάζοντας τις παραμέτρους του. Μπορεί να παρατηρηθεί ότι οι χ^2 πυρήνες έχουν καλύτερη επίδοση από τους γραμμικούς, η Ανάλυση Κυρίαρχης Χωροχρονικής Ενέργειας (Max) δίνει καλύτερα αποτελέσματα από την Υπέρθεση Ενεργειών (Sum), όπως και η Teager-Kaiser Ενέργεια από την Τετραγωνική Ενέργεια.

Από τα αποτελέσματα των πειραμάτων μπορούμε να παρατηρήσουμε μια γενικότερη υπεροχή κάποιων επιλογών παραμέτρων έναντι άλλων. Φαίνεται κατ' αρχήν πως η αναγνώριση με τη χρήση Teager-Kaiser ενέργειας έναντι Τετραγωνικής Ενέργειας είναι αποδοτικότερη. Αυτό που παρατηρήσαμε είναι ότι ο τελεστής ΤΚΕΟ δίνει μια αναπαράσταση ενέργειας που είναι πιο "τραχιά" συγκριτικά με την απλή τετραγωνική, οδηγώντας σε περισσότερες ανιχνεύσεις (3D τοπικά μέγιστα) πάνω στην περιοχή εκτέλεσης της δράσης, και επομένως πληρέστερη αναπαράσταση του ιστογράμματος.

Για τον ίδιο λόγο η μέθοδος της υπέρθεσης ενεργειών μειονεκτεί σε σχέση με την ανάλυση κυρίαρχης ενέργειας, που σε όλες τις περιπτώσεις δίνει καλύτερα αποτελέσματα. Σαφή διαφορά παρατηρούμε και στο είδος του πυρήνα που χρησιμοποιείται στα SVMs. Επιβεβαιώνουμε τα λεγόμενα της βιβλιογραφίας [60], ότι δηλαδή ο χ^2 πυρήνας είναι αυτός που υπερτερεί στο διαχωρισμό ιστογραφικών χαρακτηριστικών. Διευκρινίζουμε ότι τα πειράματα έγιναν με περιγραφητή HOG3D. Το αποτέλεσμα που παραθέσαμε στον Πίνακα. 5.1 για τον αλγόριθμο Gabor3D είναι αυτό που προκύπτει από το συνδυασμό των επιλογών παραμέτρων που υπερέχουν.

Στο Σχήμα. 5.4 βλέπουμε τον τρόπο με τον οποίο επιδρά η αλλαγή του αριθμού των οπτικών λέξεων του ιστογράμματος Bag-of-Features στην αναγνώριση του συστήματος για τον αλγόριθμο Gabor3D. Παρατηρούμε ότι ο βέλτιστος αριθμός λέξεων, όπως προτείνεται και στο [60] είναι $V = 4000$, με πολύ μικρή διαφορά. Αυτό που συμβαίνει είναι ότι όσο αυξάνουμε τις λέξεις του λεξιλογίου, τόσο περισσότερη λεπτομέρεια αποδίδουμε στο σύστημα αναγνώρισης. Ωστόσο, αν οι λέξεις αυξηθούν υπερβολικά, το σύστημα υπερεκπαιδεύεται σε λεπτομέρειες (overtrain) και η απόδοσή του μειώνεται.



Σχήμα 5.4: Ακρίβεια αναγνώρισης για τον Gabor3D ανιχνευτή αλλάζοντας τον αριθμό των λέξεων της αναπαράστασης BoF.

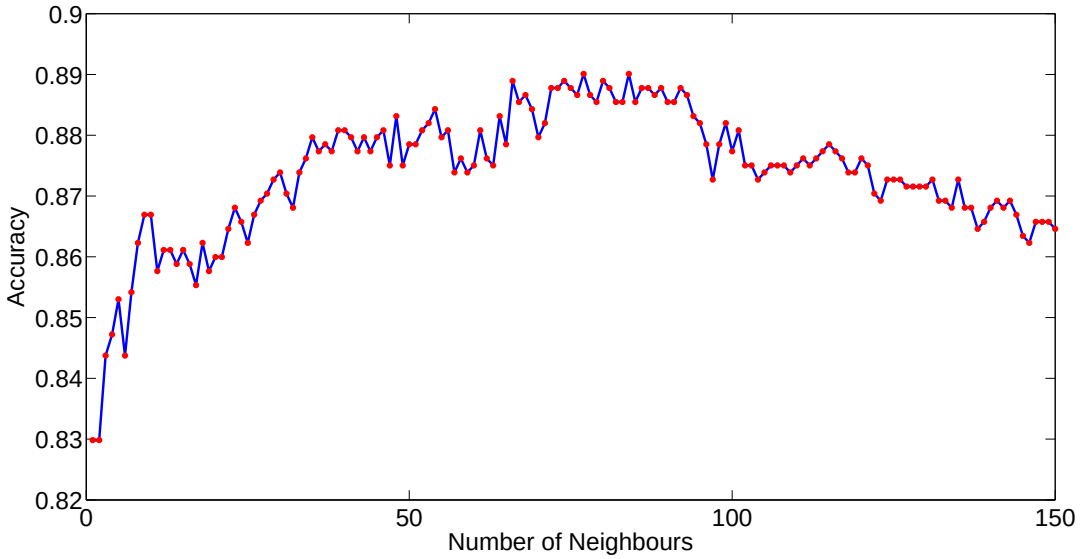
Σημειώνουμε ότι η παράμετρος C που αφορά την τιμώρία των λάθος ταξινομήσεων των δεδομένων του Training Set μετά από πειραματισμό επιλέχτηκε στην τιμή $C = 2$ που έδωσε και τα βέλτιστα αποτελέσματα.

Στο Σχήμα. 5.5 παραθέτουμε τα αποτελέσματα για τον αλγόριθμο Gabor3D με τη χρήση kNN ταξινομητή, αλλάζοντας τον αριθμό των γειτόνων που λαμβάνονται υπ᾽ όψη για να γίνει η ταξινόμηση.

Παρατηρούμε ότι η ακρίβεια ταξινόμησης ανεβαίνει μέχρι ένα σημείο προσθέτοντας γείτονες, ωστόσο δε φτάνει τα ποσοστά αναγνώρισης με τη χρήση SVMs. Οι παράμετροι που χρησιμοποιήθηκαν είναι αυτοί που οδήγησαν στο καλύτερο αποτέλεσμα με χρήση SVMs. Ως μετρικό απόστασης από τους γείτονες χρησιμοποιήθηκε η Ευκλείδια απόσταση.

5.2.2 Πειράματα Ταξινόμησης με τη χρήση HMMs

Στην KTH Action Dataset πραγματοποιήθηκαν πειράματα με τη χρήση ταξινομητών HMM, με τη διαδικασία που περιγράφηκε στην υποενότητα 5.1.2. Τα χαρακτηριστικά που εξάγονται από κάθε video εισόδου σαρώνονται από ένα κινούμενο χρονικό παράθυρο το οποίο κατασκευάζει ένα ιστόγραμμα BoF για κάθε χρονική στιγμή από τα χαρακτηριστικά εντός του παραθύρου. Αν V η διάσταση του ιστογράμματος (μέγεθος λεξιλογίου) και W_T το σύνολο των χρονικών στιγμών στις οποίες κεντράρουμε το χρονικό παράθυρο, τότε έχουμε έναν πίνακα $W_T \times V$



Σχήμα 5.5: Ακρίβεια αναγνώρισης με χρήση kNN Classifier, αλλάζοντας τον αριθμό των γειτόνων. Η μέγιστη τιμή της ακρίβειας είναι 89.04%, για 77 γείτονες.

προς ταξινόμηση για κάθε δράση.
Ωστόσο στα αρχικά μας πειράματα παρατηρήσαμε ότι το σύστημά μας δεν μπορούσε να εκπαιδευτεί, παρά μόνο εαν μειώναμε κατά πολύ το μέγεθος του λεξιλογίου, καθώς τα ιστογράμματα BoF αποτελούν μια πολύ αραιή αναπαράσταση (sparse representation) των χαρακτηριστικών, με πολλά μηδενικά σε ένα σχετικά μεγάλο διάνυσμα χαρακτηριστικών προς παρακολούθηση (ο βέλτιστος αριθμός των λέξεων του λεξιλογίου είναι $V = 4000$ [60]). Η εκαίδευση του μοντέλου δε μπορούσε να επιτευχθεί λόγω της μεγάλης διάστασης του διανύσματος παρακολούθησης, αλλά και λόγω της αραιής φύσης του. Για να αποφευχθεί το παραπάνω πρόβλημα πρέπει να μειώσουμε τη διάσταση του διανύσματος προς παρακολούθηση, με τέτοιο τρόπο ώστε να διατηρείται η πληροφορία του. Η μείωση του αριθμού των λέξεων είναι μια λύση που ωστόσο "θολώνει" την πληροφορία που εμπεριέχεται σε αυτές, οδηγώντας σε μη ικανοποιητικά αποτελέσματα.

Η προσέγγιση που ακολουθήσαμε πηγάζει από το πεδίο της Ψηφιακής Επεξεργασίας Σήματος. Συγκεκριμένα, προσπαθήσαμε να κωδικοποιήσουμε το κάθε ιστόγραμμα με Linear Predictive Coding (LPC) συντελεστές και τους μετασχηματισμούς του, δηλαδή συντελεστές PARCOR και LogArea [50]. Η συγκεκριμένη προσέγγιση επιτυγχάνεται εαν υποθέσουμε ότι το ιστόγραμμά μας είναι το φάσμα ισχύος ενός πραγματικού σήματος συσχέτισης, για το οποίο προσεγγίζουμε τους συντελεστές Γραμμικής Πρόβλεψης. Απαιτείται ωστόσο ένα στάδιο προεπεξεργασίας, λόγω της φύσης των ιστογραμμάτων. Η κάθε λέξη του

λεξιλογίου αποτελεί το κέντρο κάποιας συστάδας που προκύπτει από τον αλγόριθμο K-means. Γειτονικές ράβδοι αποτελούν διαφορετικές λέξεις που η θέση τους στο ιστόγραμμα δεν έχει καμία φυσική σημασία, οπότε τελικά προσπαθούμε να προσεγγίσουμε ένα ιστόγραμμα που έχει απότομες μεταβολές από ράβδο σε ράβδο.

Το στάδιο προεπεξεργασίας που υλοποιήσαμε είχε ως στόχο να μειώσουμε όσο γίνεται τις απότομες αυτές μεταβολές ανάμεσα σε γειτονικές λέξεις του ιστογράμματος. Η ιδέα ήταν να συγκεντρώσουμε τα bins που συμμεταβάλλονται στην ίδια γειτονιά ώστε να είναι πιο ομαλή η μορφή των ιστογραμμάτων που θέλουμε να προσεγγίσουμε. Κατασκευάζουμε λοιπόν τον πίνακα συνδιακύμανσης M_C των δεδομένων εκπαίδευσης, ο οποίος έχει διάσταση $V \times V$. Με Singular Value Decomposition ο C_V μπορεί να γραφεί ως εξής:

$$M_C = U_C \cdot S_C \cdot V_C^T \tag{5.2}$$

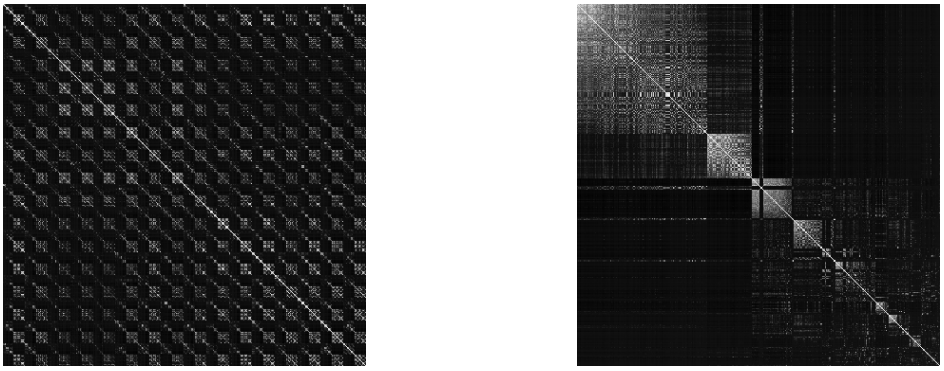
όπου S_C είναι ένας διαγώνιος πίνακας $(V \times V)$ του οποίου τα διαγώνια στοιχεία είναι ιδιοτιμές του πίνακα M_C, U_C είναι ένας ορθογώνιος πίνακας $(V \times V)$ του οποίου οι στήλες είναι αριστερά ιδιοδιανύσματα του πίνακα M_C και αντίστοιχα ο V_C είναι ένας ορθογώνιος πίνακας $(V \times V)$ του οποίου οι στήλες είναι δεξιά ιδιοδιανύσματα του M_C.

$$M_C = \begin{pmatrix} U_{1,1} & U_{1,2} & \cdots & U_{1,V} \\ U_{2,1} & U_{2,2} & \cdots & U_{2,V} \\ \vdots & \vdots & \ddots & \vdots \\ U_{V,1} & U_{V,2} & \cdots & U_{V,V} \end{pmatrix} \begin{pmatrix} S_{1,1} & 0 & \cdots & 0 \\ 0 & S_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & S_{V,V} \end{pmatrix} \begin{pmatrix} V_{1,1} & V_{1,2} & \cdots & V_{1,V} \\ V_{2,1} & V_{2,2} & \cdots & V_{2,V} \\ \vdots & \vdots & \ddots & \vdots \\ V_{V,1} & V_{V,2} & \cdots & V_{V,V} \end{pmatrix}$$

με $S_{1,1} \geq S_{2,2} \geq, \cdots, S_{V,V} \geq 0$.

Έαν θεωρήσουμε τον πίνακα U_C ως έναν πίνακα βαρών τότε εξετάζοντας τα στοιχεία του μπορούμε να κρατήσουμε τις γραμμές (ράβδους) του πίνακα για τις οποίες ο πίνακας $U_C \cdot S_C$ έχει μεγάλα βάρη σε κάθε ιδιοτιμή, και να τις τοποθετήσουμε δίπλα δίπλα στο ιστόγραμμα BoF. Έτσι, εξετάζουμε πρώτα τις γραμμές που έχουν μεγάλα βάρη για την ιδιοτιμή $S_{1,1}$, ύστερα για την $S_{2,2}$ εξαντλητικά, μέχρι να περάσουμε από όλες τις ράβδους του αρχικού ιστογράμματος. Στο Σχήμα. 5.6 φαίνεται το αποτέλεσμα της διαδικασίας που περιγράφηκε για ιστογράμματα μεγέθους 500 λέξεων των δεδομένων εκπαίδευσης της KTH Action Dataset. Στο Σχήμα. 5.6α′ φαίνεται ο αρχικός πίνακας συνδιακύμανσης, ενώ στο Σχήμα. 5.6β′ φαίνεται ο πίνακας συνδιακύμανσης των δεδομένων με αναδιατεταγμένες ράβδους bins των ιστογραμμάτων, με τη μέθοδο SVD.

Φαίνεται ότι ο πίνακας συνδιακύμανσης έχει πιο συγκεντρωμένες τις σχετικά μεγάλες τιμές γύρω από τη διαγώνιο, πράγμα που σημαίνει ότι bins των ιστογραμμάτων που συμμεταβάλλονται έχουν τοποθετηθεί σε κοινή γειτονιά. Αυτό



(α') Αρχικός πίνακας συνδιακύμανσης

(β') Τελικός πίνακας συνδιακύμανσης με αναδιατεταγμένα τα bins του αρχικού ιστογράμματος

Σχήμα 5.6: Πίνακες συνδιακύμανσης πριν και μετά την εφαρμογή της μεθόδου SVD με βάρη. Αρχίζοντας από την πρώτη ιδιοτιμή κρατάμε τα bins που έχουν μεγάλο βάρος στη στγκεκριμένη ιδιοτιμή, μέχρι να διατρέξουμε όλα τα bins. Η τιμή των βαρών κατωφλιώνεται στο 30% της μέγιστης για την κάθε ιδιοτιμή.

διευκολύνει το επόμενο βήμα της τεχνικής που εφαρμόζουμε, που είναι η προσέγγιση των ιστογραμμάτων με συντελεστές γραμμικής πρόβλεψης. Αν $H(k)$ είναι το ιστόγραμμα που θέλουμε να προσεγγίσουμε, κατασκευάζουμε το άρτιας συμμετρίας σήμα:

$$S(k) = \begin{cases} H(k) & \text{για } 0 \leq k \leq V \\ H(-k) & \text{για } -V \leq k < 0 \end{cases} \tag{5.3}$$

το οποίο θεωρούμε ως πυκνότητα φάσματος ενός πραγματικού - για αυτό και η άρτια συμμετρία - σήματος συσχέτισης $r(\tau)$ το οποίο προέρχεται από τον αντίστροφο μετασχηματισμό DFT του σήματος $S(k)$.

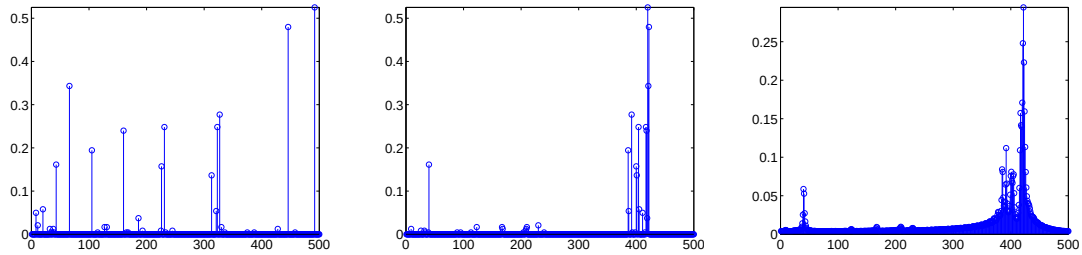
$$r(\tau) = IDFT\{S(k)\} \tag{5.4}$$

Στο προκύπτον σήμα $r(\tau)$ υπολογίζουμε μέσω του αλγόριθμου Levinson-Durbin τους συντελεστές LPC $a_k, k = 1, \cdots, P$, όπου P η τάξη του προβλέπτη και το κέρδος G. Το αρχικό ιστόγραμμα μπορεί να ανακατασκευαστεί από τους συντελεστές LPC και το κέρδος G ως:

$$H_{LPC}(\omega) = \frac{G}{|1 - \sum_{l=1}^{P} a_k \cdot e^{-j\omega l}|}, \quad \text{όπου } \omega = \frac{2\pi k}{N_{DFT}} \tag{5.5}$$

με N_{DFT} ίσο με το μέγεθος του σήματος $S(k)$ [50]. Στο Σχήμα. 5.7 φαίνεται η διαδικασία που ακολουθήθηκε για τη σχετικά ομαλή ανακατασκευή ενός

ιστογράμματος 500 λέξεων με γραμμικό προβλέπτη τάξης 140. Τα bins του αρχικού ιστογράμματος αναδιατάχτηκαν ώστε να τοποθετηθούν γειτονικά τιμές των ιστογραμμάτων που συμμεταβάλλονται. Το ιστόγραμμα που προκύπτει φαίνεται να προσεγγίζει με ικανοποιητική ακρίβεια το αρχικό ιστόγραμμα. Με



(α') Αρχικό ιστόγραμμα μεγέθους 500 λέξεων.

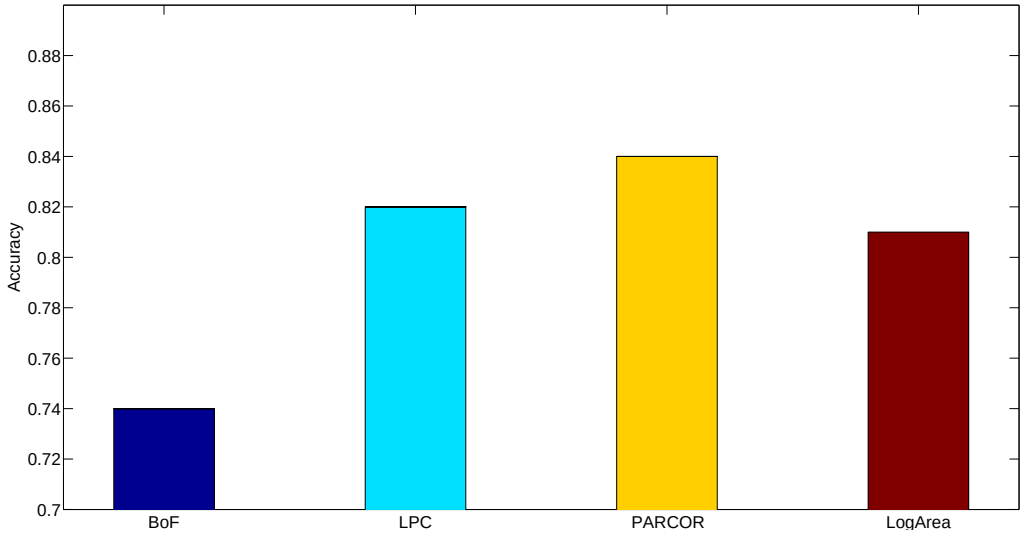
(β') Αναδιατεταγμένο ιστόγραμμα με τη χρήση της τεχνικής SVD.

(γ') Ανακατασκευασμένο ιστόγραμμα από γραμμικό προβλέπτη (LPC) τάξης 140

Σχήμα 5.7: Ανακατασκευή ιστογράμματος με γραμμική πρόβλεψη (LPC), αφού αναδιαταχθούν οι ράβδοι του ώστε να ομαλοποιηθεί η μορφή του. Χρησιμοποιήθηκε προβλέπτης τάξης 140 για ιστόγραμα μεγέθους 500.

την τεχνική αυτή καταφέραμε να συμπιέσουμε την πληροφορία που εμπεριέχεται σε ιστόγραμμα 500 λέξεων σε ένα διάνυσμα LPC συντελεστών μεγέθους 141 (μαζί με το κέδρος G), στο συγκεκριμένο παράδειγμα. Χρησιμοποιώντας τα μειωμένου μεγέθους αυτά διανύσματα έναντι των BoF ιστογραμμάτων επιτυγχάνεται η διαδικασία εκπαίδευσης με τη χρήση HMMs. Εκτός από τους συντελεστές γραμμικής πρόβλεψης χρησιμοποιήθηκαν και μετασχηματισμοί αυτών, όπως είναι οι συντελεστές PARCOR και οι συντελεστές LogArea.

Στο Σχήμα. 5.8 παρατίθεται το βέλτιστο αποτέλεσμα που επιτύχαμε με τη χρήση καθεμίας από τις διαφορετικές εναλλακτικής. Παρατηρούμε ότι για το απλό ιστόγραμμα BoF η ακρίβεια είναι αρκετά μειωμένη διότι για να εκπαιδεύσουμε το σύστημά μας αναγκαστήκαμε να μειώσουμε τον αριθμό των λέξεων τόσο που το λεξιλόγιό μας δεν ήταν επαρκές ώστε να γίνονται διακρίσεις χαρακτηριστικών.



Σχήμα 5.8: Ακρίβεια αναγνώρισης που επιτεύχθηκε με τη χρήση HMMs. Γίνεται σύγκριση των ιστογραμμάτων BoF με τους συντελεστές LPC και τους μετασχηματισμούς τους.

Παρατηρούμε επίσης ότι οι 3 διαφορετικές αναπαραστάσεις των συντελεστών LPC δίνουν μια αύξηση στα ποσοστά αναγνώρισης κατά περίπου10%, με μια μικρή υπεροχή των συντελεστών PARCOR. Ωστόσο δεν καταφέραμε να αγγίξουμε τα υψηλά ποσοστά ακρίβειας που επιτεύχθηκαν με τους ταξινομητές kNN και SVM. Αυτό έχει τη λογική εξήγηση ότι όλοι αυτοί οι μετασχηματισμοί που εφαρμόσαμε για να συμπιέσουμε την πληροφορία των ιστογραμμάτων αλλοίωσαν κατά κάποιο τρόπο τη μορφή τους. Φαίνεται επίσης από τα πειραματικά αποτελέσματα ότι η αραιή φύση και η μεγάλη σχετικά διαστασιμότητα των ιστογραμμάτων δε συνεργάζεται καλά με μοντέλα που έχουν πολλές παραμέτρους προς εκπαίδευση, όπως είναι τα HMMs.

5.3 Πειράματα Ταξινόμησης Ανθρώπινων Δράσεων στη Βάση Δεδομένων *Hollywood2*

Η Hollywood2 Action Dataset[3] αποτελεί μια συλλογή δειγμάτων video με ανθρώπινες δράσεις που αποτελούν αποσπάσματα από ταινίες του Hollywood. Οι δράσεις εκτυλίσσονται σε φυσικές ρεαλιστικές σκηνές και υπάρχει μεγάλη μεταβλητότητα στα δείγματα που ανήκουν στην ίδια κατηγορία δράσης ως προς το στήσιμο του σκηνικού, τις κινήσεις της κάμερας, τις γωνίες λήψης, την ταχύτητα

[3] `http://www.di.ens.fr/~laptev/actions/hollywood2`

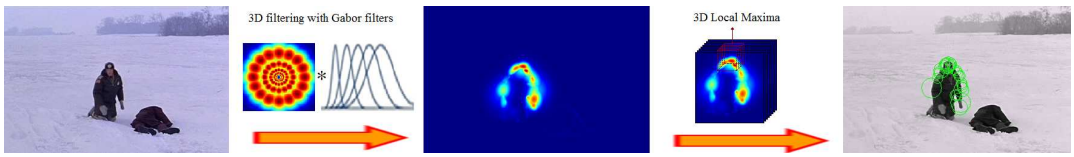
εναλλαγής των σκηνών, το φωτισμό, την ύπαρξη οπτικών εμποδίων και θορύβου ή την εμφάνιση των εικονιζόμενων προσώπων. Πρόκειται για μια από τις πιο απαιτητικές βάσεις δεδομένων για το πρόβλημα της αναγνώρισης ανθρώπινων δράσεων σε video.

Στην Hollywood2 Action Dataset περιέχονται δείγματα video από 69 διαφορετικές ταινίες. Υπάρχουν 12 ανθρώπινες δράσεις: SitDown, StandUp, SitUp, AnswerPhone, DriveCar, Eat, FightPerson, GetOutCar, HandShake, HugPerson, Kiss και Run. Συνολικά η βάση περιέχει 1707 δείγματα video, 823 εκ των οποίων αποτελούν το Training Set και 884 το Test Set. Τα πρώτα αποτελέσματα ταξινόμησης δράσεων στη συγκεκριμένη βάση έγιναν από τους Marszalek et al. [41]. Οι συγγραφείς προτείνουν ως μετρικό αναγνώρισης τον υπολογισμό της μέσης ακρίβειας (average precision) για κάθε κατηγορία δράσης ξεχωριστά και την αναφορά της μέσης τιμής της σε όλες τις δράσεις (mean Average Precision - mAP), τεχνική που έχει καθιερωθεί στη συγκεκριμένη βάση [60, 59, 17]. Ο δείκτης Precision δίνεται ως:

$$precision = \frac{\#TruePositives}{\#TruePositives + \#FalsePositives} \tag{5.6}$$

Το μέγεθος των frames ποικίλει με μια μέση τιμή πλάτους περί τα 600-800 pixels και ύψους περί τα 400 pixels, ενώ κανόνας για τη διάρκεια δεν υπάρχει καθώς υπάρχουν και δείγματα που υπερβαίνουν τα 2 λεπτά. Τα δείγματα video με πολλαπλές επισημειώσεις δράσεων (πχ. Kiss και HugPerson) εξαιρέθηκαν από τα δεδομένα εκπαίδευσης των ταξινομητών.

Η διαδικασία ταξινόμησης που χρησιμοποιήθηκε στη συγκεκριμένη βάση αφορά αποκλειστικά SVM ταξινομητές και είναι η ίδια με αυτή που περιγράψαμε για τα αντίστοιχα πειράματα της βάσης KTH στην Ενότητα 5.2. Η μόνη διαφορά είναι ότι τα video υποδειγματοληπτούνται στη μισή ανάλυση στις χωρικές διαστάσεις, πράγμα που μειώνει τον όγκο τους στο $1/4$ του αρχικού. Στο Σχήμα. 5.9 φαίνεται η διαδικασία εύρεσης χωροχρονικών σημείων ενδιαφέροντος με τον αλγόριθμο Gabor3D για ένα video της Hollywood2.



Σχήμα 5.9: Διαδικασία εύρεσης χωροχρονικών σημείων ενδιαφέροντος με τον αλγόριθμο Gabor3D για τη βάση δεδομένων Hollywood2.

Λόγω του τεράστιου όγκου δεδομένων της συγκεκριμένης βάσης η πειραματική διαδικασία εξελίχτηκε σταδιακά, προσθέτοντας δράσεις, ενώ ο πειραματισμός με μέθοδους της βιβλιογραφίας περιορίστηκε σε πειράματα 2 και 5 δράσεων. Για πειράματα σε ολόκληρη τη βάση δεδομένων (12 δράσεις) αρκεστήκαμε σε σύγκριση των μεθόδων μας με τα αποτελέσματα που παρατίθενται στη βιβλιογραφία, χωρίς να εκτελέσουμε εκ νέου πειράματα με τους ήδη υπάρχοντες ανιχνευτές. Θα παραθέσουμε λοιπόν τα πειράματά μας με τη χρονολογική σειρά που έγιναν, προσθέτοντας σταδιακά περισσότερες δράσεις.

Το πρώτο πείραμα που έγινε ήταν το πείραμα ταξινόμησης των δράσεων SitDown και StandUp. Υπάρχουν στη βάση συνολικά 475 video για τις συγκεκριμένες δράσεις. Τα δείγματα χωρίστηκαν για κάθε δράση σε 70% δείγματα εκπαίδευσης και 30% δείγματα αξιολόγησης και χρησιμοποιήθηκε η τεχνική Repeated random sub-sampling validation με 20 επαναλήψεις [16], σε καθεμία από τις οποίες τα δεδομένα χωρίστηκαν με τυχαίο τρόπο σε εκπαίδευσης και αξιολόγησης. Εφαρμόσαμε τα παραπάνω για 4 αλγόριθμους, τους DCA3D [17], TDE, Harris3D [34] και Gabor3D [37]. Στο Σχήμα. 5.10 παρατίθενται τα αποτελέσματα και των 4 μεθόδων. Το κατώφλι απόρριψης των μη μεγίστων για τους ανιχνευτές TDE και Gabor3D τέθηκε στο 8% της μέγιστης τιμής ενέργειας που ανιχνεύτηκε στο video. Φαίνεται ότι οι ανιχνευτές Harris3D, TDE και Gabor3D δείχνουν παραπλήσια αποτελέσματα με τον τελευταίο να διαχωρίζει ελαφρώς καλύτερα τα δεδομένα. Για το παραπάνω πείραμα χρησιμοποιήθηκε ο περιγραφητής HOG3D, και ταξινόμηση με SVMs με χ^2 πυρήνα. Παρατηρούμε επίσης ότι αντίθετα με την περίπτωση της βάσης KTH Action Dataset στην Hollywood2 ο αλγόριθμος TDE δεν ξεπερνά τον Harris3D, πράγμα που θα παρατηρηθεί και σε επόμενα πειράματα. Το γεγονός αυτό οφείλεται στην κίνηση της κάμερας, στην οποία ο αλγόριθμος TDE φαίνεται να είναι πολύ ευαίσθητος, οδηγώντας σε ανιχνεύσεις που δεν είναι εστιασμένες στη δράση αλλά στο background (false alarms).

Ενσωματώνοντας και άλλες δράσεις στο σύστημα αναγνώρισης προχωρήσαμε σε πείραμα 5 δράσεων. Επιλέξαμε τις δράσεις SitDown,StandUp,Kiss,HandShake και FightPerson. Συνολικά υπάρχουν 1008 video προς ταξινόμηση, με 436 video για το Training Set και 572 video για το Test Set. Χρησιμοποιήσαμε το διαχωρισμό σε δεδομένα εκπαίδευσης και ταξινόμησης που δίνεται, και ταξινόμηση με SVMs χωρίς να εκτελέσουμε κάποιας μορφής cross validation. Πειραματιστήκαμε με τους αλγόριθμους TDE, Gabor3D, ενώ χρησιμοποιήθηκε και ο Harris3D για να γίνει η σύγκριση των μεθόδων μας. Ως δείκτης επιτυχίας χρησιμοποιήθηκε ο προτεινόμενος mean Average Precision - (mAP).

Δράσεις	SitDown	StandUp
SitDown	65.3%	34.7%
StandUp	37.5%	62.5%

(α') Ανιχνευτής DCA3D. Μέση ακρίβεια Αναγνώρισης: 63.2%

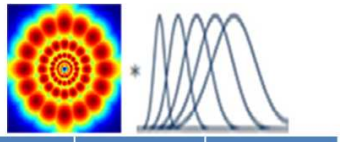
$$M = g * \begin{pmatrix} L_x^2 & L_x L_y & L_x L_t \\ L_x L_y & L_y^2 & L_y L_t \\ L_x L_t & L_y L_t & L_t^2 \end{pmatrix}$$

Δράσεις	SitDown	StandUp
SitDown	76.9%	23.1%
StandUp	20.2%	79.8%

(β') Ανιχνευτής Harris3D. Μέση ακρίβεια Αναγνώρισης: 78.6%

Δράσεις	SitDown	StandUp
SitDown	74.1%	25.9%
StandUp	19.9%	80.1%

(γ') Ανιχνευτής TDE. Μέση ακρίβεια Αναγνώρισης: 78.1%



Δράσεις	SitDown	StandUp
SitDown	78.41%	21.59%
StandUp	18.41%	81.59%

(δ') Ανιχνευτής Gabor3D. Μέση ακρίβεια Αναγνώρισης: **80.21%**

Σχήμα 5.10: Αποτελέσματα ταξινόμησης 2 δράσεων (SitDown & StandUp) με 4 διαφορετικούς ανιχνευτές.

Ανιχνευτής	Harris3D	TDE	Gabor3D
mAP	70.3%	65.7%	**73.4**%

Πίνακας 5.5: Μέση τιμή precision (mAP) διάφορων μεθόδων για το πείραμα 5 δράσεων της βάσης Hollywood2.

Παρατηρούμε και σε αυτό το πείραμα ότι ο TDE αλγόριθμος δε δίνει καλύτερα αποτελέσματα από τον Harris3D, λόγω της κινούμενης κάμερας και των πολλών λανθασμένων ανιχνεύσεων στο background. Αντίθετα, ο αλγόριθμος Gabor3D, ο οποίος παρουσιάζει γενικότερα μια πιο ομαλή μορφή ενέργειας που προέρχεται από το φιλτράρισμα με πολλαπλά φίλτρα, ανιχνεύει λιγότερα σημεία στο background ενός βιντεο με κινούμενη κάμερα, και οδηγεί σε ικανοποιητικά αποτελέσματα.

Προχωράμε τελικά σε πείραμα ταξινόμησης και των 12 ανθρώπινων δράσεων

της συγκεκριμένης βάσης. Ακολουθώντας την πειραματική διαδικασία που ακολουθήθηκε στο [60] και την οποία περιγράψαμε, προχωράμε στα πειραματικά αποτελέσματα. Έγινε σύγκριση του αλγορίθμου Gabor3D με τους ανιχνευτές Cuboids, Hessian και Harris3D, τα αποτελέσματα των οποίων χρησιμοποιήσαμε κατευθείαν από το [60].

Ανιχνευτής	Cuboids [12]	Hessian [63]	Harris3D [34]	Gabor3D [37]
mAP	46.2%	46.0%	45.2%	**47.7**%

Πίνακας 5.6: Μέση τιμή precision (mAP) διάφορων μεθόδων της Hollywood2 βάσης δεδομένων. Γίνεται σύγκριση με διάφορες μεθόδους της βιβλιογραφίας, με την πειραματική διαδικασία που περιγράψαμε. Χρησιμοποιείται ο HOG3D ως περιγραφητής για τον αλγόριθμο Gabor3D.

Παρατηρούμε ότι ο ανιχνευτής μας δίνει το μέγιστο αποτέλεσμα σε σύγκριση με τις άλλες μεθόδους, που σημαίνει ότι οδηγεί γενικότερα σε πολύ αξιόπιστα σημεία ενδιαφέροντος. Τα χαμηλά ποσοστά που δεν ξεπερνούν το 50% για όλες τις μεθόδους δείχνουν το μέγεθος των δυσκολιών που αντιμετωπίζονται τόσο στο κομμάτι της Όρασης Υπολογιστών όσο και της Αναγνώρισης Προτύπων για την ταξινόμηση δεδομένων της βάσης αυτής.

Κεφάλαιο 6

Πειράματα Ταξινόμησης και Αναγνώρισης Ανθρώπινων Δράσεων στη Βάση Δεδομένων *MOBOT*

Εκτός από τις βάσεις ανθρώπινων δράσεων KTH και Hollywood2 είχαμε στη διάθεσή μας μία ακόμα πολυτροπική, αλλά και πολυ-αισθητηριακή βάση χειρονομιών. Η βάση αυτή εξυπηρετεί τους σκοπούς του ερευνητικού προγράμματος MOBOT[1], για την κατασκευή μίας ρομποτικής πλατφόρμας υποστήριξης ηλικιωμένων ατόμων, η οποία θα ενσωματώνει και χαρακτηριστικά πολυτροπικής επικοινωνίας. Για το λόγο αυτό, το σύνολο των χρηστών της αποτελείται από άτομα μεγάλης ηλικίας με κινητικά, και πολλές φορές, και διανοητικά προβλήματα. Στη συγκεκριμένη βάση επεκτείνανε τον πειραματισμό μας και σε αναγνώριση συνεχόμενων δράσεων, εκτός από απλή ταξινόμησή τους.

6.1 Περιγραφή της Βάσης Δεδομένων

Στη συγκεκριμένη υποενότητα κάνουμε μια σύντομη παρουσίαση και της MOBOT βάσης, στα πλαίσια που αφορούν την αναγνώριση ανθρώπινων δράσεων. Αν και ο πειραματισμός στη συγκεκριμένη βάση είναι μάλλον πρώιμος, και παρουσιάζεται μικρό κομμάτι ερευνητικών αποτελεσμάτων, έχει ενδιαφέρον να παρατηρήσουμε ότι οι στόχοι του προγράμματος MOBOT αποτελούν μια ρεαλιστική απόδειξη της σημασίας της έρευνας στο πεδίο της αναγνώρισης ανθρώπινων δράσεων αλλά και των πρακτικών εφαρμογών που μπορούν να υλοποιηθούν.

Στην εν λόγω βάση δεδομένων υπήρχε η δυνατότητα, αλλά και η ανάγκη να καταγραφεί ένα πλήθος διαφορετικών τροπικοτήτων, από μια σειρά από διαφορετικούς αισθητήρες. Στη συνέχεια απαριθμούμε το σύνολο των

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

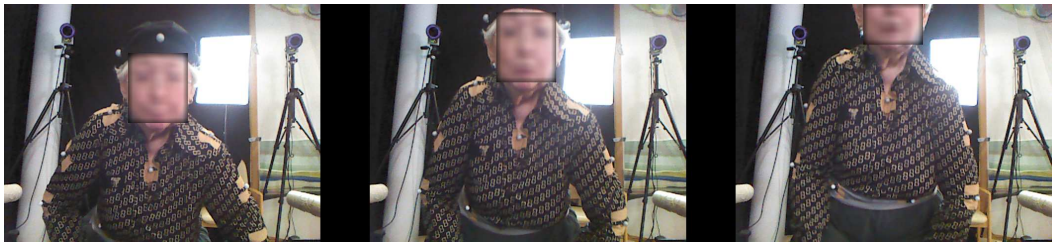
διαφορετικών αισθητήρων, μαζί με κάποια σύντομα σχόλια για τα δεδομένα που λαβαμε από καθένα από αυτούς:

- Άνω Kinect: Αισθητήρας Kinect τοποθετημένος πάνω στη ρομποτική πλατφόρμα, με σκοπό την καταγραφή του άνω μέρος του σώματος του χρήστη. Παρείχε δεδομένα RGB και βάθους.
- Κάτω Kinect: Αισθητήρας Kinect τοποθετημένος πάνω στη ρομποτική πλατφόρμα, με σκοπό την καταγραφή του κάτω μέρος του σώματος του χρήστη. Παρείχε δεδομένα RGB και βάθους.
- Κάμερα GoPro: Κάμερα ευρείας γωνίας λήψης τοποθετημένη πάνω στη ρομποτική πλατφόρμα, με σκοπό την καταγραφή του άνω μέρους του σώματος του χρήστη. Παρείχε δεδομένα RGB υψηλής ευκρίνειας.
- Συστοιχία μικροφώνων MEMS: Σειρά μικροφώνων τεχνολογίας MEMS τοποθετημένα σε γραμμική συστοιχία πάνω στη ρομποτική πλατφόρμα. Παρέχουν πολυκαναλική (8 κανάλια) καταγραφή δεδομένων ήχου.
- Δύο κάμερες υψηλής ευκρίνειας: Σταθερές κάμερες υψηλής ευκρίνειας, για την καταγραφή ολόκληρου του πεδίου δράσης. Παρέχουν δεδομένα RGB υψηλής ευκρίνειας.
- Οπτικό σύστημα καταγραφής σκελετου: Σύστημα τύπου Qualisys Motion Capture System που επιτρέπει την καταγραφή του σκελετού του χρήστη μεσω ειδικών markers που είναι τοποθετημένοι σε συγκεκριμένα σημεία στο σώμα του χρήστη.

6.2 Πειράματα στο σενάριο 3.b της βάσης δεδομένων MOBOT

Το σενάριο 3.b της βάσης δεδομένων MOBOT είναι ένα υποσύνολο της βάσης που περιέχει συνεχόμενες εκτελέσεις δράσεων από 6 υποκείμενα. Οι δράσεις που εκτελούνται είναι τρεις: SitDown, StandUp και Walking. Εκτελούνται κατά συνεχόμενο τρόπο, ενώ εκτελούνται ενδιάμεσα και διάφορες χειρονομίες που δεν ανήκουν στο λεξιλόγιο των δράσεων που επιθυμούμε να αναγνωρίσουμε. Όλες οι λήψεις που επεξεργαζόμαστε καταγράφονται από την Άνω Kinect που είναι τοποθετημένη στη ρομποτική πλατφόρμα, και έτσι αντίθετα με τις μέχρι τώρα βάσεις δεδομένων η κάμερα ακολουθεί το χρήστη. Στο Σχήμα. 6.1 φαίνονται frames από τις εκτελέσεις των 3 δράσεων για ένα υποκείμενο.

Η δυσκολία της βάσης αυτής έγκειται αφ' ενός στο ότι τα υποκείμενα πραγματοποιούν τις δράσεις με έναν νωθρό τρόπο, λόγω της προχωρημένης ηλικίας



(α') Δράση StandUp

(β') Δράση Walking

(γ') Δράση SitDown

Σχήμα 6.1: Οι δράσεις του σεναρίου 3.b του MOBOT

και τα κινητικά τους προβλήματα, και αφ' ετέρου στο γεγονός ότι στο πεδίο ορατότητας της κάμερας υπάρχει μεγάλος οπτικός θόρυβος που οφείλεται κυρίως σε βοηθητικό προσωπικό (carers) που βοηθούν τους ηλικιωμένους να εκτελούν δράσεις, καθώς και σε προσωπικό υπεύθυνο για τις λήψεις της συγκεκριμένης βάσης. Στο Σχήμα. 6.1γ' φαίνεται μια κλασική περίπτωση οπτικού θορύβου που προέρχεται από την carer η οποία μπαίνει εντός του πεδίου λήψης της κάμερας για να βοηθήσει στην εκτέλεση της δράσης SitDown.

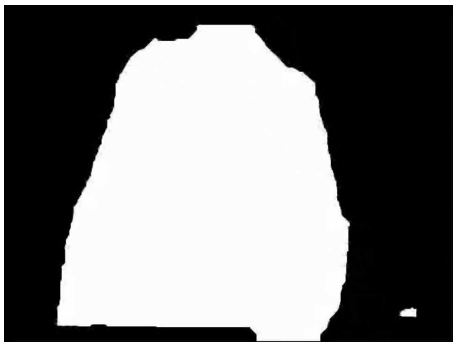
6.2.1 Χρήση του καναλιού βάθους (depth) για κατάτμηση του χρήστη

Στην προσέγγισή μας στο πρόβλημα της αναγνώρισης και ταξινόμησης ανθρώπινων δράσεων χρησιμοποιήθηκαν το RGB κανάλι και η πληροφορία βάθους (depth) από τον αισθητήρα Άνω Kinect. Εκτός από την εκμετάλλευση της RGB πληροφορίας με τον τρόπο που εφαρμόζουμε σε όλα τα πειράματα της παρούσας διπλωματικής εργασίας, έχουμε στη διάθεσή μας και την πληροφορία του βάθους. Το βάθος (depth) είναι ένα κανάλι της κάμερας Kinect το οποίο δίνει την θέση των εικονιζόμενων αντικειμένων στις 3 διαστάσεις, με τη χρήση υπέρυθρης ακτινοβολίας. Η χρήση του depth είναι πολύ διαδεδομένη σε προβλήματα αναγνώρισης χειρονομιών [46, 43, 30, 45] που μπορούν να ειδωθούν σαν μια λεπτομερέστερη περίπτωση των ανθρώπινων δράσεων. Στην προσέγγισή μας στο πρόβλημα της αναγνώρισης και ταξινόμησης ανθρώπινων δράσεων στη συγκεκριμένη βάση δεδομένων, αξιοποιήσαμε την πληροφορία του βάθους της εικόνας κάνοντας κατάτμηση του χρήστη, πράγμα που οδηγεί σε πιο εστιασμένες ανιχνεύσεις χωροχρονικών σημείων ενδιαφέροντος.

Η ανάλυση που έγινε στο κανάλι του βάθους είχε σκοπό να κρατηθεί το υποκείμενο που εκτελεί τις δράσεις, και να εστιαστούν οι ανιχνεύσεις χωροχρονικών σημείων ενδιαφέροντος πάνω του, χωρίς να δημιουργούνται false alarms από την κινούμενη κάμερα αλλά και από το θόρυβο που υπάρχει μέσα στα βίντεο. Στο Σχήμα. 6.2 φαίνεται η διαδικασία που ακολουθήθηκε για να γίνει η κατάτμηση του χρήστη.

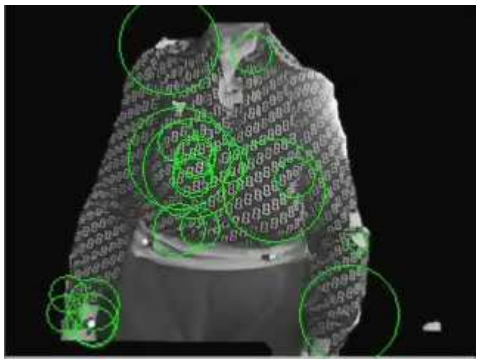
Η επεξεργασία του depth που χρησιμοποιήθηκε ήταν μια κατωφλιοποίηση σε συνδυασμό με ένα μορφολογικό φιλτράρισμα του κάθε frame. Πιο συγκεκριμένα η εικόνα κατωφλιώνεται στο 35% της μέγιστης δυνατής τιμής της και κρατούνται τα pixels που η τιμή τους είναι μικρότερη από το κατώφλι (απορρίπτονται φυσικά οι μηδενικές τιμές της εικόνας, για τις οποίες τα σημεία έχουν ξεφύγει από την εμβέλεια δράσης της κάμερας Kinect). Η κατωφλιοποίηση με αυτόν τον τρόπο μεταφράζεται σε απόρριψη των δεδομένων της κάμερας που εμφανίζονται πίσω από το επίπεδο που βρίσκεται ο ασθενής. Εφαρμόζεται επίσης ένα closing με δίσκο ακτίνας 20, αφού κρατηθούν πρώτα οι 2 μέγιστες συνεκτικές συνιστώσες της κάθε εικόνας. Ο λόγος που κρατήσαμε τις δύο μέγιστες συνεκτικές συνιστώσες και όχι τη μια είναι ότι πολλές φορές ο βοηθός του χρήστη μπαίνει μέσα στο πλάνο και η segmented εικόνα περιέχει τη μάσκα του και όχι τη μάσκα του χρήστη. Θεωρώντας πως η ύπαρξη του χρήστη είναι σημαντική σε κάθε εικόνα αποφασίσαμε να κάνουμε αυτόν τον συμβιβασμό. Παρακάτω παρουσιάζουμε αποτελέσματα πειραμάτων που έχουν γίνει τόσο με τη χρήση της πληροφορίας του βάθους για κατάτμηση του χρήστη όσο και χωρίς.

(α') Εικόνα Βάθους



(β') Μάσκα Χρήστη από την επεξεργασία της εικόνας βάθους

(γ') Αρχικά σημεία ενδιαφέροντος



(δ') Σημεία ενδιαφέροντος με κατάτμηση

Σχήμα 6.2: Διαδικασία επιλογής εστιασμένων στο χρήστη χωροχρονικών σημείων ενδιαφέροντος μέσω κατάτμησής του με την εικόνα βάθους.

6.2.2 Πειράματα Ταξινόμησης Ανθρώπινων Δράσεων

Για τα πειράματα ταξινόμησης ανθρώπινων δράσεων χρησιμοποιήθηκε η ίδια διαδικασία με τη βάση KTH Action Dataset (Ενότητα 5.2). Χρησιμοποιήθηκε ο ανιχνευτής Gabor3D για την εύρεση των σημείων ενδιαφέροντος, και ο HOG3D για την περιγραφή τους. Το πλήθος των λέξεων του ιστογράμματος BoF τέθηκε στις $V = 4000$ λέξεις.Η αξιολόγηση της μεθόδου έγινε χρησιμοποιώντας SVMs με χ^2-πυρήνα και Leave-One-User-Out Cross Validation. Από τους συνολικά N_U χρήστες επιλέγεται ο ένας για την αξιολόγηση του συστήματος και οι υπόλοιποι για την εκπαίδευσή του. Η διαδικασία επαναλαμβάνεται N_U φορές, μία για κάθε διαφορετικό χρήστη. Ως μετρικό αξιολόγησης χρησιμοποιήθηκε ο δείκτης mean accuracy, ο οποίος δίνεται ως ο μέσος όρος των accuracy των N_U επαναλήψεων.

Στο Σχήμα. 6.3 φαίνεται ο πίνακας σύγχυσης για τη μέθοδο που περιγράφηκε, με τη χρήση της πληροφορίας του depth αλλά και χωρίς. Μέσω της μάσκας χρήστη που εξάγουμε από την πληροφορία του βάθους επιλέγουμε μόνο τα σημεία ενδιαφέροντος που έχουν ανιχνευθεί και βρίσκονται εντός της μάσκας, με σκοπό την κατασκευή ενός ιστογράμματος Bag-of-Features που αποτελείται από μη θορυβώδη χαρακτηριστικά.

Δράση	SitDown	Walking	StandUp
SitDown	0.83	0.17	0
Walking	0.17	0.83	0
StandUp	0	0	1.00

(α') Πίνακας Σύγχυσης χωρίς χρήση depth πληροφορίας. MeanAccuracy: 89.0%

Δράση	SitDown	Walking	StandUp
SitDown	0.83	0.17	0
Walking	0	1.00	0
StandUp	0	0	1.00

(β') Πίνακας Σύγχυσης με χρήση depth πληροφορίας. MeanAccuracy: 94.3%

Σχήμα 6.3: Πίνακες Σύγχυσης για πείραμα ταξινόμησης 3 δράσεων στη βάση MOBOT.

Παρατηρούμε πως με την προσθήκη της πληροφορίας του depth παρατηρείται κάποια βελτίωση στην ακρίβεια ταξινόμησης των δράσεων. Λόγω του μικρού αριθμού των δεδομένων που χρησιμοποιήθηκε δεν είμαστε σε θέση να εξάγουμε ασφαλή συμπεράσματα, αφού η βελτίωση γίνεται από μία δράση που ταξινομήθηκε σωστά με τη χρήση του depth, ωστόσο υπάρχει η λογική ότι μια πιο εστιασμένη στο χρήστη προσέγγιση θα δώσει και καλύτερα αποτελέσματα.

6.2.3 Πειράματα Αναγνώρισης συνεχόμενων Ανθρώπινων Δράσεων

Το πρόβλημα της αναγνώρισης συνεχόμενων ανθρώπινων δράσεων είναι ένα πρόβλημα εντελώς διαφορετικής φύσης από το πρόβλημα της ταξινόμησης. Σε όλα τα πειράματα που έχουν περιγραφεί παραπάνω, το ζητούμενο ήταν να ταξινομηθούν τα βίντεο μιας βάσης δεδομένων σε μια από τις κλάσεις-δράσεις του λεξιλογίου μας, δεδομένων των χρονικών ορίων για την οποία εκτελείται. Αντίθετα, η πρόκληση στο πρόβλημα της αναγνώρισης είναι αφ' ενός ο σωστός εντοπισμός της ανθρώπινης δράσης σε μια συνεχή ροή βίντεο, και αφ' ετέρου η σωστή ταξινόμησή της. Δεν είναι δεδομένες οι χρονικές στιγμές στις οποίες συμβαίνουν οι ανθρώπινες δράσεις, και προφανώς δεν είναι δεδομένη και η φύση τους (στα δεδομένα αξιολόγησης). Για όλους αυτούς τους λόγους μπορεί να γίνει κατανοητό ότι το πρόβλημα της αναγνώρισης (recognition) είναι πολύ δυσκολότερο, απαιτεί συστήματα εκπαιδευμένα με πιο έξυπνο τρόπο, ενώ υπάρχουν περισσότερα "ανεξερεύνητα νερά" στην έρευνα της περιοχής.

Η προσέγγισή μας, η οποία βασίστηκε στην παραλλαγή των ιδεών που αφορούν την ταξινόμηση ανθρώπινων δράσεων και την προσαρμογή τους στο πρόβλημα της αναγνώρισης συνεχόμενων δράσεων, αποτελείται από τα εξής διακριτά βήματα:

- Για ένα βίντεο εισόδου εκτελούμε τον αλγόριθμο Gabor3D και εξάγουμε τα χωροχρονικά σημεία ενδιαφέροντος με τη μέθοδο που περιγράφηκε στην Ενότητα 3.3.

- Εξάγονται περιγραφητές από τα ανιχνευμένα σημεία ενδιαφέροντος.

- Ένα κινούμενο χρονικό παράθυρο συγκεντρώνει τους περιγραφητές που έχουν εξαχθεί από τα σημεία ενδιαφέροντος εντός του παραθύρου, και φτιάχνει από αυτά ένα ιστόγραμμα BoF $V = 4000$ λέξεων.

- Εκτός από τις ετικέτες των δράσεων προς αναγνώριση, εισάγουμε στο σύστημά μας ένα Background Model (BM) όπου αναθέτουμε όλα τα frames τα οποία δεν ανήκουν σε κάποια από τις δράσεις του λεξιλογίου μας.

- Τα δεδομένα χωρίζονται σε δεδομένα εκπαίδευσης και δεδομένα αξιολόγησης, επιλέγοντας κυκλικά έναν χρήστη για αξιολόγηση και τους υπόλοιπους για την εκπαίδευση του συστήματος. Εκτελούμε δηλαδή Unseen User πειράματα. Τα δεδομένα εισάγονται σε έναν ταξινομητή SVM, όχι ωστόσο για να ταξινομηθούν, αλλά για να εξαχθούν οι πιθανότητες να ανήκουν σε καθεμία από τις λέξεις του λεξιλογίου [47].

- Γίνεται μια ομαλοποίηση των πιθανοτήτων, τόσο με τη χρήση Γκαουσιανού πυρήνα όσο και με τη χρήση median φιλτραρίσματος.

- Τέλος, οι πιθανότητες (μία για κάθε δράση, σε κάθε χρονική στιγμή) δίνονται ως Observation Likelihoods στον αλγόριθμο Viterbi, στον πίνακα μεταβάσεων του οποίου έχει επιβληθεί μια ποινή μεταβάσεων μεταξύ καταστάσεων [18].

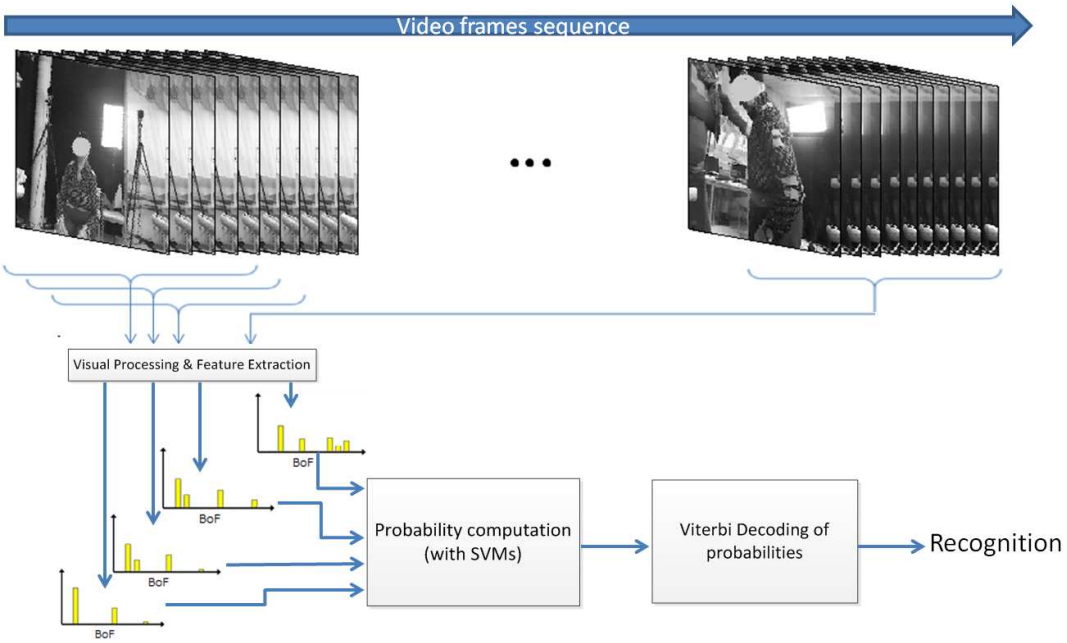
Η χρήση των ταξινομητών SVM με πιθανοτικές εξόδους (probabilistic outputs) έγινε λόγω του ότι τα αραιά ιστογράμματα BoF δεν έδειξαν καλή συμπεριφορά με τα μοντέλα HMMs, όπως αποδείχτηκε από τα πειράματα στη βάση KTH (υποενότητα 5.2.2). Αντίθετα, η χρήση των SVMs έχει δείξει την καλύτερη επίδοση ανάμεσα στους ταξινομητές που χρησιμοποιήθηκαν, σε συνδυασμό με τα ιστογράμματα BoF, και για αυτό προτιμήθηκε η επιλογή τους.

Η χρήση του Background Model (BM) ως μια επιπλέον δράση έγινε για να αναθέσουμε σε αυτό τα κομμάτια του κάθε Video του Training Set στα οποία δεν εκτελείται κάποια δράση. Η ιδέα είναι ότι σε αυτήν την κλάση θα ανατίθενται όλα τα γνωστά μέχρι τώρα κομμάτια που δεν αποτελούν τις προς έλεγχο δράσεις, αλλά οτιδήποτε διαφορετικό (πχ. χειρονομίες, σιγή, αυθόρμητη κίνηση υποκειμένου

κτλ). Ο χειρισμός της κλάσης BM είναι ακριβώς ίδιος με το χειρισμό των υπόλοιπων δράσεων, δηλαδή τα frames ταξινομούνται σε αυτήν αφού γίνει επεξεργασία των πιθανοτικών εξόδων των SVMs.

Η μέχρι τώρα διαδικασία ταξινόμησης που χρησιμοποιήθηκε έκανε έλεγχο των probabilistic outputs των one-vs-rest SVMs και ανέθετε το δείγμα αξιολόγησης στην κλάση με τη μεγαλύτερη πιθανότητα. Στην αναγνώριση συνεχόμενων δράσεων επεκτείνουμε αυτήν την προσέγγιση χρησιμοποιώντας τον αλγόριθμο Viterbi. Συγκεκριμένα οι πιθανοτικές έξοδοι των SVMs περνάνε μια διαδικασία smoothing από τον αλγόριθμο Viterbi ο οποίος αναλαμβάνει να βρει το βέλτιστο μονοπάτι καταστάσεων από το οποίο πέρασε η χρονική εξέλιξη ενός φαινομένου. Το φαινόμενο που έχουμε επιλέξει να παρακολουθούμε στην περίπτωση αυτή είναι το ιστόγραμμα Bag-of-Features. Χειριζόμαστε τις διαφορετικές ανθρώπινες δράσεις σαν να ήταν καταστάσεις (states) ενός μοντέλου HMM, τα Observation Likelihoods του οποίου θεωρούμε ότι είναι οι πιθανότητες που έχουν προκύψει για κάθε frame να ανήκει σε κάθε δράση. Ο Viterbi αναλαμβάνει να αποκωδικοποιήσει αυτά τα Observation Likelihoods και μέσω ενός πίνακα μετάβασης και μιας πιθανότητας αρχικής κατάστασης να αποφανθεί για την κατάσταση (δράση) στην οποία βρίσκεται το βίντεο σε κάθε frame. Ο πίνακας μεταβάσεων που χρησιμοποιήθηκε δεν επιτρέπει συχνές μεταβάσεις από κατάσταση σε κατάσταση, με τη λογική ότι η κάθε δράση θέλει μερικά δευτερόλεπτα για να ολοκληρωθεί οπότε δε γίνεται να αλλάζουν οι δράσεις από frame σε frame. Η μέθοδος αυτή χρησιμοποιήθηκε από τους Giannoulis et al. [18] για την ανίχνευση ακουστικών γεγονότων. Με την ίδια λογική επέβαλλαν ένα penalty στη διαγώνιο του πίνακα μεταβάσεων του αλγορίθμου ώστε να δυσχαιρένονται συχνές μεταβάσεις μεταξύ Silence και Acoustic Events. Η ίδια λογική ακολουθήθηκε και στην παρούσα διπλωματική εργασία, αυτή τη φορά για τις μεταβάσεις μεταξύ δράσεων. Η χρήση του Viterbi επίσης έχει το πλεονέκτημα ότι μπορεί να σβηστούν μεταβάσεις μεταξύ δράσεων που δεν έχουν λογική (πχ. μετά τη δράση SitDown δε μπορεί να εντοπιστεί η δράση Walking για το ίδιο υποκείμενο). Διευκρινίζεται ότι ο πίνακας μετάβασης του αλγόριθμου Viterbi δε χρειάζεται να περιέχει πιθανότητες (μπορούν βέβαια να κανονικοποιηθούν οι τιμές), καθώς ο αλγόριθμος χρησιμοποιείται για σκοράρισμα (scoring) κάποιας ακολουθίας. Οι πιθανότικές έξοδοι του SVM ταξινομητή παρατηρήσαμε ότι είναι πιο αποτελεσματικό να περνάνε ένα στάδιο Γκαουσιανού ή median φιλτραρίσματος στο χρόνο, πριν δωθούν ως είσοδος στον αλγόριθμο Viterbi. Στο Σχήμα. 6.4 απεικονίζεται σχηματικά το σύστημα αναγνώρισης συνεχόμενων ανθρώπινων δράσεων που περιγράφηκε.

Στον Πίνακα. 6.1 παραθέτουμε τα αποτελέσματα του πειραματισμού μας, με χρήση της πληροφορίας του βάθους και χωρίς, όπως ακριβώς στα πειράματα ταξινόμησης. Οι πιθανοτικές έξοδοι έχουν υποστεί στο χρόνο median φιλτράρισμα με παράθυρο διάστασης 21 pixels, ενώ για τη σάρωση των χαρακτηριστικών



Σχήμα 6.4: Η διαδικασία αναγνώρισης συνεχόμενων ανθρώπινων δράσεων με χρήση πιθανοτήτων από SVMs σε συνδυασμό με τον αλγόριθμο Viterbi για την τελική απόφαση των καταστάσεων-δράσεων.

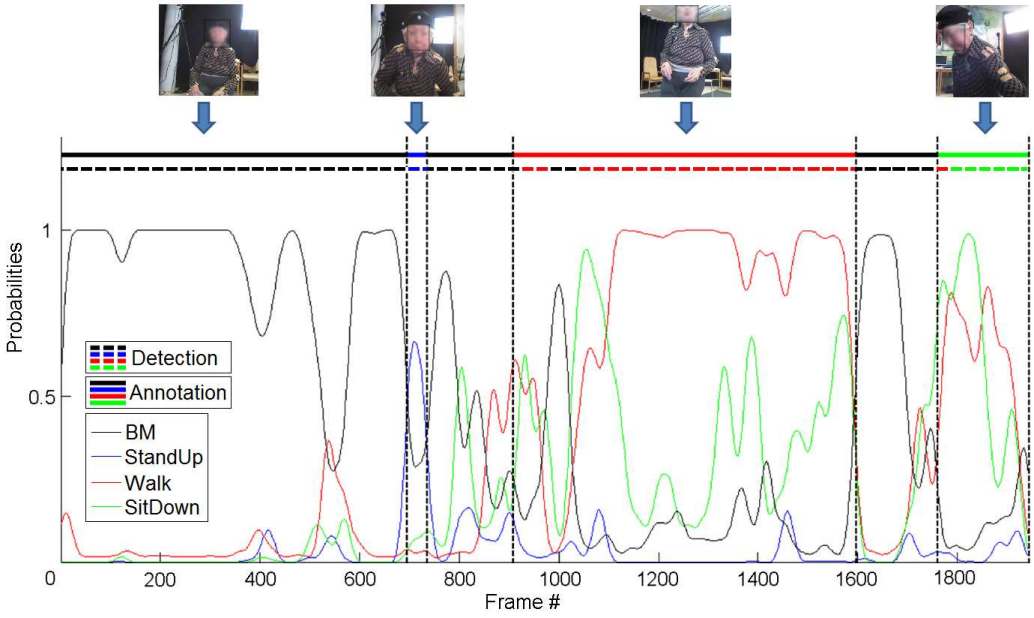
χρησιμοποιήθηκε παράθυρο μεγέθους 10 pixels. Για τον αλγόριθμο Viterbi χρησιμοποιήθηκε ομοιόμορφος πίνακας αρχικών πιθανοτήτων καταστάσεων και ομοιόμορφος πίνακας μεταβάσεων (εκτός της διαγωνίου του). Μηδενίστηκε η πιθανότητα μετάβασης από τη δράση SitDown στη δράση Walk. Ως δείκτης χρησιμοποιήθηκε το accuracy της αναγνώρισης για κάθε χρήστη.

Χρήστης	Ακρίβεια χωρίς χρήση Depth	Ακρίβεια με χρήση του Depth
p1	92.12	93.98
p3	57.98	70.40
p4	74.82	77.20
p5	80.34	85.25
p6	91.47	90.00
p18	81.80	83.35

Πίνακας 6.1: Ακρίβεια αναγνώρισης (accuracy) δράσεων για κάθε υποκείμενο της βάσης δεδομένων MOBOT 3.b για 3 δράσεις + Background Model, χωρίς και με τη χρήση της πληροφορίας του depth. Μέση ακρίβεια αναγνώρισης χωρίς τη χρήση του depth: 79.76%. Μέση ακρίβεια αναγνώρισης με τη χρήση του depth: **83.36%**.

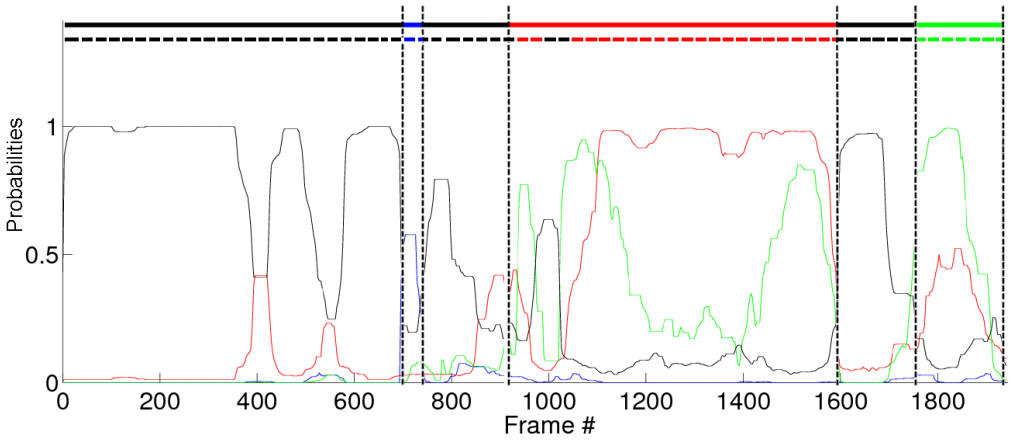
Παρατηρούμε ότι με τη χρήση της depth πληροφορίας έχουμε μια μικρή αύξηση της επίδοσης του συστήματος. Ενώ θα περίμενε κανείς ωστόσο η αύξηση στα ποσοστά αναγνώρισης να είναι ουσιώδης, μιας και απομονώθηκε ο χρήστης από ένα τόσο θορυβώδες background, αυτό δε φαίνεται να συμβαίνει (τόσο στο παρόν πείραμα αναγνώρισης συνεχόμενων δράσεων όσο και στο πείραμα ταξινόμησης της συγκεκριμένης βάσης). Αντί αυτού, παρατηρούμε σχετικά μικρές αυξήσεις στην επίδοση του συστήματος ενώ η πληροφορία εισόδου είναι σαφώς καθαρότερη, αφού έχει επιτευχθεί κατάτμηση του χρήστη-υποκειμένου. Η εξήγηση είναι ότι όταν φτιάχνονται ιστογραφικές αναπαραστάσεις του video, όπως το BoF, ακόμα και τα σημεία που ανιχνεύονται στο παρασκήνιο εμπεριέχουν πληροφορία για το είδος της δράσης. Για παράδειγμα, στη συγκεκριμένη βάση δεδομένων, το ότι το background κινείται κατά την εκτέλεση της δράσης Walking παρέχει πληροφορία που τη χαρακτηρίζει. Συμπερασματικά, χωρίς τη χρήση της πληροφορίας του βάθους από τη μία εισάγεται θόρυβος από μη εστιασμένες στο χρήστη ανιχνεύσεις αλλά από την άλλη μέσα σε αυτή τη θορυβώδη πληροφορία υπάρχει κάποια χρησιμότητα που χαρακτηρίζει ορισμένες δράσεις. Όπως βλέπουμε και στο αποτέλεσμα, οι δράσεις του ασθενούς p6 αναγνωρίζονται πιο αποδοτικά χωρίς τη χρήση του depth καναλιού. Ορισμένοι χρήστες (p3, p4) η αναγνώριση παρουσιάζει αρκετά χαμηλά ποσοστά λόγω της φύσης των βίντεο (θόρυβος από βοηθητικό προσωπικό, μη κατανόηση οδηγιών κτλ). Σε ορισμένες περιπτώσεις ο χρήστης βγαίνει εξ' ολοκλήρου εκτός εμβέλειας της κάμερας, πράγμα που κάνει την αναγνώριση δράσεων πολύ απαιτητική. Στο Σχήμα. 6.5 φαίνονται οι πιθανοτικές έξοδοι για κάθε δράση, αφού έχουν υποστεί ένα Γκαουσιανό φιλτράρισμα με παράθυρο 25 σημείων, οι ετικέτες που δώθηκαν καθώς και οι ετικέτες που ανιχνεύτηκαν για κάθε frame. Αντίστοιχα, στο Σχήμα. 6.6 αντί για Γκαουσιανό φιλτράρισμα χρησιμοποιήθηκε ένα median φίλτρο ίδιου μεγέθους.

Οι φιλτραρισμένες πιθανότητες αυτές δίνονται ως είσοδο στον αλγόριθμο Viterbi. Φαίνεται ότι στα περισσότερα σημεία μεγιστοποιείται η πιθανότητα της δράσης



Σχήμα 6.5: Τα probabilistic outputs που προκύπτουν από τον ταξινομητή SVM για κάθε κλάση, αφού έχουν φιλτραριστεί με ένα Γκαουσιανό φίλτρο μεγέθους 25 frames.

η οποία έχει επισημειωθεί. Οι πιθανότητες αυτές περνάνε ακόμα ένα στάδιο εξομάλυνσης μέσω του αλγόριθμου Viterbi, για να βγει τελικά η απόφαση δράσης για κάθε frame, όπως περιγράψαμε παραπάνω.



Σχήμα 6.6: Τα probabilistic outputs που προκύπτουν αντίστοιχα για median φίλτρο ίδιου μεγέθους.

Κεφάλαιο 7

Συμπεράσματα

7.1 Συμβολή της Διπλωματικής Εργασίας

Στην παρούσα διπλωματική εργασία ασχοληθήκαμε με το πρόβλημα της ταξινόμησης και αναγνώρισης ανθρώπινων δράσεων καθώς και με το πρόβλημα της αναγνώρισης συνεχόμενων δράσεων σε βίντεο υπό το πρίσμα των τοπικών χωροχρονικών χαρακτηριστικών, που έχουν δείξει μεγάλη επιτυχία τα τελευταία χρόνια για το συγκεκριμένο πρόβλημα. Ασχοληθήκαμε διεξοδικά με την ανίχνευση εύρωστων χωροχρονικών σημείων ενδιαφέροντος καθώς και των κλιμάκων τους. Εξετάσαμε και πειραματιστήκαμε με διάφορους επιτυχημένους ανιχνευτές σημείων ενδιαφέροντος της βιβλιογραφίας. Μελετήσαμε και εκτελέσαμε πειράματα χρησιμοποιώντας δύο πολύ γνωστούς περιγραφητές της βιβλιογραφίας, τους HOG/HOF και HOG3D, καθώς και το συνδυασμό τους με την τεχνική Bag-of-Features για να κωδικοποιηθεί η στατιστική κατανομή τους για κάθε βίντεο. Πραγματοποιήσαμε πειράματα εκμάθησης και ταξινόμησης ανθρώπινων δράσεων με διαφορετικά frameworks, με τη χρήση SVMs, k-NN Classifiers και HMMs. Αναπτύξαμε δύο ανιχνευτές χωροχρονικών σημείων ενδιαφέροντος που στηρίζονται σε βιολογικά υποστηριγμένα φίλτρα Gabor καθώς και σε Ανάλυση Κυρίαρχης Ενέργειας. Τα πειραματικά μας αποτελέσματα έδειξαν ότι ο ένας εκ των δύο ανιχνευτών που εκτελεί ένα πιο ολοκληρωμένο φιλτράρισμα στο χωροχρόνο ξεπερνά σε όλα τα πειράματα που εκτελέσαμε τις ήδη υπάρχουσες μεθόδους σε ακρίβεια αναγνώρισης. Ο μαζικός πειραματισμός μας περιείχε δύο πολύ γνωστές βάσεις ανθρώπινων δράσεων, μία με εκτελέσεις απλών δράσεων (KTH Action Dataset) και μία με απαιτητικές δράσεις σε Χολυγουντιανές ταινίες (Hollywood2 Action Dataset). Τα πειράματα και στις δύο βάσεις δείχνουν πως οι ανιχνευτές μας δείχνουν καλύτερη συμπεριφορά από τους υπάρχοντες ανιχνευτές (Harris3D, DCA3D, Cuboids, Hessian). Επεκτείναμε τον πειραματισμό μας σε μια νέα βάση δεδομένων, την πολυτροπική και πολυαισθητηριακή βάση δεδομένων MOBOT,

όπου εκτός από το πρόβλημα αναγνώρισης ανθρώπινων δράσεων αντιμετωπίσαμε και το πρόβλημα αναγνώρισης συνεχόμενων ανθρώπινων δράσεων, επεκτείνοντας τις τεχνικές που χρησιμοποιούμε.

Συμπερασματικά , η συμβολή της διπλωματικής εργασίας για το πρόβλημα της αναγνώρισης δράσεων σε βίντεο συνοψίζεται στα ακόλουθα σημεία :

- Εισαγωγή στο πρόβλημα της Αναγνώρισης Ανθρώπινων Δράσεων σε Βίντεο, αναφορά στις σχετικές προσεγγίσεις και τις κυριότερες εφαρμογές του προβλήματος.

- Μελέτη της σχετικής έρευνας για χρήση τοπικών χαρακτηριστικών στο πρόβλημα της αναγνώρισης δράσεων. Ανασκόπηση ανιχνευτών χωροχρονικών σημείων ενδιαφέροντος της βιβλιογραφίας, παρουσίαση του εννοιολογικού μαθηματικού υποβάθρου τους καθώς και της φιλοσοφίας σχεδιασμού τους.

- Ανάπτυξη δύο αλγόριθμων ανίχνευσης χωροχρονικών σημείων ενδιαφέροντος που στηρίζονται σε Gabor φίλτρα, Ανάλυση Κυρίαρχης Teager-Kaiser Ενέργειας (TDE και Gabor3D). Παρουσίαση του μαθηματικού υποβάθρου τους και της διαδικασίας επιλογής διάφορων παραμέτρων που τους απαρτίζουν. Προσπάθεια μείωσης της πολυπλοκότητας και του χρόνου εκτέλεσης του αλγορίθμου Gabor3D.

- Συνοπτική παρουσίαση των εργαλείων που χρησιμοποιήθηκαν, εισαγωγή στην τεχνική Bag-of-Features και το πως συνδυάστηκαν για να προχωρήσουμε σε ολοκληρωμένα πειράματα ταξινόμησης και αναγνώρισης. Χρήση μιας μεθόδου εμπνευσμένης από την ψηφιακή επεξεργασία σήματος για τη μείωση της διάστασης των Bag-of-Features ιστογραμμάτων.

- Μεγάλης κλίμακας πειράματα στη βάση δεδομένων KTH με τη χρήση SVMs, kNN, HMMs και σύγκριση με τη βιβλιογραφία.

- Μεγάλης κλίμακας πειράματα στη βάση δεδομένων Hollywood2 με τη χρήση SVMs.

- Πειράματα στη βάση δεδομένων MOBOT όπου επεκταθήκαμε σε τεχνικές αναγνώρισης συνεχόμενων δράσεων σε βίντεο.

- Δημοσίευση μερικών από τα αποτελέσματα της παρούσας διπλωματικής εργασίας στο IEEE International Conference on Image Processing - ICIP 2014 [37].

7.2 Μελλοντικές Κατευθύνσεις

Από τις ιδέες που αναπτύχθηκαν και από τα ερευνητικά αποτελέσματα που προέκυψαν στα πλαίσια της παρούσας διπλωματικής εργασίας αναδύονται πολλαπλές πιθανές προεκτάσεις για μελλοντική έρευνα. Η μελέτη των τοπικών χαρακτηριστικών έδειξε πως αυτά αποτελούν μια αρκετά αποδοτική προσέγγιση για το πρόβλημα της αναγνώρισης ανθρώπινων δράσεων σε βίντεο. Η ανάπτυξη των μεθόδων μας και η διεξοδική μελέτη των τεχνικών που χρησιμοποιούνται στη βιβλιογραφία μας έδειξαν τις δυνατότητες που υπάρχουν αλλά και τους περιορισμούς μας. Σίγουρα η έρευνα σε ένα τόσο ενδιαφέρον πεδίο έχει πολλά περιθώρια ακόμα να αναπτυχθεί προς πάρα πολλές κατευθύνσεις. Ας σταθούμε ωστόσο στις μελλοντικές κατευθύνσεις που πυροδοτούνται από την παρούσα διπλωματική εργασία, οι οποίες συνοψίζονται στα εξής:

- Περεταίρω επιτάχυνση του αλγόριθμου Gabor3D για εφαρμογές πραγματικού χρόνου με τη χρήση παράλληλης επεξεργασίας από GPUs. Το πολυκαναλικό φιλτράρισμα είναι μια διαδικασία που μπορεί να παραλληλοποιηθεί και να χρησιμοποιηθεί σε real time εφαρμογές. Παρόλη την προσπάθεια που έχει γίνει για μείωση της πολυπλοκότητας του αλγόριθμου σίγουρα υπάρχουν ακόμα περιθώρια βελτίωσης προς αυτήν την κατεύθυνση.

- Χρήση top-down πληροφορίας από μεθόδους εκτίμησης πόζας (pose estimation) και συνδυασμός τους με μεθόδους αναγνώρισης δράσεων για καλύτερη απομόνωση της περιοχής ενδιαφέροντος γύρω από τον άνθρωπο και καλύτερη εκτίμηση των κινήσεών του.

- Ορθότερη χρήση των μοντέλων HMMs για το πρόβλημα της αναγνώρισης δράσεων. Αυτό σημαίνει καλύτερη μοντελοποίηση της χρονικής εξέλιξης των δράσεων με τρόπο που να μπορούν να εκπαιδευτούν αποδοτικά τα HMMs

- Θα ήταν ενδιαφέρον να μελετηθούν τεχνικές εκτίμησης της κίνησης της κάμερας. Έχουν γίνει ήδη προσπάθειες προς αυτήν την κατεύθυνση, όπως περιγράφηκε στο Κεφάλαιο 2, οι οποίες είναι πολλά υποσχόμενες σε βάσεις με πολύ έντονη κίνηση της κάμερας, όπως η Hollywood2.

Βιβλιογραφία

[1] E. H. Adelson and J. R. Bergen. Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Amer.*, 2(2):284–299, 1985.

[2] N. S. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

[3] V. Bettadapura, G. Schindler, T. Plötz, and I. Essa. Augmenting bag-of-words: Data-driven discovery of temporal and structural information for activity recognition. In *Proc. IEEE Conf. CVPR*, 2013.

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

[5] N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. IEEE Conf. CVPR*, 2005.

[7] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *Computer Vision–ECCV 2006*, pages 428–441. Springer, 2006.

[8] J. Daugman. Two-dimensional spectral analysis of cortical receptive field profiles. *Vision Research*, 20(10):847–856, 1980.

[9] J. Daugman. Uncertainty Relation for Resolution in Space, Spatial Frequency and Orientation Optimized by Two-Dimensional Visual Cortical Filters. *J. Opt. Soc. Amer.*, 2(7):1160–1169, 1985.

[10] K. G. Derpanis, M. Sizintsev, K. Cannons, and R. P. Wildes. Efficient action spotting based on a spacetime oriented structure representation. In *Proc. IEEE Conf. CVPR*, 2010.

[11] D. Dimitriadis and P. Maragos. Continuous energy demodulation methods and application to speech analysis. *Speech communication*, 48(7):819–837, 2006.

[12] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Proc. IEEE Int'l Workshop on VS-PETS*, 2005.

[13] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification*. John Wiley & Sons,, 1999.

[14] W. T. Freeman and E. H. Adelson. The Design and Use of Steerable Filters. *IEEE Trans. PAMI*, 13(6):891–906, 1991.

[15] D. Gabor. Theory of Communication. *IEE Journal (London)*, 93:429–457, 1946.

[16] S. Geisser. *Predictive inference*, volume 55. CRC Press, 1993.

[17] C. Georgakis, P. Maragos, G. Evangelopoulos, and D. Dimitriadis. Dominant spatio-temporal modulations and energy tracking in videos: Application to interest point detection for action recognition. In *Proc. ICIP*, 2012.

[18] P. Giannoulis, G. Potamianos, A. Katsamanis, and P. Maragos. Multi-microphone fusion for detection of speech and acoustic events in smart places. In *Proc. Eusipco*, 2014. (to appear).

[19] C. Harris and M. Stephens. A combined corner and edge detector. In *Proc. Alvey Vision Conf.*, 1988.

[20] Z. S Harris. Distributional structure. *Word*, 1954.

[21] J. Havlicek, A. Bovik, and D. Chen. Am-fm image modeling and gabor analysis. *Visual Communication & Image Processing*, 64:343–386, 1999.

[22] J. P. Havlicek, D. S. Harding, and A. C. Bovik. Multidimensional quasi-eigenfunction approximations and multicomponent am-fm models. *IEEE Trans. Image Processing*, 9(2):227–242, 2000.

[23] D. J. Heeger. Model for the extraction of image flow. *J. Opt. Soc. Amer.*, 4(8):1455–1471, 1987.

[24] D. J. Heeger. Optical flow using spatiotemporal filters. *Int'l J. Comp. Vision*, 1(4):279–302, 1988.

[25] J. F. Kaiser. On a simple algorithm to calculate the energy'of a signal. In *Proc. IEEE Conf. ICASSP*, 1990.

[26] A. Kläser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3D-Gradients. In *Proc. BMVC*, 2008.

[27] H. Knutsson and M. T. Andersson. What's so good about quadrature filters? In *Proc. ICIP*, 2003.

[28] J. J. Koenderink and A.J. van Doorn. Representation of local geometry in the visual system. *Biol. Cybern.*, 55:367–375, 1987.

[29] P. Koutras, A. Katsamanis, and P. Maragos. Predicting eyes' fixations in movie videos: Visual saliency experiments on a new eye-tracking database. In *Engineering Psychology and Cognitive Ergonomics*, pages 183–194. Springer, 2014.

[30] A. Kurakin, Z. Zhang, and Z. Liu. A real time system for dynamic hand gesture recognition with a depth sensor. In *Proc. Eusipco*, pages 1975–1979, 2012.

[31] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005.

[32] I. Laptev and T. Lindeberg. Space-time interest points. In *Proc. ICCV*, 2003.

[33] I. Laptev and T. Lindeberg. Local descriptors for spatio-temporal recognition. In *Proc. SCVMA*, 2004.

[34] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Proc. IEEE Conf. CVPR*, 2008.

[35] D. G. Lowe. Object recognition from local scale-invariant features. In *Proc. Int'l Conf. on Computer Vision*, 1999.

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

[37] 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 *Proc. ICIP*, 2014. (to appear).

[38] P. Maragos. *Image Analysis and Computer Vision*. NTUA, 2011.

[39] P. Maragos and A. C. Bovik. Image demodulation using multidimensional energy separation. *J. Opt. Soc. Amer.*, 12(9):1867–1876, 1995.

[40] P. Maragos, J. F. Kaiser, and T. F. Quatieri. Energy separation in signal modulations with application to speech analysis. *IEEE Trans. Signal Processing*, 41(10):3024–3051, 1993.

[41] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *Proc. IEEE Conf. CVPR*, 2009.

[42] D. R. Martin, C. C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Trans. PAMI*, 26(5):530–549, 2004.

[43] K. Nandakumar, K. W. Wan, S. Chan, W. Ng, J. G. Wang, and W. Y. Yau. A multi-modal gesture recognition system using audio, video, and skeletal joint data. In *ICMI*, pages 475–482, 2013.

[44] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *Int'l J. Comp. Vision*, 79(3):299–318, 2008.

[45] O. Oreifej and L. Zicheng. HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences. In *CVPR*, pages 716–723, 2013.

[46] G. Pavlakos, S. Theodorakis, V. Pitsikalis, S. Katsamanis, and P. Maragos. Kinect-based multimodal gesture recognition using a two-pass fusion scheme. In *Proc. ICIP*, 2014. (to appear).

[47] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in large margin classifiers*. Citeseer, 1999.

[48] R. Poppe. A survey on vision-based human action recognition. *Image and vision computing*, 28(6):976–990, 2010.

[49] L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[50] L. Rabiner and R. Schafer. *Digital processing of speech signals*, volume 100. Prentice-hall Englewood Cliffs, 1978.

[51] L. R. Rabiner and B. H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[52] S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. In *Proc. IEEE Conf. CVPR*, 2012.

[53] C. Schüldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In *Proc. ICPR*, 2004.

[54] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proc. Int'l Conf. on Multimedia*, 2007.

[55] H. Teager and S. Teager. Evidence for nonlinear sound production mechanisms in the vocal tract. In *Speech production and speech modelling*, pages 241–261. Springer, 1990.

[56] C. Theriault, N. Thome, and M. Cord. Dynamic scene classification: Learning motion descriptors with slow features analysis. In *Proc. IEEE Conf. CVPR*, 2013.

[57] Tinne Tuytelaars. Dense interest points. In *Proc. IEEE Conf. CVPR*, pages 2281–2288. IEEE, 2010.

[58] H. Wang, A. Kläser, C. Schmid, and C. Liu. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, 103(1):60–79, 2013.

[59] H. Wang and C. Schmid. Action recognition with improved trajectories. In *Proc. ICCV*, pages 3551–3558. IEEE, 2013.

[60] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *Proc. BMVC*, 2009.

[61] Y. Wang and G. Mori. Human action recognition by semilatent topic models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(10):1762–1774, 2009.

[62] J. Willamowski, D. Arregui, G. Csurka, C. R. Dance, and L. Fan. Categorizing nine visual classes using local appearance descriptors. In *Proc. ICPR Workshop on Learning for Adaptable Visual Systems*, 2004.

[63] G. Willems, T. Tuytelaars, and L. Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *Computer Vision–ECCV 2008*, pages 650–663. Springer, 2008.

[64] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. *The HTK book*, volume 2. Entropic Cambridge Research Laboratory Cambridge, 1997.

[65] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision*, 73(2):213–238, 2007.

[66] Z. Zhang and D. Tao. Slow feature analysis for human action recognition. *IEEE Trans. PAMI*, 34(3):436–450, 2012.