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

Μέθοδοι Εξαγωγής και Κωδικοποίησης Χαρακτηριστικών για την Αναγνώριση Ανθρώπινων Δράσεων και Χειρονομιών με εφαρμογές στην Επικοινωνία Ανθρώπου-Μηχανής

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Νικολάου Α. Κάρδαρη

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2015

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

Μέθοδοι Εξαγωγής και Κωδικοποίησης Χαρακτηριστικών για την Αναγνώριση Ανθρώπινων Δράσεων και Χειρονομιών με εφαρμογές στην Επικοινωνία Ανθρώπου-Μηχανής

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Νικολάου Α. Κάρδαρη

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 30η Οκτωβρίου 2015.

.....	.....	.....
Πέτρος Μαραγκός	Γεώργιος Παπαβασιλόπουλος	Γεράσιμος Ποταμιάνος
Καθηγητής	Καθηγητής	Αναπληρωτής Καθηγητής
Ε.Μ.Π	Ε.Μ.Π	Παν/μίου Θεσσαλίας

Αθήνα, Οκτώβριος 2015

.....
Νικόλαος Α. Κάρδαρης
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Ευχαριστίες

Αρχικά θα ήθελα να ευχαριστήσω θερμά τον Καθ. Πέτρο Μαραγκό, για την ευκαιρία που μου έδωσε να εκπονήσω την παρούσα διπλωματική και την καθοδήγησή του όλο το προηγούμενο διάστημα. Οι διαλέξεις του διαδραμάτισαν καταλυτικό ρόλο στη διαμόρφωση των επιστημονικών μου ενδιαφερόντων και η αγάπη του για το πεδίο της Όρασης Υπολογιστών αποτέλεσε έμπνευση για μένα. Ήταν μεγάλη μου χαρά να συνεργαστώ τόσο με τον ίδιο, αλλά και τη δραστήρια ερευνητική του ομάδα. Θα ήθελα επίσης να ευχαριστήσω το Βασίλη Πιτσικάλη για τις συμβουλές, το ενδιαφέρον του, τη συνεχή ενθάρρυνση καθώς και τη συμβολή του στη διεύρυνση των ερευνητικών μου οριζόντων. Ιδιαίτερη αναφορά θα ήθελα να κάνω στην Έφη Μαυρουδή και τον Ισίδωρο Ροδομαγουλάκη, με τους οποίους είχα τη χαρά να έχω μια γόνιμη κι ευχάριστη συνεργασία. Επίσης, ευχαριστώ όλα τα υπόλοιπα τα μέλη του εργαστηρίου Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων για τη βοήθειά τους σε ο,τιδήποτε χρειάστηκε, αλλά κι επειδή έκαναν τη διπλωματική αυτή μια όμορφη εμπειρία.

Δεδομένου πως η παρούσα εργασία σηματοδοτεί και το τέλος των φοιτητικών μου χρόνων, θα ήθελα να εκφράσω τις ευχαριστίες μου σε όλους φίλους και συμφοιτητές μου με τους οποίους περάσαμε μαζί τα χρόνια αυτά, μοιραστήκαμε εμπειρίες και ανταλλάξαμε ιδέες.

Τέλος, θα ήθελα να εκφράσω την ευγνωμοσύνη μου στους γονείς μου και τον αδερφό μου για την αστείρευτη αγάπη και φροντίδα τους, την υπομονή τους, τη συνεχή στήριξη και τη συμπαράσταση που μου έδειξαν όλα τα προηγούμενα χρόνια.

Περίληψη

Η παρούσα διπλωματική εργασία πραγματεύεται το πρόβλημα της αυτόματης ταξινόμησης και αναγνώρισης ανθρώπινων δράσεων και χειρονομιών σε βίντεο καθώς την εφαρμογή τους κυρίως στην αλληλεπίδραση ανθρώπου-μηχανής. Αντιμετωπίζουμε την αναγνώριση δράσεων και χειρονομιών υπό κοινό πρίσμα, θεωρώντας τις τελευταίες ως ειδικές περιπτώσεις των πρώτων. Για την εξαγωγή χαρακτηριστικών βασιζόμαστε σε ευρέως χρησιμοποιούμενα χαρακτηριστικά: τα χωρο-χρονικά σημεία ενδιαφέροντος και τις πυκνές τροχιές. Εν συνεχεία πειραματιζόμαστε εκτενώς με μια σειρά αναπαραστάσεων βίντεο, όπως οι Bag-of-Words, χωρο-χρονικές πυραμίδες, VLAD και διάνυσμα Fisher. Για την αξιολόγηση των διαφορετικών μεθόδων πραγματοποιούμε τα πειράματά μας σε μια σειρά βάσεων δεδομένων, οι οποίες είναι διαδεδομένες στη διεθνή βιβλιογραφία. Επίσης, αξιολογούμε την επίδοση των χρησιμοποιούμενων μεθόδων στην ερευνητική βάση χειρονομιών MOBOT, η οποία περιέχει ένα ευρύ σύνολο χειρονομιών που εκτελούνται από ηλικιωμένους και ως εκ τούτου παρουσιάζει ιδιαίτερες προκλήσεις. Τέλος, περιγράφουμε την εφαρμογή των παραπάνω στη δημιουργία ενός διαδραστικού περιβάλλοντος επικοινωνίας ανθρώπου-ρομπότ μέσω οπτικής αναγνώρισης χειρονομιών. Για το σκοπό αυτό αναπτύξαμε ένα σύστημα εντοπισμού των των χειρονομιών, το οποίο μας επιτρέπει την αναγνώρισή τους σε συνεχή ροή βίντεο.

Λέξεις κλειδιά: αναγνώριση ανθρώπινων δράσεων, αναγνώριση ανθρώπινων χειρονομιών, αναπαράσταση βίντεο, χωροχρονικά σημεία ενδιαφέροντος, πυκνές τροχιές, Bag-Of-Words, VLAD, διάνυσμα Fisher, Μηχανές Διανυσμάτων Υποστήριξης, συσταδοποίηση, on-line αναγνώριση χειρονομιών

Abstract

The aim of this thesis is to deal with the problem of automatic human action and gesture recognition, as well as its application in human-computer interaction. We examine action and gesture recognition from a common point of view, with the latter considered as a special case of the former. We employ "state-of-the-art" visual feature extraction methods, such as spatiotemporal interest points, and dense trajectories. We further experiment with widely used video representation methods, such as Bag-of-Words, spatiotemporal pyramids, VLAD and Fisher vector. Different methods are evaluated on popular action databases which constitute standard benchmarks in the literature. We also evaluate our system on the challenging MOBOT dataset, which contains various gestures performed by elderly people. Last but not least, we describe the basic steps towards the development of an on-line gesture recognition system for the purpose of human-robot interaction. We developed a gesture activity detection system which allows gesture recognition from a continuous video steam, by localizing gestures in time.

Keywords: human action recognition, human action recognition, video representation, spatiotemporal interest points, dense trajectories, Bag-Of-Words, VLAD, Fisher vector, Support Vector Machines, clustering, on-line gesture recogntiion

Περιεχόμενα

Κατάλογος σχημάτων

Κατάλογος πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Γενικά περί Όρασης Υπολογιστών

Η Όραση Υπολογιστών αποτελεί ένα σύγχρονο και συνεχώς αναπτυσσόμενο κλάδο της επιστήμης και της τεχνολογίας. Όπως μαρτυρά και η σημασία των λέξεων, σκοπός της Όρασης Υπολογιστών είναι να επιτρέψει στις μηχανές να "δουν", να αντιληφθούν και να ερμηνεύσουν τον κόσμο μέσω οπτικών ερεθισμάτων. Για τον άνθρωπο η ικανότητα αυτή είναι εγγενής και αυτόματη: όταν για παράδειγμα βλέπουμε ένα αντικείμενο ή ένα πρόσωπο, το αναγνωρίζουμε αμέσως, ακόμα κι αν το έχουμε δει ελάχιστες φορές, με διαφορετική μορφή, υπό άλλες συνθήκες κτλ. Η "ευκολία" αυτή του ανθρώπου είναι φυσικά αποτέλεσμα εξελικτικής πορείας χιλιετιών. Από αυτήν την ικανότητα εμπνέεται η Όραση Υπολογιστών και αναλογιζόμενοι τη νεότητα του συγκεκριμένου τομέα θα λέγαμε πως η σχετική έρευνα έχει παρουσιάσει έως σήμερα σημαντικά αποτελέσματα.

Διατυπώνοντάς το πιο φορμαλιστικά, η Όραση Υπολογιστών συνιστά ένα σύνολο μεθόδων για την εξαγωγή συμβολικής πληροφορίας μέσω της ανάλυσης εικόνων κι εν γένει πολυδιάστατων σημάτων του φυσικού κόσμου. Σκοπός, δηλαδή, είναι η επίλυση του αντίστροφου προβλήματος σε σχέση με άλλους συγγενείς κλάδους, όπως τα γραφικά υπολογιστών, που προσπαθούν από τη συμβολική περιγραφή του κόσμου να κατασκευάσουν εικόνες. Για την ανάλυση και την κατανόηση της οπτικής πληροφορίας η Όραση Υπολογιστών συνδυάζει έννοιες και μεθόδους από πολλούς τομείς, όπως η επεξεργασία σημάτων, η αναγνώριση προτύπων, η μηχανική μάθηση (machine learning), τα εφαρμοσμένα μαθηματικά, η φυσική, η νευροβιολογία και ο αυτόματος έλεγχος. Η προς ανάλυση οπτική πληροφορία μπορεί να έχει τη μορφή μιας απλής ψη-

φιακής εικόνας, μιας ακολουθίας εικόνων (βίντεο), μιας εικόνας βάθους, ενός πολυδιάστατου βιοϊατρικού σήματος κ.α. Οι πρακτικές εφαρμογές της Όρασης Υπολογιστών είναι αμέτρητες και καλύπτουν μια πληθώρα διαφορετικών τομέων. Κάποιες αντιπροσωπευτικές είναι οι εξής:

- Επικοινωνία ανθρώπου-μηχανής μέσω διαδραστικών περιβαλλόντων, όπως π.χ. ενός συστήματος αυτόματης αναγνώρισης χειρονομιών.
- Ανάκτηση και οργάνωση πληροφοριών, π.χ. ανάλυση του περιεχομένου και κατηγοριοποίηση των εικόνων και βίντεο του διαδικτύου για τη δημιουργία βάσης δεδομένων και την εύκολη αναζήτηση και ανάκλησή τους βάσει του περιεχομένου.
- Αυτόματη επιτήρηση και επισκόπηση π.χ. σε μια αλυσίδας παραγωγής για τον εντοπισμό ελαττωματικών προϊόντων.
- Αυτόματη πλοήγηση μηχανών, όπως π.χ. αυτοκινήτων ή αυτόνομων ρομποτικών διατάξεων μέσω της οπτικής ανάλυσης του περιβάλλοντός τους.
- Βιοϊατρικές εφαρμογές, όπως συστήματα υποβοηθούμενης διάγνωσης για την αυτόματη διάγνωση από βιοϊατρικά δεδομένα.

1.2 Τα προβλήματα της αναγνώρισης ανθρώπινων δράσεων και χειρονομιών

Ορισμός Το πρόβλημα της αυτόματης αναγνώρισης δράσεων έχει γίνει ιδιαίτερα δημοφιλές τα τελευταία χρόνια, με την έρευνα πάνω στο αντικείμενο να σημειώνει αλματώδη πρόοδο. Υπάρχουν πολλοί τρόποι να ορίσει κανείς μια δράση, ανάλογα με τη μέθοδο που υιοθετείται και την εφαρμογή για στην οποία αναφέρεται. Η πιο διαδεδομένη προσέγγιση [1] ορίζει τις ανθρώπινες δράσεις σε τρία επίπεδα: τις *πρωταρχικές δράσεις*, τις *δράσεις* και τις *δραστηριότητες*. Μια *πρωταρχική δράση* (action primitive) αποτελεί μια μεμονωμένη κίνηση που σχετίζεται συνήθως με κάποιο μέλος του ανθρώπινου σώματος, π.χ. η κίνηση του ποδιού προς τα εμπρός. Μια *δράση* απαρτίζεται από μια σειρά διαδοχικών πρωταρχικών δράσεων και περιλαμβάνει συνήθως κινήσεις ολόκληρου του σώματος, π.χ. "τρέχω". Μια *δραστηριότητα* περιλαμβάνει επιμέρους δράσεις και περιγράφει μια πιο γενική και περίπλοκη ασχολία. Μπορεί επίσης να περιλαμβάνει παραπάνω από έναν ανθρώπους, καθώς

και αντικείμενα τα οποία αλληλεπιδρούν. Για παράδειγμα, η δραστηριότητα "παίζω ποδόσφαιρο" περιλαμβάνει τις δράσεις "τρέχω", "κλοτσάω τη μπάλα", "κυνηγάω τον αντίπαλο" κ.α. Η συντριπτική πλειονότητα των μεθόδων που έχουν αναπτυχθεί στοχεύουν στην αναγνώριση των δεύτερων, "μεσαίου" επιπέδου δράσεων, μιας και αποτελούν τις μικρότερες αυτόνομες και σημασιολογικά συνεκτικές οντότητες.

Απ' την άλλη πλευρά, ο ορισμός των χειρονομιών είναι πιο συγκεκριμένος και συνήθως νοούνται ως κινήσεις του χεριού με αρχή και τέλος. Η χειρονομίες μπορεί να είναι αυθόρμητες, όπως αυτές που εκτελούμε καθώς μιλάμε, ή προκαθορισμένες, οι οποίες π.χ. μπορεί να αντιστοιχούν σε κάποια εντολή προς μια μηχανή. Στο άκρο βρίσκεται η νοηματική γλώσσα, στην οποία οι χειρονομίες είναι καθορισμένες και μεταφέρουν πλήρες, σαφές και δομημένο σημασιολογικό περιεχόμενο. Στην περίπτωση αυτή μιλάμε για "νοήματα" (signs), τα οποία δεν είναι απλές "χονδροειδείς" κινήσεις, αλλά λεπτομερώς καθορισμένες κινήσεις, των οποίων η νοηματοδοσία εμπλέκει τη χειρομορφή, τη στάση του σώματος, τη σειρά με την οποία γίνονται κ.α. Στην παρούσα διπλωματική, ακολουθώντας και τις τάσεις της διεθνούς βιβλιογραφίας, αντιμετωπίζουμε τις χειρονομίες ως μια υποκατηγορία δράσεων και εφαρμόζουμε τις ίδιες μεθόδους για την ανάλυση και την αναγνώρισή τους.

Όταν αναφερόμαστε στην αναγνώριση δράσεων η είσοδος μπορεί να είναι είτε ένα βίντεο που απεικονίζει μια δράση και έχει μαγνητοσκοπηθεί και αποθηκευθεί σε προηγούμενο χρόνο, είτε μια συνεχής ροή από καρέ που λαμβάνονται on-line ή σε πραγματικό χρόνο από κάποιον αισθητήρα. Τα έγχρωμα δεδομένα των καρέ μπορεί ακόμη να συνοδεύονται και από εικόνες ή αλλιώς "χάρτες" βάθους, όπως συμβαίνει με αισθητήρα Kinect.

Συναφή και επιμέρους προβλήματα Το πρόβλημα της αναγνώρισης ανθρώπινων δράσεων συνήθως παίρνει την απλή μορφή της *ταξινόμησης* ενός βίντεο σε μια κατηγορία που αντιστοιχεί στη δράση την οποία περιέχει. Για παράδειγμα, αν έχουμε ένα βίντεο στο οποίο ένας άνθρωπος περπατά θέλουμε με αυτόματο τρόπο να του αποδώσουμε την ετικέτα (label) "περπατώ". Ένα συναφές πρόβλημα είναι αυτό του *χωρικού* και *χρονικού εντοπισμού* των δράσεων εντός ενός βίντεο (action localization). Στο προηγούμενο παράδειγμα, δηλαδή, θα θέλαμε να βρούμε σε κάθε καρέ το μικρότερο ορθογώνιο που περιβάλλει τον άνθρωπο (bounding box), καθώς και τη χρονική στιγμή που αρχίζει και τελειώνει το περπάτημα. Ο συνδυασμός των δύο παραπάνω αποτελεί ένα πιο γενικό και αρκετά δυσκολότερο πρόβλημα αναγνώρισης δράσεων

σε συνεχή ροή βίντεο, στην οποία δε γνωρίζουμε κατά πόσο υπάρχει ή πού εντοπίζεται κάποια δράση. Τυπικά, στο πρώτο αντιστοιχεί ο όρος *ταξινόμηση δράσεων* και στο τελευταίο ο όρος *αναγνώριση δράσεων*. Ωστόσο στη βιβλιογραφία οι δύο αυτοί όροι συχνά συγχέονται και αναφέρονται και οι δύο συνήθως στην ταξινόμηση δράσεων. Τη σύμβαση αυτή ακολουθούμε και στην παρούσα διπλωματική, ορίζοντας ρητά το πρόβλημα στις περιπτώσεις που δεν είναι σαφές από τα συμφραζόμενα.

Οι προκλήσεις του προβλήματος της αναγνώρισης ανθρώπινων δράσεων Παρ' όλη την εξέλιξη που έχει σημειωθεί τα τελευταία χρόνια, η αυτόματη αναγνώριση δράσεων παραμένει ένα δυσεπίλυτο πρόβλημα. Αυτό απορρέει τόσο από την εγγενή πολυπλοκότητά του, όσο και από τις ιδιαίτερες συνθήκες κάτω από τις οποίες εκτελούνται και καλούνται να αναγνωριστούν οι διάφορες δράσεις. Ενδεικτικά, οι σημαντικότερες προκλήσεις που καλείται να αντιμετωπίσει ένα σύστημα αναγνώρισης δράσεων είναι οι εξής:

- *Μεταβολές στη γωνία λήψης και την κλίμακα* Η γωνία λήψης διαδραματίζει ίσως το σημαντικότερο ρόλο στην στη μορφή που παίρνει μια δράση και στη δυνατότητα αναγνώρισής της και μπορεί να είναι αιτία σφαλμάτων ακόμα κι από τον άνθρωπο. Χαρακτηριστικό παράδειγμα είναι ο διαχωρισμός των δράσεων "περπατώ" και "τρέχω" όταν έχουν ληφθεί *en face* (με την κάμερα μπροστά από τον άνθρωπο). Επίσης, η απόσταση από την κάμερα και η ανάλυση του βίντεο μπορεί να μεταβάλει σημαντικά την εμφάνιση και καθορίζει το πόσο λεπτομερώς απεικονίζεται μια δράση. Είναι, ακόμη, αρκετά συνηθισμένο να μεταβάλλονται τόσο η γωνία λήψης όσο και η απόσταση της κάμερας διάρκεια του βίντεο, π.χ. η κάμερα μπορεί να κινείται μαζί με τον άνθρωπο, να περιστρέφεται να κάνει "ζουμ" κτλ. Τέτοιες άσχετες κινήσεις οδηγούν σε μια περίπλοκη συνισταμένη κίνηση, από την οποία είναι δύσκολο να απομονωθεί αυτή που αντιστοιχεί στον άνθρωπο.

- *Συνθήκες λήψης δεδομένων* Το περιβάλλον στο οποίο εκτελείται μια δράση έχει μεγάλο αντίκτυπο στη δυνατότητα αναγνώρισή της. Σε δυναμικά και ρεαλιστικά περιβάλλοντα μπορεί να έχουμε επικαλύψεις, όταν π.χ. ένα μέρος του ανθρώπου κρύβεται πίσω από κάποιο αντικείμενο ή ανθρώπους τα οποία ίσως κινούνται, ή φεύγει στιγμιαία έξω από τα όρια του καρέ (shot boundaries). Επίσης ο φωτισμός και οι σκιάσεις έχουν μεγάλη επίδραση στην εμφάνιση μιας δράσης.

- *Μεταβλητότητα μεταξύ δράσεων εντός της ίδιας κλάσης και μεταξύ των διαφορικών κλάσεων* Η εκτέλεση της ίδιας δράσεις μπορεί να διαφέρει σημαντικά από άνθρωπο σε άνθρωπο ως προς την ταχύτητα εκτέλεσης, το μοτίβο κτλ. Επίσης, πολλές κατηγορίες δράσεων παρουσιάζουν μεγάλες ομοιότητες, όπως για παράδειγμα οι κλάσεις "περπατώ" και "τρέχω". Έτσι, μπορεί για παράδειγμα ένας άνθρωπος που περπατάει γρήγορα να μοιάζει περισσότερο με κάποιον άλλο που τρέχει αργά και λιγότερο με κάποιον του περπατά αλλά είναι ψηλότερος και έχει μεγαλύτερη δρασκελιά. Αυτό αναφέρεται συνήθως ως inter-class και intra-class variation.

- *Συλλογή δεδομένων* Ένα από τα σημαντικότερα προβλήματα είναι η κατασκευή αντιπροσωπευτικών βάσεων δεδομένων. Η διαδικασία βιντεοσκόπησης δράσεων είναι δύσκολη και χρονοβόρα και είναι πρακτικά αδύνατο να συγκεντρωθεί χειροκίνητα μεγάλος αριθμός δεδομένων που να καλύπτει ένα ευρύ φάσμα δράσεων. Γι' αυτό οι περισσότερες βάσεις δεδομένων έχουν προκύψει από ήδη διαθέσιμα βίντεο, όπως τηλεοπτικές μεταδόσεις 1.3.2, διαδικτυακές υπηρεσίες 1.3.4 (π.χ. youtube), ταινίες του Χόλυγουντ 1.3.3 κ.α. Ακόμη και σε αυτές τις περιπτώσεις όμως οι επισημείωσεις (annotations) των βίντεο που συγκεντρώνονται είναι εξίσου επίπονη και χρονοβόρα και οι αυτόματες επισημειώσεις [2] εισάγουν αρκετά σφάλματα. Επίσης, όπως έδειξαν οι [3]. οι κατηγορίες δράσεων που συγκεντρώνονται με αυτόν τον τρόπο κατά κανόνα δεν είναι αντιπροσωπευτικές των πραγματικών δράσεων που εκτελεί ένας άνθρωπος στην καθημερινότητά του.

Εφαρμογές της αναγνώρισης ανθρώπινων δράσεων Αναμφίβολα, η έντονη ερευνητική δραστηριότητα στο πεδίο της αναγνώρισης δράσεων αντανακλά την πληθώρα πρακτικών εφαρμογών που βρίσκει και το ενδιαφέρον που υπάρχει γι' αυτές. Ενδεικτικά αναφέρουμε τις εξής πιο χαρακτηριστικές:

- Αυτόματη αρχειοθέτηση και ανάκτηση δεδομένων βίντεο. Με τον αριθμό των διαθέσιμων οπτικοακουστικών δεδομένων στο διαδίκτυο να αυξάνεται ραγδαία, η αυτόματη σημασιολογική ανάλυσή τους και η οργάνωσή τους σε βάσεις δεδομένων αποτελεί μια μεγάλη πρόκληση. Αυτό θα επιτρέψει π.χ. στους χρήστες να εντοπίζουν δεδομένα βίντεο βάσει του πραγματικού περιεχομένου τους κι όχι άλλων "μεταδεδομένων".

- Ανάπτυξη "'εξυπνων", διαδραστικών περιβαλλόντων, όπως π.χ. έξυπνα σπίτια, τα οποία θα αναγνωρίζουν τη δραστηριότητα του χρήστη αντιδρώντας κατάλληλα και τα οποία θα μπορούν να λαμβάνουν εντολές μέσω χειρονομιών. Για παράδειγμα ένα "έξυπνο γραφείο" ή χώρος εργασίας θα μπορεί να αναγνωρίζει να αναγνωρίζει κινδύνους για την αποφυγή εργατικών ατυχημάτων, κ.α.

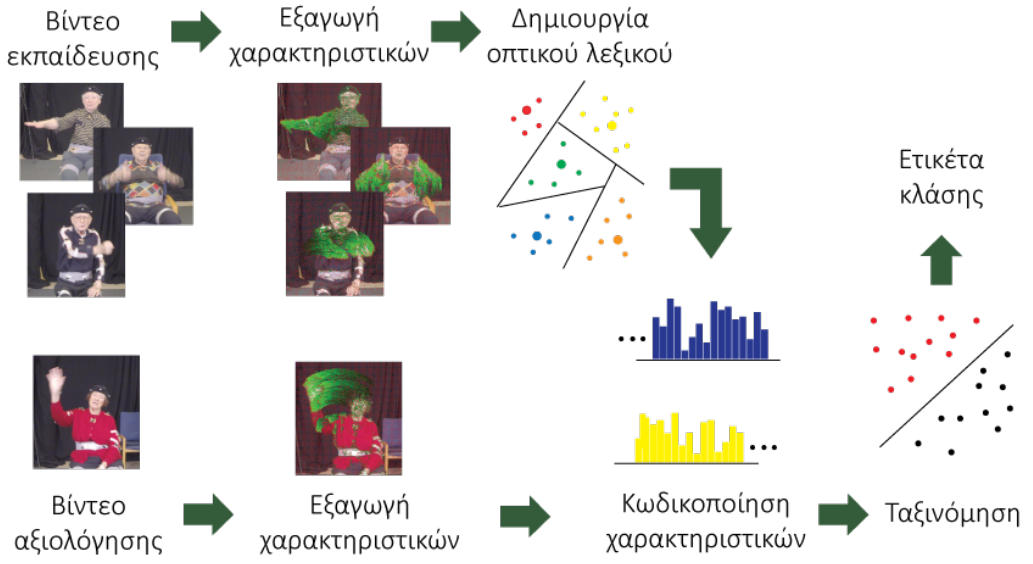
- Δημιουργία διαδραστικών βιντεοπαιχιδιών που θα απεικονίζουν τις πραγματικές δράσεις του χρήστη σε αντίστοιχες εντός του παιχνιδιού, προσομοιάζοντας έτσι εμπειρία εικονικής πραγματικότητας.

- Διάγνωση παθολογικής δραστηριότητας από δεδομένα ιατρικών απεικονιστικών διατάξεων, π.χ. αναγνώριση παθολογικής δραστηριότητας του εγκεφάλου από δεδομένα λειτουργικής μαγνητικής τομογραφίας (fMRI).

- Ανάπτυξη ρομποτικών διατάξεων που θα υποβοηθούν ηλικιωμένους ή άτομα με αναπηρίες, αναγνωρίζοντας τις προθέσεις τους (π.χ. πότε θέλουν να σηκωθούν, πότε χρειάζονται βοήθεια) προλαμβάνοντας πιθανά ατυχήματα, ή εντοπίζοντας πιθανές παθολογίες [4].

Σημαντικό ρόλο στην ολοένα μεγαλύτερη πρακτική εφαρμογή μεθόδων αναγνώρισης δράσεων και χειρονομιών έχει διαδραματίσει η ανάπτυξη αισθητήρων μεγάλης ευκρίνεια και χαμηλού κόστους όπως ο αισθητήρας Kinect. Αυτοί επιτρέπουν τη λήψη και αξιοποίηση δεδομένων πολλών τροπικοτήτων (modalities), δηλαδή διαφορετικών καναλιών πληροφορίας, όπως το έγχρωμο βίντεο, το βάθος, ο σκελετός, ο ήχος κτλ. Έχει γίνει έτσι εφικτή η ανάπτυξη πολυαισθητηριακών διατάξεων που πραγματοποιούν *πολυτροπική επεξεργασία* και σύμμειξη (multimodal fusion) μιας πληθώρας οπτικών αλλά και άλλων δεδομένων (ακουστικών, laser, κ.α.).

Γενική περιγραφή ενός συστήματος αναγνώρισης ανθρώπινων δράσεων Στη γενικότερή τους μορφή, τα περισσότερα και δημοφιλέστερα συστήματα αναγνώρισης δράσεων που συναντά κανείς στη βιβλιογραφία ακολουθούν όλα ή κάποια από τα επιμέρους στάδια που φαίνονται στο Σχήμα 1.1 και αντιστοιχούν σε διαφορετικά υποσυστήματα. Κατά κανόνα τα διαθέσιμα δεδομένα χωρίζονται σε δεδομένα εκπαίδευσης και αξιολόγησης, συνήθως με ποσοστά 80% και 20% αντίστοιχα. Τα πρώτα χρησιμοποιούνται για την εκπαίδευση και την εξαγωγή ενός



Σχήμα 1.1: Επισκόπηση ενός γενικού συστήματος ταξινόμησης δράσεων, όπου απεικονίζονται τα βασικά βήματα επεξεργασίας του βίντεο.

μοντέλου για κάθε μια από τις κλάσεις, ενώ τα δεύτερα για την αξιολόγηση της επίδοσης του συστήματος. Στη συνέχεια δίνουμε μια σύντομη περιγραφή του καθενός από τα επιμέρους βήματα που αναφέρθηκαν:

- **Εξαγωγή Χαρακτηριστικών** Η εξαγωγή χαρακτηριστικών αποτελεί το πρώτο βήμα και το μοναδικό που επεξεργάζεται τα τρισδιάστατα δεδομένα βίντεο στην αρχική τους μορφή. Συνήθως αποτελείται από δύο επιμέρους βήματα: την καθεαυτό εξαγωγή των χαρακτηριστικών και την περιγραφή τους μέσω κατάλληλων *περιγραφητών*. Τα χαρακτηριστικά είναι οι μετρήσεις ή παρατηρήσεις χαμηλού επιπέδου στην εικόνα, οι οποίες ιδανικά συμπυκνώνουν το σημασιολογικό περιεχόμενο που θέλουμε να εξάγουμε. Εισάγουν ένα επίπεδο αφαίρεσης, αφού απεικονίζουν το "ογκώδες" και πολύπλοκο σήμα του βίντεο σε ένα σχετικά μικρό σύνολο αριθμών. Ενδεικτικά παραδείγματα χαρακτηριστικών είναι τα χωροχρονικά σημεία ενδιαφέροντος 2.1 ή η σιλουέτα του ανθρώπου 2.2. Κατά κανόνα η εξαγωγή των χαρακτηριστικών ακολουθείται από την "περιγραφή" τους, συνήθως μέσω διαφόρων μετρήσεων στο σήμα του βίντεο. Οι μετρήσεις αυτές ονομάζονται "περιγραφητές", και υπολογίζονται συνήθως σε μια γειτονιά γύρω από τα χαρακτηριστικά. Χαρακτηριστικά παραδείγματα περιγραφητών είναι

οι HoG και HoF (βλ. Ενότητες 3.3.1, 3.3.2). Συνήθως στη βιβλιογραφία ο όρος "χαρακτηριστικά" αναφέρεται τόσο στα καθεαυτό χαρακτηριστά όσο και στους περιγραφητές. Στην παρούσα εργασία θα διαχωρίζουμε ρητά τις δύο έννοιες όταν απαιτείται. Αντίθετα, όπου είναι σαφές από τα συμφραζόμενα και δεν υπάρχει κίνδυνος σύγχυσης, θα χρησιμοποιούμε τον όρο χαρακτηριστικά χωρίς διάκριση.

- **Προεπεξεργασία δεδομένων** Συνήθως οι περιγραφητές που έχουν εξαχθεί από το προηγούμενο βήμα έχουν αρκετά μεγάλη διάσταση και παρουσιάζουν συσχέτιση μεταξύ τους. Πολλές φορές προκύπτει η ανάγκη μείωσης της διάστασής και αποσυσχέτισής τους. Αυτό γίνεται κατά κανόνα με *ανάλυση σε πρωτεύουσες συνιστώσες* (Principal Component Analysis, εν συντομία PCA).

- **Υπολογισμός του οπτικού λεξικού** Σκοπός του βήματος αυτού είναι η διαμέριση του χώρου των χαρακτηριστικών προκειμένου να βρεθούν ομάδες ή "συστάδες" χαρακτηριστικών που έχουν κοινά στοιχεία και αντιπροσωπεύουν παρεμφερείς δομές στο βίντεο. Αυτό γίνεται συνήθως μέσω του αλγορίθμου K-means, ή της εκπαίδευση ενός μίγματος γκαουσιανών κατανομών (Gauusian Mixture Model).

- **Κωδικοποίηση χαρακτηριστικών** Τα χαρακτηριστικά που εξήχθησαν από το πρώτο στάδιο δεν είναι άμεσα αξιοποιήσιμα για την ταξινόμηση του βίντεο εισόδου. Το βήμα της κωδικοποίησής τους υπολογίζει μια κατάλληλη και ενιαία αναπαράσταση των χαρακτηριστικών και κατά συνέπεια του βίντεο, η οποία μπορεί να χρησιμοποιηθεί από κάποιον ταξινομητή. Συνήθως η κωδικοποίηση συνίσταται στον υπολογισμό στατιστικών μεγεθών των χαρακτηριστικών με τη χρήση του οπτικού λεξικού, όπως π.χ. στην περίπτωση του Bag-of-Words (βλ. Ενότητα 3.4).

- **Ταξινόμηση** Στο βήμα αυτό χρησιμοποιείται ένας ταξινομητής για την κατηγοριοποίηση ενός βίντεο εισόδου σε μια από τις προκαθορισμένες κλάσεις, βάσει της αναπαράστασης που έχει προκύψει στο προηγούμενο βήμα. Στη φάση της εκπαίδευσης ο ταξινομητής χρησιμοποιεί τα κωδικοποιημένα βίντεο εκπαίδευσης από το προηγούμενο βήμα και την πληροφορία της κλάσης στην οποία ανήκει καθένα από αυτά, για τον υπολογισμό ενός μοντέλου για κάθε διαφορετική κατηγορία δράσης. Στη φάση της αξιολόγησης τα μοντέλα αυτά χρησιμοποιούνται για να κατατάξουν αυτόματα

Σχήμα 1.2: Ενδεικτικά καρέ από τη βάση KTH. Από αριστερά προς τα δεξιά, αυτά αντιστοιχούν στις κλάσεις "walking", "jogging", "running", "boxing", "hand waving" και "hand clapping".

τα "άγνωστα" βίντεο αξιολόγησης (τα οποία ο ταξινομητής δεν έχει "συναντήσει" πιο πριν) σε μια κατηγορία. Συνήθως επιλέγονται "διαχωριστικοί" (discriminative) ταξινομητές όπως οι Μηχανές Διανυσμάτων Υποστήριξης [5] (Support Vector Machines, εν συντομία SVM, βλ. Ενότητα 3.5) έναντι "αναγεννητικών", όπως τα Κρυφά Μαρκοβιανά Μοντέλα (Hidden Markov Models, εν συντομία HMMs) (Βλ. Ενότητα 2.3).

1.3 Βάσεις δεδομένων ανθρώπινων δράσεων και χειρονομιών

Όπως αναφέραμε παραπάνω, η εξέλιξη στο πεδίο της αυτόματης αναγνώρισης δράσεων είναι άρρηκτα συνδεδεμένη με τις βάσεις δεδομένων που χρησιμοποιούνται και αποτελούν κοινό μέτρο σύγκρισης για τις περισσότερες εργασίες. Στην παρούσα διπλωματική εργασία πειραματιζόμαστε σε μια σειρά βάσεων δεδομένων που καλύπτουν ένα ευρύ φάσμα δυσκολίας. Αυτές περιγράφονται αναλυτικά στη συνέχεια.

1.3.1 Η βάση ανθρώπινων δράσεων KTH

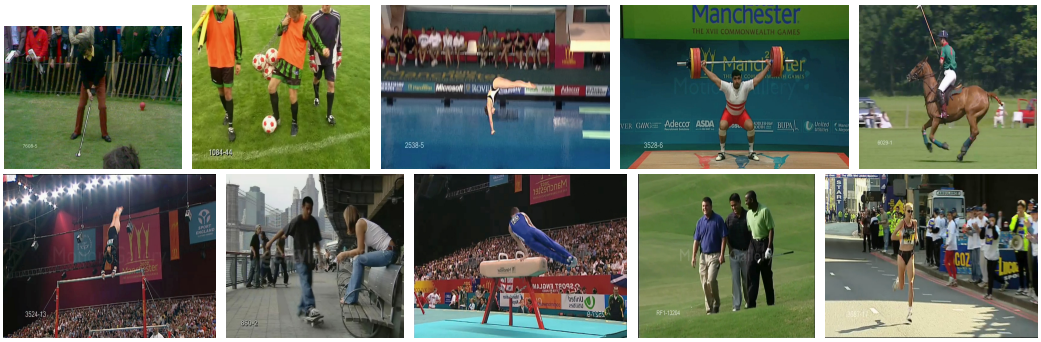
Η KTH [2] αποτελεί μια από τις πρώτες βάσεις δεδομένων ανθρώπινων δράσεων, η οποία παραμένει μέχρι σήμερα αρκετά δημοφιλής. Τα βίντεο που περιλαμβάνει έχουν μαγνητοσκοπηθεί χειροκίνητα και σε ελεγχόμενο περιβάλλον γι' αυτό το σκοπό, οπότε η εκτέλεση των δράσεων είναι "καθαρή", χωρίς επικαλύψεις ή ιδιαίτερες αποκλίσεις μεταξύ των διαφορετικών εκτελέσεων. Γι' αυτό παραμένει μέχρι σήμερα δημοφιλής, μιας και καθιστά ευκολότερη την ανάλυση και την εις βάθος επισκόπηση των διαφόρων μεθόδων.

Συνολικά περιλαμβάνει 2391 βίντεο, καθένα από τα οποία ανήκει σε μια από τις εξής κατηγορίες: *walking*, *jogging*, *running*, *boxing*, *hand waving* και *hand clapping*. Ενδεικτικά καρέ από κάθε κατηγορία φαίνονται στο

Σχήμα 1.2. Κάθε δράση εκτελείται περίπου 4 φορές από 25 διαφορετικά άτομα και υπό 4 διαφορετικές συνθήκες: σε εξωτερικό χώρο, σε εξωτερικό χώρο με μεταβλητή κλίμακα (ζουμ κατά τη διάρκεια του βίντεο), σε εξωτερικό χώρο με διαφορετικά ρούχα και σε εσωτερικό χώρο. Η λήψη όλων των βίντεο έγινε στατική κάμερα και με ρυθμό 25 καρέ ανά δευτερόλεπτο (fps). Στη συνέχεια, τα βίντεο υποδειγματοληπτήθηκαν ώστε να έχουν σταθερή ανάλυση 160×120. Τα δεδομένα χωρίζονται σε σύνολο εκπαίδευσης και σύνολο αξιολόγησης, όπως προτάθηκε από τους δημιουργούς της βάσης. Το δεύτερο περιλαμβάνει τα βίντεο των δειγμάτων (ατόμων) 2,3,5,6,7,8,9,10 και 22, ενώ το πρώτο αποτελείται από τα βίντεο όλων των υπολοίπων δειγμάτων. Αξιολογούμε τις διάφορες μεθόδους σε αυτή τη βάση υπολογίζοντας το ποσοστό σωστής ταξινόμησης, ή αλλιώς *ακρίβεια* (accuracy), δηλαδή το λόγο των σωστά προς τα λάθος ταξινομημένα βίντεο,

1.3.2 Η βάση ανθρωπίνων δράσεων UCF Sports

Η βάση UCF Sports [6] αποτελεί ένα σύνολο δράσεων που έχουν συλλεχθεί από αθλητικούς αγώνες, οι οποίοι έχουν μεταδοθεί από μεγάλα τηλεοπτικά δίκτυα (όπως π.χ. το BBC). Περιλαμβάνει συνολικά 150 βίντεο ανάλυσης 720×480, ρυθμού δειγματοληψίας 25 fps και μέσης διάρκειας $6,39$ sec. Η δράση που απεικονίζει κάθε βίντεο ανήκει σε μία από τις εξής 10 κατηγορίες: *Diving*, *Golf Swing*, *Kicking*, *Lifting*, *Riding Horse*, *Running*, *SkateBoarding*, *Swing-Bench*, *Swing-Side*, *Walking*.



Σχήμα 1.3: Ενδεικτικά καρέ από τη βάση UCF Sports. Από αριστερά προς τα δεξιά, πρώτη σειρά: "Golf Swing", "Kicking", "Diving", "Lifting", "Riding Horse", δεύτερη σειρά: "Swing-Side", "SkateBoarding", "Swing-Bench", "Walking", "Running".

Παρ' όλο που τα βίντεο είναι μαγνητοσκοπημένα σε σχετικά ελεγχόμενο περιβάλλον, η UCF Sports παρουσιάζει αρκετές δυσκολίες όπως

η κίνηση της κάμερας, κάποιες επικαλύψεις, διαφορετικές οπτικές γωνίες λήψης και μεγάλη μεταβλητότητα στην εκτέλεση των δράσεων. Οι δημιουργοί της βάσης πρότειναν την αξιολόγηση πάνω σε αυτή τη βάση με το σχήμα *leave-one-out*, κατά το οποίο ένα βίντεο χρησιμοποιείται για αξιολόγηση και όλα τα υπόλοιπα για εκπαίδευση, το οποίο επαναλαμβάνεται για κάθε βίντεο. Ωστόσο, οι Lan et al. [7] παρατήρησαν πως πολλές δράσεις λάμβαναν χώρα υπό το ίδιο ακριβώς σκηνικό (π.χ. στο ίδιο γήπεδο). Γι' αυτό πρότειναν ένα συγκεκριμένο διαχωρισμό σε σύνολο εκπαίδευσης και αξιολόγησης, τον οποίο ακολουθούμε κι εμείς στα πειράματα αυτής της βάσης. Ομοίως με την KTH, στα πειράματά μας σε αυτή τη βάση υπολογίζουμε την ακρίβεια ταξινόμησης.

1.3.3 Η βάση ανθρωπίνων δράσεων Hollywood2

Η βάση ανθρωπίνων δράσεων Hollywood2 [8] προέρχεται από 69 διαφορετικές ταινίες του Hollywood. Αποτελείται από 1707 κλιπ, τα οποία περιέχουν τις εξής κατηγορίες δράσεων: "AnswerPhone", "DriveCar", "Eat", "FightPerson", "GetOutCar", "HandShake", "HugPerson", "Kiss", "Run", "SitDown", "SitUp", "StandUp". Στο Σχήμα 1.4 απεικονίζονται κάποια αντιπροσωπευτικά καρέ.

Τα βίντεο έχουν εξαχθεί από τους δημιουργούς της βάσης με αυτόματο τρόπο, αντιστοιχίζοντας τον ήχο της ταινίας με το κείμενο του σεναρίου. Λόγω της προέλευσης τους, οι δράσεις παρουσιάζουν πολύ μεγάλη μεταβλητότητα ως προς την εκτέλεση, το φόντο, τη γωνία λήψης και το φωτισμό, καθώς και έντονη ή/και μη-ομαλή κίνηση της κάμερας και μεγάλες επικαλύψεις. Επίσης, λόγω της μεθόδου συλλογής τους σε πολλές περιπτώσεις η διάρκεια κάθε κλιπ είναι αρκετά μεγαλύτερη από τη διάρκεια της δράσης που απεικονίζεται, ή μπορεί κάποια δράση να συμβαίνει αλλά να μην απεικονίζεται, αλλά να υπονοείται (π.χ. ένα αυτοκίνητο καταφθάνει και μετά από λίγο εμφανίζεται ο ηθοποιός, οπότε το βίντεο ανήκει στην κατηγορία "GetOutCar").

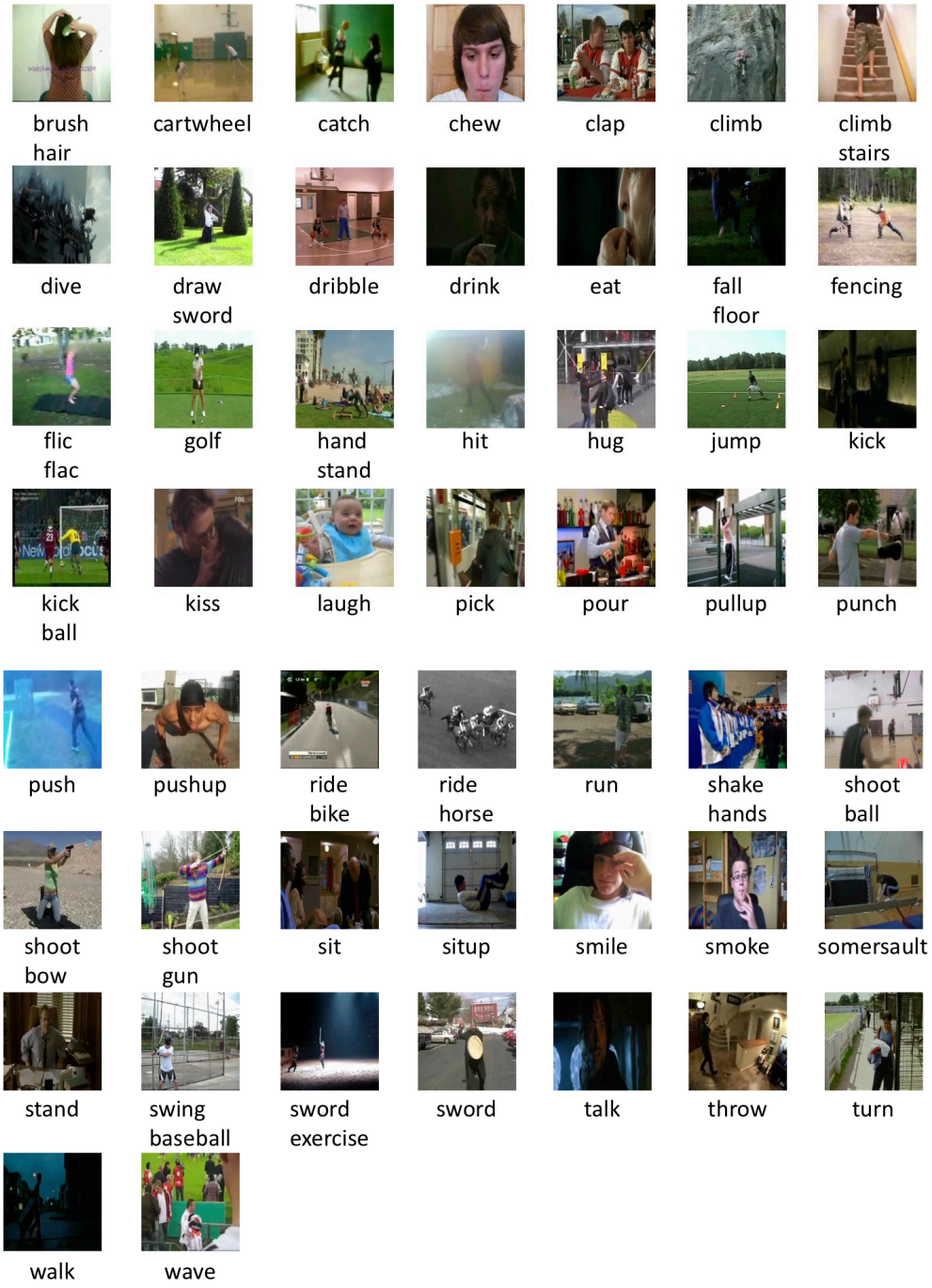
Για τον πειραματισμό σε αυτή τη βάση ακολουθούμε το διαχωρισμό σε σύνολα εκπαίδευσης και αξιολόγησης που προτείνουν οι συγγραφείς [8] (823 και 884 βίντεο αντίστοιχα), με το πρώτο να προέρχεται από διαφορετικές ταινίες απ' ότι το δεύτερο. Λόγω του ότι ορισμένα βίντεο περιέχουν παραπάνω από μια δράσεις από τις παραπάνω κατηγορίες, η αξιολόγηση γίντει υπολογίζοντας τη μέση ακρίβεια (Average Precision - AP) για κάθε κατηγορία και στη συνέχεια βρίσκοντας το μέσο όρο των AP όλων των κατηγοριών (mean average precision - mAP).

Σχήμα 1.4: Ενδεικτικά καρέ από τη βάση Hollywood2. Από αριστερά προς τα δεξιά, πρώτη σειρά: *GetOutCar*, *Run*, *HugPerson*, *SitUp*, δευτερη σειρά: *DriveCar*, *Eat*, *FightPerson*, *StandUp*, τρίτη σειρά: *AnswerPhone*, *HandShake*, *SitDown*, *Kiss*.

1.3.4 Η βάση ανθρωπίνων δράσεων HMDB51

Η HMDB51 [9] είναι μια από τις νεότερες και πιο απαιτητικές βάσεις δεδομένων ανθρώπινων δράσεων με ευρεία χρήση στη βιβλιογραφία. Τα βίντεο που την απαρτίζουν προέρχονται από διαδικτυακές υπηρεσίες (Prelinge, YouTube, Google) και έχουν κλιμακωθεί κατάλληλα έτσι ώστε να έχουν ύψος 240 pixels και εύρος κατάλληλο έτσι ώστε να διατηρείται η αναλογία διαστάσεων (aspect ratio) του κάθε βίντεο. Ο ρυθμός δειγματοληψίας είναι σταθερός για κάθε βίντεο και ίσος με 30 καρέ ανά δευτερόλεπτο. Συνολικά περιέχει 6766 βίντεο, καθένα από τα οποία ανήκει σε μία από τις 51 διαφορετικές κλάσεις δράσεων. Κάθε κλάση αντιστοιχεί περιέχει τουλάχιστον 101 βίντεο. Οι κλάσεις αυτές, οι οποίες απεικονίζονται στο Σχήμα 1.5 προέρχονται από 6 ευρύτερες κατηγορίες:

- Δράσεις που σχετίζονται με εκφράσεις του προσώπου: *smile*, *laugh*, *chew*, *talk*.
- Δράσεις που σχετίζονται με το προσώπου και περιλαμβάνουν αλληλεπίδραση με αντικείμενα: *smoke*, *eat*, *drink*.



Σχήμα 1.5: Ενδεικτικά καρέ από τις 51 κλάσεις της HMDB51.

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

- Γενικές κινήσεις του σώματος που περιλαμβάνουν αλληλεπίδραση με αντικείμενα: *brush hair*,*catch*, *draw sword*, *dribble*, *golf*, *hit something*, *kick ball*, *pick*, *pour*, *push something*, *ride bike*, *ride horse*, *shoot ball*, *shoot bow*, *shoot gun*, *swing baseball bat*, *sword exercise*, *throw*.

- Γενικές κινήσεις του σώματος που περιλαμβάνουν αλληλεπίδραση με ανθρώπους: *fencing*, *hug*, *kick someone*, *kiss*, *punch*, *shake hands*, *sword fight*.

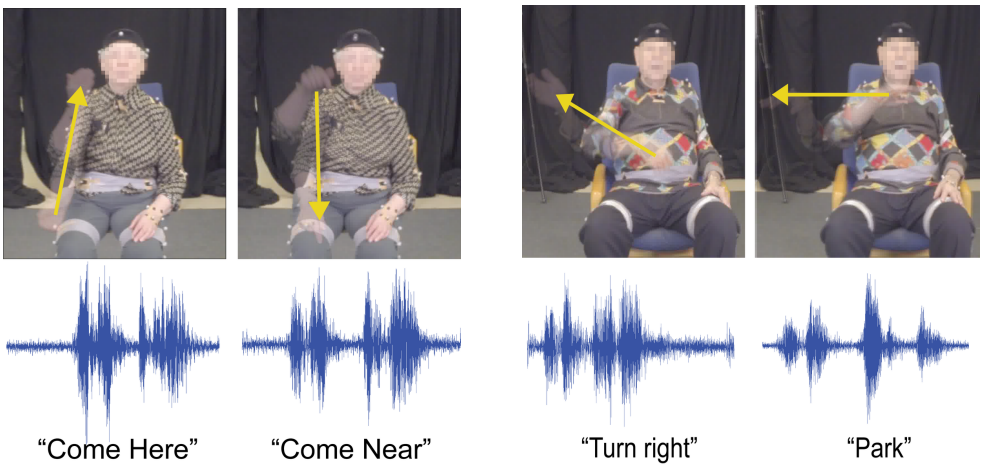
Η HMDB51 είναι ιδιαίτερα απαιτητική, αφ' ενός λόγω του μεγάλου αριθμού των κλάσεων που περιλαμβάνει και εφ' ετέρου λόγω του μη ελεγχόμενων συνθηκών στις οποίες έχουν ληφθεί τα βίντεο και των δυσκολιών που απορρέουν από αυτό. Συγκεκριμένα, οι επικαλύψεις, η μη ομαλή κίνηση της κάμερας και οι διαφορετικές κλίμακες κ.τ.λ. αποτελούν συχνά φαινόμενα.

Για την αξιολόγηση ακολουθούμε το πειραματικό πλαίσιο που προτείνουν οι συγγραφείς και χρησιμοποιούμε τρία "splits", δηλαδή τρεις διαφορετικές διαμερίσεις των δεδομένων σε σύνολα training και testing. Κάθε τέτοιο split περιέχει 70 βίντεο εκπαίδευσης και 30 βίντεο αξιολόγησης από κάθε κλάση. Η ακρίβεια προκύπτει ως ο μέσος όρος της ακρίβειας ταξινόμησης που υπολογίζεται για κάθε split.

1.3.5 Η βάση χειρονομιών MOBOT-6.a

Η βάση χειρονομιών MOBOT-6.a προέρχεται από ένα ευρύτερο σύνολο δεδομένων (βάση MOBOT) που συλλέχθηκαν στα πλαίσια του ερευνητικού προγράμματος MOBOT [1]. Το πρόγραμμα αυτό αποσκοπεί στη δημιουργία μιας ρομποτικής πλατφόρμας που θα υποβοηθά ηλικιωμένους με κινητικές ή/και διανοητικές δυσκολίες προκειμένου να ανταπεξέρχονται στις στις καθημερινές τους δραστηριότητες [10]. Η ρομποτική πλατφόρμα περιλαμβάνει μια σειρά από αισθητήρες που βοηθούν τον εντοπισμό και την επικοινωνία με τον ασθενή, την κίνηση της πλατφόρμας, κ.α. Η βάση MOBOT είναι *πολυτροπική* (multimodal), δηλαδή αποτελείται από δεδομένα πολλών διαφορετικών αισθητήρων που προσφέρουν συμπληρωματικές *τροπικότητες* (modalities), όπως ήχος, βίντεο, χάρτες βάθους, δεδομένα λέιζερ, κ.α. Στο Σχήμα 1.8 φαίνεται

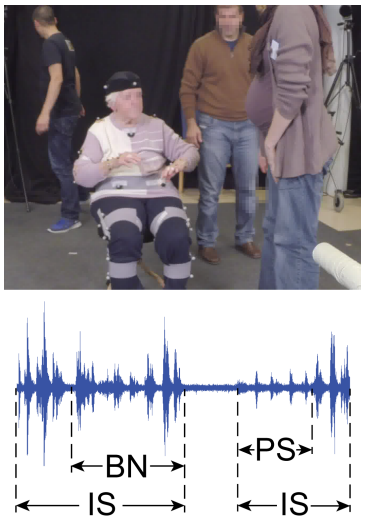
[1] `http://www.mobot-project.eu`



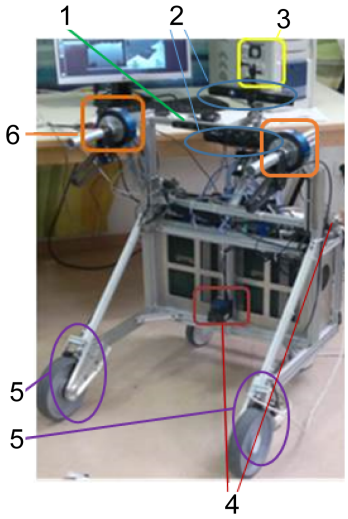
Σχήμα 1.6: Δείγματα εντολών από το λεξικό της βάσης MOBOT-6.a. Αυτές περιλαμβάνουν τόσο φωνητικές εντολές, όσο και χειρονομίες που υποβοηθούν την επικοινωνία του ασθενή με τη ρομποτική πλατφόρμα.

το πρωτότυπο δείγμα της ρομποτικής πλατφόρμας. Λόγω της φύσης των δεδομένων, η βάση MOBOT είναι ιδιαίτερα απαιτητική και αντικατοπτρίζει σε μεγάλο βαθμό τις ρεαλιστικές δυσκολίες που καλούνται να αντιμετωπίσουν τα επιμέρους συστήματα που αναπτύσσονται στο πλαίσιο αυτό (Σχήμα 1.7). Το σενάριο 6.a της πολυτροπικής βάσης δεδομένων MOBOT περιλαμβάνει 19 διαφορετικές χειρονομίες που έχουν επιλεχθεί προκειμένου να διευκολύνουν την επικοινωνία των ασθενών με τη ρομποτική πλατφόρμα (Σχήμα 1.6). Κάθε μία από αυτές συνοδεύεται από μια φωνητική εντολή και αντιστοιχεί σε μια συγκεκριμένη εντολή προς την πλατφόρμα. Στη συνέχεια παραθέτουμε τις διαφορετικές εντολές. Το υποσύνολο αυτών που χρησιμοποιούμε στα πειράματά μας φαίνεται στο Σχήμα 1.9, όπου παρατίθενται κάποια αντιπροσωπευτικά καρέ. Στα πλαίσια της παρούσας διπλωματικής αναφερόμαστε στη βάση MOBOT-6.a εννοώντας μόνο τα έγχρωμα δεδομένα βίντεο που περιλαμβάνει το σενάριο 6.a και προέρχονται από τον αισθητήρα Kinect.

1. “Help”
2. “I want to stand up”
3. “I want to perform a task”
4. “I want to sit down”
5. “Come here”
6. “Come closer”
7. “Go straight”
8. “Park”
9. “Stop”
10. “Go away”
11. “Let’s go”
12. “Turn left”
13. “Turn Right”



Σχήμα 1.7: Δυσκολίες που παρουσιάζει η βάση MOBOT (οπτικοακουστικός θόρυβος, επικαλύψεις, κ.α). IS: ομιλία νοσοκόμου, PS: ομιλία ασθενούς, BN: ακουστικός θόρυβος.



Σχήμα 1.8: Ρομποτική πλατφόρμα: 1. Συστοιχία μικροφώνων MEMS, 2. Αισθητήρας Kinect, 3. Κάμερα GoPro, 4. Αισθητήρε laser, 5. Κωδικοποιητές, 6. Αισθητήρες ροπής/δύναμης.

14. "Avoid obstacle"
15. "Go through the door"
16. "Yes"
17. "No"
18. "Where am I?"
19. "What time is it?"

Η λήψη των δεδομένων έχει γίνει μέσω του αισθητήρα Kinect με ανάλυση 640×480 και με μέσο ρυθμό δειγματοληψίας περίπου 13 καρέ το δευτερόλεπτο (fps), το οποίο παρουσιάζει διακυμάνσεις. Κατά τη διάρκεια της μαγνητοσκόπησης ο ασθενής παραμένει καθιστός σε απόσταση περίπου 2 μέτρων από τη ρομποτική πλατφόρμα, η οποία παραμένει ακίνητη. Το φόντο δε μεταβάλλεται ιδιαίτερα, με εξαίρεση ορισμένες φιγούρες ανθρώπων που κινούνται στη σκηνή. Παρ' όλο που το περιβάλλον στο οποίο έχουν ληφθεί τα δεδομένα είναι σχετικά ελεγχόμενο και δεν εισάγει μεγάλα εμπόδια στην αναγνώριση των χειρονομιών, η συγκεκριμένη βάση παρουσιάζει αρκετές προκλήσεις. Κατ' αρχάς είναι σύνηθες στις βάσεις χειρονομιών να παρατηρείται μεταβλητότητα στην εκτέλεση της ίδιας χειρονομίας μεταξύ διαφορετικών χρηστών. Ωστόσο, στην περίπτωση της βάσης MOBOT-6.a, τα κινητικά προβλήματα ορισμένων ασθενών δυσχεραίνουν τη σωστή εκτέλεση των χειρονομιών. Έτσι, οι διαφορετικές και αποκλίνουσες εκτελέσεις αποτε-

(a) “Help” (b) “StandUp” (c) “PerformTask” (d) “SitDown”

(e) “ComeCloser” (f) “ComeHere” (g) “Stop” (h) “GoStraight”

Σχήμα 1.9: Ενδεικτικά καρέ από τις εντολές του λεξικού της βάσης MOBOT-6.a.

λούν τον κανόνα. Επίσης, ορισμένοι ασθενείς παρουσιάζουν διανοητικές δυσκολίες, με αποτέλεσμα οι διαδοχικές εκτελέσεις της ίδιας χειρονομίας από τον ίδιο ασθενή να διαφέρουν σημαντικά. Για τους παραπάνω λόγους θεωρούμε πως η βάση MOBOT-6.a περιλαμβάνει ένα αντιπροσωπευτικό δείγμα των πραγματικών προκλήσεων που θα καλούταν να αντιμετωπίσει ένα σύστημα αναγνώρισης χειρονομιών υπό ρεαλιστικές συνθήκες χρήσης από ηλικιωμένα άτομα.

1.4 Διάρθρωση της διπλωματικής εργασίας

Στην παρούσα διπλωματική εργασία μελετάμε το πρόβλημα της ταξινόμησης και αναγνώρισης ανθρώπινων δράσεων και χειρονομιών, αντιμετωπίζοντάς τα ως ένα ενιαίο πρόβλημα. Αναλύουμε και συγκρίνουμε τις διαφορετικές προσεγγίσεις για τα κυριότερα στάδια της επεξεργασίας, δηλαδή την εξαγωγή και κωδικοποίηση χαρακτηριστικών. Πιο συγκεκριμένα, η παρούσα εργασία είναι οργανωμένη σε κεφάλαια ως εξής:

Στο Κεφάλαιο 2 συνοψίζουμε τη σχετική έρευνα που έχει γίνει μέχρι σήμερα στον τομέα αναφέροντας περιληπτικά κάποιες αντιπροσωπευτικές μεθόδους που έχουν εμφανιστεί σε δημοσιευμένες εργασίες. Πραγματευόμαστε κυρίως τις *τοπικές* (local) τεχνικές, μιας και έχουν τη μεγαλύτερη συμβολή στην ανάπτυξη του κλάδου.

Στο Κεφάλαιο 3 αναλύουμε διεξοδικά τις μεθόδους που χρησιμοποιήσαμε και με της οποίες πειραματιστήκαμε στα πλαίσια της παρούσας

εργασίας. Στις Ενότητες 3.1 και 3.2 περιγράφουμε δύο ευρέως χρησιμοποιούμενες τεχνικές εξαγωγής χαρακτηριστικών: τα χωρο-χρονικά σημεία ενδιαφέροντος και τις πυκνές τροχιές. Στην Ενότητα 3.4, μελετάμε τις διαφορετικές μεθόδους αναπαράστασης βίντεο: το μοντέλο συνόλου οπτικών λέξεων (Bag-of-Words), τις χωρο-χρονικές πυραμίδες (spatio-temporal pyramids), το VLAD και το διάνυσμα Fisher. Η Ενότητα 3.5 πραγματεύεται την ταξινόμηση των βίντεο με Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines).

Το Κεφάλαιο 4 παρουσιάζει και αναλύει τα πειραματικά αποτελέσματα των παραπάνω μεθόδων. Τα πειράματά μας πραγματοποιούνται σε μια πληθώρα βάσεων δεδομένων ανθρώπινων δράσεων, καθώς και στη βάση χειρονομιών MOBOT-6a.

Το Κεφάλαιο 5) περιγράφει τα βασικά βήματα που ακολουθήσαμε στην κατεύθυνση υλοποίησης ενός on-line συστήματος αναγνώρισης χειρονομιών. Η ανάπτυξη του συστήματος αυτού έγινε στα πλαίσια του ερευνητικού προγράμματος MOBOT με σκοπό τη δημιουργία ενός διαδραστικού περιβάλλοντος επικοινωνίας ανθρώπου-ρομπότ.

Τέλος, στο Κεφάλαιο 6 συνοψίζουμε τη συνεισφορά της παρούσας διπλωματικής, την τρέχουσα εργασία που γίνεται ως προέκτασή της και προτείνουμε κατευθύνσεις για μελλοντική έρευνα.

Κεφάλαιο 2

Σχετικές Εργασίες

Όπως αναφέρθηκε νωρίτερα, η βιβλιογραφία σχετικά με την αναγνώριση ανθρώπινων δράσεων είναι εκτενέστατη και συνεχίζει να μεγεθύνεται με ταχείς ρυθμούς. Η σχετική έρευνα τα τελευταία χρόνια έχει επηρεαστεί σε μεγάλο βαθμό από την αντίστοιχη έρευνα στο πεδίο της ταξινόμησης εικόνων. Καταλυτικό ρόλο είχε η εισαγωγή των τοπικών (local) αναπαραστάσεων [11], οι οποίες περιγράφουν ένα βίντεο ως μια συλλογή "εξεχόντων" (salient) σημείων, ή πιο πρόσφατα, τροχιών [12]. Τελευταία, επίσης, ακολουθώντας και τις γενικές τάσεις στην Όραση Υπολογιστών, έχουν αρχίσει να υιοθετούνται βαθιές αρχιτεκτονικές νευρωνικών δικτύων [13].

Έχουν προκύψει αρκετές κατηγοριοποιήσεις των διαφόρων δημοσιευμένων εργασιών και τεχνικών που έχουν αναπτυχθεί. Μια συνηθισμένη [14], την οποία ακολουθούμε στην ενότητα αυτή, διαχωρίζει τις διάφορες μεθόδους σε *τοπικές* (local), *ολικές* (global) και παραλλαγές τους, βάσει του τρόπου που μοντελοποιείται η χαμηλού επιπέδου πληροφορία της εικόνας. Θα επικεντρωθούμε στις local προσεγγίσεις, μιας και είναι ευρέως χρησιμοποιούμενες, έχουν επηρεάσει σημαντικά την έρευνα στο πεδίο αυτό, όπως και σε συγγενή πεδία, κι εν γένει παρουσιάζουν τις καλύτερες επιδόσεις. Φυσικά είναι αδύνατον να παρουσιάσουμε μεγάλο ποσοστό των δημοσιευμένων εργασιών, γι' αυτό στην παρούσα ενότητα προσπαθούμε να σταχυολογήσουμε ορισμένες από αυτές, τις οποίες θεωρούμε αντιπροσωπευτικές.

2.1 Τοπικές (local) προσεγγίσεις

Οι τοπικές (local) μέθοδοι περιγράφουν ένα βίντεο ως μια συλλογή χωρο-χρονικών σημείων ή περιοχών ενδιαφέροντος, τα οποία είναι "εξέ-

χουσης σημασίας" (salient) για μια συγκεκριμένη δράση και μπορούν να τη χαρακτηρίσουν. Τα σημεία αυτά αναφέρονται στη βιβλιογραφία ως STIP[1] (Spatio-Temporal Interest Points) και οι τεχνικές που εξάγουν αυτού του τύπου τα χαρακτηριστικά είχαν κυριαρχήσει και συνεχίζουν να χρησιμοποιούνται κατά κόρον στο συγκεκριμένο πεδίο.

Η πρώτη σχετική εργασία δημοσιεύθηκε από τους Laptev και Lindeberg [11], οι οποίοι, εμπνευσμένοι από αντίστοιχες δουλειές στο πεδίο της αναγνώρισης αντικειμένων σε εικόνες [15], επέκτειναν τον ανιχνευτή γωνιών Harris στο πεδίο του χρόνου (Harris3D). Εντοπίζουν, έτσι, σημεία που η εικόνα αλλάζει απότομα ως προς τις τρεις κατευθύνσεις που ορίζει ο όγκος του βίντεο. Για την περιγραφή ενός σημείου ενδιαφέροντος, χρησιμοποίησαν local jets (δηλαδή παραγώγους μεγάλης τάξης) σε μια μικρή χωροχρονική γειτονιά του. Πρότειναν, ταυτόχρονα και την αυτόματη επιλογή κλίμακας, την οποία ωστόσο εγκατέλειψαν αργότερα [16], χρησιμοποιώντας αντ' αυτού προκαθορισμένες κλίμακες για την ανίχνευση των STIP. Στην ίδια εργασία χρησιμοποίησαν την τεχνική Bag-of-Words σε συνδυασμό με μη-γραμμικά SVM για την κωδικοποίηση των STIP και την ταξινόμηση των βίντεο, μέθοδοι που μέχρι σήμερα χρησιμοποιούνται κατά κόρον στη βιβλιογραφία για την αναγνώριση δράσεων. Αργότερα, οι Laptev et al. [2] χρησιμοποιούν τον περιγραφητή HoG για να περιγράψουν τις γειτονιές γύρω από κάθε σημείο ενδιαφέροντος. Ο HoG βασίζεται σε τοπικά ιστογράμματα της κλίσης της εικόνας και σημείωνε εξαιρετικά αποτελέσματα στο πρόβλημα της αναγνώρισης αντικειμένων σε εικόνες. Στην ίδια εργασία εισάγουν και τον περιγραφητή HoF, ο οποίος, όμοια με τον HoG, βασίζεται σε τοπικά ιστογράμματα της οπτικής ροής[2]. Την επέκταση του HoG στον τρισδιάστατο χώρο, ονόματι HoG3D, ανέπτυξαν λίγο αργότερα οι Klaser et al. [17].

Παράλληλα, οι Dollar et al. [18] πρότειναν τη χρήση φίλτρων Gabor για την ανίχνευση σημείων ενδιαφέροντος. Η ανίχνευση γίνεται φιλτράροντας το σήμα του βίντεο στο χώρο και στο χρόνο ανεξάρτητα, με ένα ζευγάρι (quadrature pair) μονοδιάστατων φίλτρων και μεταβάλλοντας τον αριθμό των ανιχνευθέντων σημείων ενδιαφέροντος αλλάζοντας την κλίμακα στην οποία τα σημεία ενδιαφέροντος αναζητούνται. Ο συγκεκριμένος ανιχνευτής δίνει έμφαση σε περιοδικές δομές που εμφανίζονται σε βίντεο, αγνοώντας π.χ. μια απλή κίνηση με σταθερή ταχύτητα. Για την περιγραφή των σημείων ενδιαφέροντος εισήγαγαν τα Cuboids, τα οποία αποτελούν προκαθορισμένες τρισδιάστατες περιοχές γύρω από κάθε σημείο, εντός των οποίων υπολογίζονται περιγραφητές όμοιοι με τον SIFT [19], και συγκεκριμένα ιστογράμματα των τιμών έντασης των pixels, της παράγωγου και της οπτική ροής. Σημειώνουμε πως ο δη-

μοφιλής περιγραφητής SIFT επεκτάθηκε στον τρισδιάστατο χώρο από τους Scovanner et al. [20].

Οι Willems et al. [21] επεκτείνουν το μετρικό σημαντικότητας (saliency measure) Hessian που χρησιμοποιείται για τον εντοπισμό δομών "σταγόνας" (blobs) σε εικόνες. Υπολογίζουν, λοιπόν, την ορίζουσα του τρισδιάστατου πίνακα Hessian ενός βίντεο, την οποία χρησιμοποιούν για την εύρεση σημείων ενδιαφέροντος. Στην ίδια εργασία, εισάγουν τον περιγραφητή eSURF, ο οποίος αποτελεί επέκταση του περιγραφητή SURF [22] στο πεδίο πεδίο του χρόνου και βασίζεται στον υπολογισμό κυματιδίων Haar εντός μιας ορθογώνιας περιοχής γύρω από κάθε σημείο ενδιαφέροντος.

Οι Maninis et al. [23], εξελίσσοντας τις ιδέες των Georgakis et al. [24], φιλτράρουν το σήμα του βίντεο με τρισδιάστατα φίλτρα Gabor σε πολλαπλές συχνοτικές ζώνες και χρησιμοποιούν τον τελεστή ενέργειας Teager-Kaiser για να δημιουργήσουν ένα ενεργειακό χάρτη "σημαντικότητας". Τα μέγιστα αυτού αντιστοιχούν σε σημεία ενδιαφέροντος, γύρω από τα οποία υπολογίζονται διαφορετικοί περιγραφητές.

Σε μια πιο πρόσφατη εργασία, οι Wang et al. [25] συγκρίνουν διάφορους ανιχνευτές σημείων ενδιαφέροντος σε μια σειρά από βάσεις δεδομένων και καταλήγουν στο συμπέρασμα πως μια απλή *πυκνή δειγματοληψία* (dense sampling) σημείων ενδιαφέροντος σε ένα σταθερό χωροχρονικό πλέγμα στο βίντεο έχει σημαντικά καλύτερη απόδοση από την υπολογιστικά "ακριβή" αναζήτηση συγκεκριμένων σημείων.

Ένα βασικό μειονέκτημα των χωρο-χρονικών σημείων ενδιαφέροντος είναι πως εν γένει αντιμετωπίζουν το χώρο και χρόνο με ενιαίο τρόπο. Εξελίσσοντας την ιδέα των STIP, πολλές εργασίες ενσωματώνουν πληροφορία σχετικά με την κίνηση στο βίντεο, παρακολουθώντας τα σημεία ενδιαφέροντος στο χρόνο και χρησιμοποιώντας τις *τροχιές* τους για να περιγράψουν μια δράση. Πρώτοι οι Messing et al. [26] εξάγουν τροχιές από βίντεο εντοπίζοντας σημεία ενδιαφέροντος με τον ανιχνευτή Harris και παρακολουθώντας τα στο χρόνο με τον Kanade-Lucas-Tomashi (KLT) tracker. Στη συνέχεια χρησιμοποιούν τις σχετικές ταχύτητες των σημείων που έχουν παρακολουθηθεί ως χαρακτηριστικά. Οι Sun et al. [27] παρακολουθούν σημεία ενδιαφέροντος υπολογίζοντας την ομοιότητα μεταξύ διαδοχικών καρέ μέσω περιγραφητών SIFT (SIFT matching) και χρησιμοποιούν στατιστικά μεγέθη των εξαχθέντων τροχιών για την αναπαράσταση ενός βίντεο. Οι Sun et al. [28] συνδυάζουν τις δύο παραπάνω μεθόδους και εξάγουν τροχιές τροχιών μεγάλης διάρκειας. Χρησιμοποιούν, επίσης το μέτρο της τοπικής κλίσης και τη μεταβλητότητα των τιμών φωτεινότητας της εικόνας ως κριτήριο σημαντικότητας (saliency) και βάσει αυτού ρυθμίζουν την πυκνότητα των

σημείων ενδιαφέροντος που παρακολουθούνται. Συνδυάζοντας τις ιδέες της πυκνής δειγματοληψίας και της χρήσης τροχιών για την αναπαράσταση των δράσεων, οι Wang et al. [12] εξάγουν *πυκνές τροχιές* (dense trajectories). Αυτές υπολογίζονται πραγματοποιώντας πυκνή δειγματοληψία στα καρέ του βίντεο και παρακολουθώντας τα σημεία που προκύπτουν με χρήση ενός πυκνού πεδίου οπτικής ροής. Στην ίδια εργασία χρησιμοποιούν μεταξύ άλλων τον περιγραφητή MBH λόγω της ευρωστίας του στην κίνηση της κάμερας. Η μέθοδος των πυκνών τροχιών, η οποία χρησιμοποιείται στην παρούσα διπλωματική εργασία, αναλύεται με λεπτομέρεια στην ενότητα 3.2. Λίγο αργότερα, οι Wang et al. αναπτύσσουν μια βελτιωμένη εκδοχή της εν λόγω μεθόδου, ονόματι *βελτιωμένες τροχιές* [29] (impoved trajectories), οι οποίες βασίζονται στην αντιστάθμιση της κίνησης της κάμερας για τον υπολογισμό της οπτικής ροής.

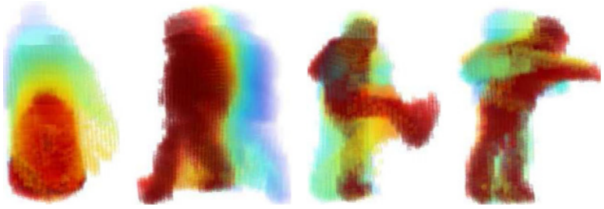
2.2 Ολικές (global) και Part-based προσεγγίσεις

Οι *ολικές* (global) προσεγγίσεις περιγράφουν μια δράση χρησιμοποιώντας ολόκληρο το βίντεο, ή μια ολόκληρη περιοχή ενδιαφέροντος εντός αυτού. Συνήθως εντοπίζεται η σιλουέτα του ανθρώπου που κινείται στο βίντεο και χρησιμοποιείται η περιοχή που περικλείεται από από αυτή για τον υπολογισμό περιγραφητών. Συχνά αυτές οι μέθοδοι συνδυάζονται με απλούς ταξινομητές, όπως SVM ή Nearest Neighbour.

Σε μια από τις παλιότερες σχετικές εργασίες, οι Darel et al. [30] υπολογίζουν την ομοιότητα μεταξύ βίντεο χειρονομιών απ' ευθείας μέσω της συσχέτισης ανάμεσα στα καρέ τους, χωρίς εξαγωγή χαρακτηριστικών. Μια άλλη χαρακτηριστική δουλειά είναι αυτή των Bobik και Davis [31], οι οποίοι απομονώνουν τη σιλουέτα του ανθρώπου σε κάθε καρέ και συσσωρεύουν τις εικόνες που προκύπτουν σε μια τελική εικόνα. Προκύπτουν, έτσι, οι λεγόμενες Motion History Image (MHI) και Motion Energy Image (MEI) οι οποίες χρησιμοποιούνται σαν "τεμπλέτες" (templates) και συγκρίνονται βάσει των "Hu moments" [32] τους. Οι Weiland et al. [33] γενικεύουν την ιδέα αυτή στον τρισδιάστατο χώρο και

[1]Σημειώνουμε πως σε αρκετές δημοσιευμένες εργασίες χρησιμοποιείται ο όρος STIP αναφερόμενος συγκεκριμένα στον Harris3D. Στην παρούσα διπλωματική χρησιμοποιούμε τον όρο STIP γενικά, εννοώντας όλες τις εργασίες που εξάγουν σημεία ενδιαφέροντος για την περιγραφή μιας δράσης.

[2]ιδέα με την οποία είχαν πειραματιστεί και νωρίτερα οι Laptev και Lindberg [5].



Σχήμα 2.1: Motion Energy Volumes (Σχήμα από [33]).

υπολογίζουν τους λεγόμενους "Motion Energy Volumes" (Σχήμα 2.1). Αυτοί συγκρίνονται χρησιμοποιώντας περιγραφητές βασισμένους στο μετασχηματισμό Fourier τους σε κυλινδρικές συντεταγμένες. Οι Gorelick et al. [34] αντιμετωπίζουν το περίγραμμα του ανθρώπου ως ένα ενιαίο τρισδιάστατο σχήμα που παρακολουθούν στο χρόνο, και στο οποίο υπολογίζουν μια σειρά από χαρακτηριστικά ("Plateness", "Stickness" κ.α.). Οι Yilmaz και Shah [35] υπολογίζουν ένα παρόμοιο χωρο-χρονικό σχήμα από το περίγραμμα του ανθρώπου και υπολογίζουν χαρακτηριστικά εμπνευσμένα από έννοιες διαφορικής γεωμετρίας. Φυσικά, όλες οι τεχνικές αυτές δεν κωδικοποιούν πληροφορία σχετικά με την κίνηση εντός του βίντεο και επίσης προϋποθέτουν στατικό ή ομοιόμορφο background. Για την αντιμετώπιση του προβλήματος αυτού, οι Efros et al. [36]. εξάγουν την οπτική ροή από βίντεο αθλητικών αγώνων στα οποία η σκηνή είναι κεντραρισμένη στον άνθρωπο για να υπολογίζουν χαρακτηριστικά σχετικά με την κίνηση. Τέλος, μια ενδιαφέρουσα εργασία είναι αυτή των Junejo et al. [37], οι οποίοι παρατηρούν πως ο πίνακας "αυτό-ομοιότητας" (self-similarity matrix) είναι χαρακτηριστικός μια δράσης ανεξαρτήτως της οπτικής γωνίας που έχει ληφθεί το βίντεο. Αντιμετωπίζουν αυτούς τους πίνακες ως εικόνες, στις οποίες υπολογίζουν περιγραφητές. Στη συνέχεια σχηματίζουν ιστογράμματα των περιγραφητών και συγκρίνουν μια σειρά από ταξινομητές.

Σημειώνουμε πως οι global μέθοδοι παρουσιάζουν σημαντικά ελαττώματα, όπως ευαισθησία σε οπτικό θόρυβο, επικαλύψεις (occlusions) και αλλαγές στην οπτική γωνία λήψης. Συχνά, επίσης, επαφίενται σε αλγορίθμους "αφαίρεσης" του φόντου (background substruction). Ως εκ τούτου, δεν έχουν πολύ καλές αποδόσεις σε απαιτητικές βάσεις δεδομένων με ρεαλιστικές συνθήκες εκτέλεσης των δράσεων. Γι' αυτό η χρήση τους έχει ως ένα βαθμό εγκαταλειφθεί.

Μια σειρά από ενδιαφέρουσες τεχνικές βασίζονται σε low-level πληροφορία, ή σε local χαρακτηριστικά, τα οποία χρησιμοποιούν για να εντοπίσουν μεγαλύτερες κινούμενες δομές ("parts") εντός του βίντεο.

Τελικά, η δράση περιγράφεται από τη "μεσαίου επιπέδου" (mid-level) πληροφορία που αφορά τη σχέση μεταξύ των διαφόρων parts.

Χαρακτηριστική είναι η εργασία των Raptis et al. [38], οι οποίοι εξάγουν dense trajectories [12] από το βίντεο, τις οποίες ομαδοποιούν βάσει ενός κριτηρίου χωρο-χρονικής εγγύτητας. Έτσι, διαιρούν τον όγκο του βίντεο σε μέρη (parts) που ανήκουν στον ίδιο κινούμενο άνθρωπο ή στο ίδιο αντικείμενο. Στη συνέχεια, υπολογίζουν γνωστούς περιγραφητές (HoG, HoF, κ.α.) για κάθε part και χρησιμοποιούν ένα Markov Random Field (MRF) για να μοντελοποιήσουν τις δράσεις. Κάθε testing βίντεο ταξινομείται στη δράση για την οποία το αντίστοιχο MRF έχει τη μεγαλύτερη ενέργεια. Οι Yang et al. [39] υιοθετούν ένα σχήμα ιεραρχικής ομαδοποίησης (hierarchical clustering) για να εντοπίσουν "υπο-δράσεις" (sub-actions), ή αλλιώς "action primitives". Συγκεκριμένα, αφού κεντράρουν την ανθρώπινη φιγούρα στην εικόνα, εξάγουν την οπτική ροή u, v σε κάθε σημείο της (x, y), και ομαδοποιούν τα διανύσματα (x, y, u, v) με τον αλγόριθμο K-means. Κάθε cluster αντιστοιχεί και σε μια εκτέλεση ενός primitive. Στη συνέχεια χρησιμοποιούν γραφοθεωρητικές μεθόδους για να ομαδοποιήσουν τα clusters που προέκυψαν και εξάγουν μια εννιαία περιγραφή για κάθε primitive. Εν τέλει μια δράση αναπαρίσταται από μια ακολουθία από action primitives. Οι συγγραφείς πειραματίζονται με διάφορες μεθόδους ταξινόμησης, όπως αλγορίθμους string matching, HMMs κ.α.

2.3 Άλλες προσεγγίσεις

Αξίζει να αναφέρουμε και κάποιες εργασίες που προσπαθούν να μοντελοποιήσουν ρητά τη χρονική-δυναμική πληροφορία (temporal dynamics), χρησιμοποιώντας κατά κανόνα δυναμικά μοντέλα. Οι Bhattacharya et al. [40] αξιοποιούν την ιδέα των "υπο-δράσεων" (sub-actions) και χρησιμοποιούν ένα σχετικά μικρό κυλιόμενο χρονικό παράθυρο, εξάγοντας σκορ από κάθε μικρό τμήμα του βίντεο σχετικά με την υπο-δραση που υπάρχει σε αυτό. Η σχέση και η αλληλουχία των υπο-δράσεων μοντελοποιείται με ένα γραμμικό δυναμικό μοντέλο (Linear Dynamical System - LDS). Οι Kuehne et al. [9] ακολουθούν ένα παρόμοιο σκεπτικό: ταξινομούν μικρά τμήματα του βίντεο εντός ενός κυλιόμενου παραθύρου σε μια υπο-δράση (χρησιμοποιούν τον όρο "action unit") και αναπαριστούν κάθε τέτοιο τμήμα με ένα ιστόγραμμα Bag-of-Features. Στη συνέχεια εκπαιδεύουν ένα Κρυφό Μαρκοβιανό Μοντέλο (Hidden Markov Model - HMM), κάθε κατάσταση του οποίου αντιστοιχεί σε ένα κυλιόμενο παράθυρο και το έχει το διάνυσμα παρατήρησής της είναι αντίστοιχο BoF

ιστόγραμμα. Σε μια πιο πρόσφατη εργασία, οι Fernando et al. [41] κωδικοποιούν την χρονική εξέλιξη μιας δράσης κατά τη διάρκεια του βίντεο εκπαιδεύοντας ένα σύνολο γραμμικών μηχανών "κατάταξης" (ranking machines). Αυτές μοντελοποιούν τη χρονική διάταξη των frames ως εξής: για δύο διαδοχικά frames δίνουν μικρότερο "σκορ" σε αυτό που προηγείται χρονικά και μεγαλύτερο σε αυτό που έπεται, βασιζόμενοι σε έναν περιγραφητή που εξάγεται από το εκάστοτε frame. Για κάθε βίντεο εκπαιδεύουν μια τέτοια μηχανή και χρησιμοποιούν τις παραμέτρους της για να το αναπαραστήσουν. Η ταξινόμηση γίνεται κι εδώ με SVM.

Μια ιδιαίτερα ενδιαφέρουσα εργασία προέρχεται από τους Jain et al. [42], οι οποίοι χρησιμοποιούν έναν ανιχνευτή αντικειμένων βασισμένο σε βαθιά νευρωνικά δίκτυα για να εντοπίσουν αντικείμενα σε κάθε frame ενός βίντεο. Ο ανιχνευτής αυτός είναι εκπαιδευμένος να εντοπίζει αντικείμενα από 15000 διαφορετικές κλάσεις της μεγάλης βάσης δεδομένων ImageNet [1]. Οι συγγραφείς δείχνουν μέσα από διεξοδικά πειράματα πως η πληροφορία σχετικά με τα αντικείμενα που υπάρχουν σε ένα βίντεο είναι σε μεγάλο βαθμό ενδεικτική της δράσης που εκτελείται και μπορεί να βελτιώσει το αποτέλεσμα της ταξινόμησης δράσεων έως και $9,8\%$.

Τέλος, αξίζει να αναφέρουμε τη χαρακτηριστική εργασία των Simonyan & Zisserman [13], οι οποίοι επιτυγχάνουν υψηλά ποσοστά ταξινόμησης με βαθιά νευρωνικά δίκτυα. Οι συγγραφείς επικαλούνται μια θεωρία στη νευροβιολογία, σύμφωνα με την οποία η οπτική πληροφορία ακολουθεί δύο ξεχωριστά "μονοπάτια" επεξεργασίας στον οπτικό φλοιό του ανθρώπινου εγκεφάλου, με το πρώτο να δίνει έμφαση στα αντικείμενα και το δεύτερο στην κίνηση. Χρησιμοποιούν, λοιπόν, δύο ανεξάρτητα συνελικτικά δίκτυα (ConvNets). Το πρώτο βασίζεται στη στατική εμφάνιση του βίντεο και δέχεται ένα υποσύνολο των frames του ως είσοδο. Το δεύτερο δέχεται την οπτική ροή και βασίζεται στην κίνηση που παρατηρείται στο βίντεο. Το αποτέλεσμα των δύο αυτών συνδυάζεται για να προκύψει το τελικό αποτέλεσμα.

[1] `http://image-net.org`

Κεφάλαιο 3

Θεωρητικό Υπόβαθρο

Στο παρόν Κεφάλαιο αναλύονται διεξοδικά οι θεωρητικές πτυχές όλων των επιμέρους σταδίων και μεθόδων που χρησιμοποιούμε για την αναγνώριση ανθρώπινων δράσεων. Στις Ενότητες 3.1 και 3.2 παρουσιάζονται δύο δημοφιλείς κατηγορίες χαρακτηριστικών που εξάγονται από βίντεο: τα χωρο-χρονικά σημεία ενδιαφέροντος και οι πυκνές τροχιές. Τα χαρακτηριστικά αυτά αναπαριστώνται από κατάλληλους *περιγραφητές*, οι οποίοι κωδικοποιούν πληροφορία σχετικά με την εμφάνιση και την κίνηση εντός μιας γειτονιάς γύρω από τα σημεία ενδιαφέροντος ή τις τροχιές. Οι διάφοροι περιγραφητές που χρησιμοποιούμε αναλύονται στην ενότητα 3.3. Τα χαρακτηριστικά και οι αντίστοιχοι περιγραφητές τους αποτελούν τη "χαμηλού επιπέδου" πληροφορία που εξάγουμε από το βίντεο και δεν είναι άμεσα αξιοποιήσιμα για την ταξινόμηση διαφορετικών κλάσεων από δράσεις. Για το σκοπό αυτό έχουν προταθεί αρκετές αναπαραστάσεις βίντεο που βασίζονται στην *κωδικοποίηση* των χαρακτηριστικών και αναλύονται στην ενότητα 3.4. Τέλος, στην ενότητα 3.5 περιγράφουμε την ταξινόμηση των βίντεο με τη χρήση Μηχανών Διανυσμάτων Υποστήριξης.

3.1 Χωρο-χρονικά σημεία ενδιαφέροντος (STIP)

Τα χωρο-χρονικά σημεία ενδιαφέροντος (Spatio-Temporal Interest Points - STIP) αποτελούν μια από τις πρώτες κατηγορίες χαρακτηριστικών που αναπτύχθηκαν για το σκοπό της αυτόματης αναγνώρισης ανθρώπινων δράσεων. Βασίζονται στην ιδέα της αναπαράστασης ενός βίντεο από ένα σύνολο "σημαντικών" ή "εξέχοντων" (salient) σημείων του.

3.1.1 Ο ανιχνευτής Harris3D

Ο ανιχνευτής Harris3D είναι ίσως ο δημοφιλέστερος ανιχνευτής χωροχρονικών σημείων ενδιαφέροντος και χρησιμοποιείται κατά κόρον στη βιβλιογραφία. Αναπτύχθηκε από τους Laptev και Lindberg [11], με αφορμή τη επιτυχία του ανιχνευτή γωνιών Harris [43], του οποίου αποτελεί επέκταση στον τρισδιάστατο χώρο.

Η βασική ιδέα του ανιχνευτή γωνιών Harris σε εικόνες είναι η ανίχνευση των σημείων εκείνων στα οποία οι τιμές των πίξελ μεταβάλλονται πάνω από ένα κατώφλι και στις δύο κατευθύνσεις. Ομοίως, ο Harris3D ανιχνεύει σημεία οι τιμές των οποίων μεταβάλλονται επαρκώς και στις τρεις διαστάσεις. Τα σημεία αυτά παρουσιάζουν μη σταθερή κίνηση εντός μιας χωρο-χρονικής γειτονιάς. Βασική διαφορά της τρισδιάστατης έκδοσης είναι η διαφορετική αντιμετώπιση του χρόνου: τα σημεία ενδιαφέροντος αναζητούνται σε διαφορετικές κλίμακες στο χρόνο από ότι στο χώρο. Συγκεκριμένα, έστω $f : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$ το σήμα βίντεο. Η αναπαράστασή του σε πολλαπλές κλίμακες (scale-space representation) κατασκευάζεται συνελίσσοντας το σήμα βίντεο με έναν τρισδιάστατο ανισοτροπικό γκαουσιανό πυρήνα με διαφορετική μεταβλητότητα (variance) στο πεδίο του χώρου και του χρόνου (σ_l και τ_l αντίστοιχα):

$$L(\cdot; \sigma_l^2, \tau_l^2) = g(\cdot; \sigma_l^2, \tau_l^2) * f(\cdot), \tag{3.1}$$

όπου $g(\cdot)$ ο τρισδιάστατος διαχωρίσιμος γκαουσιανός πυρήνας:

$$g(x, y, t; \sigma_l^2, \tau_l^2) = \frac{\exp(-(x^2 + y^2)/2\sigma_l^2 - t^2/2\tau^2)}{\sqrt{(2\pi)^3 \sigma_l^4 \tau_l^2}} \tag{3.2}$$

Για την εύρεση των χωρο-χρονικών σημείων ενδιαφέροντος κατασκευάζεται ο 3×3 πίνακας των δεύτερων "στιγμών" (second moments) της L, αποτελούμενος από τις πρώτες παραγώγους της και σταθμίζεται με ένα γκαουσιανό πυρήνα $g(x, y, t; \sigma_i^2, \tau_i^2)$:

$$M = g(x, y, t; \sigma_i^2, \tau_i^2) * \begin{pmatrix} L_x^2 & L_x L_y & L_x Lt \\ L_x L_y & L_y^2 & L_y L_t \\ L_x L_t & L_y L_t & L_t^2 \end{pmatrix}, \tag{3.3}$$

όπου $\sigma_i^2 = s\sigma_l^2$, $\tau_i^2 = s\tau_l^2$ η χωρική και χρονική κλίμακα ολοκλήρωσης. Οι πρώτες παράγωγοι L_u ορίζονται ως εξής: $L_u(\cdot; \sigma_l^2, \tau_l^2) = \partial_u (f * g)$. Οι ιδιοτιμές $\lambda_1, \lambda_2, \lambda_3$ του πίνακα M χαρακτηρίζουν τη μεταβολή του βίντεο στις τρεις διευθύνσεις. Τα σημεία, λοιπόν, στα οποία μεγιστοποιούνται και οι τρεις ιδιοτιμές αντιστοιχούν σε σημεία ενδιαφέροντος, σημεία δηλαδή, που έχουμε έντονη μεταβολή τόσο στο χώρο όσο και στο χρόνο.

Ομοίως με το δισδιάστατο ανιχνευτή Harris, οι συγγραφείς προτείνουν ένα κριτήριο "γωνιότητας" για την αποφυγή του υπολογισμού των ιδιοτιμών, ο οποίος έχει μεγάλη υπολογιστική πολυπλοκότητα, δεδομένων και των συνήθων διαστάσεων του πίνακα M. τα σημεία ενδιαφέροντος ορίζονται, λοιπόν, ως τα σημεία τοπικού μεγίστου της παρακάτω ποσότητας, τα οποία αποδεικνύουν πως αντιστοιχούν σε σημεία μεγιστοποίησης και των τριών ιδιοτιμών:

$$H = \det(M) - k \cdot \text{trace}^3(M) = \lambda_1\lambda_2\lambda_3 - k(\lambda_1 + \lambda_2\lambda_3)^3 .$$

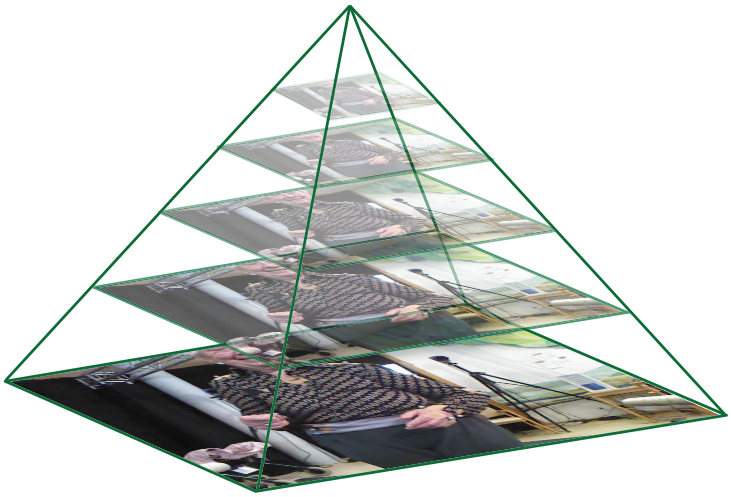
Οι συγγραφείς προτείνουν και έναν επαναληπτικό αλγόριθμο αυτόματης επιλογής της κλίμακας (scale selection) τον οποίο όμως αργότερα εγκαταλείπουν [2] μεταξύ άλλων λόγω του υπολογιστικού του κόστους και των artifacts που πολλές φορές δημιουργεί. Αντί αυτού προτείνουν τον εντοπισμό σημείων ενδιαφέροντος σε πολλαπλές προκαθορισμένες κλίμακες $\sigma_i = 2^{(1+i)/2}, i = 1 \ldots 6$ και $\tau_j = 2^{j/2}, j = 1, 2$.

3.1.2 Πυκνή δειγματοληψία

Σε μια μια εργασία των Wang et al. το 2009 [25], στην οποία αξιολογούνταν οι δημοφιλέστερες τεχνικές εξαγωγής χωρο-χρονικών σημείων ενδιαφέροντος, παρατηρήθηκε πως υπό τις ίδιες συνθήκες όλες οι τεχνικές παρουσιάζουν παρεμφερή αποτελέσματα, με τον Cuboids ανιχνευτή να έχει ελαφρώς καλύτερες επιδόσεις στις πιο απαιτητικές βάσεις δεδομένων. Στην ίδια εργασία, οι συγγραφείς πειραματίστηκαν δειγματοληπτώντας σημεία ενδιαφέροντος από το βίντεο σε ένα σταθερό χωροχρονικό πλέγμα, εξάγοντας περιγραφητές σε προκαθορισμένες χωροχρονικές γειτονιές γύρω από αυτά. Έδειξαν λοιπόν, πως η απλή αυτή, όσον αφορά τη σύλληψη και την υπολογιστική πολυπλοκότητα, μέθοδος, οδηγούσε σε καλύτερα αποτελέσματα συγκριτικά με κάθε προηγούμενη μέθοδο.

3.2 Πυκνές Τροχιές

Η μέθοδος των Πυκνών Τροχιών [29] (Dense Trajectories) αποτελεί τη δημοφιλέστερη μέθοδο εξαγωγής χαρακτηριστικών λόγω της απλότητας και των εξαιρετικών αποτελεσμάτων που δίνει σε απαιτητικές βάσεις δεδομένων. Η μέθοδος είναι επηρεασμένη αφ' ενός από την πυκνή δειγματοληψία, η οποία είχε δειχθεί πως υπερτερεί έναντι άλλων τεχνικών



Σχήμα 3.1: Αναπαράσταση των διαδοχικών κλιμάκων στις οποίες γίνεται η πυκνή δειγματοληψία. Η εικόνα σε κάθε επίπεδο της πυραμίδας αποτελεί την υποδειγματοληπτημένη κατά $\sqrt{2}$ έκδοση της εικόνας στο ακριβώς προηγούμενο επίπεδο.

εντοπισμού χωρο-χρονικών σημείων ενδιαφέροντος 3.1.2, και αφ' ετέρου από μεθόδους που περιγράφουν την κίνηση εντός των βίντεο μέσω των τροχιών που ακολουθούν τα σημεία ενδιαφέροντος (π.χ. [28]).

Η βασική μεθοδολογία, λοιπόν, συνίσταται στην πυκνή δειγματοληψία σημείων ενδιαφέροντος σε ένα ορθογώνιο πλέγμα στο χώρο της εικόνας και την παρακολούθησή τους (tracking) κατά τη διάρκεια του βίντεο μέσω της οπτικής ροής. Η παρακολούθηση των σημείων ενδιαφέροντος γίνεται ανεξάρτητα σε 8 κλίμακες και για κάθε τροχιά που σχηματίζεται υπολογίζονται 5 διαφορετικοί περιγραφητές. Στη συνέχεια περιγράφουμε αναλυτικά τα επιμέρους βήματα που περιλαμβάνει η εξαγωγή Πυκνών Τροχιών.

3.2.1 Σημεία ενδιαφέροντος

Το αρχικό βήμα της μεθόδου είναι η δειγματοληψία σημείων της εικόνας σε ένα ορθογώνιο πλέγμα με βήμα W pixels, το οποίο επιλέγεται έτσι ώστε να υπάρχει επαρκής κάλυψη της εικόνας. Η δειγματοληψία γίνεται ανεξάρτητα σε 8 κλίμακες το πολύ, ανάλογα με την ανάλυση του βίντεο, οι οποίες διαφέρουν μεταξύ τους κατά έναν παράγοντα $1/\sqrt{2}$ (βλ. Σχήμα 3.1 και 3.3β΄).

Ένα σημαντικό βήμα είναι η απόρριψη των σημείων του πλέγματος που ανήκουν σε ομοιογενείς περιοχές της εικόνας. Τα σημεία αυτά εξαι-

ρούνται από την παρακολούθηση, μιας και είναι αδύνατο να ακολουθηθεί η τροχιά τους μέσω της οπτικής ροής. Η επιλογή τους γίνεται με το κριτήριο των Shi & Tomashi [44], το οποίο βασίζεται στον ανιχνευτή γωνιών Harris [43]. Ο τελευταίος υπολογίζει έναν πίνακα "δεύτερων στιγμών" (second moments) ή πίνακα αυτοσυσχέτισης $M(x, y)$ όμοιο με τον 3.3 (η διαφορά συνίσταται στη μη ύπαρξη της διάστασης του χρόνου), ο οποίος ποσοτικοποιεί το κατά πόσο μεταβάλλεται η εικόνα σε κάθε pixel ως προς τις δύο κατευθύνσεις. Βάσει του πίνακα αυτού, οι Shi & Tomashi προτείνουν το εξής κριτήριο "γωνιότητας":

$$H\left(x, y\right) = \min\left(\lambda_1, \lambda_2\right), \tag{3.4}$$

όπου λ_1, λ_2 οι ιδιοτιμές του πίνακα $M(x, y)$. Τα υποψήφια σημεία για παρακολούθηση στο χρόνο είναι εκείνα τα σημεία του πλέγματος στα οποία η $H(x, y)$ έχει τιμή μεγαλύτερη από ένα κατώφλι, δηλαδή:

$$H\left(x, y\right) > \kappa \cdot \max_{x,y} H\left(x, y\right). \tag{3.5}$$

Για την τιμή του κατωφλίου, οι συγγραφείς επιλέγουν πειραματικά $\kappa = 0.001$.

3.2.2 Οπτική ροή

Η τροχιά που διαγράφει καθένα από τα παραπάνω σημεία στο χρόνο παρακολουθείται με τη χρήση της *οπτικής ροής*.

Η οπτική ροή είναι το μέγεθος εκείνο που περιγράφει τη σχετική κίνηση μεταξύ της κάμερας και της σκηνής που απαθανατίζεται σε ένα βίντεο ή μια αλληλουχία εικόνων. Σε διακριτές εικόνες, η οπτική ροή ποσοτικοποιεί ουσιαστικά τη "στιγμιαία" ταχύτητα κάθε pixel, ή αλλιώς τη σχετική μετατόπιση που υφίσταται μεταξύ δύο διαδοχικών καρέ I_t και I_{t+1}. Υπάρχει μια πληθώρα μεθόδων που κάνουν εκτίμηση της οπτικής ροής, με αυτήν των Lukas & Kanade [45] να αποτελεί την πιο διαδεδομένη.

Εν τω προκειμένω, για την παρακολούθηση των σημείων χρησιμοποιείται η μέθοδος που αναπτύχθηκε από τον Gunnar Farnebäck [46], η οποία υπολογίζει ένα πυκνό πεδίο οπτικής ροής $\mathbf{d}_t = (u_t, v_t)$, όπου u_t, v_t το οριζόντιο και κατακόρυφο μέρος της. Η μέθοδος αυτή εκτιμά τη μετατόπιση κάθε pixel μεταξύ δύο διαδοχικών καρέ με τη χρήση πολυωνυμικών αναπτυγμάτων. Στη συνέχεια περιγράφουμε τα βασικά σημεία της μεθόδου.

3.2.2.1 Η μέθοδος του Farnebäck

Η κεντρική ιδέα του συγκεκριμένου αλγορίθμου είναι η προσέγγιση της τιμής φωτεινότητας ενός pixel με ένα πολυωνυμικό ανάπτυγμα. Αν λοιπόν $I(\mathbf{x})$, όπου $\mathbf{x} = (x, y)$ είναι το σήμα της εικόνας, έχουμε:

$$I(\mathbf{x}) \approx \mathbf{x}^\intercal \mathbf{A} \mathbf{x} + \mathbf{b}^\intercal \mathbf{x} + c, \tag{3.6}$$

όπου οι όροι $\mathbf{A}$ και $\mathbf{B}$ υπολογίζονται με τη μέθοδο των ελαχίστων τετραγώνων. Στη συνέχεια αναλύουμε την επίδραση που θα έχει μια μετατόπιση των pixel της γειτονιάς κατά τη διεύθυνση του διανύσματος $\mathbf{d}$. Θεωρούμε το ακριβές πολυώνυμο:

$$I_1(\mathbf{x}) = \mathbf{x}^\intercal \mathbf{A}_1 \mathbf{x} + \mathbf{b}_1^\intercal \mathbf{x} + c_1, \tag{3.7}$$

και κατασκευάζουμε τη μετατοπισμένη έκδοση του $I_1(\mathbf{x})$ κατά $\mathbf{d}$:

$$\begin{aligned} I_2(\mathbf{x}) = I_1(\mathbf{x} - \mathbf{d}) &= (\mathbf{x} - \mathbf{d})^\intercal \mathbf{A}_1^\intercal (\mathbf{x} - \mathbf{d}) + \mathbf{b}_1^\intercal (\mathbf{x} - \mathbf{d}) + c_1 \\ &= \mathbf{x}^\intercal \mathbf{A}_1 \mathbf{x} + (\mathbf{b}_1 - 2\mathbf{A}_1 \mathbf{d})^\intercal + \mathbf{d}^\intercal \mathbf{A}_1 \mathbf{d} - \mathbf{b}_1^\intercal \mathbf{d} + c_1 \\ &= \mathbf{x}^\intercal \mathbf{A}_2 \mathbf{x} + \mathbf{b}_2^\intercal \mathbf{x} + c_2. \end{aligned} \tag{3.8}$$

Εξισώνοντας τους όμοιους όρους στις (3.7), (3.8) προκύπτει:

$$\mathbf{A}_2 = \mathbf{A}_1 \tag{3.9}$$

$$\mathbf{b}_2 = \mathbf{b}_1 - 2\mathbf{A}_1 \mathbf{d} \tag{3.10}$$

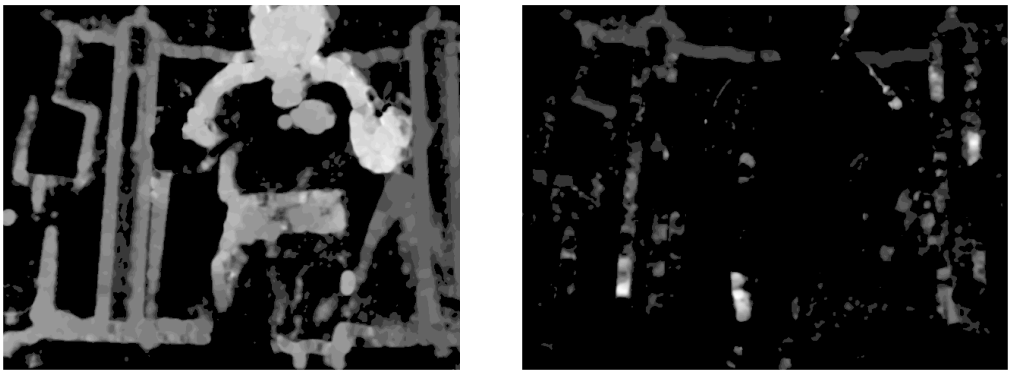
$$c_2 = \mathbf{d}^\intercal \mathbf{A}_1 \mathbf{d} - \mathbf{b}_1^\intercal \mathbf{d} + c_1. \tag{3.11}$$

Επομένως αν ο $\mathbf{A}_1$ είναι αντιστρέψιμος, μπορεί να υπολογιστεί η μετατόπιση μέσω της (3.11):

$$\begin{aligned} 2\mathbf{A}_1 \mathbf{d} = \mathbf{b}_1 - \mathbf{b}_2 \Rightarrow \\ \mathbf{d} = -\frac{1}{2} \mathbf{A}_1^{-1} (\mathbf{b}_2 - \mathbf{b}_1). \end{aligned} \tag{3.12}$$

Ωστόσο, η υπόθεση ότι μια εικόνα ή μια ολόκληρη περιοχή της περιγράφεται από ένα πολυώνυμο είναι μη ρεαλιστική. Γι' αυτό αντικαθιστούμε τα πολυωνυμικά αναπτύγματα με *σημείο προς σημείο* αναπτύγματα, δηλαδή το συντελεστή $\mathbf{A}_1$ με $\mathbf{A}_1(\mathbf{x})$, τον $\mathbf{b}_1$ με $\mathbf{b}_1(\mathbf{x})$, κοκ. Επίσης, η (3.10) πρακτικά δεν ικανοποιείται, γι' αυτό χρησιμοποιούμε τον κοινό πίνακα:

$$\mathbf{A}(\mathbf{x}) = \frac{\mathbf{A}_1(\mathbf{x}) + \mathbf{A}_2(\mathbf{x})}{2}. \tag{3.13}$$



Σχήμα 3.2: Η οριζόντια (αριστερά) και κατακόρυφη (δεξιά) συνιστώσα της οπτικής ροής που προκύπτει με τη μέθοδο του Farnebäck. Η οπτική ροή έχει υπολογιστεί μεταξύ του frame του Σχήματος 3.3α΄ και του επόμενού του στο χρόνο.

Έτσι, η (3.12) γράφεται:

$$\mathbf{A}(\mathbf{x})\mathbf{d}(\mathbf{x}) = \mathbf{\Delta b}(\mathbf{x}), \tag{3.14}$$

όπου $\mathbf{\Delta b}(\mathbf{x}) = {}^{(\mathbf{b}_1(\mathbf{x}) - \mathbf{b}_2(\mathbf{x}))}/_2$. Παρ' όλο που η (3.14) λύνεται σημείο-προς-σημείο, ο δημιουργός της μεθόδου υποστηρίζει πως τα αποτελέσματα είναι ιδιαίτερα θορυβώδη. Γι' αυτό, θεωρούμε πως το διάνυσμα $\mathbf{d}(\mathbf{x})$ μεταβάλλεται αργά, οπότε αναζητούμε το $\mathbf{d}(\mathbf{x})$ που περιγράφει με βέλτιστο τρόπο τη μετατόπιση όλων των pixels εντός μιας γειτονιάς $\mathcal{N}$ του $\mathbf{x}$. Βάσει της (3.14) αυτό επιτυγχάνεται ελαχιστοποιώντας την παρακάτω ποσότητα:

$$\sum_{\mathbf{\Delta x}\in\mathcal{N}} = w\left(\mathbf{\Delta x}\right) \left\|\mathbf{A}\left(\mathbf{x}+\mathbf{\Delta x}\right) - \mathbf{\Delta b}\left(\mathbf{x}+\mathbf{\Delta x}\right)\right\|^2, \tag{3.15}$$

όπου $w\left(\mathbf{\Delta x}\right)$ μια συνάρτηση που σταθμίζει τα pixel εντός της γειτονιάς (συνήθως γκαουσιανός πυρήνας). Διαφορίζοντας και επιλύοντας την παραπάνω εξίσωση ως προς $\mathbf{d}$ προκύπτει:

$$\mathbf{d}(\mathbf{x}) = \left(\sum_{\mathcal{N}} w\mathbf{A}^\intercal\mathbf{A}\right)^{-1} \sum_{\mathcal{N}} w\mathbf{A}^\intercal\mathbf{\Delta b}. \tag{3.16}$$

Στην περίπτωση που έχουμε μια πρότερη (prior) γνώση ή εκτίμηση για τις μετατοπίσεις των pixels, μπορούμε να την ενσωματώσουμε, αναζητώντας στη συνέχεια τη σχετική μετατόπιση μεταξύ των διαδοχικών καρέ,

η οποία πιθανότατα θα παρουσιάζει μικρότερο σφάλμα. Αυτό γίνεται θεωρώντας πως το δεύτερο καρέ είναι σε διαφορετικό σύστημα συντεταγμένων σε σχέση με το πρώτο. Αν $\tilde{\mathbf{d}}(\mathbf{x})$ η prior γνώση της μετατόπισης των pixel του καρέ I_2 σε σχέση με το I_1, αντικαθιστούμε $\tilde{\mathbf{x}} = \mathbf{x} + \tilde{\mathbf{d}}(\mathbf{x})$ στις (3.13), (3.11) έχουμε:

$$\mathbf{A}(\mathbf{x}) = \frac{\mathbf{A}_1(\mathbf{x}) + \mathbf{A}_2(\tilde{\mathbf{x}})}{2} \tag{3.17}$$

$$\mathbf{b}_2(\tilde{\mathbf{x}}) = \mathbf{b}_1(\mathbf{x}) - 2\mathbf{A}(\mathbf{x})\left(\mathbf{d}(\mathbf{x}) + \tilde{\mathbf{d}}(\mathbf{x})\right) \Rightarrow$$

$$\boldsymbol{\Delta}\mathbf{b}(\mathbf{x}) = -\frac{1}{2}\left(\mathbf{b}_2\left(\tilde{\mathbf{x}}\right) - \mathbf{b}_1\left(\mathbf{x}\right)\right) + \mathbf{A}(\mathbf{x})\tilde{\mathbf{d}}(\mathbf{x}). \tag{3.18}$$

Η επιπλέον μετατόπιση κάθε pixel αντιστοιχεί στους όρους $\mathbf{b}_1(\mathbf{x})$ και $\mathbf{b}_2(\tilde{\mathbf{x}})$, οπότε υπολογίζοντάς τα και αντικαθιστώντας στην (3.12) προκύπτει το ζητούμενο. Προφανώς θέτοντας $\tilde{\mathbf{d}}(\mathbf{x}) = 0$ καταλήγουμε στις προηγούμενες εξισώσεις. Εκμεταλλευόμενοι αυτό το αποτέλεσμα, μπορούμε να "κλείσουμε το βρόχο" και να θεωρήσουμε έναν επαναληπτικό αλγόριθμο, ο οποίος για τον υπολογισμό του $\mathbf{d}_i(\mathbf{x})$ σε κάθε βήμα i θα χρησιμοποιεί την εκτίμηση της μετατόπισης του προηγούμενου βήματος $\mathbf{d}_{i-1}(\mathbf{x})$ ως prior γνώση. Κάθε βήμα αντιστοιχεί και σε μειούμενη χωρική κλίμακα, δηλαδή αρχικά υπολογίζουμε τη μετατόπιση για τη μεγαλύτερη κλίμακα (υποδειγματοληπτημένη εικόνα), τη χρησιμοποιούμε ως prior για τον υπολογισμό της μετατόπισης στη μικρότερη κλίμακα (λιγότερο υποδειγματοληπτημένη εικόνα), κ.ο.κ. Συνήθως πραγματοποιούνται $1-3$ επαναλήψεις του αλγορίθμου. Στο Σχήμα 3.2 φαίνεται ένα παράδειγμα εξαγωγής της οπτικής ροής με τη συγκεκριμένη μέθοδο.

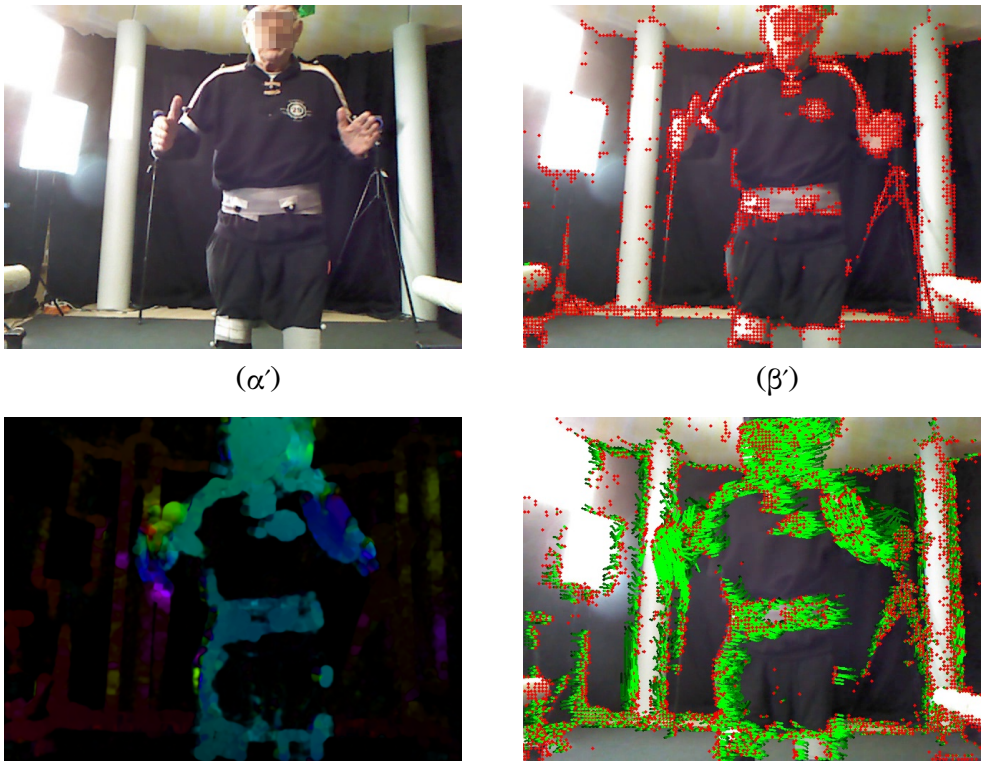
3.2.3 Σχηματισμός και αναπαράσταση τροχιών

Αφού, λοιπόν, υπολογιστεί η οπτική ροή $\mathbf{d}(x, y) = (d_x(x, y), d_y(x, y))$ με τη μέθοδο που περιγράφηκε, ομαλοποιείται με ένα φίλτρο διαμέσου (median) $\mathbf{M}$ διάστασης 3×3. Δοσμένου, λοιπόν, ενός σημείου $\mathbf{P}_t = (x_t, y_t)$ από αυτά που προέκυψαν από το βήμα της πυκνής δειγματοληψίας στο καρέ I_t, τότε αυτό ακολουθείται στο επόμενο καρέ I_{t+1} ως εξής:

$$\mathbf{P}_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (\mathbf{M} * \mathbf{d}_t)|_{(x_t, y_t)}. \tag{3.19}$$

Μια *τροχιά* (trajectory) αποτελείται τις θέσεις από τις οποίες περνά ένα σημείο κατά μήκος διαδοχικών καρέ: $(\mathbf{P}_t, \mathbf{P}_{t+1}, \mathbf{P}_{t+2}, \ldots)$. Επειδή οι τροχιές έχουν την τάση να ολισθαίνουν κατά τη διάρκεια ενός βίντεο, οι

συγγραφείς περιορίζουν το μήκος τους σε $L = 15$ σημεία, οπότε όταν η τροχιά ενός σημείου ξεπεράσει το προκαθορισμένο αυτό μήκος το σημείο παύει να παρακολουθείται. Σε κάθε frame, αν δεν παρακολουθείται κανένα σημείο εντός μιας γειτονιάς $W \times W$ σημείων, τότε δειγματοληπτείται ένα νέο σημείο και προστίθεται στη διαδικασία του tracking. Τροχιές που είναι στατικές αφαιρούνται στη συνέχεια, μιας και δεν περιέχουν καμία πληροφορία που σχετίζεται με κίνηση εντός του βίντεο. Τέλος, τροχιές που μεταβάλλονται σε δύο διαδοχικά καρέ περισσότερο από το 70% της ολικής μετατόπισης της τροχιάς απορρίπτονται.



(α′) (β′)

(γ′) (δ′)

Σχήμα 3.3: Βήματα εξαγωγής των πυκνών τροχιών: (α) Αρχικό καρέ της βάσης MOBOT, (β) Πυκνή δειγματοληψία, (γ) Χρωματική αναπαράσταση του πλάτους της οπτικής ροής μεταξύ του καρέ της εικόνας 3.3α' και του επόμενού στο χρόνο, (δ) Οι τροχιές που προκύπτουν για τα δειγματοληπτημένα σημεία.

Για την αναπαράσταση κάθε τροχιάς, χρησιμοποιούνται 5 διαφορετικοί περιγραφητές: περιγραφητής τροχιάς (Trajectory Descriptor εν συντομία T.D.), HoG, HoF, MBHx, MBHy καθώς και ο MBH, η ένωση

(concatenation) δηλαδή, των δύο τελευταίων. Αυτοί υπολογίζονται εντός χωρο-χρονικών όγκων διάστασης $N \times N \times L$ ευθυγραμμισμένων με την τροχιά. Για να ενσωματώσουμε πληροφορία σχετικά με τη δομή της τροχιάς ο όγκος κατά μήκος της τροχιάς διαιρείται σε ένα πλέγμα $n_\sigma \times n_\sigma \times n_\tau$ επιμέρους χωρο-χρονικών χωρίων (βλ. σχήμα 3.4β′). Κάθε περιγραφητής υπολογίζεται στους επιμέρους όγκους και ο τελικός περιγραφητής της τροχιάς προκύπτει από την ένωση (concatenation) των περιγραφητών των επιμέρους όγκων. Για τους HoG, MBHx, MBHy χρησιμοποιούνται 8 bins (θέσεις στο ιστόγραμμα) και για τον HoF 9 bins. Το μέγεθός λοιπόν των τελικών περιγραφητών μιας τροχιάς είναι $8 \times 2 \times 2 \times 3 = 96$ και $9 \times 2 \times 2 \times 3 = 108$ αντίστοιχα. Φυσικά, ο MBH έχει μέγεθος $2 \times 96 = 192$ bins. Ο Trajectory Descriptor είναι ένας απλός περιγραφητής που κωδικοποιεί το σχήμα της τροχιάς. Απαρτίζεται από τις σχετικές μετατοπίσεις των σημείων της τροχιάς, κανονικοποιημένες με το άθροισμά τους. Αν $(\mathbf{\Delta P}_t, \mathbf{\Delta P}_{t+1}, \mathbf{\Delta P}_{t+2}, \ldots, \mathbf{\Delta P}_{t+L-1})$ οι σχετικές μετατοπίσεις, όπου $\mathbf{\Delta P}_t = \mathbf{P}_{t+1} - \mathbf{P}_t$, τότε ο Trajectory Descriptor ορίζεται ως εξής:

$$T = \frac{(\mathbf{\Delta P}_t, \mathbf{\Delta P}_{t+1}, \ldots, \mathbf{\Delta P}_{t+L-1})}{\sum_{i=t}^{i=t+L-1} \|\mathbf{\Delta P}_i\|}$$

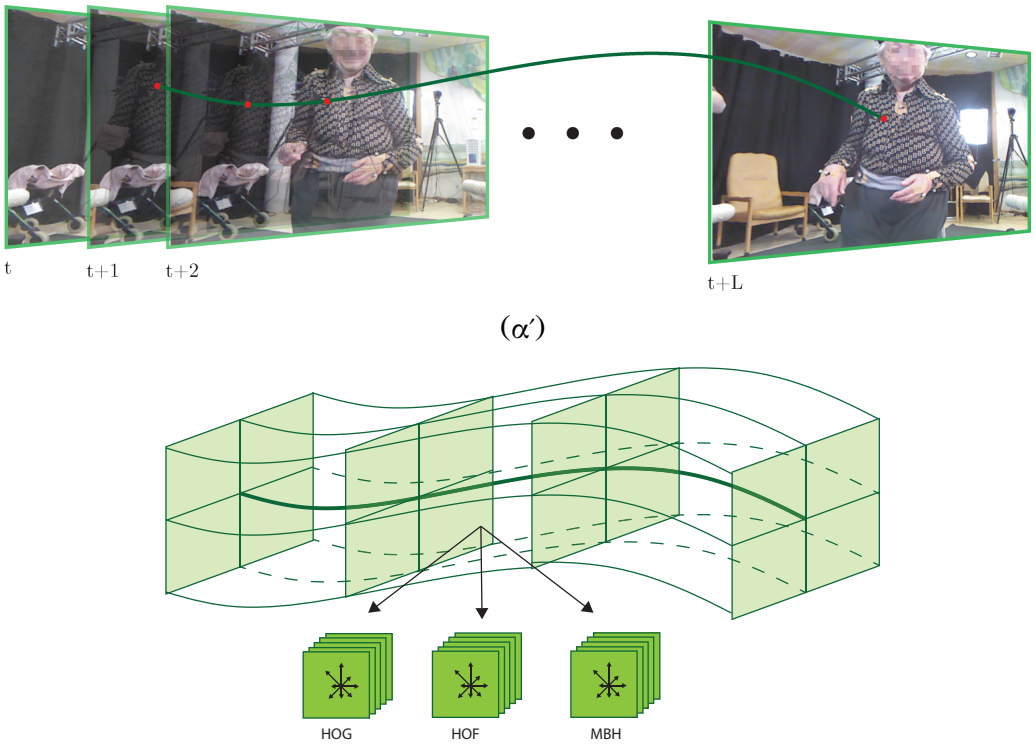
Εφ' όσον $\mathbf{\Delta P}_t = (x_{t+1} - x_t, y_{t+1} - y_t)$, ο Trajectory Descriptor έχει μέγεθος $2 \times L$.

Ο HoG περιγράφει το σχήμα και την εμφάνιση γύρω από κάθε τροχιά. Ο HoF και ο MBH κωδικοποιούν την κίνηση, με τον τελευταίο να αντισταθμίζει εν μέρει την κίνηση της κάμερας, επομένως περιμένουμε να είναι πιο εύρωστος από τον HoF. Για μια πιο διεξοδική ανάλυση, βλ. Ενότητα 3.3. Είναι φανερό πως οι περιγραφητές αυτοί κωδικοποιούν διαφορετικά κανάλια πληροφορίας που χαρακτηρίζουν μια δράση. Εκμεταλλευόμαστε, λοιπόν, την εν δυνάμει αυτή συμπληρωματικότητά τους συνδυάζοντάς τους σε έναν τελικό "Συνδυαστικό" περιγραφητή (Combined descriptor). Αυτό γίνεται στο επίπεδο του ταξινομητή, αθροίζοντας τους επιμέρους πυρήνες τους, όπως περιγράφεται στην ενότητα 3.5.2.

3.3 Περιγραφητές

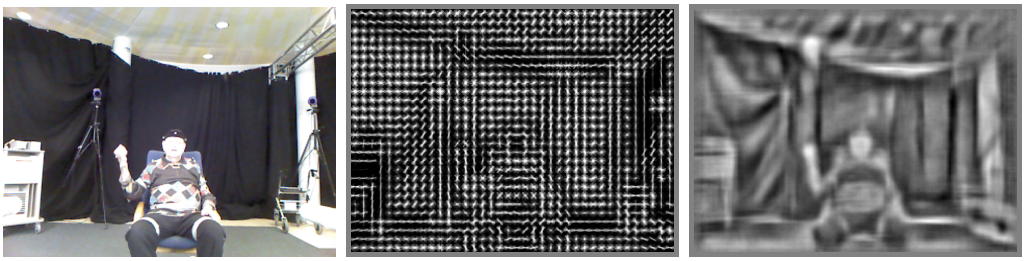
3.3.1 Ιστογράμματα κατευθυνόμενων παραγώγων

Τα ιστογράμματα κατευθυνόμενων παραγώγων (Histograms of Oriented Gradiends, εν συντομία HoG), αναπτύχθηκαν το 2005 από τους Dalal



Σχήμα 3.4: Σχηματισμός τροχιάς μήκους L και υπολογισμός περιγραφητών κατά μήκος της: (α) Μια τροχιά αποτελείται από τις διαδοχικές θέσεις ενός σημείου που παρακολουθείται, (β) Ο όγκος κατά μήκος της τροχιάς διαιρείται σε ένα πλέγμα $n_\sigma \times n_\sigma \times n_\tau$ επιμέρους χωροχρονικών χωρίων. Κάθε περιγραφητής υπολογίζεται στους επιμέρους όγκους και ο τελικός περιγραφητής της τροχιάς προκύπτει από την ένωση (concatenation) των περιγραφητών των επιμέρους όγκων.

και Triggs [47] για το πρόβλημα της αναγώρισης αντικειμένων σε εικόνες. Έκτοτε η χρήση τους έχει επεκταθεί σε μια πληθώρα προβλημάτων της Όρασης Υπολογιστών. Στο πρόβλημα της αναγνώρισης ανθρωπίνων δράσεων εφαρμόστηκαν πρώτη φορά από από τους Laptev et al. [2]. Η βασική ιδέα στην οποία βασίζεται ο HoG είναι πως το σχήμα και η εμφάνιση των απεικονιζόμενων αντικειμένων σε μια μικρή περιοχή της εικόνας μπορεί να περιγραφεί ικανοποιητικά από την κατανομή των τοπικών κλίσεων (gradients), δηλαδή των κατευθύνσεων των ακμών στην περιοχή αυτή. Φυσικά, σε ψηφιακές εικόνες το παραπάνω αντιστοιχεί στον υπολογισμό της (διακριτής) παραγώγου της εικόνας σε μια γει-



Σχήμα 3.5: Οπτικοποίηση του περιγραφητή HoG. Η μεσαία εικόνα δείχνει την κεντρική ιδέα του HoG, δηλαδή τον υπολογισμό των τοπικών κλίσεων σε μικρές γειτονιές της εικόνας. Η αριστερή εικόνα προκύπτει μέσω αντιστροφής του HoG και οπτικοποιεί την πληροφορία που κωδικοποιείται από το συγκεκριμένο περιγραφητή, δηλαδή το σχήμα και την εμφάνιση των αντικειμένων της εικόνας. Η εικόνα παράχθηκε βάσει της εργασίας και της αντίστοιχης υλοποίησης των Vondric et al. [48].

τονιά και την κατασκευή του ιστογράμματος των κλίσεων που προκύπτουν. Πολλοί περιγραφητές που βασίζονταν στην ίδια ιδέα υπήρχαν και πριν τη δημιουργία του HoG (π.χ. SIFT), ωστόσο θα λέγαμε πως ο HoG αποτελεί το "σημείο ωριμότητας" αυτής της οικογένειας περιγραφητών.

Πιο συγκεκριμένα, ο υπολογισμός του HoG ακολουθεί τα παρακάτω βήματα, στα οποία κατέληξαν οι δημιουργοί του πραγματοποιώντας εξαντλητικά πειράματα πάνω στις παραμέτρους που εμπλέκονται:

- Πραγματοποιείται διόρθωση εκθετικού κανόνα ή αλλιώς γάμμα (gamma correction) στην εικόνα.
- Υπολογίζεται η κλίση της εικόνας ως προς τους άξονες x και y συνελίσσοντάς την με τον πυρήνα $\begin{bmatrix} -1 & 0 & 1 \end{bmatrix}$ και τον ανάστροφό του.
- Αν $I_x(x, y)$ και $I_y(x, y)$ οι κλίση της εικόνας ως προς τον x και y άξονα αντίστοιχα, υπολογίζεται για κάθε pixel η διεύθυνση και το μέτρο της κλίσης:

$$s(x, y) = \sqrt{I_x^2 + I_y^2}$$
$$\theta(x, y) = arctan\left(\frac{I_y}{I_x}\right).$$

  Η κατεύθυνση κβαντίζεται συνήθως σε 8 ή 9 στάθμες (bins), από $0°$ έως $180°$ (χωρίς πρόσημο), ή $0°$ έως $360°$.

- Η περιοχή της εικόνας που θέλουμε να περιγράψουμε χωρίζεται σε *κελιά* διάστασης 8×8 pixel συνήθως. Για τα pixel εντός κάθε κελιού υπολογίζεται το ιστόγραμμα των διευθύνσεων της κλίσης, σταθμισμένο με το αντίστοιχο πλάτος της κλίσης. Κάθε θέση (bin) του ιστογράμματος αντιστοιχεί σε ένα μικρό εύρος διευθύνσεων, π.χ. αν κβαντίζουμε όλο το εύρος $0°$ έως $360°$ σε 8 στάθμες, το πρώτο bin αντιστοιχεί σε γωνίες από $0°$ έως $45°$, το δεύτερο από $45°$ έως $90°$ κ.ο.κ.

- Τα κελιά ομαδοποιούνται σε μπλοκ (συνήθως 2×2 κελιών) και τα ιστογράμματα των επιμέρους κελιών ενώνονται (concatenation) σε έναν ενιαίο περιγραφητή κάθε μπλοκ ο οποίος κανονικοποιείται, συνήθως με την $L2$ νόρμα. Η κανονικοποίηση γίνεται ανά μπλοκ έτσι ώστε να αντισταθμιστούν αλλαγές στη φωτεινότητα και την αντίθεση της εικόνας. Έτσι π.χ. ένα pixel μπορεί να είναι "σημαντικό" σε ένα κελί (να συνεισφέρει αρκετά στο ιστόγραμμα), αλλά τελικά σε επίπεδο μπλοκ να είναι "ασήμαντο" συγκριτικά με τα pixels των γειτονικών κελιών (δηλαδή τα τελευταία παρουσιάζουν κατά μέτρο σημαντικά μεγαλύτερη κλίση).

- Αν η περιοχή της εικόνας στην οποία εξάγουμε HoG περιλαμβάνει παραπάνω από ένα μπλοκ, ο τελικός HoG αποτελείται από την ένωση όλων των επιμέρους περιγραφητών του κάθε μπλοκ. Τα μπλοκ μπορούν να είναι και επικαλυπτόμενα, οπότε κάθε pixel θα συνεισφέρει εν τέλει παραπάνω από μια φορά στο τελικό διάνυσμα του περιγραφητή.

Στην περίπτωση που έχουμε βίντεο, η μόνη διαφορά είναι πως τα κελιά είναι χωρο-χρονικές (τρισδιάστατες) γειτονιές και ενώνονται σχηματίζοντας ένα τρισδιάστατο πλέγμα $n_x \times n_y \times n_t$ κελιών, δηλαδή ένα χωρο-χρονικό μπλοκ.

Το κύριο πλεονέκτημα του HoG είναι ότι παραμένει αναλλοίωτος (invariant) σε φωτομετρικές και γεωμετρικές μεταβολές, κυρίως επειδή υπολογίζεται σε μικρές γειτονιές - κελιά και λόγω των κανονικοποιήσεων που υφίσταται. Συγκεκριμένα, παραμένει σχετικά αμετάβλητος σε μετατοπίσεις (translations) μικρότερες από το μέγεθος του κελιού, σε περιστροφές (rotations) μικρότερες της διαφοράς δύο διαδοχικών στάθμεων της κλίσης, καθώς και σε αλλαγές του φωτισμού, σκιές κτλ.

3.3.2 Ιστογράμματα οπτικής ροής

Τα ιστογράμματα οπτικής ροής (Histograms of Optical Flow, εν συντομία HoF) εισήχθησαν από τους Laptev et al. [2] και ουσιαστικά είναι συνέχεια μιας παλαιότερης εργασίας των Laptev και Lindeberg [5], στην οποία πειραματίζονταν με διάφορες παραλλαγές ιστογραμμάτων οπτικής ροής. Ο HoF περιγράφει την κίνηση σε ένα μικρό χωρο-χρονικό όγκο του βίντεο ακολουθώντας το ίδιο πνεύμα με τον HoG και τον SIFT.

Συγκεκριμένα, υπολογίζεται αρχικά η οπτική ροή ως προς τους δύο άξονες και κατόπιν το μέτρο και η διεύθυνσή της. Η διεύθυνση κβαντίζεται σε 9 συνήθως στάθμες με την ένατη να προορίζεται για τα pixels εκείνα στα οποία η οπτική ροή είναι κατά μέτρο μικρότερη από κάποιο κατώφλι. Στη συνέχεια, η περιοχή που θέλουμε να περιγράψουμε υποδιαιρείται σε ένα πλέγμα $n_x \times n_y \times n_t$ κελιών, για κάθε ένα από τα οποία κατασκευάζεται το ιστόγραμμα των διευθύνσεων της οπτικής ροής, όπως ακριβώς και στον HoG. Τέλος, τα επιμέρους ιστογράμματα ενώνονται (concatenation) και κανονικοποιούνται σχηματίζοντας το τελικό διάνυσμα που περιγράφει την κίνηση εντός του αρχικού όγκου.

Ο HoF χρησιμοποιείται ευρέως σε προβλήματα ανάλυσης και κωδικοποίησης της κίνησης σε βίντεο. Φυσικά, η διακριτική του ικανότητα εξαρτάται άμεσα από την ακρίβεια της οπτικής ροής.

3.3.3 Ιστογράμματα περιγράμματος κίνησης

Τα ιστογράμματα περιγράμματος κίνησης (Motion Boundary Histograms, εν συντομία MBH), επινοήθηκαν από τους Dalal et al. [47], προκειμένου να αντιμετωπιστεί η επίπτωση της κίνησης της κάμερας στη διακριτική ικανότητα άλλων περιγραφητών που χρησιμοποιούνταν για την περιγραφή της κίνησης. Συγκεκριμένα, κατά τη λήψη ορισμένων βίντεο είναι συνηθισμένο η κάμερα να κινείται ομαλά, π.χ. παρακολουθώντας κάποιον κινούμενο άνθρωπο ή αντικείμενο, το οποίο αποτυπώνεται στο βίντεο ως κίνηση του παρασκηνίου (background). Περιγραφητές όπως ο HoF[1] κωδικοποιούν την κίνηση αυτή, δίνοντας μικρότερη σημασία, ή ακόμα και αγνοώντας την πραγματική κίνηση που λαμβάνει χώρα.

Η ιδέα, λοιπόν, του MBH είναι η περιγραφή της κίνησης μέσω της παραγώγου της οπτικής ροής προκειμένου να αντισταθμιστεί η κίνηση της κάμερας. Η οπτική ροή $\omega = (u, v)$ χωρίζεται στο οριζόντιο και κατακόρυφο μέρος της και υπολογίζεται η παράγωγος για καθένα από αυτά. Η συνέχεια είναι παρόμοια με αυτή του HoG: η κλίση σε κάθε pixel κβαντίζεται σε (8 συνήθως) στάθμες και εντός ενός κελιού κατασκευάζεται

το ιστόγραμμα των κλίσεων σταθμισμένες με το μέτρο τους. Οι δύο περιγραφητές που προκύπτουν (MBHx και MBHy) κανονικοποιούνται ξεχωριστά με την $L2$ νόρμα.

Ο MBH είναι αρκετά εύρωστος στην κίνηση της κάμερας, αφού όταν η κίνηση είναι ομαλή, η οπτική ροή περιέχει μια σταθερή συνιστώσα, η οποία αφαιρείται κατά τον υπολογισμό της παραγώγου. Έτσι, κωδικοποιούνται μόνο οι μεταβολές στο πεδίο της οπτικής ροής. Το όνομά του προέρχεται από το γεγονός πως η παράγωγος της οπτικής ροής απεικονίζει τις "ακμές" της κίνησης.

3.4 Κωδικοποίηση χαρακτηριστικών

Τα χαρακτηριστικά που εξάγονται από ένα βίντεο περιγράφουν τη δράση που απεικονίζεται σε αυτό. Πολλές, φυσικά, δράσεις παρουσιάζουν ομοιότητες, όπως για παράδειγμα η δράση "σηκώνομαι" και "ανεβαίνω σκάλες", οι οποίες περιλαμβάνουν παρόμοια κίνηση προς τα πάνω. Ιδανικά, η ομοιότητα αυτή αντανακλάται και στα χαρακτηριστικά, δηλαδή οι δύο κινήσεις που αναφέρθηκαν θα "μοιράζονται" κάποια χαρακτηριστικά (υπό την έννοια της σχετικά μικρής ευκλείδειας απόστασης μεταξύ των αντίστοιχων περιγραφητών τους), αλλά διαφέρουν σε άλλα. Επίσης, ο αριθμός των χαρακτηριστικών που εντοπίζονται μπορεί να διαφέρει σημαντικά μεταξύ διαφορετικών δράσεων ή διαφορετικών εκτελέσεων της ίδιας δράσης, αρχίζοντας από μερικές δεκάδες και φτάνοντας έως εκατοντάδες χιλιάδες για βίντεο ίδιας διάρκειας. Είναι λοιπόν φανερό, πως για να είναι σε θέση ένας ταξινομητής να διαχωρίσει τις διαφορετικές -πλην όμως εν δυνάμει όμοιες τμηματικά- δράσεις, θα πρέπει να ορίσουμε μια ενιαία αναπαράσταση που να κωδικοποιεί το *σύνολο* των χαρακτηριστικών που εντοπίζονται σε ένα βίντεο. Προς αυτήν την κατεύθυνση έχουν επικρατήσει μέθοδοι που περιγράφουν την *κατανομή* των τοπικών χαρακτηριστικών ενός βίντεο, με χαρακτηριστικότερη τη μέθοδο *Bag of Words*. Τέτοιες μέθοδοι είναι εμπνευσμένες από αντίστοιχες μεθόδους στα πεδία της επεξεργασίας φυσικής γλώσσας και ανάκτησης πληροφορίας (information retrieval) και χρησιμοποιούνται κατά κόρον στην Όραση Υπολογιστών.

[1]Σημειώνουμε πως ο HoF δεν είχε δημιουργηθεί ακόμα, τουλάχιστον στην τρέχουσα μορφή του, και αναφέρεται απλά ως παράδειγμα.

3.4.1 Το οπτικό λεξικό

Το πρώτο βήμα για την αναπαράσταση μιας συλλογής χαρακτηριστικών που περιγράφουν ένα βίντεο είναι η δημιουργία του λεγόμενου "οπτικού λεξικού" (visual vocabulary). Το οπτικό λεξικό περιέχει "οπτικές λέξεις" κάθε μια από τις οποίες εκπροσωπεί ένα σύνολο παρόμοιων ή "κοντινών" χαρακτηριστικών. Όπως και σε ένα πραγματικό λεξικό, αποτελείται, από όλα τα χαρακτηριστικά που είναι δυνατό να εμφανιστούν σε κάποια παραλλαγή τους στο βίντεο. Μια οπτική λέξη μπορεί να αντιπροσωπεύει κάποια ιδιότητα που είναι κατανοητή από τον άνθρωπο. Παραδείγματος χάριν, ένα συγκεκριμένο σχήμα που εμφανίζεται με κάποιες παραλλαγές σε κάποια βίντεο ιδανικά αντιστοιχεί σε μια οπτική λέξη που αναπαριστά το "μέσο" αυτό σχήμα. Ωστόσο, οι οπτικές λέξεις συνήθως δεν είναι άμεσα ερμηνεύσιμες.

Δοθείσας, λοιπόν, μιας συλλογής N χαρακτηριστικών εκπαίδευσης $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, όπου $\mathbf{x}_i \in \mathbb{R}^D$ και D η διάσταση κάθε χαρακτηριστικού, αναζητούμε ένα αντιπροσωπευτικό υποσύνολο $\mathcal{D} = \{\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_K\}$ του $\mathbf{X}$. Το $\mathcal{D}$ θα είναι το οπτικό μας λεξικό το οποίο θα χρησιμοποιηθεί για να υπολογίσουμε στατιστικά μεγέθη για τα χαρακτηριστικά που εξάγονται από κάθε βίντεο. Συνήθως υπολογίζεται με κάποιον αλγόριθμο ομαδοποίησης (clustering), με δημοφιλέστερους τον αλγόριθμο K-μέσων (K-means clustering) και τον αλγόριθμο ομαδοποίησης με Μείγμα Γκαουσιανών Κατανομών (Gaussian Mixture Model clustering, εν συντομία GMM), οι οποίοι χρησιμοποιούνται ευρέως στην Όραση Υπολογιστών και την Αναγνώριση Προτύπων. Στη συνέχεια παρουσιάζουμε μια σύντομη επισκόπηση των δύο αυτών αλγορίθμων.

K-means clustering Ο K-means είναι ίσως ο δημοφιλέστερος αλγόριθμος διανυσματικής κβάντισης (vector quantization). Χρησιμοποιείται για να διαιρέσει τον αρχικό χώρο $\mathbf{X}$ των χαρακτηριστικών σε K περιοχές, κάθε μια από τις οποίες αντιπροσωπεύεται από ένα διάνυσμα $\mathbf{d}_k \in \mathcal{D} = \{\mathbf{d}_1, \ldots, \mathbf{d}_K\}$, όπου $\mathcal{D}$ τα "κεντροειδή" του K-means, δηλαδή το οπτικό λεξικό. Για τον υπολογισμό των κεντρων $\mathcal{D}$, ο K-means χρησιμοποιεί ένα σύνολο βοηθητικών μεταβλητών $r_{i,k} \in \{0, 1\}$, $i = 1, \ldots, N$, $k = 1, \ldots K$ που δείχνουν αν το στοιχείο $\mathbf{x}_i$ ανήκει στην περιοχή που αντιπροσωπεύεται από το κέντρο $\mathbf{d}_k$ ή όχι ($r_{i,k} = 1$ ή 0 αντίστοιχα). Προφανώς κάθε στοιχείο ανήκει σε ένα μόνο κέντρο, δηλαδή αν $r_{i,k} = 1$ τότε $r_{j,k} = 0$ $\forall j \neq i$. Ο υπολογισμός των κεντροειδών γίνεται ελαχιστο-

ποιώντας το συναρτησιακό:

$$\min \sum_{i=1}^{N} \sum_{k=1}^{K} r_{i,k} \left\| \mathbf{x}_i - \mathbf{d}_k \right\|^2 \tag{3.20}$$

Για την ελαχιστοποίηση ακολουθείται μια επαναληπτική διαδικασία αποτελούμενη από δύο βήματα που εναλλάσσονται μέχρι τη σύγκλιση:

- θεωρώντας σταθερά $\mathbf{d}_k$ ελαχιστοποιείται το συναρτησιακό ως προς $r_{i,k}$
- υπολογίζονται τα νέα κέντρα ως η μέση τιμή των διανυσμάτων που ανήκουν στην ίδια περιοχή

Για περισσότερες πληροφορίες, παραπεμπουμε στο [49].

GMM clustering Τα μείγματα γκαουσιανών κατανομών (gaussian mixtures) αποτελούν αναγεννητικά μοντέλα (generative models) που περιγράφουν μια κατανομή σε ένα χώρο χαρακτηριστικών και χρησιμοποιούνται συχνά για την εύρεση ομάδων/συστάδων (clustering) εντός αυτού. Ένα GMM με K γκαουσιανές ορίζεται ως εξής:

$$p\left(\mathbf{x}, \boldsymbol{\theta}\right) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right), \tag{3.21}$$

όπου $\boldsymbol{\theta} = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$ είναι οι παράμετροι των γκαουσιανών $\mathcal{N}\left(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$, $k = 1 \ldots K$. Δοσμένων των χαρακτηριστικών εκπαίδευσης $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_N\}$, υπολογίζονται οι παράμετροι θ του GMM μέσω των οποίων περιγράφεται καλύτερα το σύνολο $\mathbf{X}$. Εν συνεχεία, το clustering γίνεται αντιστοιχίζοντας σε κάθε διάνυσμα $\mathbf{x}$ του χώρου μια posterior πιθανότητα $p\left(k|\mathbf{x}\right)$, η οποία ποσοτικοποιεί το κατά πόσο το $\mathbf{x}$ "ανήκει" ή περιγράφεται από τη γκαουσιανή k. Κάθε γκαουσιανή του GMM αντιστοιχεί, λοιπόν και σε ένα "κέντρο" και ορίζει έμμεσα ένα cluster του χώρου των χαρακτηριστικών. Η "ανάθεση" ενός διανύσματος $\mathbf{x}$ σε κάθε cluster μπορεί να γίνει είτε με "αυστηρό" τρόπο, αναθέτοντας δηλαδή το $\mathbf{x}$ στο cluster της γκαουσιανής γκαουσιανή που μεγιστοποιεί την $p\left(k|\mathbf{x}\right)$, ή με "χαλαρό" τρόπο, αναθέτοντας το $\mathbf{x}$ σε όλες τις γκαουσιανές, με βάρος $p\left(k|\mathbf{x}\right)$. Η πρώτη περίπτωση είναι ισοδύναμη με τον K-means, με τη διαφορά ότι στην περίπτωση του GMM το σχήμα κάθε cluster είναι μεταβλητό και καθορίζεται από τον πίνακα συμμεταβλητότητας $\boldsymbol{\Sigma}_k$ της αντίστοιχης γκαουσιανής. Το τελευταίο, σε συνδυασμό με τη δυνατότητα "χαλαρής" ανάθεσης που αναφέρθηκε, καθιστούν τα μίγματα γκαουσιανών

ένα ευέλικτο εργαλείο σε προβλήματα clustering. Για περισσότερες λεπτομέρειες σχετικά με την ομαδοποίηση με μείγματα γκαουσιανών και την εκπαίδευσή τους παραπέμπουμε στο [49].

3.4.2 Τεχνικές κωδικοποίησης

Όπως αναφέρθηκε, για να ταξινομήσουμε ένα βίντεο, θα πρέπει να το αναπαραστήσουμε με κάποιο τρόπο που να το καθιστά διαχωρίσιμο από τα υπόλοιπα. Αυτό γίνεται υπολογίζοντας στατιστικά μεγέθη πάνω στα χαρακτηριστικά που έχουν εξαχθεί από αυτό, χρησιμοποιώντας το οπτικό λεξικό. Η διαδικασία αυτή αναφέρεται στη βιβλιογραφία ως *κωδικοποίηση χαρακτηριστικών* (feature encoding) και αποτελείται συνήθως από δύο βήματα:

- την εξαγωγή του *κώδικα* (code) $\mathbf{s}_i$ κάθε χαρακτηριστικού $\mathbf{x}_i \in \mathbf{X}$ του βίντεο, όπου υπολογίζεται η σχέση του συγκεκριμένου χαρακτηριστικού με τις οπτικές λέξεις του λεξικού,
- τη συσσώρευση (pooling) των επιμέρους κώδικων $\mathbf{s}_i$, $i = 1 \dots N$ σε ένα ενιαίο διάνυσμα $\boldsymbol{\mathcal{S}}$.

Σε αυτή τη βάση, παρακάτω περιγράφουμε τις τεχνικές κωδικοποίησης που χρησιμοποιούνται στα πλαίσια της παρούσας διπλωματικής.

3.4.2.1 Μοντέλο συνόλου οπτικών λέξεων (Bag-of-Words)

Το μοντέλο συνόλου οπτικών λέξεων [16] (Bag-of-Words, ή BoW εν συντομία) είναι η πιο απλή μέθοδος κωδικοποίησης και είναι εμπνευσμένη από τα πεδία της επεξεργασίας φυσικής γλώσσας (natural language processing) και ανάκτησης πληροφορίας (information retrieval). Η γενική ιδέα της μεθόδου είναι η περιγραφή του βίντεο μέσω της κατανομής που ακολουθούν τα χαρακτηριστικά του. Το κάθε χαρακτηριστικό αντικαθίσταται από την κοντινότερη οπτική του λέξη από το λεξικό και το βίντεο αναπαρίσταται από τη συχνότητα με την οποία εμφανίζεται η κάθε οπτική λέξη. Συνήθεις τιμές για το μέγεθος του οπτικού λεξικού είναι $K = 500 - 4000$ κέντρα.

Πιο φορμαλιστικά, για κάθε χαρακτηριστικό $\mathbf{x}_i \in \mathbf{X}$, υπολογίζεται ένας κώδικας $\mathbf{s}_i$, με κάθε στοιχείο $\mathbf{s}_i(k)$ να δίνεται από την σχέση:

$$\mathbf{s}_i(k) = \begin{cases} 1 & \text{εάν } k = \underset{k}{\operatorname{argmin}} \|\mathbf{x}_i - \mathbf{d}_k\| \\ 0 & \text{αλλιώς,} \end{cases} \tag{3.22}$$

όπου $k = 1 \dots K$. Όπως είναι φανερό, ο κώδικας κάθε χαρακτηριστικού έχει διάσταση K και περιέχει μονάδα στη θέση που αντιστοιχεί στην κοντινότερή του οπτική λέξη και μηδέν στις υπόλοιπες. Ουσιαστικά, ο κώδικας κβαντίζει (κάνει "αυστηρή ανάθεση") κάθε χαρακτηριστικού σε μια οπτική λέξη χρησιμοποιώντας ευκλείδεια απόσταση. Στη συνέχεια, οι κώδικες $\mathbf{s}_i$ όλων των χαρακτηριστικών $\mathbf{x}_i$ αθροίζονται (συσσωρεύονται με την πράξη της άθροισης - sum pooling) και το προκύπτον διάνυσμα κανονικοποιείται με το συνολικό πλήθος τους. Προκύπτει λοιπόν:

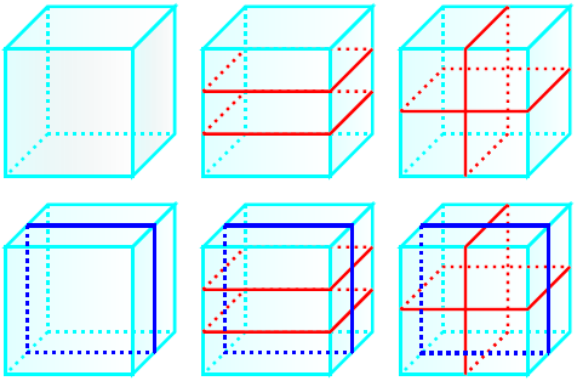
$$\mathbf{p} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{s}_i, \tag{3.23}$$

όπου κάθε θέση του $\mathbf{p}$ περιέχει τη σχετική συχνότητα εμφάνισης της αντίστοιχης οπτικής λέξης μέσα στο βίντεο. Τέλος, το $\mathbf{p}$ κανονικοποιείται ξανά, χρησιμοποιώντας την $L1$, ή συνηθέστερα, την $L2$ νόρμα. Επί της ουσίας, λοιπόν, το μοντέλο Bag-of-Words αποτελεί ένα (συνήθως αραιό) ιστόγραμμα εμφάνισης των οπτικών λέξεων, κατάλληλα κανονικοποιημένο ώστε να εξαλειφθεί η εξάρτησή της από τον αριθμό των χαρακτηριστικών που ανιχνεύθηκαν. Κωδικοποιεί μια αρκετά απλοϊκή στατιστική πληροφορία σχετικά με τα χαρακτηριστικά ενός βίντεο (συχνότητα εμφάνισης), αγνοώντας τόσο τη χωρική όσο και τη χρονική διάταξή τους. Αυτό αποτελεί και τη βασικότερη αδυναμία αυτής της οικογένειας αναπαραστάσεων. Αξίζει, τέλος, να σημειωθεί πως υπάρχουν παραλλαγές του μοντέλου Bag-of-Words που κάνουν "χαλαρή ανάθεση" κάθε χαρακτηριστικού σε μια ή περισσότερες οπτικές λέξεις π.χ. [50].

3.4.2.2 Χωρο-χρονικές πυραμίδες (spatio-temporal pyramids)

Οι χωρο-χρονικές πυραμίδες (spatio-temporal pyramids) εισήχθησαν από τους Lazebnik et al. [51] για το πρόβλημα της ταξινόμησης εικόνων (image classification) προκειμένου να βελτιώσουν τη διακριτική ικανότητα (discriminative power) του μοντέλου Bag-of-Words ενσωματώνοντας πληροφορία σχετικά με τη χωρική διάταξη των χαρακτηριστικών εντός των εικόνων. Οι Schuldt at al. [16] γρήγορα προσάρμοσαν τη συγκεκριμένη μέθοδο στο πρόβλημα της αναγνώρισης ανθρώπινων δράσεων, επεκτείνοντάς την στη διάσταση του χρόνου.

Όπως αναλύθηκε παραπάνω, το μοντέλο Bag-of-Words κωδικοποιεί με αμφίσημο τρόπο τα τοπικά χαρακτηριστικά ενός βίντεο, αφού δεν περιέχει πληροφορία σχετικά με σχετική τους διάταξη. Έτσι, π.χ. δύο χαρακτηριστικά που αντιστοιχούν στην ίδια οπτική λέξη αλλά εμφανίζονται σε τελείως διαφορετική θέση ή με μεγάλη χρονική καθυστέρηση,



Σχήμα 3.6: Σχηματική αναπαράσταση χωρο-χρονικών πυραμίδων (Σχήμα από [29]).

θα κωδικοποιηθούν με την ίδια οπτική λέξη. Οι χωρο-χρονικές πυραμίδες ενσωματώνουν πληροφορία σχετικά με τη διάταξη των χαρακτηριστικών, διαιρώντας τον τρισδιάστατο όγκο ενός βίντεο σε μικρότερους υπο-όγκους/κελιά και υπολογίζοντας ένα ξεχωριστό ιστόγραμμα (Bag-of-Words) για κάθε τέτοιο τρισδιάστατο κελί. Τα επιμέρους διανύσματα συνενώνονται (concatenation) σε ένα ενιαίο διάνυσμα, το οποίο κανονικοποιείται αρχικά με το συνολικό πλήθος των χαρακτηριστικών και εν συνεχεία με την $L2$ νόρμα, σχηματίζοντας την τελική αναπαράσταση του βίντεο (βλ. Σχήμα 3.6). Σημειώνουμε πως τα ιστογράμματα των επιμέρους κελιών υπολογίζονται χρησιμοποιώντας το ίδιο οπτικό λεξικό.

3.4.2.3 Διάνυσμα τοπικά συσσωρευμένων περιγραφητώ (VLAD)

Το διάνυσμα τοπικά συσσωρευμένων περιγραφητών (Vector of Locally Aggregated Descriptors - VLAD) αναπτύχθηκε από τους Jegou et al. [52] ως προέκταση του μοντέλου Bag-of-Words και έχει στόχο να προσεγγίσει μια "μέση λύση" μεταξύ του BoW και του διανύσματος Fisher (βλ. επόμενη παράγραφο) ως προς τη διάσταση και τη διακριτική ικανότητα.

Για τον υπολογισμό του VLAD χρησιμοποιείται κι εδώ ένα οπτικό λεξικό $\mathcal{D} = \{\mathbf{d}_1, \ldots, \mathbf{d}_K\}$, που προκύπτει συνήθως από τον K-means. Κάθε διάνυσμα ανατίθεται στην κοντινότερη οπτική λέξη του οπτικού λεξικού. Αντί ωστόσο, να καταγράφεται η συχνότητα εμφάνισης κάθε οπτικής λέξης εντός του βίντεο, το VLAD αθροίζει τις διαφορές κάθε οπτικής λέξης από τα χαρακτηριστικά που έχουν ανατεθεί σε αυτή. Έστω, λοιπόν, $C(k)$ το σύνολο των χαρακτηρτιστικών που έχουν ανατε-

θεί στο κέτρο k:

$$\mathcal{C}_k = \mathbf{x} \text{ τέτοια ώστε } \underset{k}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{d_k}\| = k. \tag{3.24}$$

Αν $\mathbf{v_k}$ το κομμάτι του VLAD που αντιστοιχεί στην οπτική λέξη k και D η διάσταση των χαρακτηριστικών μας, κάθε στοιχείο του v_k υπολογίζεται ως εξής:

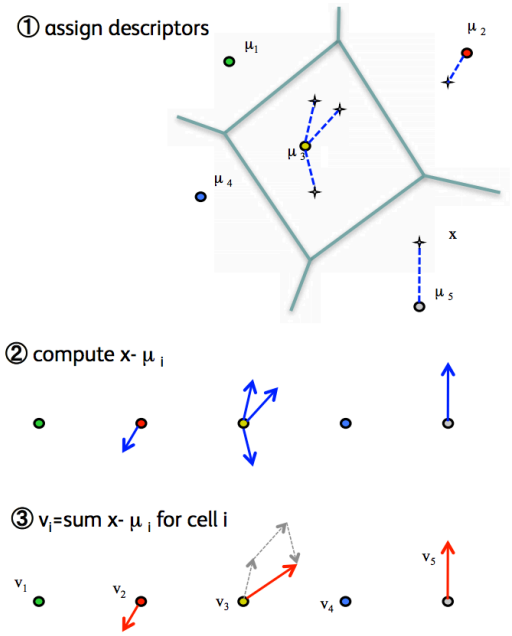
$$v_k(i) = \sum_{\mathbf{x} \in \mathcal{C}_i} \|\mathbf{x}(i) - \mathbf{d_k}(i)\|, \ i = 1 \dots D. \tag{3.25}$$

Το VLAD σχηματίζεται συνενώνοντας τα επιμέρους κομμάτια όλων των οπτικών λέξεων $k = 1 \dots K$:

$$\mathbf{v} = \begin{bmatrix} \mathbf{v}_1^T & \mathbf{v}_2^T & \cdots & \mathbf{v}_K^T \end{bmatrix}^T \tag{3.26}$$

και κανονικοποιώντας το διάνυσμα $\mathbf{v}$. Αρχικά οι συγγραφείς πρότειναν $L2$ κανονικοποίηση, ωστόσο στη συνέχεια προτάθηκε επιπλέον κανονικοποίηση σεσημασμένης τετραγωνικής ρίζας [53] (Signed Square Root - SSR), δηλαδή κανονικοποιήση κάθε στοιχείου $v(i)$ με την ποσότητα $sign(v(i))\sqrt{|v(i)|}$. Η τελευταία αντιμετωπίζει περιπτώσεις στις οποίες οι τιμές του VLAD που αντιστοιχούν σε κάποια οπτική λέξη έχουν δυσανάλογα μεγάλη τιμή ("bursty features"). Αυτό συμβαίνει συνήθως λόγω της επαναλαμβανόμενης παρουσίας κάποιας συγκεκριμένης δομής στο βίντεο (π.χ. κάποιο μοτίβο όπως ριγέ κτλ) και έχει ως αποτέλεσμα να υποβαθμίζεται η συνεισφορά των υπόλοιπων οπτικών λέξεων στην αναπαράσταση. Τέλος, οι Arandjelovic et al. [54] πρότειναν τη λεγόμενη "Intra Normaization", σύμφωνα με την οποία κάθε επιμέρους κομμάτι v_k που αντιστοιχεί μία οπτική λέξη k κανονικοποιείται αυτόνομα με την $L2$ νόρμα του προτού γίνει το concatenation (3.26) και το συνενωμένο διάνυσμα κανονικοποιείται ξανά με την $L2$ νόρμα. Έδειξαν πως αυτή η τεχνική αντιμετωπίζει πιο αποτελεσματικά το φαινόμενο των bursty features.

Όπως είναι φανερό, το VLAD κωδικοποιεί τα χαρακτηριστικά ενός βίντεο με ένα διάνυσμα διάστασης $K \cdot D$, η οποία είναι αρκετά μεγαλύτερη από την αντίστοιχη του BoW, οπότε καταλήγει σε αυξημένες απαιτήσεις μνήμης. Γι' αυτό, συχνά προηγείται προεπεξεργασία των χαρακτηριστικών με PCA. Από την άλλη πλευρά, το VLAD αποτελεί μια πιο συμπαγή αναπαράσταση από το BoW, μιας και αποθηκεύει περισσότερη πληροφορία σχετικά με κάθε οπτική λέξη. Ως εκ τούτου, συνήθως χρησιμοποιείται μικρότερο λεξικό σε σχέση με το BoW, συνήθως μεγέθους 256 οπτικών λέξεων.



Σχήμα 3.7: Υπολογισμός του VLAD (Σχήμα από `http://prateekvjoshi.com`). Αρχικά τα χαρακτηριστικά που εξάγονται ανατίθενται σε μια οπτική λέξη. Στη συνέχεια υπολογίζεται η απόσταση κάθε χαρακτηριστικού από την οπτική λέξη στην οποία ανατέθηκε και τέλος, για αθροίζονται οι αποστάσεις αυτές για κάθε cluster.

3.4.2.4 Διάνυσμα Fisher

Το διάνυσμα Fisher εισήχθη από τους Perronnin et al. [55] αποτελεί μια μέθοδο αναπαράστασης εμπνευσμένη από τον μετρικό Fisher (Fisher Information metric) που συνδυάζει πλεονεκτήματα τόσο των generative όσο και των discriminative μοντέλων. Η κεντρική ιδέα είναι η μοντελοποίηση της κατανομής (συνάρτησης πυκνότητας πιθανότητας) που ακολουθούν τα χαμηλού επιπέδου χαρακτηριστικά ενός βίντεο με ένα παραμετρικό μοντέλο και η χρήση της κλίσης (gradient) της λογαριθμικής πιθανοφάνειας (log-likelihood) ως προς τις παραμέτρους της για την αναπαράσταση του βίντεο. Επί της ουσίας η αναπαράσταση αυτή μας δίνει την κατεύθυνση στο χώρο των παραμέτρων προς την οποία πρέπει να "κινηθεί" η το μοντέλο μας προκειμένου να ταιριάξει καλύτερα στα δεδομένα.

Συγκεκριμένα, έστω ότι γνωρίζουμε πως τα χαρακτηριστικά μας αποτελούν τυχαίες μεταβλητές με συνάρτηση πυκνότητας πιθανότητας (probability density function - pdf) $p(\mathbf{X}|\lambda)$ με παραμέτρους λ. Ένα σύ-

νολο χαρακτηριστικών διάστασης D που έχουν εξαχθεί από ένα βίντεο μπορούν να χαρακτηριστούν από το κανονικοποιημένο διάνυσμα κλίσης:

$$\mathbf{F}_{\lambda}^{-1/2}\nabla_{\lambda}\log p\left(\mathbf{X}|\lambda\right),$$

όπου $\mathbf{F}_{\lambda}$ είναι μια προσέγγιση του "πίνακα Fisher" (Fisher information matrix) της $p\left(\mathbf{X}|\lambda\right)$, που χρησιμοποιείται ως παράγοντας κανονικοποίησης ώστε κάθε διάσταση της πιθανοφάνειας να έχει μεταβλητότητα 1 (whitening factor).

Θεωρώντας πως το παραμετρικό μοντέλο μας είναι ένα μίγμα γκαουσιανών κατανομών (GMM) με K γκαουσιανές και υποθέτοντας πως τα χαρακτηριστικά $\mathbf{x}$ είναι στατιστικά ανεξάρτητα, προκύπτει η πιθανοφάνεια (log-likelihood):

$$\mathcal{L}\left(\mathbf{X}|\lambda\right)=\frac{1}{N}\sum_{i=1}^{N}\log p\left(\mathbf{x}_{i}|\lambda\right),$$

όπου η $p\left(\mathbf{x}_{i}|\lambda\right)$ δίνεται από:

$$p\left(\mathbf{x}_{i}|\lambda\right)=\sum_{k=1}^{K}w_{k}p_{k}\left(\mathbf{x}_{i}|\lambda\right)$$

και όπου $p_{k}\left(\mathbf{x}_{i}|\lambda\right)$ η k-οστή γκαουσιανή του GMM:

$$p_{k}\left(\mathbf{x}|\lambda\right)=\frac{1}{(2\pi)^{L/2}|\mathbf{\Sigma}|^{1/2}}exp\left\{-\frac{1}{2}\left(\mathbf{x}-\mu_{k}\right)^{T}\mathbf{\Sigma}_{k}^{-1}\left(\mathbf{x}-\mu_{k}\right)\right\}.$$

Τα βάρη w_{k} υπόκεινται στον περιορισμό:

$$\sum_{k=1}^{K}w_{k}=1.$$

Το σύνολο των γκαουσιανών $p_{k},k=1\ldots K$ αποτελεί το οπτικό μας λεξικό, του οποίου οι παράμετροι $\lambda=\{w_{k},\mu_{k},\mathbf{\Sigma}_{k},k=1\ldots K\}$ υπολογίζονται με maximum likelihood, όπου $w_{k},\mu_{k},\mathbf{\Sigma}_{k}$ το βάρος, το διάνυσμα της μέσης τιμής και ο πίνακας συμεταβλητότητας της k-οστής γκαουσιανής αντίστοιχα. Έστω $\gamma_{i}\left(k\right)$ η "ευθύνη" (responsibility) της γκαουσιανής k για το χαρακτηριστικό $\mathbf{x}_{i}$:

$$\gamma_{i}\left(k\right)=p\left(k|\mathbf{x}_{i},\lambda\right)=\frac{w_{k}p_{k}\left(\mathbf{x}_{i}|\lambda\right)}{\sum_{j=1}^{K}w_{j}p_{j}\left(\mathbf{x}_{i}|\lambda\right)}.$$

Η $\gamma_i(k)$ αποτελεί ουσιαστικά την πιθανότητα του $\mathbf{x}_i$ να έχει "παραχθεί" από την k-οστή γκαουσιανή, δηλαδή ο βαθμός συμμετοχής ή η αλλιώς η "*χαλαρή ανάθεση*" (soft assignment) του χαρακτηριστικού $\mathbf{x}_i$ στην k-οστή γκαουσιανή. Θεωρώντας διαγώνιο πίνακα συμεταβλητότητας και υπολογίζοντας το διάνυσμα κλίσης της log-likelihood $\mathcal{L}(\mathbf{X}|\lambda)$ ως προς $w_k, \mu_k, \mathbf{\Sigma}_k$ προκύπτει:

$$\mathcal{G}^{\mathbf{X}}_{\mathbf{w},k} = \frac{\partial \mathcal{L}(\mathbf{X}|\lambda)}{\partial \mathbf{w_k}} = \sum_{i=1}^{N} \left[\frac{\gamma_i(k)}{w_k} - \frac{\gamma_i(1)}{w_1} \right] \text{ για } k \geqslant 2, \tag{3.27}$$

$$\mathcal{G}^{\mathbf{X}}_{\mu,k} = \frac{\partial \mathcal{L}(\mathbf{X}|\lambda)}{\partial \mu_k} = \frac{1}{N\sqrt{w_k}} \sum_{i=1}^{N} \gamma_i(k) \mathbf{\Sigma}_k^{-1/2} (\mathbf{x}_i - \mu_k), \tag{3.28}$$

$$\mathcal{G}^{\mathbf{X}}_{\sigma,k} = \frac{\partial \mathcal{L}(\mathbf{X}|\lambda)}{\partial \sigma_k} = \frac{1}{T\sqrt{2w_k}} \sum_{i=1}^{N} \gamma_i(k) \left[(\mathbf{x}_i - \mu_k)^N \mathbf{\Sigma}_k^{-1/2} (\mathbf{x}_i - \mu_k) - 1 \right]. \tag{3.29}$$

Το διάνυσμα Fisher $\mathcal{G}^{\mathbf{X}}_{\lambda}$ σχηματίζεται συνενώνοντας (concatenation) όλα τα ζευγάρια $\mathcal{G}^{\mathbf{X}}_{\mu,k}$ και $\mathcal{G}^{\mathbf{X}}_{\sigma,k}$ για $k = 1 \dots K$:

$$\text{FV}: \quad \mathcal{G}^{\mathbf{X}}_{\lambda} = \left[\left(\mathcal{G}^{\mathbf{X}}_{\mu,1}\right)^T \quad \left(\mathcal{G}^{\mathbf{X}}_{\sigma,1}\right)^T \quad \cdots \quad \left(\mathcal{G}^{\mathbf{X}}_{\mu,K}\right)^T \quad \left(\mathcal{G}^{\mathbf{X}}_{\sigma,K}\right)^T \right]^T$$

και κανονικοποιώντας με την L_2 νόρμα καθώς και με ύψωση σε δύναμη (power normaliztaion) και συγκεκριμένα σεσημασμένη τετραγωνική ρίζα (Signed Square Root - SSR). Η εν λόγω κανονικοποίηση διαδραματίζει σημαντικό ρόλο, όπως έδειξαν οι συγγραφείς αργότερα [53] και ανέδειξε την αποτελεσματικότητα του Fisher vector. Η κλίση ως προς τα βάρη w_k συνήθως παραλείπεται, μιας και δε συνεισφέρει στη διακριτική ικανότητα της μεθόδου.

Όπως φαίνεται από την τελευταία σχέση, η τελική αναπαράσταση με το διάνυσμα Fisher έχει διάσταση $2 \cdot D \cdot K$, η οποία είναι πολύ μεγαλύτερη από την αντίστοιχη αναπαράσταση Bag-of-Features. Ως εκ τούτου, συχνά εφαρμόζεται PCA προκειμένου να μειωθεί η υπολογιστική πολυπλοκότητα που απορρέει από τη μεγαλύτερη διάσταση. Επίσης, το επιπλέον υπολογιστικό κόστος μετριάζεται εν μέρει από το γεγονός ότι το διάνυσμα Fisher συνδυάζεται συνήθως με γραμμικά SVMs. Η αναπαράσταση με Bag-of-Words και με το διάνυσμα Fisher συνδέονται από το γεγονός πως η πρώτη λαμβάνει υπ' όψη στατιστικά μεγέθη "μηδενικού βαθμού" (συχνότητα εμφάνισης), ενώ η δεύτερη υπολογίζει

και διαφορές πρώτης και δεύτερης τάξης (σχέσεις (3.28) και (3.29)). Εναλλακτικά το ιστόγραμμα Bag-of-Words μπορεί να θεωρηθεί ως ειδική περίπτωση του διανύσματος Fisher αν συμξπεριληφθούν μόνο οι παράγωγοι ως προς τα βάρη w_k του GMM (σχέση (3.27)).

3.5 Ταξινόμηση με Μηχανές Διανυσμάτων Υποστήριξης

3.5.1 Επισκόπηση

Στην ενότητα αυτή δίνουμε μια αδρή περιγραφή της μεθόδου ταξινόμησης με Μηχανές Διανυσμάτων Υποστήριξης. Για τη διατήρηση της ροής του κειμένου, η ανάλυση εδώ είναι κάθε άλλο παρά εξαντλητική και έχει σκοπό την υπενθύμιση κάποιων βασικών εννοιών. Παραπέμπουμε τον ενδιαφερόμενο αναγνώστη στα [56], [57].

Οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines, εν συντομία SVM) είναι δημοφιλείς ταξινομητές με αμέτρητες εφαρμογές σε προβλήματα επιβλεπόμενης μάθησης. Η ευρεία χρήση τους οφείλεται κυρίως στη δυνατότητα γενίκευσης που προσφέρουν. Συγκεκριμένα, θεωρούμε δυαδικό (binary) πρόβλημα ταξινόμησης με ένα σύνολο D-διάστατων γραμμικά διαχωρίσιμων δεδομένων εκπαίδευσης $\mathcal{X} = \{\mathbf{x}_i, y_i\}$, $\mathbf{x}_i \in \mathbb{R}^D$, $y_i \in \{-1, 1\}$, όπου y_i η ετικέτα (label) του $\mathbf{x}_i$ που μας πληροφορεί αν ανήκει στην πρώτη ($y_i = -1$) ή τη δεύτερη ($y_i = 1$) κατηγορία. Η εκπαίδευση του SVM συνίσταται στην εύρεση του βέλτιστου υπερεπιπέδου στο χώρο $\mathbb{R}^D$ το οποίο διαχωρίζει τα δεδομένα. Αν $\mathbf{w} \cdot \mathbf{x} + b$ το βέλτιστο αυτό υπερεπίπεδο, θα πρέπει να ισχύει:

$$\begin{aligned} \mathbf{w} \cdot \mathbf{x}_i + b < 0 \text{ για } y_i < 0, \\ \mathbf{w} \cdot \mathbf{x}_i + b > 0 \text{ για } y_i > 0. \end{aligned} \tag{3.30}$$

Μετά την εκπαίδευση του SVM προκύπτουν τα λεγόμενα "διανύσματα υποστήριξης" (support vectors) τα οποία είναι ένα "σημαίνον" υποσύνολο των δεδομένων εκπαίδευσης, υπό την έννοια ότι βρίσκονται πιο κοντά στο βέλτιστο υπερεπίπεδο. Η απόσταση ενός διανύσματος από το υπερεπίπεδο (το πρόσημο της οποία δείχνει την κατηγορία στην οποία ανήκει) μπορεί να εκφραστεί ως γραμμικός συνδυασμός εσωτερικών γινομένων μεταξύ των διανυσμάτων υποστήριξης,

Στην συνήθη περίπτωση στην οποία τα δεδομένα δεν είναι γραμμικά διαχωρίσιμα, εισάγονται οι λεγόμενες "μεταβλητές χαλάρωσης" (slack

variebles) στο δεξί μέλος των ανισοτήτων 3.30, επιτρέποντας έτσι ορισμένα "λάθη" στον ταξινομητή. Το υπερεπίπεδο που προκύπτει, είναι αυτό που ελαχιστοποιεί αυτά τα λάθη.

Ένας επιπλέον λόγος της ευρείας αποδοχής των SVM, είναι ότι μπορούν να βρίσκουν *μη-γραμμικές* διαχωριστικές επιφάνειες. Αυτό γίνεται απεικονίζοντας τα δεδομένα $\mathbf{x}_i$ από τον ευκλείδειο χώρο σε έναν άλλο χώρο εσωτερικού γινομένου μέσω μιας συνάρτησης πυρήνα $\Phi(\mathbf{x}_i)$. Η $\Phi(\cdot)$ δεν είναι αναγκαίο να είναι καθορισμένη, αρκεί να γνωρίζουμε πώς ορίζεται το εσωτερικό γινόμενο στο νέο χώρο. Έτσι, τα εσωτερικά γινόμενα μεταξύ των διανυσμάτων που απαιτούνται για τον προσδιορισμό του βέλτιστου υπερεπιπέδου δίνονται μέσω του *πυρήνα*: $\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j) = \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)$.

Για την ταξινόμηση δεδομένων που βασίζονται σε ιστογράμματα, όπως η αναπαράσταση Bag-of-Words, χρησιμοποιείται συχνά ο πυρήνας χ^2. Ο πυρήνας αυτός υπολογίζει την "απόσταση" μεταξύ δύο ιστογραμμάτων $\mathbf{h}_i, \mathbf{h}_j$ ως εξής:

$$D\left(\mathbf{h}_i, \mathbf{h}_j\right) = \sum_{k=1}^{K} \frac{\left(\mathbf{h}_i^k - \mathbf{h}_j^k\right)^2}{\mathbf{h}_i^k + \mathbf{h}_j^k}, \tag{3.31}$$

όπου $\mathbf{h}_i^k$ το k - οστό στοιχείο του ιστογράμματος $\mathbf{h}_i$.

Όπως είδαμε, το SVM στην απλή μορφή του εκτελεί δυαδική (binary) ταξινόμηση. Έχουν προταθεί αρκετές εναλλακτικές για την ταξινόμηση σε πολλαπλές κλάσεις (βλ. ενότητα 5.3 του [49]), όπως οι μέθοδοι "ένας εναντίον ενός" (One-against-One) ή "ένας εναντίον όλων" (One-against-All). Εμείς χρησιμοποιούμε την τελευταία, η οποία συνίσταται στην εκπαίδευση ενός ταξινομητή για κάθε κλάση. Κάθε τέτοιος ταξινομητής "αποφασίζει" κατά πόσο ένα δεδομένο-δείγμα ανήκει στην αντίστοιχη κλάση, υπολογίζοντας τη σχετική πιθανότητα. Έτσι, όταν θέλουμε να ταξινομήσουμε το δείγμα σε μια από τις πολλαπλές κλάσεις, υπολογίζουμε με την πιθανότητα που έχει να ανήκει σε κάθε μια ξεχωριστά μέσω του αντίστοιχου SVM και διαλέγουμε την κλάση με τη μεγαλύτερη πιθανότητα.

3.5.2 Σύμμειξη πολλαπλών καναλιών πληροφορίας

Όπως έχει γίνει σαφές, ένα βίντεο μπορεί να αναπαρασταθεί με πολλούς τρόπους (π.χ. διαφορετικούς περιγραφητές, διαφορετική κωδικοποίηση, κ.α.), αξιοποιώντας πολλές φορές διαφορετική πληροφορία που υπάρχει στο βίντεο. Μπορούμε, λοιπόν, να εκμεταλλευτούμε την εν δυνάμει συμπληρωματικότητα των διαφόρων καναλιών πληροφορίας, προ-

κειμένου να επιτύχουμε μια πληρέστερη αναπαράσταση και ένα καλύτερο αποτέλεσμα ταξινόμησης. Χαρακτηριστική περίπτωση είναι αυτή των διαφόρων περιγραφητών, όπως οι HoG και HoF, οι οποίοι κωδικοποιούν πληροφορία σχετικά με την εμφάνιση (σχήμα) και την κίνηση αντίστοιχα. Αρκετές μέθοδοι έχουν χρησιμοποιηθεί στη βιβλιογραφία για το συνδυασμό διαφορετικών περιγραφητών, όπως οι παρακάτω:

1. Συνδυασμός στο επίπεδο των περιγραφητών (early fusion): συνιστά απλή συνένωση (concatenation) των περιγραφητών. Ο περιγραφητής που προκύπτει από τη συνένωση χρησιμοποιείται ως ξεχωριστός περιγραφητής και δίνεται ως είσοδος στο επόμενο στάδιο της επεξεργασίας. Χρησιμοποιείται συχνά για το συνδυασμό των MBHx και MBHy.

2. Συνδυασμός στο επίπεδο της αναπαράστασης (representation level fusion): αφού κωδικοποιηθούν τα χαρακτηριστικά (π.χ. με BoW), οι διαφορετικές αναπαραστάσεις συνδυάζονται με απλή συνένωση. Η μέθοδος αυτή χρησιμοποιείται για το συνδυασμό των BoW ιστογραμμάτων που υπολογίζονται στα επιμέρους κελιά στις χωροχρονικές πυραμίδες (βλ. ενότητα 3.4.2.2), καθώς και για το συνδυασμό των αναπαραστάσεων VLAD και Fisher Vector που αντιστοιχούν σε διαφορετικούς περιγραφητές.

3. Σύμμειξη τελικού σταδίου (late fusion): τα σκορ προκύπτουν από την ταξινόμηση ενός βίντεο με SVM και αντιστοιχούν σε διαφορετικά κανάλια (π.χ. διαφοερικούς περιγραφητές) συνδυάζονται, συνήθως με κάποιο γραμμικό συνδυασμό, ώστε να γίνει η τελική απόφαση.

4. Συνδυασμός στο επίπεδο του μοντέλου: το κάθε κανάλι πληροφορίας χρησιμοποιείται για τον υπολογισμό των αποστάσεων μεταξύ των βίντεο ή ενός πυρήνα του SVM και οι επιμέρους αποστάσεις ή πυρήνες συνδυάζονται σχηματίζοντας τον τελικό πυρήνα που χρησιμοποιείται. Πιο συγκεκριμένα, αν $\mathbf{x}_i^c, \mathbf{x}_j^c$ οι αναπαραστάσεις (π.χ. BoW) δύο διαφορετικών βίντεο i και j υπολογισμένες στο c-οστό κανάλι (π.χ. με τον c-οστό descriptor), ο συνδυαστικός πυρήνας προκύπτει ως εξής:

$$K(\mathbf{x}_i^c, \mathbf{x}_j^c) = \exp\left(-\sum_{c=1}^{N_c} \frac{1}{A^c} D\left(\mathbf{x}_i^c, \mathbf{x}_j^c\right)\right), \tag{3.32}$$

όπου N_c ο αριθμός των διαφορετικών καναλιών και $D(\cdot,\cdot)$ το μετρικό που χρησιμοποιείται για την υπολογιστεί η "ομοιότητα" με-

ταξύ δύο βίντεο (π.χ. χ^2 στην περίπτωση του BoW). Στην παραπάνω σχέση, οι αποστάσεις μεταξύ $\mathbf{x}_i$ και $\mathbf{x}_j$ για τα διάφορα κανάλια αθροίζονται κι εν συνεχεία υπολογίζεται ο πυρήνας του SVM. Εναλλακτικά, η άθροιση μπορεί να γίνει και μετά τον υπολογισμό του πυρήνα, δηλαδή:

$$K(\mathbf{x}_i^c, \mathbf{x}_j^c) = \sum_{c=1}^{N_c} \exp\left(-\frac{1}{A^c} D\left(\mathbf{x}_i^c, \mathbf{x}_j^c\right)\right). \tag{3.33}$$

Στην παρούσα διπλωματική χρησιμοποιούμε την (3.32) για το συνδυασμό των επιμέρους περιγραφητών. Επίσης βασιζόμαστε στην (3.33) για να συνδυάσουμε με βέλτιστο τρόπο τους επιμέρους πυρήνες μέσω της *Μάθησης πολλαπλών πυρήνων* (Multiple Kernel Learning - MKL), η οποία αναλύεται στη συνέχεια.

3.5.3 Εκμάθηση πολλαπλών πυρήνων (Multiple Kernel Learning)

Η εκμάθηση πολλαπλών πυρήνων (Multiple Kernel Learning - MKL) στοχεύει στην εύρεση του κατάλληλου συνδυασμού μεταξύ διαφορετικών πυρήνων του SVM. Η ιδέα αυτή εδράζεται στην υπόθεση των συμπληρωματικών (complementary) καναλιών πληροφορίας, τα οποία συνδυαζόμενα κατάλληλα μπορούν να αυξήσουν την επίδοση του SVM. Εκτός αυτού, η MKL μας προσφέρει έναν αυτόματο τρόπο να αξιολογήσουμε τη τη διακριτική ικανότητα κάθε καναλιού. Για παράδειγμα, ένα μεγαλύτερο βάρος στο κανάλι του περιγραφητή HoG μας πληροφορεί πως η πληροφορία σχετικά με την εμφάνιση και το σχήμα βοηθά στο διαχωρισμό μεταξύ κάποιων κατηγοριών. Στην παρούσα ενότητα θα παρουσιάσουμε συνοπτικά δύο αντιπροσωπευτικές τεχνικές που έχουν αναφερθεί στη βιβλιογραφία, διατυπώνοντας μόνο τις κύριες σχέσεις που βοηθούν στην κατανόηση της κεντρικής ιδέας, δεδομένου πως οι μέθοδοι αυτές ανήκουν σε διαφορετικό επιστημονικό πεδίο. Παραπέμπουμε, λοιπόν, τον ενδιαφερόμενο αναγνώστη στις αντίστοιχες δημοσιεύσεις, καθώς και στα [56], [57] για το σχετικό θεωρητικό υπόβαθρο των SVM.

Σκοπός μας, λοιπό, είναι να βρούμε έναν πυρήνα $k(\mathbf{x}_i, \mathbf{x}_j)$ που αποτελεί γραμμικό συνδυασμό N_c διαφορετικών πυρήνων:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sum_{c=1}^{N_c} \mathcal{K}_c(\mathbf{x}_i, \mathbf{x}_j). \tag{3.34}$$

Στην ενότητα 4.17 του [57] βλέπουμε πως η απόσταση από το το βέλτιστο υπερεπίπεδο του SVM για ένα διάνυσμα $\mathbf{x}$ στη μη-γραμμική περίπτωση δίνεται από την παρακάτω σχέση, την οποία επαναδιατυπώνουμε:

$$g(x) = \mathbf{w}^T \Phi(\mathbf{x}) + b = f(\mathbf{x}) + b = \sum_{i=1}^{N_s} \alpha_i \mathcal{K}(\mathbf{x}_i, \mathbf{x}) + b, \tag{3.35}$$

όπου $\mathbf{x}_i$ τα διανύσματα υποστήριξης και $\mathcal{K}(\mathbf{x}_i, \mathbf{x})$ η συνάρτηση πυρήνα που ικανοποιεί το θεώρημα Mercer και συνδέεται με την απεικόνιση $\Phi(\cdot)$. Το "πρωταρχικό" (primal) πρόβλημα βελτιστοποίησης γράφεται ως εξής:

$$\begin{aligned} \min_{f,b,\xi} \quad & \frac{1}{2} \|f\|^2 + C \sum_{i=1}^{N_s} \xi_i \\ \text{s.t.} \quad & y_i \left(f(\mathbf{x}_i) + b \right) \geqslant 1 - \xi_i, \quad \forall i \\ & \xi_i \geqslant 0, \quad \forall i. \end{aligned} \tag{3.36}$$

Στο πλαίσιο του MKL, αναζητούμε μια συνάρτηση απόφασης της μορφής:

$$g(\mathbf{x}) = \sum_{c=1}^{N_c} f_c(\mathbf{x}) + b = \sum_{c=1}^{N_c} \sum_{i=1}^{N_s} \alpha_i \mathcal{K}_c(\mathbf{x}_i, \mathbf{x}) + b, \tag{3.37}$$

όπου f_c οι συναρτήσεις απόφασης που αντιστοιχούν στον c-οστό πυρήνα $\mathcal{K}_c$.

Simple MKL Η μέθοδος αυτή εισήχθη το 2008 από τους Rakotomamonjy et al. [58], οι οποίοι επαναδιατύπωσαν το πρόβλημα βελτιστοποίησης (3.36) ως εξής:

$$\begin{aligned} \min_{f,b,\xi} \quad & \frac{1}{2} \sum_{c=1}^{N_c} \frac{1}{d_c} \|f_c\|^2 + C \sum_{i=1}^{N_s} \xi_i \\ \text{s.t.} \quad & y_i \left(\sum_{c=1}^{N_c} f_c(\mathbf{x}_i) + b \right) \geqslant 1 - \xi_i, \quad \forall i \\ & \xi_i \geqslant 0, \quad \forall i \\ & \sum_{i=1}^{N_c} d_c = 1, \quad \forall i. \end{aligned} \tag{3.38}$$

Ο τελευταίος περιορισμός επιβάλει μια συνθήκη κυρτότητας στα d_c. Όπως όμως υποστηρίζουν οι συγγραφείς, το δυϊκό (dual) πρόβλημα του

(3.38) είναι δύσκολο να λυθεί, γι' αυτό διατυπώνουν το το (3.38) διαφορετικά:

$$\min_{d} J(\mathbf{d}) \quad \text{s.t.} \quad \sum_{c=1}^{N_c} d_c = 1, \quad d_c \geqslant 0, \tag{3.39}$$

όπου:

$$J(\mathbf{d}) = \begin{cases} \min\limits_{g,b,\xi} & \frac{1}{2}\sum\limits_{c=1}^{N_c} \frac{1}{d_c} \|f_c\|^2 + C\sum\limits_{i=1}^{N_s} \xi_i \\ \text{s.t.} & y_i \left(\sum\limits_{c=1}^{N_c} f_c(\mathbf{x}_i) + b \right) \geqslant 1 - \xi_i, \quad \forall i \\ & \xi_i \geqslant 0, \quad \forall i \end{cases} \tag{3.40}$$

Παρατηρούμε από την (3.40) πως υπολογισμός της τιμής της συνάρτησης $J(\mathbf{d})$ αποτελεί ουσιαστικά ένα πρόβλημα ελαχιστοποίησης της αντικειμενικής συνάρτησης ενός SVM, για τον υπολογισμό της οποίας μπορεί να χρησιμοποιηθεί οποιοσδήποτε σχετικός αλγόριθμος. Έτσι, οι συγγραφείς προτείνουν τον υπολογισμό των βέλτιστων $\mathbf{d}_c$ σε δύο βήματα: πρώτα την εύρεση της αντικειμενικής συνάρτησης d και μετά τη βελτιστοποίηση ως προς τα βάρη $\mathbf{d}_c$ με τη χρήση μεθόδου "απότομης κατάβασης" (gradient descent), την οποία αναπτύσσουν γι' αυτό το σκοπό, αποδεικνύοντας τη σύγκλιση της μεθόδου.

GMKL Μια ενδιαφέρουσα εργασία είναι αυτή των Varma & Babu [59], οι οποίοι γενίκευσαν την ιδέα του MKL σε περιπτώσεις μη-γραμμικού (πολυωνυμικού κ.α.) συνδυασμού των επιμέρους πυρήνων, ονομάζοντας τη μέθοδό τους Generalized Multiple Kernel Learning (GMKL). Αρχικά, λοιπόν, εκφράζουν το πρόβλημα (3.36) ως εξής:

$$\begin{aligned} \min_{f,b,\mathbf{d}} \quad & \frac{1}{2}\|f\|^2 + \sum_{i=1}^{N} l\left(y_i, f(\mathbf{x}_i)\right) + r\left(\mathbf{d}\right) \\ \text{s.t.} \quad & \mathbf{d} \geqslant 0, \end{aligned} \tag{3.41}$$

όπου ο παράγοντας ομαλοποίησης $r(\mathbf{d})$ μπορεί να είναι οποιαδήποτε συνεχώς παραγωγίσιμη συνάρτηση ως προς $\mathbf{d}$ και $l(\cdot)$ είναι μια συνάρτηση ποινής. Όπως φαίνεται, οι περιορισμοί ως προς τα βάρη $\mathbf{d}$ είναι αρκετά χαλαροί και μπορούν εν δυνάμει να προσαρμοστούν σε κάποια πρότερη (prior) πληροφορία που μπορεί να υπάρχει σχετικά με τους πυρήνες. Ακόμα και η απαίτηση να είναι θετικά τα βάρη μπορεί να αρθεί,

αρκεί να εξασφαλίζεται πως ο πυρήνας που προκύπτει είναι θετικά ορισμένος. Όπως και στην προηγούμενη μέθοδο, οι συγγραφείς εκφράζουν το πρόβλημα (3.41) ως δύο διακριτά προβλήματα βελτιστοποίησης:

$$\min_{\mathbf{d}} J(\mathbf{d}) \quad \text{s.t.} \quad \mathbf{d} \geqslant 0, \tag{3.42}$$

όπου:

$$J(\mathbf{d}) = \min_{f,b} \quad \frac{1}{2}\|f\|^2 + \sum_{i=1}^{N} l\left(y_i, f(\mathbf{x}_i)\right) + r\left(\mathbf{d}\right), \tag{3.43}$$

Στη συνέχεια, οι συγγραφείς αποδεικνύουν πως η $J(\mathbf{d})$ είναι παραγωγίσιμη ως προς $\mathbf{d}$, οπότε η τιμή της $J(\mathbf{d})$ μπορεί να γίνει με οποιονδήποτε αλγόριθμο εκπαίδευσης SVM. Έτσι, προτείνουν κι εδώ έναν επαναληπτικό αλγόριθμο, κατά τον οποίο εναλλάσσονται τα βήματα της εύρεσης των παραμέτρων του SVM και της βελτιστοποίησης ως προς $\mathbf{d}$ με gradient descend. Σημειώνουμε πως μια βελτιωμένη έκδοση του GMKL παρουσίασαν οι Jain et al. [60], ονόματι SPG-GMKL, στην οποία χρησιμοποιούν έναν εναλλακτικό gradient αλγόριθμο βελτιστοποίησης (Spectral Projection Algorithm), προσεγγίζοντας ταχύτερα και με μεγαλύτερη ακρίβεια το ελάχιστο της $J(\mathbf{d})$.

Κεφάλαιο 4

Πειραματικά αποτελέσματα

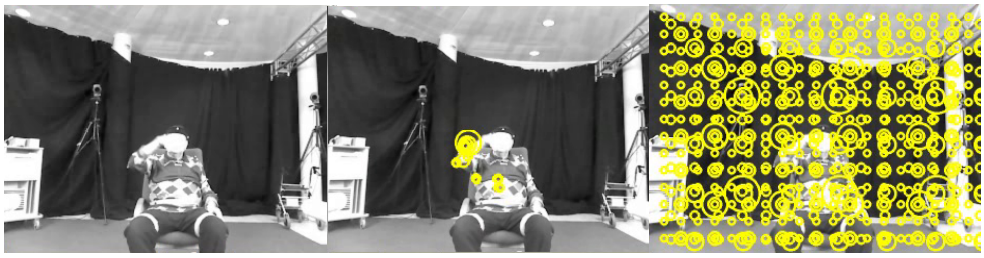
Στο κεφάλαιο αυτό παρουσιάζουμε τα πειραματικά αποτελέσματα ταξινόμησης ανθρώπινων δράσεων και χειρονομιών. Οι πρώτες δύο ενότητες αφορούν πειράματα με διαφορετικά χαρακτηριστικά, τα χωροχρονικά σημεία ενδιαφέροντος και τις πυκνές τροχιές. Στην τρίτη ενότητα αξιολογούμε την επίδραση των χωρο-χρονικών πυραμίδων στο αποτέλεσμα της αναγνώρισης. Στην τέταρτη ενότητα συγκρίνουμε μεταξύ εναλλακτικών μεθόδων υπολογισμού του οπτικού λεξικού και κωδικοποίησης.

4.1 Χωρο-χρονικά σημεία ενδιαφέροντος

Στην ενότητα αυτή περιγράφουμε τις λεπτομέρειες και τις επιμέρους παραμέτρους που χρησιμοποιήσαμε για την ταξινόμηση δράσεων και χειρονομιών με χωρο-χρονικά σημεία ενδιαφέροντος. Στη συνέχεια παρουσιάζουμε και αναλύουμε τα αποτελέσματα που προκύπτουν σε διάφορες βάσεις δεδομένων.

4.1.1 Πειραματικό Πλαίσιο

Για την εξαγωγή χωρο-χρονικών σημείων ενδιαφέροντος με τον ανιχνευτή Harris3D χρησιμοποιήσαμε τις default παραμέτρους των συγγραφέων. Συγκεκριμένα, αναζητήθηκαν σημεία σε 3 χωρικές κλίμακες, ξεκινώντας από τη μηδενική (δηλαδή με την αρχική ανάλυση) και υπολογίστηκαν οι περιγραφητές HoG, HoF, καθώς και το η συνένωσή τους HoG/HoF, σε μια χωρο-χρονική γειτονιά γύρω από τα εντοπισμένα σημεία. Αν σ και τ η κλίμακα στην οποία εντοπίστηκε κάποιο σημείο, το μέγεθος της γειτονιάς γύρω του είναι $\Delta_x(\sigma) \times \Delta_y(\sigma) \times \Delta_t(\tau)$, όπου



(α′) Αρχικό frame από τη βάση MOBOT-6a (β′) Harris3D (γ′) Πυκνή δειγματοληψία

Σχήμα 4.1: Παραδείγματα εξαγωγής χωρο-χρονικών σημείων ενδιαφέροντος στη βάση MOBOT-6a. Το μέγεθος κάθε κύκλου συναρτάται της κλίμακας στην οποία εντοπίστηκε. Η τρίτη εικόνα δίνεται μόνο για λόγους επισκόπησης, καθώς στην παρούσα διπλωματική δεν παρουσιάζονται αποτελέσματα με χρήση πυκνής δειγματοληψίας.

$\Delta_x(\sigma) = \Delta_y(\sigma) = 18\sigma$ και $\Delta_t(\tau) = 8\tau$. Κάθε τέτοια γειτονιά χωρίζεται σε ένα πλέγμα $n_x \times n_y \times n_t$ κελιών, εντός των οποίων υπολογίζονται τα ιστογράμματα ο HoG και ο HoF με 4 και 5 bins αντίστοιχα. Ο τελικός περιγραφητής κάθε γειτονιάς προκύπτει από τη συνένωση (concatenation) των περιγραφητών των επιμέρους κελιών. Χρησιμοποιούμε $n_x = n_y = 3$ και $n_t = 2$ οπότε το μέγεθος του HoG προκύπτει $4 \cdot 3 \cdot 3 \cdot 2 = 72$ και του HoF $5 \cdot 3 \cdot 3 \cdot 2 = 90$. Επίσης, συνενώνουμε τους δύο αυτούς περιγραφητές σε έναν ενιαίο HoG/HoF. Η εξαγωγή των STIP γίνεται με χρήση της υλοποίησης των συγγραφέων [1] Στο Σχήμα 4.1β′ φαίνεται ένα παράδειγμα ανίχνευσης σημείων ενδιαφέροντος με τον ανιχνευτή Harris3D.

Τα χαρακτηριστικά και οι περιγραφητές που υπολογίστηκαν κωδικοποιήθηκαν με τη μέθοδο Bag-of-Words. Η εξαγωγή του λεξικού έγινε ομαδοποιώντας 100000 τυχαία δειγματοληπτημένα διανύσματα με τον αλγόριθμο K-means, χρησιμοποιώντας την υλοποίηση της βιβλιοθήκης yael [2]. Επιλέξαμε $K = 4000$ κέντρα, αριθμός που χρησιμοποιείται συνήθως στη βιβλιογραφία και έχει δείξει να δίνει καλά αποτελέσματα. Για την ταξινόμηση με SVM χρησιμοποιήθηκε η βιβλιοθήκη LibSVM [61].

4.1.2 Αποτελέσματα

Στον Πίνακα 4.1 παρατίθενται τα αποτελέσματα ταξινόμησης σε 3 διαφορετικές βάσεις δεδομένων. Βλέπουμε πως παρά την απλότητα της

[1] `https://www.di.ens.fr/~laptev/interestpoints.html`

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

Πίνακας 4.1: Αποτελέσματα ταξινόμησης με χωρο-χρονικά σημεία ενδιαφέροντος σε 3 βάσεις δεδομένων. Στη Hollywood2 αναφέρουμε τη μέτρηση mean Average Precision (mAP), στην KTH το ποσοστό σωστής ταξινόμησης στο σύνολο εκπαίδευσης, ενώ στη MOBOT-6.a το μέσο όρο του ποσοστού σωστής ταξινόμησης για όλους του ασθενείς (βλ. επίσης Ενότητα 1.3).

	KTH	Hollywood2	MOBOT-6.a
HoG	80.8	38.7	51.1
HoF	89.3	43.6	46.4
HoG/HoF	89.3	47.4	55.6

μεθόδου, τα αποτελέσματα είναι ενθαρρυντικά. Για απλές δράσεις που εκτελούνται σε ελεγχόμενο περιβάλλον, όπως αυτό της KTH, η μέθοδος καταφέρνει να κωδικοποιήσει πληροφορία δεκάδων καρέ μέσα σε μερικές εκατοντάδες σημεία. Από το confusion matrix (Σχήμα 4.2) διαπιστώνουμε πως τα σφάλματα ταξινόμησης παρουσιάζονται κυρίως μεταξύ όμοιων στην εμφάνιση κλάσεων, όπως 'jogging'' και ''runnung'', οι οποίες διαφέρουν κυρίως ως προς την ταχύτητα εκτέλεσης. Όπως εί-

	walking	jogging	running	boxing	handwaving	handclapping
walking	0.88	0.03	0.01	0.08	0	0
jogging	0.02	0.74	0.24	0	0	0
running	0	0.12	0.88	0	0	0
boxing	0	0	0	0.99	0	0.01
handwaving	0	0	0	0	0.99	0.01
handclapping	0	0	0	0	0.09	0.91

Πίνακας 4.2: Confusion Matrix για τη βάση KTH. Η ταξινόμηση έγινε με χωρο-χρονικά σημεία ενδιαφέροντος και την περιγραφητή HoG/HoF.

ναι αναμενόμενο, τα αποτελέσματα στη Hollywood δείχνουν πως τα χωρο-χρονικά σημεία ενδιαφέροντος δεν είναι εύρωστα σε δυναμικά περιβάλλοντα και μη ελεγχόμενες συνθήκες λήψης. Είναι λογικό όταν έχουμε κινούμενα αντικείμενα και κίνηση της κάμερας, οι χωριχρονικές ''γωνίες'' που εντοπίζει ο Harris3D να μην αντιστοιχούν αναγκαία στην κίνηση του ανθρώπου. Στη βάση MOBOT-6.a, δεδομένου ότι το περιβάλλον είναι σχετικά ελεγχόμενο, θα περίμενε κανείς αποτελέσματα συγκρίσιμα με αυτά της KTH. Ωστόσο, οι χειρονομίες αποτελούν μια

κατηγορία πιο "λεπτών" ("fine grained") δράσεων και κατά συνέπεια απαιτούν πιο λεπτομερείς αναπαραστάσεις που να συλλαμβάνουν μικρές/αμυδρές διαφορές στην εμφάνιση και την κίνηση. Παρατηρούμε, επίσης, πως η επιλογή του καλύτερου περιγραφητή εξαρτάται από το τις εκάστοτε συνθήκες. Μπορούμε έτσι να συμπεράνουμε σε αδρές γραμμές τι είδους πληροφορία παίζει σημαντικότερο ρόλο κάθε φορά: για παράδειγμα στην KTH η κίνηση είναι περισσότερο ικανή να διαχωρίσει τις δράσεις μεταξύ τους, εξ' ου και η μεγαλύτερη ακρίβεια του HoF. Αντίθετα, η στατική εμφάνιση γύρω από τα σημεία ενδιαφέροντος, που κωδικοποιείται μέσω του HoG είναι σημαντική για τις χειρονομίες. Είναι λογικό ο συνδυασμός των δύο αυτών δίνει τα καλύτερα αποτελέσματα, αφού έτσι κωδικοποιείται πλουσιότερη πληροφορία.

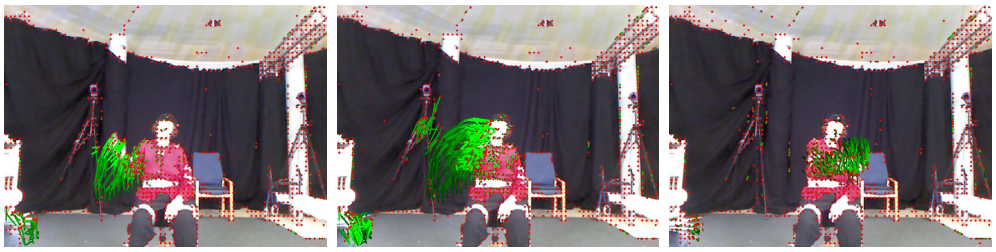
4.2 Πυκνές τροχιές

Στην ενότητα αυτή περιγράφουμε τις λεπτομέρειες υλοποίησης του συστήματος ταξινόμησης δράσεων και χειρονομιών με πυκνές τροχιές. οι τροχιές αυτές εξάγονται δειγματοληπτώντας το κάθε καρέ σε ένα σταθερό πλέγμα και παρακολουθώντας (tracking) τα σημεία αυτά στο χρόνο. Στη δεύτερη υποενότητα παραθέτουμε και αναλύουμε τα πειραματικά μας αποτελέσματα.

4.2.1 Πειραματικό Πλαίσιο

Για την εξαγωγή πυκνών τροχιών χρησιμοποιούμε την υλοποίηση των συγγραφέων [3] με το default μήκος τροχιάς $L = 15$. Οι περιγραφητές υπολογίζονται ενός ενός χωρο-χρονικού όγκου μεγέθους $N \times N \times L$ κατά μήκος κάθε τροχιάς, όπου $N = 32$. Ο όγκος αυτός χωρίζεται σε ένα πλέγμα $n_s \times n_s \times n_t$ κελιών εντός των οποίων υπολογίζεται κάθε περιγραφητής. Ο τελικός περιγραφητής που αντιστοιχεί σε κάθε τροχιά προκύπτει από τη συνένωση (concatenation) των περιγραφητών των επιμέρους κελιών. Χρησιμοποιούμε τις τιμές $n_s = 2$, $n_t = 3$. Συνολικά υπολογίζουμε 6 διαφορετικούς περιγραφητές: Περιγραφητής Τροχιάς (Trajectory Descriptor - T.D.), HoG, HoF, MBHx, MBHy, MBH. Ο MBH αποτελεί τη συνένωση των MBHx, MBHy. Επίσης, συνδυάζουμε τους επιμέρους περιγραφητές προκειμένου να εκμεταλλευτούμε την εν δυνάμει συμπληρωματικότητά τους. Ένα παράδειγμα εξαγωγής πυκνών τροχιών φαίνεται στο Σχήμα 4.2.

[3] `https://lear.inrialpes.fr/people/wang/dense_trajectories`



Σχήμα 4.2: Παράδειγμα εξαγωγής πυκνών τροχιών για την ασθενή 16 της βάσης χειρονομιών MOBOT-6a

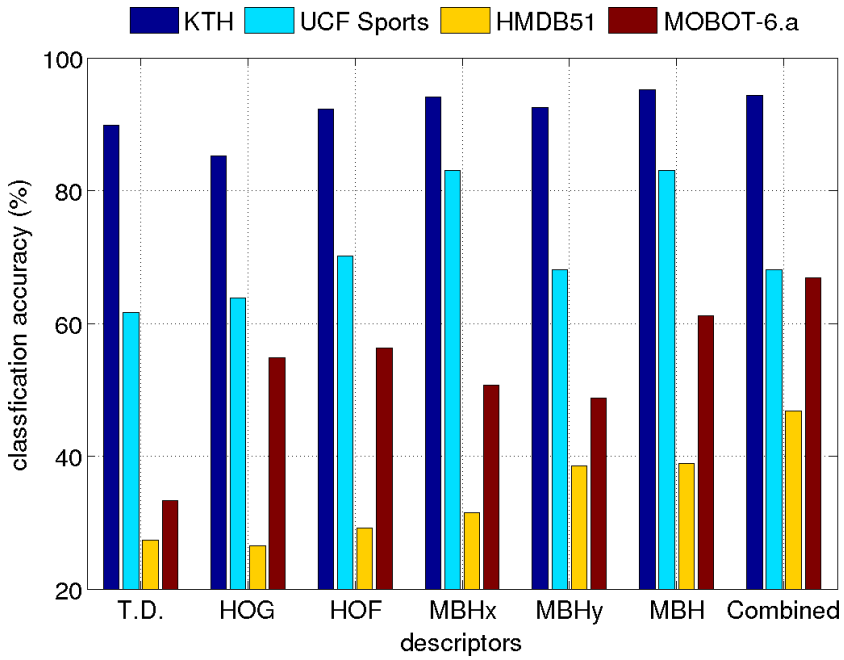
4.2.2 Αποτελέσματα

Στο Σχήμα 4.3 φαίνονται τα αποτελέσματα ταξινόμησης με τη χρήση πυκνών τροχιών σε τρεις διαφορετικές βάσεις δεδομένων: KTH, HMDB51, UCF Sports και MOBOT-6.a. Για την τελευταία έχει υπολογιστεί ο μέσος όρος της ακρίβειας ταξινόμησης όλων των ασθενών. Στον Πίνακα 4.3 φαίνονται αναλυτικά τα αποτελέσματα στη βάση MOBOT-6.a πιο για κάθε ασθενή ξεχωριστά. Στα σχήματα αναφερόμαστε στον περιγραφητή τροχιάς με το ακρωνύμιο "T.D.", ενώ στο συνδυαστικό περιγραφητή με την ονομασία "Combined". Βάσει των αποτελεσμάτων συμπεραίνουμε τα εξής: Οι συνθήκες υπό τις οποίες εκτελείται μια δράση επιδρά κατα-

Πίνακας 4.3: Πειραματικά αποτελέσματα ταξινόμησης με πυκνές τροχιές και διαφορετικούς περιγραφητές για όλους τους ασθενείς MOBOT-6.a. Η τελευταία γραμμή ("Comb.") δείχνει τα αποτελέσματα για το συνδυαστικό περγραφητή.

	p1	p4	p7	p10	p12	p13	p16	p18	avg
T.D.	37.5	40.0	38.2	33.3	29.2	28.1	41.7	21.9	33.7
HoG	56.3	32.0	38.2	78.8	62.5	21.9	70.8	78.1	54.8
HoF	59.4	64.0	35.3	69.7	58.3	31.3	66.7	65.6	56.3
MBHx	40.6	40.0	38.2	57.6	58.3	40.6	70.8	59.4	50.7
MBHy	53.1	60.0	20.6	42.4	50.0	34.4	66.7	62.5	48.7
MBH	**68.8**	64.0	44.1	**66.7**	58.3	43.8	75.0	68.8	61.2
Comb,	65.6	**72.0**	**47.1**	63.6	**75.0**	**46.9**	**83.3**	**81.3**	**66.8**

λυτικά στη δυνατότητα του συστήματος τα ταξινομήσει σωστά τις δράσεις. Βλέπουμε πως στην απλούστερη βάση δεδομένων KTH, η ακρίβεια ταξινόμησης πολύ μεγάλη και τα σφάλματα είναι κυρίως μεταξύ των δράσεων "jogging" και "running", τις οποίες ο άνθρωπος πιθανώς θα δυσκολευόταν να διακρίνει. Αντίθετα, η UCF Sports απεικονίζει δράσεις



Σχήμα 4.3: Πειραματικά αποτελέσματα ταξινόμησης με πυκνές τροχιές και διαφορετικούς περιγραφητές σε 3 βάσεις δεδομένων. Το ακρωνύμιο T.D. αντιστοιχεί στον περιγραφητή τροχιάς (Tajectory Descriptor) και ο "Combined" στο συνδυαστικό περιγραφητή.

υπό ρεαλιστικές συνθήκες, με σκιάσεις, επικαλύψεις κ.α. Αυτό αντικατοπτρίζεται στα αρκετά χαμηλότερα αποτελέσματα, παρ' όλου που περιέχει παρόμοιο αριθμό κλάσεων με την KTH 1.3.2.

Ο MBH σαφώς υπερέχει σαφώς έναντι των άλλων περιγραφητών. Η ικανότητά του να αντισταθμίζει την ομαλή κινηση της κάμερας αναδεικνύεται κυρίως στη UCF Sports. Στην πρώτη είναι σύνηθες φαινόμενο να κινείται η κάμερα καθώς παρακολουθεί κάποιον αθλούμενο. Παρ' όλη την ευρωστία του, ο MBH δεν αρκεί για να αντιμετωπίσει φαινόμενα μη-ομαλής και απότομης κίνησης της κάμερας, όπως πολλές φορές συμβαίνει στην HMDB51.

Ο συνδυαστικός περιγραφητής έχει συνεπή καλά αποτελέσματα. Αυτό συμβαίνει ακόμα κι αν κάποιοι από τους επιμέρους περιγραφητές παρουσιάζουν χαμηλά ποσοστά. Το γεγονός αυτό δείχνει πως ο συνδυασμός των περιγραφητών αξιοποιεί εν μέρει τη συμπληρωματικότητα των περιγραφητών. Αυτό είναι περισσότερο εμφανές στην HMDB51, όπου ο συνδυαστικός περιγραφητής έχει περίπου 8% διαφορά από τον αμέσως καλύτερο (MBHx). Αυτό οφείλεται στο γεγονός πως η αναγνώριση περίπλοκων δράσεων υποβοηθείται ιδιαίτερα από πληροφορία

όπως τα αντικείμενα και η σκηνή. Έτσι, τα συμπληρωματικά κανάλια, όπως π.χ. η στατική εμφάνιση που κωδικοποιείται από το HoG μπορούν να εμπλουτίσουν την τελική αναπαράσταση.

Όσον αφορά τη βάση MOBOT-6.a (Πίνακας 4.3), είναι εμφανές πως η μεγάλη μεταβλητότητα εκτέλεσης μεταξύ των χρηστών έχει σημαντικό αντίκτυπο στην ακρίβεια, το οποίο διαπιστώνει κανείς παρατηρώντας και τα δεδομένα (π.χ. ο ασθενής p13 είναι αριστερόχειρας, ο ασθενής 10 εκτελεί πολλές από τις χειρονομίες με διαφορετικό τρόπο, κ.α.). Όπως περιμέναμε, ο MBH έχει τα καλύτερα αποτελέσματα με εξαίρεση τον ασθενή p12. Επίσης, το γεγονός πως η οριζόντια συνιστώσα του MBH (MBHx) έχει καλύτερα αποτελέσματα από την κατακόρυφη (MBHy), δείχνει πως η κίνηση προς τον άξονα x διαδραματίζει σημαντικότερο ρόλο για το διαχωρισμό μεταξύ των κλάσεων. Συμπερασματικά, το πειραματικό framework της ταξινόμησης δράσεων είναι σε θέση να δώσει ικανοποιητικά αποτελέσματα σε ένα πρόβλημα ταξινόμησης χειρονομιών.

4.3 Η επίδραση διαφορετικών κωδικοποιήσεων

4.3.1 Πειραματικό Πλαίσιο

Όπως έχει αναφερθεί, η κωδικοποίηση Bag-of-Words αποτελεί μια σχετικά απλή αναπαράσταση που υπολογίζει ένα απλό στατιστικό μέγεθος των χαρακτηριστικών, χωρίς καμιά πληροφορία σχετικά με τη διάταξή τους. Στην ενότητα αυτή αξιολογούμε την επίδραση εναλλακτικών αναπαραστάσεων, όπως η Bag-of-Words με χωρο-χρονικές πυραμίδες, το VLAD και το διάνυσμα Fisher.

Οι χωρο-χρονικές πυραμίδες ενσωματώνουν εν μέρει πληροφορία αναφορικά με τη χωρική και χρονική σχέση των χαρακτηριστικών που έχουν εξαχθεί, χωρίζοντας το βίντεο σε επιμέρους υπο-όγκους/κελιά, υπολογίζοντας ένα ιστόγραμμα Bag-of-Words για κάθε ένα από αυτά και συνενώνοντάς τα στο τέλος για την τελική αναπαράσταση του βίντεο. Έτσι, μια θέση π.χ. του τελικού διανύσματος αντιστοιχεί στη συχνότητα που εμφανίστηκε μια οπτική λέξη σε συγκεκριμένο χωρικό εύρος και συγκεκριμένο χρονικό διάστημα. Για την πειραματική τους αξιολόγηση χρησιμοποιήσαμε 6 συχνά χρησιμοποιούμενα πλέγματα. Συγκεκριμένα, για τις δύο χωρικές διαστάσεις χρησιμοποιήθηκε ολόκληρο το βίντεο (1×1 μπλοκ), η υποδιαίρεσή του σε 3 χωρικές λωρίδες (3×1) και ένα χωρικό 2×2 πλέγμα. Στη διάσταση του χρόνου χρησιμοποιήθηκε ολόκληρο το βίντεο καθώς και η διαίρεσή του σε 2 μπλοκ. Αν

με h, v, t συμβολίζουμε τον αριθμό των υποδιαιρέσεων στον οριζόντιο άξονα, τον κατακόρυφο άξονα και το χρόνο αντίστοιχα, έχουμε συνολικά τις εξής 6 πυραμίδες, οι οποίες απεικονίζονται στο Σχήμα 3.6: $h1v1t1$, $h3v1t1$, $h2v2t1$, $h1v1t2$, $h3v1t2$, $h2v2t2$. Για την ταξινόμηση χρησιμοποιούμε, όπως και στην περίπτωση του απλού Bag-of-Words, μη γραμμικά SVM, συγκρίνοντας την ομοιότητα μεταξύ δύο βίντεο με την απόσταση χ^2.

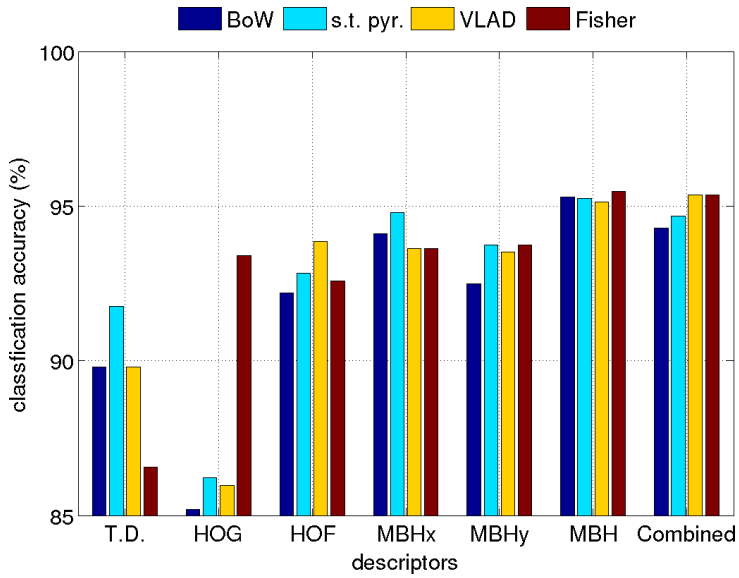
Για την ενσωμάτωση πλουσιότερης στατιστικής πληροφορίας σχετικά με τις οπτικές λέξεις, πειραματιζόμαστε με τις κωδικοποιήσεις VLAD και διάνυσμα Fisher. Για το VLAD εξάγουμε ένα οπτικό λεξικό $K = 256$ λέξεων μέσω του K-means, δειγματοληπτώντας τυχαία 100000 χαρακτηριστικά από το σύνολο εκπαίδευσης. Για το διάνυσμα Fisher εκπαιδεύουμε ένα μείγμα γκαουσιανών (GMM) με $K = 256$ γκαουσιανές με τον ίδιο αριθμό τυχαίων χαρακτηριστικών εκπαίδευσης. Η αναπαράσταση περιλαμβάνει τις κλίσεις (gradients) ως προς τη μέση τιμή και το variance. Λόγω των μεγάλων απαιτήσεων σε μνήμη που απορρέουν από τη χρήση του διανύσματος Fisher, προ-επεξεργαζόμαστε τα δεδομένα με PCA, μειώνοντας τη διάσταση του κάθε περιγραφητή κατά $1/2$, κρατώντας, δηλαδή τις μισές μόνο κύριες συνιστώσες (principal components). Παραδείγματος χάριν, η αρχική διάσταση της αναπαράστασης του HoF με διάνυσμα Fisher 3.4.2.4 είναι ίση με $2 \cdot K \cdot D = 2 \cdot 256 \cdot 108 = 55296$, ενώ η τελική $2 \cdot 256 \cdot 54 = 27648$. Για την ταξινόμηση χρησιμοποιούμε γραμμικά SVM, τόσο στην περίπτωση τόσο του VLAD όσο και του διανύσματος Fisher. Η εκπαίδευση του GMM και ο υπολογισμός του διανύσματος Fisher έγινε με την υλοποίηση της βιβλιοθήκης yael, όπως και στην περίπτωση του K-means. Για το VLAD χρησιμοποιήθηκε η βιβλιοθήκη VLFeat [4].

4.3.2 Αποτελέσματα

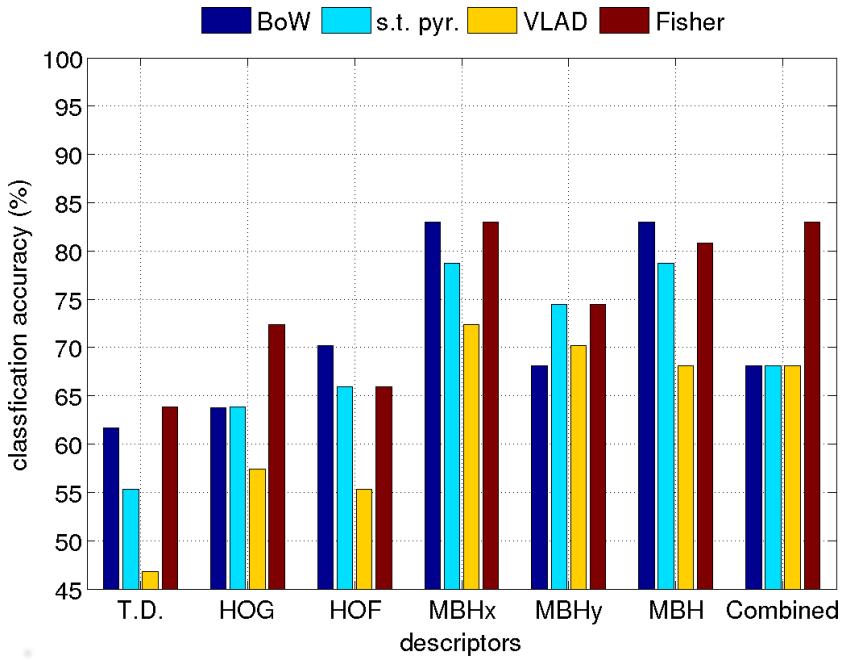
Στα Σχήματα 4.4 έως 4.7 φαίνονται τα αποτελέσματα της αναγνώρισης δράσεων για διάφορες μεθόδους κωδικοποίησης στις βάσεις KTH, UCF Sports, HMDB51 και MOBOT-6.a. Για την HMDB51 αναφέρουμε μόνο αποτελέσματα για τις χωρο-χρονικές πυραμίδες. Σε όλα τα σχήματα παρατίθενται για λόγους εύκολης σύγκρισης και τα αντίστοιχα αποτελέσματα για κωδικοποίηση με Bag-of-Words, το οποίο αποτελεί το baseline μας.

Όσον αφορά τις χωρο-χρονικές πυραμίδες, βλέπουμε πως τα αποτελέσματα είναι ελαφρώς καλύτερα εν σχέσει με αυτά που προκύπτουν

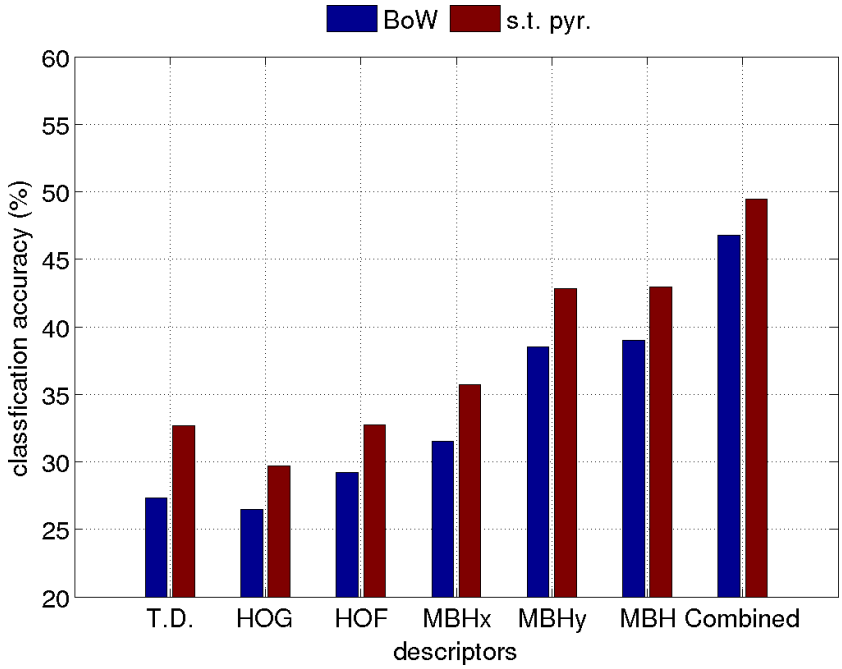
[4] `http://www.vlfeat.org`



Σχήμα 4.4: Σύγκριση 4 μεθόδων κωδικοποίησης για διαφορετικούς περιγραφητές στη βάση KTH (*s.t. pyr.*: χωρο-χρονικές πυραμίδες).



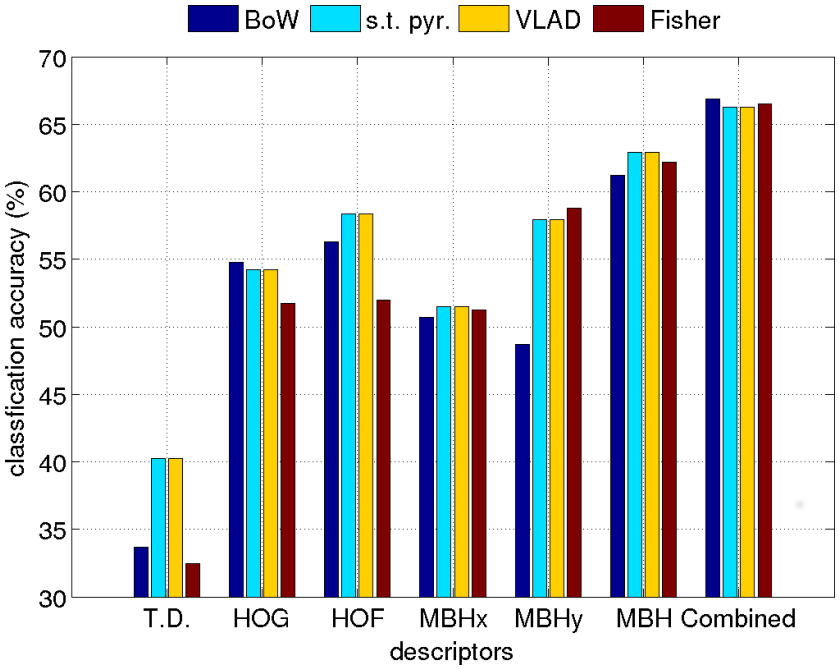
Σχήμα 4.5: Σύγκριση 4 μεθόδων κωδικοποίησης για διαφορετικούς περιγραφητές στη βάση UCF Sports (*s.t. pyr.*: χωρο-χρονικές πυραμίδες).



Σχήμα 4.6: Σύγκριση BoW και χωρο-χρονικών πυραμίδων για διαφορετικούς περιγραφητές στη βάση HMDB51 (*s.t. pyr.*: χωρο-χρονικές πυραμίδες).

από το απλό μοντέλο Bag-of-Words. Στην HMDB51 4.6 η βελτίωση αυτή φτάνει το $3,1\%$, ενώ στη UCF Sports 4.5 δεν παρατηρείται καμία διαφορά. Παρ' όλο που οι χωρο-χρονικές πυραμίδες εμπλουτίζουν την αναπαράσταση Bag-of-Words ενσωματώνοντας κάποια χωρο-χρονική δομή, δε μοντελοποιούν δυναμική πληροφορία. Αντίθετα χρησιμοποιούν προκαθορισμένες υποδιαιρέσεις του όγκου του βίντεο. Έτσι, π.χ. ίδιες δράσεις με διαφορετική διάρκεια, ή που εκτελούνται σε διαφορετική θέση στα καρέ του βίντεο, καταλήγουν να διαφέρουν σημαντικά στην αναπαράστασή τους. Γι' αυτό η επίδρασή τους είναι περιορισμένη, όπως άλλωστε επιβεβαιώνεται και στη βιβλιογραφία [62]. Αξίζει επίσης να προσθέσουμε την υπολογιστική πολυπλοκότητα που εισάγει η συγκεκριμένη μέθοδος.

Αντίθετα, η προσθήκη επιπλέον στατιστικής πληροφορίας σχετικά με τις οπτικές λέξεις, είτε με το VLAD, είτε με διάνυσμα Fisher έχει θετική επίδραση στην ακρίβεια ταξινόμησης για την KTH 4.4 και κυρίως για τη UCF Sports 4.5. Στην KTH το αποτέλεσμα είναι ήδη πολύ υψηλό, οπότε είναι δύσκολο να αναδειχθεί η επίδραση των πιο πολύπλοκων αναπαραστάσεων. Στη UCF Sports βλέπουμε πως το VLAD και κυ-



Σχήμα 4.7: Σύγκριση 4 μεθόδων κωδικοποίησης για διαφορετικούς περιγραφητές στη βάση MOBOT-6.a (*s.t. pyr.*: χωρο-χρονικές πυραμίδες).

ρίως το διάνυσμα Fisher είναι σε θέση να κωδικοποιήσουν λεπτομερείς στατιστικές σχέσεις των χαρακτηριστικών, ακόμα και με λίγες οπτικές λέξεις. Αυτό δε συμβαίνει όμως και με τη βάση MOBOT-6.a 4.7, για την οποία μόνο το VLAD παρουσιάζει καλύτερα αποτελέσματα σε σχέση με το απλό Bag-of-Words, με το Fihser Vector να δίνει περίπου 2% χαμηλότερο ποσοστό. Σημαντικό ρόλο πιθανώς έχει η προ-επεξεργασία των δεδομένων με PCA πριν τον υπολογισμό του διανύσματος FIsher. Επίσης, στην περίπτωση του VLAD χρησιμοποιούμε πιο εκλεπτυσμένη μέθοδο κανονικοποίησης, την Intra-Normalization (βλ. Ενότητα 3.4.2.3), την οποία θα μπορούσαμε να αξιοποιήσουμε και στην περίπτωση του διανύσματος Fisher.

Από τα παραπάνω συμπεραίνουμε πως τα αποτελέσματα της ταξινόμησης δράσεων επηρεάζονται από πολλές παραμέτρους και λεπτομέρειες υλοποίησης (κανονικοποίηση δεδομένων, προ-επεξεργασία) και η βέλτιστη επιλογή παραμέτρων διαφέρει μεταξύ των διαφόρων βάσεων δεδομένων.

Κεφάλαιο 5

On-line Αναγνώριση συνεχούς ροής χειρονομιών

5.1 On-line ταξινόμηση

Όπως είναι φανερό, το σύστημα που έχει περιγραφεί ως τώρα δε μπορεί να αξιοποιηθεί άμεσα για τη δημιουργία ενός διαδραστικού περιβάλλοντος επικοινωνίας ανθρώπου-υπολογιστή μέσω χειρονομιών. Τα εμπόδια είναι αρκετά: η δυνατότητα των αλγορίθμων να εκτελούνται "στον αέρα" (on-line), η υπολογιστική πολυπλοκότητα, οι συνθήκες λειτουργίας, καθώς και η δυνατότητα αξιόπιστου εντοπισμού των χειρονομιών στο χρόνο [63]. Το τελευταίο αποτελεί και το σημαντικότερο σκόπελο που θα πρέπει να υπερβεί κανείς, δεδομένου πως στη γνώση των χρονικών ορίων κάθε εκτελούμενης χειρονομίας βασίζεται η επίδοση του συστήματος και η συνολική εμπειρία του χρήστη.

Στα πλαίσια του ερευνητικού προγράμματος MOBOT έγινε μια προσπάθεια υλοποίησης [1] ενός συστήματος αναγνώρισης συνεχούς ροής χειρονομιών. Σκοπός είναι η ανάπτυξη ενός περιβάλλοντος στο οποίο ο χρήστης θα εκτελεί προκαθορισμένες χειρονομίες μπροστά από έναν αισθητήρα Kinect και το σύστημα αφ' ενός θα εντοπίζει πότε εκτελείται μια τέτοια χειρονομία και αφ' ετέρου θα την αναγνωρίζει.

Στην ενότητα αυτή παρουσιάζουμε αναλυτικά τα πρακτικά βήματα που ακολουθήθηκαν στην κατεύθυνση υλοποίησης του εν λόγω συστήματος. Αυτά αφορούν αρκετές σχεδιαστικές αποφάσεις, καθώς και την επίλυση πολλών προβλημάτων προγραμματιστικής φύσης. Εφεξής θα αναφερόμαστε στο σύστημα που έχει περιγραφεί στα προηγούμενα κε-

[1] Θα πρέπει να σημειωθεί η σημαντική συμβολή της Έφης Μαυρουδή στην ανάπτυξη της πρώτης αυτής έκδοσης του συστήματος.

φάλαια με τη φράση "off-line" σύστημα και στο νέο σύστημα με τη φράση "on-line" σύστημα. Το πρώτο εκτελεί ένα πλήρες πείραμα που περιλαμβάνει όλα τα βίντεο, τους περιγραφητές, κ.α. όπως έχει περιγραφεί ήδη. Τα δεδομένα που χρησιμοποιούνται για την ανάπτυξη του on-line συστήματος προέρχονται από τη βάση MOBOT-6.a και συγκεκριμένα από το υποσύνολο που που αναφέρουμε στην Ενότητα 1.3.5 και περιλαμβάνει των 8 κλάσεις.

Ο πυρήνας του on-line συστήματος αποτελείται ουσιαστικά από τον κάτω κλάδο του διαγράμματος 1.1. Συγκεκριμένα, για ένα τμήμα βίντεο ή αλλιώς "κλιπ" που θέλουμε να ταξινομηθεί, ακολουθούνται τα εξής βήματα: 1) εξαγωγή χαρακτηριστικών, 2) κωδικοποίηση χαρακτηριστικών, 3) ταξινόμηση, για τα οποία χρησιμοποιούνται τα εξής δεδομένα που έχουν προκύψει από την εκπαίδευση:

1. Το οπτικό λεξικό, που αποτελείται από K διανύσματα διάστασης D,
2. Τα κωδικοποιημένα χαρακτηριστικά των δεδομένων εκπαίδευσης που χρησιμοποιούνται για τον υπολογισμό του πυρήνα του SVM,
3. Ο παράγοντας κανονικοποίησης του πυρήνα του SVM,
4. Τα εκπαιδευμένα SVM μοντέλα.

Οι παράμετροι που χρησιμοποιήσαμε αρχικά είναι: dense trajectories χαρακτηριστικά με MBH περιγραφητή, Bag-of-Words κωδικοποίηση με $K = 4000$ οπτικές λέξεις και μη-γραμμικά SVM με χ^2 πυρήνα και κόστος $C = 100$.

Υπολογιστική πολυπλοκότητα Για τη μέτρηση της "καθαρής" υπολογιστικής πολυπλοκότητας της ταξινόμησης αναπτύχθηκε ένα εργαλείο που εκτελεί τα παραπάνω βήματα και μετρήθηκε ο μέσος ρυθμός επεξεργασίας για όλα τα βίντεο τριών διαφορετικών βάσεων δεδομένων. Αυτοί φαίνονται στον πίνακα 5.1, μαζί με την ανάλυση της κάθε πηγής. Παρατηρούμε εξάρτηση του χρόνου επεξεργασίας από την ανάλυση του βίντεο, το οποίο είναι αναμενόμενο, με εξαίρεση τη βάση HMDB51. Την απόκλιση αυτή την αποδίδουμε στο γεγονός πως η HMDB51 περιλαμβάνει πολλά κλιπ στα οποία οι δράσεις που απεικονίζονται καταλαμβάνουν μικρή περιοχή στην εικόνα ή/και μικρή διάρκεια σε σχέση με τη συνολική διάρκεια του κλιπ. Έτσι, ο χρόνος εξαγωγής χαρακτηριστικών, που αποτελεί και το μεγαλύτερο μέρος του συνολικού χρόνου επεξεργασίας (βλ. παρακάτω), είναι πολύ μικρότερος, μιας και οι πυκνές τροχιές εξάγονται κυρίως σε μη στατικές περιοχές της εικόνας.

Πίνακας 5.1: Μετρήσεις του ρυθμού επεξεργασίας για τρεις διαφορετικές βάσεις δεδομένων.

Βάση δεδομένων	Ανάλυση και frame rate βίντεο	Frame rate συστήματος ταξινόμησης (fps)
KTH	160×120, 24 fps	17.28
HMDB51	367×240 (μέση), 25 fps	24.32
MOBOT 6.a	640×480, 25 fps	2.15

Ενσωμάτωση στο ROS Για τη μείωση του χρόνου επεξεργασίας, αλλά λόγω της ανάγκης για ενσωμάτωση του on-line συστήματος στο κοινό πλαίσιο του ROS, ο πηγαίος κώδικας επανεγγράφηκε σε python. Πλέον, αντί της χρήσης των βίντεο της βάσης δεδομένων MOBOT 6.a, χρησιμοποιούνται τα λεγόμενα "rosbag" αρχεία, τα οποία περιέχουν αποθηκευμένα στην αρχική τους (raw) μορφή τα δεδομένα όλων των αισθητήρων που είναι τοποθετημένοι στη ρομποτική πλατφόρμα. Το πλεονέκτημα των rosbags είναι ότι μπορούν να "αναπαραχθούν" (play) στο πλαίσιο του ROS, αποστέλλοντας κάθε frame σε τακτά χρονικά διαστήματα, ακριβώς όπως λήφθηκε εξ' αρχής από το Kinect. Η αναπαραγωγή των rosbags προσομοιώνει, λοιπόν, τη λειτουργία ενός kinect σε πραγματικό χρόνο. Πλέον, θα αναφερόμαστε χωρίς διάκριση στα frames είτε του Kinect, είτε ενός rosbag, ανεξάρτητα από την πηγή από την οποία προέρχονται στην πράξη.

Δεδομένου πως τα raw frames από το Kinect δεν έχουν υποστεί μετατροπή, συμπίεση, κ.τ.λ. όπως αυτά που προκύπτουν από τα βίντεο της βάσης, αναπτύχθηκε σε python ένα εργαλείο εξαγωγής χαρακτηριστικών που αντλεί RGB δεδομένα από τα rosbags. Η εκπαίδευση πραγματοποιείται και πάλι με το off-line σύστημα, με κατάλληλη μετατροπή των δεδομένων που παράγονται και απαιτούνται για την ταξινόμηση.

Μετρήθηκε εκ νέου ο χρόνος επεξεργασίας για τη βάση MOBOT 6.a μόνο, και αξιολογήθηκαν διάφορες τεχνικές που θα μπορούσαν να αξιοποιηθούν για για την περαιτέρω μείωση την υπολογιστικής πολυπλοκότητας. Η επίδραση των εν λόγω τεχνικών στο χρόνο αλλά και στην ακρίβεια φαίνονται στον πίνακα 5.2. Συγκρίνοντας την τρίτη γραμμή του Πίνακα 5.1 και την πρώτη του Πίνακα 5.2, επιβεβαιώνουμε πως το προγραμματιστικό framework που υιοθετείται παίζει σημαντικό ρόλο στην ταχύτητα εκτέλεσης. Επίσης, είναι φανερό πως η υποδειγματοληψία μειώνει δραστικά το χρόνο εκτέλεσης, αλλά με αρνητικό αντίκτυπο στην ακρίβεια ταξινόμησης. Το τελευταίο αντισταθμίζεται ελαφρώς με

Πίνακας 5.2: Αξιολόγηση του ρυθμού επεξεργασίας και της ακρίβειας ταξινόμησης στη βάση MOBOT 6.a, χρησιμοποιώντας τον ασθενή p1 ως δοκιμαστικό δείγμα.

	Frame Rate (fps)	Ακρίβεια
Αρχικό βίντεο	3.861	71.88
Υποδειγματοληπτημένο βίντεο με συντελεστή $\sqrt{2}$	8.06	65.63
Υποδειγματοληπτημένο βίντεο με συντελεστή 2	16.45	43.75
Υποδειγματοληπτημένο βίντεο με συντελεστή 2 και γκαουσιανό φιλτράρισμα	16.51	53.13
Αρχικό βίντεο, υπολογισμός μόνο του περιγραφητή MBH	4.74	71.88
Συνδυασμός των δύο τελευταίων	20.32	53.13

το γκαουσιανό φιλτράρισμα. Στον Πίνακα 5.3 ο χρόνος επεξεργασίας αναλύεται βήμα προς βήμα για ένα βίντεο-δείγμα, προκειμένου να εντοπιστούν τα κύρια σημεία κωλύματος. Προφανώς, ο χρόνος εξαγωγής χαρακτηριστικών καταλαμβάνει με διαφορά το μεγαλύτερο μέρος του συνολικού χρόνου. Αυτό φυσικά εξηγεί και το λόγο που η υποδειγματοληψία μειώνει τόσο πολύ το συνολικό χρόνο.

Σύστημα "Press-to-Gesture" Για να παρακάμψουμε προσωρινά το πρόβλημα του χρονικού εντοπισμού των χειρονομιών, αναπτύχθηκε ένα περιβάλλον "Press-to-Gesture", στο οποίο ο χρήστης υποδεικνύει την αρχή και το τέλος της χειρονομίας με το πάτημα ενός πλήκτρου. Πιο αναλυτικά, το σύστημα αυτό περιλαμβάνει δύο προγράμματα:

1. Το keyboard listener, το οποίο διαβάζει τα καρέ που αποστέλλονται από το Kinect, τα απεικονίζει στην οθόνη και παρακολουθεί τα πλήκτρα που πατά ο χρήστης. Μόλις ένα συγκεκριμένο πλήκτρο (space-bar) πατηθεί, αποστέλλει ένα μήνυμα στον ταξινομητή.

2. Τον ταξινομητή χειρονομιών, ο οποίος αγνοεί όλα τα frames από το Kinect, έως ότου λάβει σήμα από το keyboard listener, οπότε αρχίζει να αποθηκεύει στη μνήμη τα εισερχόμενα frames. Μόλις

Πίνακας 5.3: Ανάλυση του χρόνου επεξεργασίας για ένα βίντεο-δείγμα (70 καρέ, 2.8 sec, χωρίς υποδειγματοληψία).

	time (sec)	% time
Εξαγωγή Χαρακτηριστικών (dense trajectories)	16.048	**82.4**
Parsing της εξόδου του προηγούμενου βήματος	1.817	9.3
Κωδικοποίηση χαρακτηριστικών (BoF)	1.593	8.2
Υπολογισμός πυρήνα SVM (χ^2 απόσταση)	0.127	0.1
Άλλα βήματα	$<$ 1ms	$<$ 0.01
Σύνολο	19.45	100

λάβει δεύτερο σήμα, σταματά να λαμβάνει frames και ταξινομεί την ακολουθία των καρέ που έλαβε μεταξύ των δύο μηνυμάτων.

5.2 Χρονικός εντοπισμός χειρονομιών - ο Ανιχνευτής Χειρονομιών (Gesture Activity Detector)

Είναι φανερό πως για την αναγνώριση συνεχούς ακολουθίας χειρονομιών απαιτείται ένα τρόπος αυτόματου εντοπισμού των χρονικών ορίων κάθε χειρονομίας που εκτελείται σε μια ροή βίντεο. Μια απλή μέθοδος για την αντιμετώπιση αυτού του προβλήματος είναι να θεωρήσουμε έναν δυαδικό (binary) ταξινομητή, ο οποίος επεξεργάζεται μικρά τμήματα του βίντεο και αποφαίνεται σχετικά με την ύπαρξη ή μη κάποιας (οποιασδήποτε) χειρονομίας.

Αναπτύχθηκε, λοιπόν, ο *ανιχνευτής χειρονομιών*, ο οποίος χρησιμοποιεί ένα κυλιόμενο (μη-επικαλυπτόμενο) παράθυρο μήκους 10 καρέ[2] και κατηγοριοποιεί μικρά τμήματα του βίντεο σε δύο κλάσεις: "Rest" και "NonRest". Για την ανάπτυξη του δυαδικού ταξινομητή χρησιμοποιήθηκε το υπάρχον pipeline του συστήματος αναγνώρισης δράσεων.

Επισημειώσεις Επειδή όμως δεν υπήρχαν διαθέσιμες επισημειώσεις (annotations) των βίντεο για τις κατηγορίες Rest και NonRest, δημιουρ-

[2]το οποίο αντιστοιχεί σε $0,75$ sec. περίπου, δοσμένου του μέσου frame rate του Kinect.

γήθηκαν επισημειώσεις με *αυτόματο* τρόπο, βασισμένες σε αυτές των χειρονομιών (βλ. ενότητα 1.3.5). Η διαδικασία που ακολουθήσαμε είναι η εξής:

- Διαδοχικές επαναλήψεις της ίδια;ς χειρονομίας από τον ασθενή λαμβάνεται ως ένα ενιαίο NonRest τμήμα.

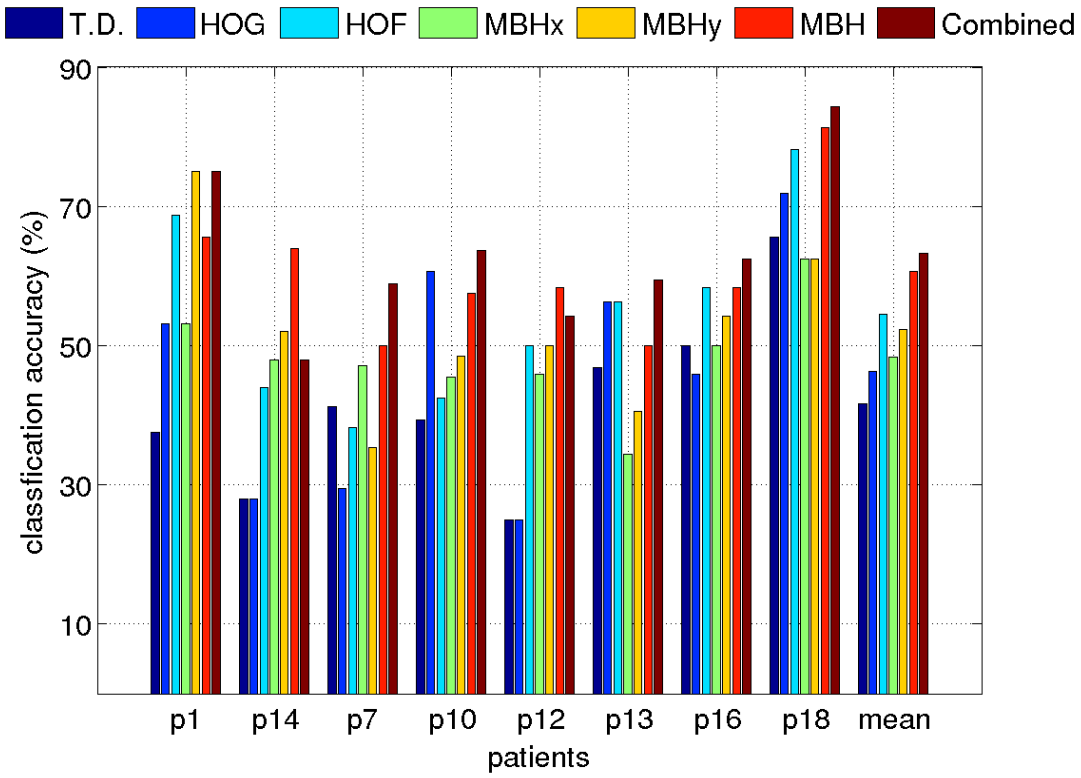
- Το ενιαίο αυτό τμήμα χωρίζεται σε μικρότερα NonRest τμήματα διάρκειας 10 καρέ και επισημειώνεται κατάλληλα.

- Θεωρούμε την ύπαρξη δύο Rest τμημάτων και δύο NonRest τμημάτων πριν και μετά από κάθε ενιαίο NonRest τμήμα.

Προφανώς οι αυτόματες επισημειώσεις που προκύπτουν από την παραπάνω διαδικασία είναι αρκετά θορυβώδεις, μιας και οι υποθέσεις μας πολλές φορές δεν επαληθεύονται. Για παράδειγμα, σε αρκετές περιπτώσεις υπάρχουν διαστήματα μεταξύ των διαδοχικών εκτελέσεων μιας χειρονομίας στα οποία ο ασθενής σχεδόν ακινητοποιείται για αρκετό χρονικό διάστημα, ενίοτε μεγαλύτερο των 10 καρέ.

Για την αξιολόγηση του ανιχνευτή χειρονομιών ακολουθήσαμε την ίδια διαδικασία με την ταξινόμηση δράσεων, δηλαδή εκπαιδεύσαμε το σύστημα με τα δεδομένα όλων των ασθενών πλην ενός, με τον οποίο έγινε στη συνέχεια το testing. Τα αποτελέσματα απεικονίζονται στο Σχήμα 5.1. Εδώ, φυσικά τα "δεδομένα" κάθε ασθενή αποτελούνται από τα Rest/NonRest που δημιουργήθηκαν αυτόματα. Ως εκ τούτου, τα αποτελέσματα αποτελούν "χονδροειδείς" εκτιμήσεις που εν πολλοίς αδικούν το σύστημα.

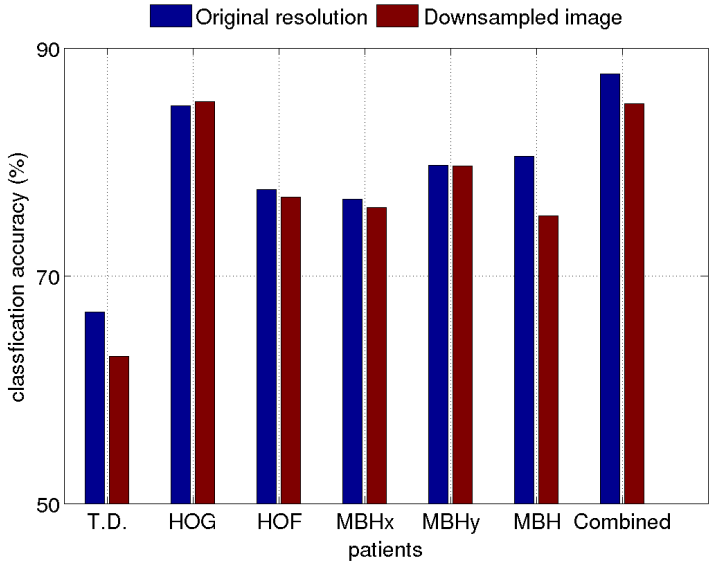
Έχοντας κατά νου πως ο συνολικός χρόνος επεξεργασίας εξαρτάται σε μεγάλο βαθμό από τον ανιχνευτή χειρονομιών, επιλέξαμε παραμέτρους που μειώνουν το χρόνο επεξεργασίας του. Το μέγεθος του οπτικού λεξικού K δεν επηρεάζει σημαντικά την ακρίβεια, όπως φαίνεται στο Σχήμα 5.3, οπότε επιλέξαμε $K = 500$, για να επιταχύνουμε το στάδιο της κβάντισης και του υπολογισμού του πυρήνα του SVM. Οι παράμετροι για την εξαγωγή πυκνών τροχιών είναι: βήμα πυκνής δειγματοληψίας, $W = 10$ και μήκος τροχιών $L = 9$. Οι υπόλοιπες παράμετροι είναι ίδιες με αυτές της ενότητας 4.2. Αξιολογήσαμε, επίσης, την επίδραση της (χωρικής) υποδειγματοληψίας κάθε καρέ στην επίδοση του ανιχνευτή (Σχήμα 5.2), που όπως είδαμε πιο πριν μειώνει σημαντικά το χρόνο επεξεργασίας (Πίνακας 5.2. Όπως είδαμε και στο προηγούμενο Κεφάλαιο, ο συνδυασμός των επιμέρους περιγραφητών δίνει τα καλύτερα αποτελέσματα. Συγκρίσιμα, ωστόσο, αποτελέσματα



Σχήμα 5.1: Ακρίβεια ταξινόμησης χρονικών παραθύρων της εισερχόμενης ροής βίντεο στις κατηγορίες "Rest" και "NonRest" για όλους τους ασθενείς και διαφορετικούς περιγραφητές.

δίνει και ο περιγραφητής HoG, ο οποίο ξεπερνά όλους τους υπόλοιπους πλην του συνδυαστικού. Αυτό υποδεικνύει τη σημασία της στατικής εμφάνισης για το διαχωρισμό μεταξύ τμημάτων που εκτελείται χειρονομία και τμημάτων σχετικής ακινησίας. Επιλέξαμε τη χρήση του HoG, μιας και ο υπολογισμός του συνδυαστικού περιγραφητή επιβαρύνει το χρόνο επεξεργασίας, προσφέροντας δυσανάλογα μικρό όφελος στην ακρίβεια. Το συνολικό σύστημα αποτελείται από δύο επιμέρους λειτουργικά μέρη: τον *ανιχνευτή χειρονομιών* και τον *ταξινομητή*. Το συνολικό σύστημα απεικονίζεται στο Σχήμα 5.4.

Όπως αναφέραμε νωρίτερα, ο ανιχνευτής λαμβάνει καρέ από το Kinect και χρησιμοποιεί ένα κυλιόμενο, μη-επικαλυπτόμενο παράθυρο για να ταξινομήσει τμήματα διάρκειας L καρέ σε μια από τις κατηγορίες Rest ή NonRest. Αυτές υποδηλώνουν την κατάσταση στην οποία

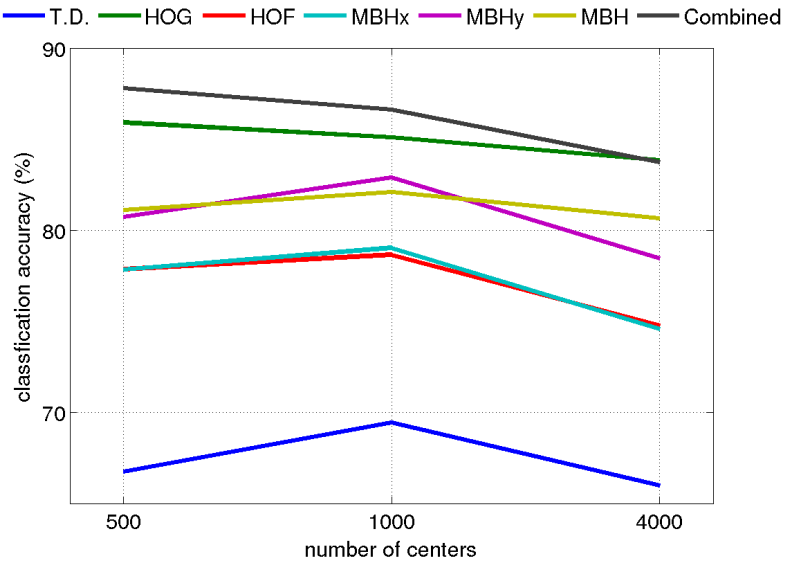


Σχήμα 5.2: Επίδραση της υποδειγματοληψίας στην απόδοση του ανιχνευτή χειρονομιών. Κάθε εισερχόμενο καρέ υποδειγματοληπτείται κατά έναν παράγοντα 2.

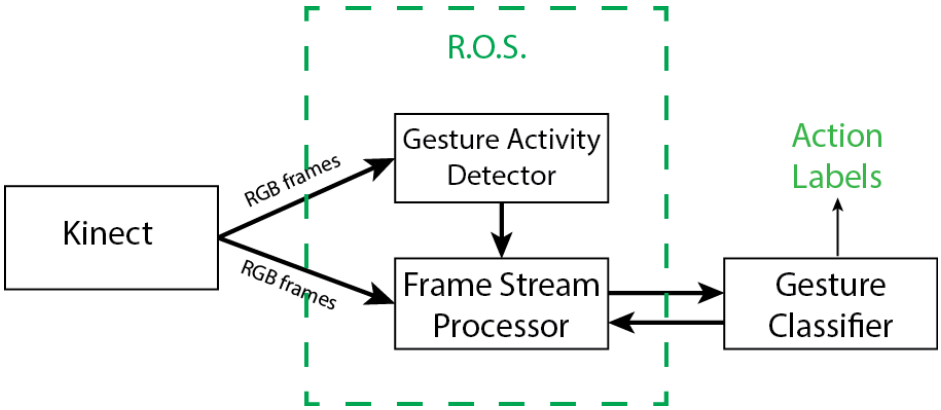
βρίσκεται ο ασθενής, αν δηλαδή εκτελεί ή όχι χειρονομίες. Όταν λάβει χώρα μια μετάβαση μεταξύ των δύο αυτών καταστάσεων, ο ανιχνευτής στέλνει ένα μήνυμα στον ταξινομητή, το οποίο περιέχει το χρόνο (timestamp) που συνέβη η μετάβαση, καθώς και το είδος της, δηλαδή την προηγούμενη και την τρέχουσα κατάσταση. Προφανώς, οι μεταβάσεις Rest → NonRest και NonRest → Rest υποδεικνύουν την αρχή και το τέλος μιας χειρονομίας.

Ο ταξινομητής λαμβάνει καρέ από το Kinect όπως και ο ανιχνευτής, τα οποία αποθηκεύει μαζί με τα timestamps τους. Όταν γίνουν δύο διαδοχικές μεταβάσεις Rest → NonRest και NonRest → Rest, ο ανιχνευτής εντοπίζει τα αντίστοιχα timestamps των μεταβάσεων, τα οποία αντιστοιχούν στην αρχή και το τέλος της χειρονομίας. Στη συνέχεια, ταξινομεί το τμήμα του βίντεο που μεσολαβεί μεταξύ των μεταβάσεων σε μια από τις 8 κατηγορίες (Σχήμα 5.5).

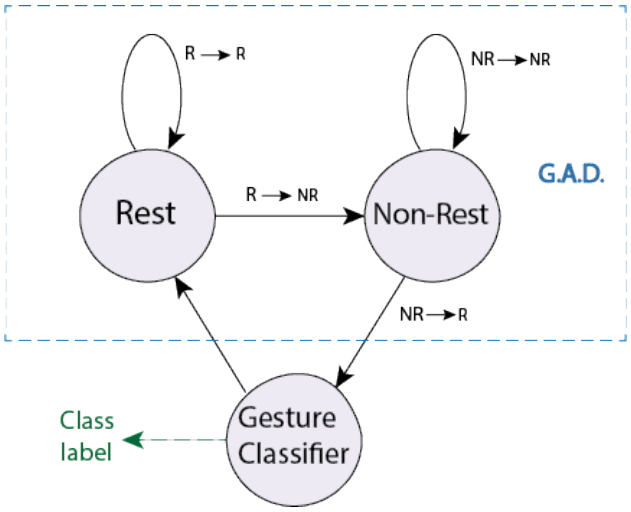
Φυσικά, το παραπάνω σύστημα απέχει πολύ από τη δημιουργία ενός πραγματικού διαδραστικού συστήματος. Κύρια εμπόδια είναι η δυσκολία εντοπισμού χειρονομιών και η ταυτόχρονη απόρριψη άλλων κινή-



Σχήμα 5.3: Επίδραση του αριθμού των κέντρων της αναπαράστασης Bag-of-Words στην απόδοση του ανιχνευτή χειρονομιών.



Σχήμα 5.4: Δομή του On-line συστήματος. Τα καρέ του Kinect λαμβάνονται από τον ανιχνευτή χειρονομιών καθώς και από μια βοηθητική διεργασία (Frame Stream Processor - FSP). Μόλις ο ανιχνευτής εντοπίσει μια χειρονομία ειδοποιεί με τα κατάλληλα μηνύματα τον FSP, ο οποίος προεπεξεργάζεται τα καρέ του κλιπ που περιέχει τη χειρονομία και τα στέλνει στον ταξινομητή.



Σχήμα 5.5: Αναπαράσταση της διασύνδεσης του GAD με τον ταξινομητή χειρονομιών. Ο GAD ανιχνεύει μεταβάσεις μεταξύ των καταστάσεων Rest και NonRest. Μόλις εντοπιστούν δύο μεταβάσεις Rest → NonRest και NonRest → Rest ο ταξινομητής αναλαμβάνει την επεξεργασία και κατηγοριοποίηση του κλιπ που μεσολάβησε μεταξύ των δύο μεταβάσεων.

σεων που ίσως μοιάζουν, αλλά δεν περιέχονται στο λεξικό των χειρονομιών. Δεδομένου πως ο σωστός και κατά το δυνατόν ακριβής εντοπισμός τους είναι βασική προϋπόθεση για την ανάπτυξη του συνολικού συστήματος, θεωρούμε πρωταρχικής σημασίας την περαιτέρω μελέτη και έρευνα στο κομμάτι αυτό.

Κεφάλαιο 6

Συμπεράσματα - Επίλογος

6.1 Συμβολή της διπλωματικής εργασίας

Στη διπλωματική αυτή πραγματευτήκαμε το πρόβλημα της ταξινόμησης και αναγνώρισης ανθρώπινων δράσεων και χειρονομιών με σκοπό την πρακτική τους ενσωμάτωση σε εφαρμογές ρομποτικής. Αντιμετωπίσαμε τις χειρονομίες ως μια ειδική περίπτωση δράσεων, το οποίο δείξαμε πως αποτελεί μια λογική και αποτελεσματική προσέγγιση.

Αρχικά αναλύσαμε από θεωρητική και πρακτική σκοπιά τις κύριες προσεγγίσεις και τάσεις της διεθνούς βιβλιογραφίας, αναδεικνύοντας τις θετικές πτυχές αλλά και τις αδυναμίες τους. Πραγματοποιήσαμε πειράματα σε μια σειρά βάσεων δεδομένων που καλύπτουν ένα μεγάλο εύρος όσον αφορά τη δυσκολία και την πολυπλοκότητα που παρουσιάζουν και εξάγαμε αποτελέσματα συγκρίσιμα με τα αυτά της βιβλιογραφίας..

Τα χωρο-χρονικά σημεία ενδιαφέροντος αποδείχτηκαν αποτελεσματικά για απλές δράσεις που εκτελούνται υπό ελεγχόμενες συνθήκες, αλλά ανεπαρκείς σε πολύπλοκα και σύνθετα περιβάλλοντα. Οι πυκνές τροχιές αποτελούν μια πολύ πιο εύρωστη επιλογή, καθώς κωδικοποιούν έμμεσα δυναμική πληροφορία μέσω της παρακολούθησης σημείων του βίντεο στο χρόνο. Τα αποτελέσματα που δίνει είναι συγκρίσιμα με τα καλύτερα της βιβλιογραφίας σε απαιτητικές βάσεις δεδομένων, με μεγαλύτερο όμως υπολογιστικό κόστος. Πειραματιστήκαμε σε μια πληθώρα από βάσεις δεδομένων, από τις πιο απλές έως τις πιο απαιτητικές και εξάγαμε αποτελέσματα συγκρίσιμα με αυτά που έχουν αναφερθεί στη βιβλιογραφία.

Στη συνέχεια, δοκιμάσαμε αρκετές από τις πιο ευρέως χρησιμοποιούμενες αναπαραστάσεις βίντεο για την κωδικοποίηση των "χαμηλού επιπέδου" χαρακτηριστικών που εξάγονται από το βίντεο. Το δη-

μοφιλές μοντέλο Bag-of-Words, λόγω της απλότητάς του, δεν μπορεί να κωδικοποιήσει την πλούσια πληροφορία που περιέχουν τα χαρακτηριστικά του βίντεο. Αντίθετα, τεχνικές όπως το VLAD και το διάνυσμα Fisher καταγράφουν πιο λεπτομερή στατιστικά μεγέθη, οδηγώντας σε πιο εκλεπτυσμένες αναπαραστάσεις του βίντεο.

Επίσης, εφαρμόσαμε πολλές από τις μεθόδους αυτές στην ερευνητική βάση χειρονομιών MOBOT-6.a με ικανοποιητικά αποτελέσματα, ενισχύοντας την αρχική μας αντιμετώπιση των χειρονομιών ως ειδικές περιπτώσεις δράσεων. Η βάση MOBOT-6.a, λόγω του ότι περιέχει χειρονομίες ηλικιωμένων, παρουσιάζει ιδιαίτερες προκλήσεις που δεν εντοπίζονται σε άλλες παρόμοιες βάσεις της βιβλιογραφίας. Επίσης, αντανακλά ένα μεγάλο μέρος των ρεαλιστικών δυσκολιών στις οποίες καλείται ανταπεξέλθει ένα σύστημα αναγνώρισης χειρονομιών σε ένα πολυαισθητηριακό περιβάλλον.

Τέλος, περιγράψαμε τις προκλήσεις που παρουσιάζει η πρακτική εφαρμογή των μεθόδων αυτών σε ένα πραγματικό, on-line σύστημα αναγνώρισης χειρονομιών με σκοπό τη δημιουργία ενός διαδραστικού περιβάλλοντος αλληλεπίδρασης ανθρώπου-ρομπότ. Οι προκλήσεις αυτές αφορούν κρίσιμα θέματα όπως η υπολογιστική πολυπλοκότητα, η ενσωμάτωση κ.α. Η σημαντικότερη όμως, είναι η ανάγκη για τον ταυτόχρονο on-line χρονικό εντοπισμό των χειρονομιών που εκτελεί ο χρήστης. Προτείναμε, λοιπόν, ένα απλό σύστημα εντοπισμού, το οποίο βασίζεται σε έναν δυαδικό (binary) ταξινομητή, ο οποίος επεξεργάζεται μικρής διάρκειας κομμάτια από την εισερχόμενη ροή του βίντεο. Δείξαμε πως θα μπορούσε να αποτελέσει μια αποτελεσματική λύση για το εν λόγω πρόβλημα.

Δεδομένου του πεπερασμένου κάθε διπλωματικής εργασίας, δεν υπήρξε ο χρόνος να μελετήσουμε και να ακολουθήσουμε πολλές από τις ερευνητικές κατευθύνσεις που θα επιθυμούσαμε, ορισμένες από τις οποίες καταγράφουμε στην επόμενη ενότητα.

6.2 Εν εξελίξει εργασία και κατευθύνσεις για μελλοντική έρευνα

Το πεδίο της αναγνώρισης δράσεων αναπτύσσεται ταχύτατα και είναι ερευνητικά ιδιαίτερα πλούσιο. Θα μπορούσε, λοιπόν, κανείς να προτείνει αμέτρητες προεκτάσεις, βελτιώσεις, αλλά και καινοτόμες ιδέες. Παρακάτω αναφέρουμε ένα μικρό μέρος αυτών, τις οποίες θεωρούμε υποσχόμενες. Σε αυτές περιλαμβάνονται και ιδέες με τις οποίες πειρα-

ματιστήκαμε κατά τη διάρκεια της παρούσας διπλωματικής, αλλά δεν έφτασαν στο κατάλληλο "σημείο ωριμότητας" ώστε να συμπεριληφθούν στο βασικό κορμό της εργασίας.

Ενσωμάτωση περαιτέρω στατιστικών μεγεθών στο στάδιο την κωδικοποίησης Όπως είδαμε, η ενσωμάτωση πιο πλούσιας στατιστικής πληροφορίας στην αναπαράσταση του βίντεο βελτιώνει σημαντικά το αποτέλεσμα ταξινόμησης. Οι υπάρχουσες μέθοδοι κωδικοποίησης αντιμετωπίζουν τα χαρακτηριστικά που εξάγονται από το το βίντεο σα μεμονωμένες παρατηρήσεις, χωρίς να εξετάζουν τη μεταξύ τους σχέση. Η σχέση αυτή όμως είναι σημαντική, για παράδειγμα στη δράση "περπατάω" η κίνηση των ποδιών συνοδεύεται από την κίνηση των άνω άκρων. Θα μπορούσε, λοιπόν, εκτός της πληροφορίας σχετικά με την *εμφάνιση* των οπτικών λέξεων στο βίντεο, να ενσωματωθεί στην αναπαράσταση περαιτέρω πληροφορία σχετικά με την *συνεμφάνισή* [64] (co-ocurrence) τους.

Ενσωμάτωση της χρονικής πληροφορίας στο στάδιο την κωδικοποίησης Όπως αναφέρθηκε, οι χρησιμοποιούμενες αναπαραστάσεις βίντεο αντιμετωπίζουν τα χαρακτηριστικά ως ένα σύνολο *χωρίς διάταξη*, αγνοώντας τη χρονική αλληλουχία τους εντός του βίντεο. Αυτή, ωστόσο, είναι σημαντική, μιας και πολλές δράσεις απαρτίζονται από παρεμφερείς χωροχρονικές δομές ή υπο-δράσεις, που όμως εμφανίζονται με διαφορετική σειρά (π.χ. στις δράσεις "σηκώνομαι" και "κάθομαι"). Η χρήση της χρονικής ακολουθίας των οπτικών λέξεων σε συνδυασμό με τις άλλες συνήθεις κωδικοποιήσεις φάνηκε, μέσω πειραματισμών που πραγματοποιήσαμε, να είναι μια πολλά υποσχόμενη μέθοδος.

Μοντελοποίηση χρονικής δυναμικής πληροφορίας Αρκετές εργασίες στη βιβλιογραφία πραγματεύονται τη χρήση δυναμικών μοντέλων για την αναγνώριση δράσεων. Το διαισθητικό κίνητρο γι' αυτού του είδους την προσέγγιση είναι πως διαφορετικές εκτελέσεις μιας δράσης διαφέρουν ως προς τη διάρκεια ή το ρυθμό με τον οποίο πραγματοποιούνται. Επίσης πολλές δράσεις περιλαμβάνουν επαναλαμβανόμενες κινήσεις, (π.χ. "τρέχω"). Τα δυναμικά μοντέλα, όπως τα Κρυφά Μαρκοβιανά Μοντέλα είναι πολύ αποτελεσματικά στη διαχείριση αυτής της χρονικής μεταβλητότητας, όπως έχει δείξει η εμπειρία από το πεδίο της αυτόματης αναγνώρισης φωνής (Automatic Speech Recognition, εν συντομία ASR). Κάποιες εργασίες έχουν επιχειρήσει να χρησιμοποιήσουν HMM και άλλα δυναμικά μοντέλα για την ταξινόμηση και αναγνώριση

Σχήμα 6.1: Εκτίμηση ανθρώπινης πόζας σε ένα ενδεικτικό καρέ της βάσης MOBOT-6.a, με τον αλγόριθμο [67] που βασίζεται παραμορφώσιμα μοντέλα (Deformable Part Models, εν συντομία DPM).

δράσεων (βλ. Ενότητα 2.3), χωρίς όμως να έχουν την επιτυχία των τοπικών (local) μεθόδων (βλ. Ενότητα 2.1). Δεν παύει όμως να ισχύει ότι η χρονική δυναμική πληροφορία των δράσεων απουσιάζει από το σύνολο σχεδόν των ευρέως χρησιμοποιούμενων αναπαραστάσεων βίντεο. Η μοντελοποίηση της, λοιπόν, αποτελεί μια μεγάλη ερευνητική πρόκληση.

Χρήση ανθρώπινης πόζας Η κίνηση του ανθρώπου είναι αναμφίβολα το πιο χαρακτηριστικό στοιχείο μιας δράσης. Επομένως, η αξιοποίηση της πληροφορίας σχετικά με την πόζα (σκελετός), δηλαδή η γνώση της θέσης στην οποία βρίσκονται τα μέλη του ανθρώπου στην εικόνα κάθε στιγμή, αποτελεί μια πλούσια πληροφορία που μπορεί να για αξιοποιηθεί την αναγνώριση δράσεων [65], [66]. Οι εργασίες που βασίζονται στην ανθρώπινη πόζα χρησιμοποιούν συνήθως δεδομένα Motion Capture. Ωστόσο, η εξέλιξη αισθητήρων όπως το Kinect έχουν πλέον επιτρέψει την εκτίμηση και παρακολούθηση της ανθρώπινης πόζας ("σκελετό") σε πραγματικό χρόνο (skeleton tracking) και με μεγάλη ευρωστία. Παράλληλα, καθώς η εκτίμηση της ανθρώπινης πόζας είναι ένα αυτόνομο ερευνητικό πρόβλημα, έχουν προκύψει πολύ αποτελεσματικές μέθοδοι [67] των οποίων τη χρήση μελετάμε (Σχήμα 6.1).

Πολυτροπική Αναγνώριση Καθώς αναφερόμαστε σε μεγάλες βάσεις δεδομένων με βίντεο και σύγχρονους αισθητήρες, δε θα πρέπει να ξεχνάμε πως τα δεδομένα που προέρχονται από αυτούς συνοδεύονται τις περισσότερες φορές από ήχο, με τη μορφή φυσικής γλώσσας, ηχητικών γεγονότων, θορύβου στο παρασκήνιο, κ.α. Όπως έχουν δείξει αρκετές εργασίες (π.χ. [68], [69]), η συμπληρωματική αυτή *τροπικότητα* μπορεί

να διαδραματίσει σημαντικό ρόλο, βελτιώνοντας τα αποτελέσματα της ταξινόμησης ή αναγνώρισης δράσεων και χειρονομιών. Ο κατάλληλος τρόπος αξιοποίησής της αποτελεί άλλο ένα πολλά υποσχόμενο πεδίο έρευνας.

On-line αναγνώριση δράσεων Η εφαρμογή των υπαρχόντων συστημάτων της βιβλιογραφίας για την αναγνώριση συνεχούς ροής δράσεων ή χειρονομιών (on-line αναγνώριση) παρουσιάζει πολλές προκλήσεις. Αυτές απορρέουν από το γεγονός πως οι περισσότερες μέθοδοι εστιάζουν στην *ταξινόμηση* δράσεων από βίντεο που έχουν ληφθεί εκ των προτέρων και είναι αποθηκευμένα. Ένα περιβάλλον αναγνώρισης δράσεων θα πρέπει να είναι σε θέση να αντιλαμβάνεται πότε ο χρήστης εκτελεί κάποια δράση, και κατά πόσο αυτή ανήκει στο σύνολο των προκαθορισμένων δράσεων τις οποίες είναι εκπαιδευμένο να αναγνωρίζει. Ο χρονικός εντοπισμός των δράσεων, αποτελεί, φυσικά, ένα αυτόνομο ερευνητικό πρόβλημα. Θα μπορούσαν, λοιπόν, να αναζητηθούν οι μέθοδοι εκείνες που θα επιτρέψουν τον εντοπισμό των χρονικών ορίων των δράσεων σε πραγματικό χρόνο.

Κρίσιμος παράγοντας είναι επίσης η χρονική πολυπλοκότητα των χρησιμοποιούμενων αλγορίθμων. Παρά την αποτελεσματικότητά τους στην ταξινόμηση βίντεο, τόσο οι "πυκνές" τοπικές μέθοδοι (όπως οι πυκνές τροχιές) αλλά και τα βαθιά νευρωνικά δίκτυα δεν αποτελούν εκ πρώτης όψεως ελκυστικές επιλογές για ένα on-line σύστημα. Το ελάττωμα αυτό αντισταθμίζεται εν μέρει από την πρόοδο στην τεχνολογία των επεξεργαστών γραφικών και των σχετικών παράλληλων αλγορίθμων.

Τελειώνοντας, αξίζει να πούμε πως η παρούσα διπλωματική αποτελεί απλά ένα μικρό "στιγμιότυπο" της δεδομένης χρονικής περιόδου κατά τη διάρκεια της οποίας εκπονήθηκε. Με την έρευνα να προχωρά με εντυπωσιακούς ρυθμούς, οι ερευνητικοί ορίζοντες που ανοίγονται στο πεδίο διευρύνονται και θα διευρύνονται συνεχώς και η πραγμάτωση τεχνολογικών καινοτομιών που πριν ορισμένα χρόνια ενέπιπταν στην κατηγορία της επιστημονικής φαντασίας, μοιάζει ολοένα και πιο εφικτή.

Bibliography

[1] T. B. Moeslund, A. Hilton, and V. Krüger, "A Survey of Advances in Vision-based Human Motion Capture and Analysis," *J. Computer Vision and Image Understanding*, vol. 104, no. 2, pp. 90–126, Nov. 2006.

[2] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning Realistic Human Actions from Movies," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition, (CVPR 2008)*, Jun. 2008, pp. 1–8.

[3] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: a Large-Scale Video Benchmark for Human Activity Understanding," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2015)*, Jun. 2012.

[4] X. S. Papageorgiou, C. S. Tzafestas, P. Maragos, G. Pavlakos, G. Chalvatzaki, G. Moustris, I. Kokkinos, A. Peer, B. Stanczyk, E.-S. Fotinea, and others, "Advances in Intelligent Mobility Assistance Robot Integrating Multimodal Sensory Processing," in *Universal Access in Human-Computer Interaction. Aging and Assistive Environments*, Springer, 2014, pp. 692–703.

[5] I. Laptev and T. Lindeberg, "Local Descriptors for Spatio-temporal Recognition," in *Spatial Coherence for Visual Motion Analysis*, ser. Lecture Notes in Computer Science 3667, W. J. MacLean, Ed., Springer Berlin Heidelberg, Jan. 2006, pp. 91–103.

[6] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action mach: a Spatio-Temporal Maximum Average Correlation Height Filter for Action Recognition," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2008)*, Jun. 2008, pp. 1–8.

[7] T. Lan, Y. Wang, and G. Mori, "Discriminative Figure-Centric Models for Joint Action Localization and Recognition," in *Proc. Int.'l Conf. on Computer Vision (ICCV 2011)*, Nov. 2011, pp. 2003–2010.

[8] M. Marszalek, I. Laptev, and C. Schmid, "Actions in Context," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,(CVPR 2009)*, Jun. 2009, pp. 2929–2936.

[9] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a Large Video Database for Human Motion Recognition," in *Proc. IEEE Int.'l Conf. on Computer Vision (ICCV 2011)*, Nov. 2011, pp. 2556–2563.

[10] G. Chalvatzaki, G. Pavlakos, K. Maninis, X. Papageorgiou, V. Pitsikalis, C. Tzafestas, and P. Maragos, "Towards an Intelligent Robotic Walker for Assisted Living using Multimodal Sensorial Data," in *Proc. 4th Int.'l Conf. on Wireless Mobile Communication and Healthcare (Mobihealth 2014)*, Nov. 2014, pp. 156–159.

[11] I. Laptev and T. Lindeberg, "Space-time Interest Points," in *Proc. IEEE Int.'l Conf. on Computer Vision (ICCV 2003)*, Oct. 2003, pp. 432–439.

[12] H. Wang, A. Klaser, C. Schmid, and C. Liu, "Action Recognition by Dense Trajectories," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2011)*, Jun. 2011, pp. 3169–3176.

[13] K. Simonyan and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos," in *Advances in Neural Information Processing Systems 27*, Curran Associates, Inc., 2014, pp. 568–576.

[14] R. Poppe, "A Survey on Vision-based Human Action Recognition," *J. Image and Vision Computing*, vol. 28, no. 6, pp. 976–990, Jun. 2010.

[15] 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.

[16] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing Human Actions: a Local SVM Approach," in *Proc. 17th Int.'l Conf. on Pattern Recognition, (ICPR 2004)*, Aug. 2004, 32–36 Vol.3.

[17] A. Kläser, M Marszałek, and C. Schmid, "A Spatio-Temporal Descriptor Based on 3D-Gradients," in *Proc. British Machine Vision Conf. (BMVC 2008)*, Sep. 2008, pp. 995–1004.

[18] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior Recognition via Sparse Spatio-Temporal Features," in *2nd Joint IEEE Int.'l Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, Oct. 2005, pp. 65–72.

[19] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int.'l J. of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[20] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional Sift Descriptor and Its Application to Action Recognition," in *Proc. 15th Int.'l Conf. on Multimedia (ACM 2007)*, Nov. 2007, pp. 357–360.

[21] G. Willems, T. Tuytelaars, and V. L. Gool, "An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector," in *10th European Conf. on Computer Vision (ECCV 2008)*, ser. Lecture Notes in Computer Science, Jan. 2008, pp. 650–663.

[22] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up Robust Features," in *Proc. European Conf. on Computer Vision (ECCV 2006)*, 2006, pp. 404–417. (visited on 10/20/2015).

[23] 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. IEEE Int.'l Conf. on Image Processing (ICIP 2014)*, Oct. 2014, pp. 1490–1494.

[24] 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. 19th IEEE Int.'l Conf. on Image Processing (ICIP 2012)*, Sep. 2012, pp. 741–744.

[25] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid, "Evaluation of Local Spatio-Temporal Features for Action Recognition," in *Proc. British Machine Vision Conf. (BMVC 2009)*, Sep. 2009, pp. 124.1–124.11.

[26] R. Messing, C. Pal, and H. Kautz, "Activity Recognition using the Velocity Histories of Tracked Keypoints," in *Proc. IEEE 12th International Conf. on Computer Vision (ICCV 2009)*, Sep. 2009, pp. 104–111.

[27] J. Sun, X. Wu, S. Yan, L. F. Cheong, T. S. Chua, and J. Li, "Hierarchical Spatio-Temporal Context Modeling for Action Recognition," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition, (CVPR 2009)*, Jun. 2009, pp. 2004–2011.

[28] J. Sun, Y. Mu, S. Yan, and L. Cheong, "Activity Recognition using Dense Long-Duration Trajectories," in *Proc. IEEE Int.'l Conf. on Multimedia and Expo (ICME 2010)*, Jul. 2010, pp. 322–327.

[29] H. Wang and C. Schmid, "Action Recognition with Improved Trajectories," in *Proc. IEEE Int.'l Conf. on Computer Vision (ICCV 2013)*, Dec. 2013, pp. 3551–3558.

[30] T. Darrell and A. Pentland, "Space-time Gestures," in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, (CVPR 1993)*, Jun. 1993, pp. 335–340.

[31] A. Bobick and J. Davis, "The Recognition of Human Movement Using Temporal Templates," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, Mar. 2001.

[32] M. K. Hu, "Visual Pattern Recognition by Moment Invariants," *IRE Transactions on Information Theory*, vol. 8, no. 2, pp. 179–187, Jan. 1962.

[33] D. Weinland, R. Ronfard, and E. Boyer, "Free Viewpoint Action Recognition using Motion History Volumes," *J. Computer Vision and Image Understanding*, vol. 104, no. 2–3, pp. 249–257, Nov. 2006, ISSN: 1077-3142.

[34] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as Space-Time Shapes," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247–2253, Dec. 2007.

[35] A. Yilmaz and M. Shah, "Actions Sketch: a Novel Action Representation," in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, (CVPR 2005)*, vol. 1, Jun. 2005, pp. 984–989.

[36] A. Efros, A. Berg, G. Mori, and J. Malik, "Recognizing Action at a Distance," in *Proc. 9th IEEE Int.'l Conf. on Computer Vision, (ICCV 2003)*, vol. 2, Oct. 2003, pp. 726–733.

[37] I. N. Junejo, E. Dexter, I. Laptev, and P. Perez, "View-Independent action Recognition from Temporal Self-Similarities," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 172–185, Aug. 2011.

[38] M. Raptis, I. Kokkinos, and S. Soatto, "Discovering Discriminative Action Parts from Mid-level Video Representations," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2012)*, Jun. 2012, pp. 1242–1249.

[39] Y. Yang, I. Saleemi, and M. Shah, "Discovering Motion Primitives for Unsupervised Grouping and One-Shot Learning of Human Actions, Gestures, and Expressions," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1635–1648, Jul. 2013.

[40] S. Bhattacharya, M. M. Kalayeh, R. Sukthankar, and M. Shah, "Recognition of complex events: Exploiting Temporal Dynamics between Underlying Concepts," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2014)*, IEEE, Jun. 2014, pp. 2243–2250.

[41] B. Fernando, E. Gavves, J. M. Oramas, A. Ghodrati, and T. Tuytelaars, "Modeling Video Evolution for Action Recognition," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2015)*, Jun. 2015.

[42] M. Jain, J. C. van Gemert, and C. G. M. Snoek, "What do 15,000 Object Categories Tell Us About Classifying and Localizing Actions?" In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2015)*, Jun. 2015.

[43] C. Harris and M. Stephens, "A Combined Corner and Edge Detector," in *Proc. Fourth Alvey Vision Conf.*, 1988, pp. 147–151.

[44] J. Shi and C. Tomasi, "Good Features to Track," in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, (CVPR 1994)*, Jun. 1994, pp. 593–600.

[45] B. D. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision," in *Proc. 7th Int.'l Joint Conf. on Artificial Intelligence*, Aug. 1981, pp. 674–679.

[46] G. Farnebäck, "Two-Frame Motion Estimation Based on Polynomial Expansion," in *Image Analysis*, ser. Lecture Notes in Computer Science 2749, Springer Berlin Heidelberg, Jun. 2003, pp. 363–370.

[47] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR 2005)*, Jun. 2005, pp. 886–893.

[48] C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba, "HOGgles: Visualizing Object Detection Features," in *Proc. IEEE International Conference on Computer Vision (ICCV 2013)*, Dec. 2013.

[49] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2nd edition)*. Wiley-Interscience, 2000, ISBN: 0471056693.

[50] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in Quantization: Improving Particular Object Retrieval in Large Scale Image Databases," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2008)*, IEEE, 2008, pp. 1–8.

[51] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR 2006)*, Jun. 2006, pp. 2169–2178.

[52] H. Jegou, M. Douze, C. Schmid, and P. Perez, "Aggregating Local Descriptors into a Compact Image Representation," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2010)*, Jun. 2010, pp. 3304–3311.

[53] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the Fisher Kernel for Large-Scale Image Classification," in *Proc. 11th European Conf. on Computer Vision (ECCV 2010)*, Sep. 2010, pp. 143–156.

[54] R. Arandjelovic and A. Zisserman, "All About VLAD," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2013)*, Jun. 2013, pp. 1578–1585.

[55] F. Perronnin and C. Dance, "Fisher Kernels on Visual Vocabularies for Image Categorization," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition, (CVPR 2007)*, Jun. 2007, pp. 1–8.

[56] V. Kecman, "Support Vector Machines – An Introduction," in *Support Vector Machines: Theory and Applications*, ser. Studies in Fuzziness and Soft Computing, L. Wang, Ed., vol. 177, Springer, 2005, pp. 1–47.

[57] S. Theodoridis and K Koutroumbas, *Pattern Recognition (3rd edition)*. Academic Press, 2006, ISBN: 9781597492720.

[58] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," *J. Machine Learning Research*, vol. 9, pp. 2491–2521, 2008.

[59] M. Varma and B. R. Babu, "More Generality in Efficient Multiple Kernel Learning," in *Proc. 26th Annual Int.'l Conf. on Machine Learning (ICML 2009)*, Jun. 2009, pp. 1065–1072.

[60] A. Jain, S. V. N. Vishwanathan, and M. Varma, "SPG-GMKL: Generalized Multiple Kernel Learning with a Million Kernels," in *Proc. ACM SIGKDD Conf. on Knowledge Discovery and Data Mining*, Aug. 2012.

[61] C. C. Chang and C. J Lin, "Libsvm: A Library for Support Vector Machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, 27:1–27:27, May 2011.

[62] M. M. Ullah, S. N. Parizi, and I. Laptev, "Improving bag-of-features Action Recognition with Non-Local Cues," in *Proc. British Machine Vision Conf. (BMVC 2010)*, Aug. 2010, pp. 95.1–95.11.

[63] S. R. Fanello, I. Gori, G. Metta, and F. Odone, "Keep it Simple and Sparse: Real-Time Action Recognition," *J. of Machine Learning Research*, vol. 14, pp. 2617–2640, 2013.

[64] P. Agustí, V. J. Traver, and F. Pla, "Bag-of-words with aggregated temporal pair-wise word co-occurrence for human action recognition," *Pattern Recognition Letters*, vol. 49, no. 1, pp. 224–230, Nov. 2014.

[65] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele, "A Database for Fine Grained Activity Detection of Cooking Activities," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2012)*, Jun. 2012, pp. 1194–1201.

[66] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards Understanding Action Recognition," in *Proc. Int.'l Conf. on Computer Vision (ICCV 2013)*, Dec. 2013, pp. 3192–3199.

[67] Y. Yang and D. Ramanan, "Articulated Human Detection with Flexible Mixtures-of-Parts," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2878–2890, 2013.

[68] V. Pitsikalis, A. Katsamanis, S. Theodorakis, and P. Maragos, "Multimodal Gesture Recognition via Multiple Hypotheses Rescoring," *J. Machine Learning Research*, vol. 16, pp. 255–284, 2015.

[69] Z. Xu, Y. Yang, and A. G. Hauptmann, "A Discriminative CNN Video Representation for Event Detection," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2015)*, Jun. 2015, pp. 1798–1807.