Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων

τομεας σηματων, ελεγχου και ρομποτικης

εργαστηριο ορασης υπολογιστων, επικοινωνιας φωνης
και επεξεργασιας σηματος

Αναγνώριση Ανθρώπινων Δράσεων σε Βίντεο

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Χρήστου Γ. Γεωργάκη

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2011

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΗΜΑΤΩΝ, ΕΛΕΓΧΟΥ ΚΑΙ ΡΟΜΠΟΤΙΚΗΣ
ΕΡΓΑΣΤΗΡΙΟ ΟΡΑΣΗΣ ΥΠΟΛΟΓΙΣΤΩΝ, ΕΠΙΚΟΙΝΩΝΙΑΣ ΦΩΝΗΣ ΚΑΙ ΕΠΕΞΕΡΓΑΣΙΑΣ ΣΗΜΑΤΟΣ

Αναγνώριση Ανθρώπινων Δράσεων σε Βίντεο

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Χρήστου Γ. Γεωργάκη

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 18η Οκτωβρίου 2011.

.....	.....	.....
Πέτρος Μαραγκός	Κωνσταντίνος Τζαφέστας	Γεράσιμος Ποταμιάνος
Καθηγητής Ε.Μ.Π.	Επίκ. Καθηγητής Ε.Μ.Π.	Ερευνητής Α, "Δημόκριτος"

Αθήνα, Οκτώβριος 2011.

.....
(**Χρήστος Γ. Γεωργάκης**)
(Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Ηλεκτρονικών Υπολογιστών)

Περίληψη

Η παρούσα διπλωματική εργασία εξετάζει το θέμα της βασισμένης στην όραση αναγνώρισης ανθρώπινων δράσεων σε βίντεο. Το πρόβλημα αυτό θεωρείται ως μια θεμελιώδης αναγκαία προϋπόθεση για την ανάλυση και ερμηνεία βίντεο και για τούτο ερευνάται και εφαρμόζεται ευρέως σε διάφορους τομείς όπως η ανάκτηση δεδομένων βίντεο, η οπτική επιτήρηση και παρακολούθηση, η ρομποτική και η αλληλεπίδραση ανθρώπου-υπολογιστή. Συγκεκριμένα, η εργασία μας επικεντρώνεται στο πρόβλημα της εκμάθησης και ταξινόμησης ανθρώπινων δράσεων χρησιμοποιώντας αναπαραστάσεις βίντεο που προκύπτουν από τοπικά χωροχρονικά χαρακτηριστικά.

Μελετάμε διεξοδικά τις υπάρχουσες προσεγγίσεις για την ανίχνευση και περιγραφή χωροχρονικών σημείων ενδιαφέροντος σε ακολουθίες βίντεο, παρέχοντας ταυτόχρονα το εννοιολογικό τους πλαίσιο και το μαθηματικό τους υπόβαθρο. Διεξάγεται μια πιο ενδελεχής ανάλυση συγκεκριμένων τοπικών ανιχνευτών και περιγραφέων αποδεδειγμένης αποτελεσματικότητας εντός της σχετικής έρευνας, συνοδευόμενη από χαρακτηριστικά αποτελέσματα σε δείγματα βίντεο που λαμβάνουμε από διαθέσιμες ή εντός της παρούσας εργασίας αναπτυγμένες υλοποιήσεις. Στη συνέχεια, δίνεται έμφαση σε αναπαραστάσεις υφής προερχόμενες από μοντέλα ανάλυσης υφής που στηρίζονται στο φιλτράρισμα σε πολλαπλές ζώνες συχνοτήτων στον χώρο και σε αλγορίθμους αποδιαμόρφωσης εικόνων. Αφού περιγράψουμε με λεπτομέρεια τη διαδικασία εξαγωγής χαρακτηριστικών, προχωρούμε στην τροποποίηση μιας υπάρχουσας υλοποίησης προκειμένου να μειώσουμε τον υπολογιστικό φόρτο και να την καταστήσουμε εφαρμόσιμη σε ακολουθίες βίντεο. Επιδιώκουμε την αξιοποίηση χαρακτηριστικών κυρίαρχης σε επίπεδο εικονοστοιχείου ενέργειας διαμόρφωσης για την ανίχνευση ή περιγραφή των χωροχρονικών σημείων, οδηγώντας τα ως στάδιο προεπεξεργασίας στην είσοδο ενός γνωστού συνδυασμού ανιχνευτή-περιγραφέα. Η αξιολόγηση της επίδρασής τους στην απόδοση του συστήματος αναγνώρισης επιτυγχάνεται μέσω πειραμάτων στα οποία μαθαίνονται και ταξινομούνται απαιτητικές κατηγορίες ανθρώπινων δράσεων σε δείγματα ταινιών από μία κοινά χρησιμοποιούμενη βάση δεδομένων καί για τα δύο σενάρια των καθιερωμένων και εναλλακτικών σχημάτων αντίστοιχα.

Τέλος, βασιζόμενοι σε πρόσφατες ιδέες σχετικά με το χωροχρονικό φιλτράρισμα και τον εντοπισμό κίνησης, σχεδιάζουμε έναν ανιχνευτή χωροχρονικών σημείων ενδιαφέροντος που εξάγει οπτικά σημαίνοντα σημεία με κριτήριο την κυρίαρχη ενέργεια διαμόρφωσης τόσο στον χώρο όσο και στον χρόνο. Ο ανιχνευτής παρουσιάζεται διεξοδικά και αξιολογείται για πρώτη φορά στο πλαίσιο της αναγνώρισης ανθρώπινων δράσεων μέσω πειραμάτων ταξινόμησης.

Παράλληλα με την παροχή μιας λεπτομερούς και εκτενούς επισκόπησης της υπάρχουσας εργασίας που μελετά την αναγνώριση δράσεων υπό το πρίσμα τοπικών αναπαραστάσεων, αυτή η διπλωματική εργασία εξέτασε περαιτέρω και ενσωμάτωσε επιπρόσθετα χαρακτηριστικά που αντλούνται από προσεγγίσεις σχετικών θεμάτων της Όρασης Υπολογιστών όπως η ανάλυση υφής και η αποδιαμόρφωση διακριτών σημάτων. Μέσα σε αυτό το πλαίσιο, ελπίζουμε να προσφέρουμε χρήσιμη γνώση και πολλαπλές κατευθύνσεις για μελλοντική εργασία.

Λέξεις-κλειδιά: όραση υπολογιστών, ανάλυση ανθρώπινης κίνησης, αναγνώριση ανθρώπινων δράσεων, τοπικά χωροχρονικά χαρακτηριστικά, ανιχνευτές και περιγραφείς χωροχρονικών σημείων ενδιαφέροντος, σάκος χαρακτηριστικών, αποδιαμόρφωση εικόνων, ανάλυση υφής, φιλτράρισμα σε πολλαπλές ζώνες συχνοτήτων στον χώρο και τον χρόνο, ενέργεια διαμόρφωσης, ανάλυση κυρίαρχων συνιστωσών

Abstract

The subject of this thesis is vision-based human action recognition in videos. It is currently a fundamental prerequisite for video analysis and interpretation and widely researched and applied in several domains such as video retrieval, visual surveillance, robotics and human-computer interaction. Our work focuses, in particular, on the issue of learning and classifying human actions using video representations in terms of local space-time features.

A thorough review of existing approaches has been initially undertaken, covering the detection and description of spatiotemporal interest points in video sequences, providing both their conceptual framework and theoretical background. A more detailed analysis of certain local detectors and descriptors of proved efficiency within the related research has also been conducted, accompanied with characteristic results on video samples yielded by available or herein developed implementations. Emphasis is next placed on texture representations procured by texture analysis paradigms which rely on spatial multiband filtering and image demodulation algorithms. After describing in detail the feature extraction procedure, we proceed by modifying an existing implementation in order to reduce the computational load and render it applicable to video sequences. We pursue the exploitation of pixel-wise dominant modulation energy features in spatiotemporal points detection or description, feeding them into a well-known detector-descriptor combination as a pre-process step. Assessment of their influence on the recognition performance is obtained by learning and classifying challenging action classes in movie clips from a commonly used dataset, for both the standard and alternative schemes.

Finally, we build upon recent ideas on spatiotemporal filtering and motion tracking in order to design a spatiotemporal interest point detector which extracts salient points in terms of dominant modulation energy in both space and time. The detector is thoroughly presented and evaluated for the first time within the action recognition framework through classification experiments.

Aside from furnishing a detailed and comprehensive review of the existing work which addresses action recognition employing local representations, this diploma thesis has further examined and integrated additional features derived from approaches in related Computer Vision topics such as texture analysis and discrete signal demodulation. Within this scope, we hope to provide useful insights and several directions for future work.

Keywords: computer vision, human motion analysis, human action recognition, local space-time features, spatiotemporal interest point detectors and descriptors, bag-of-features, image demodulation, texture analysis, multiband spatial and temporal filtering, modulation energy, dominant component analysis

Ευχαριστίες

Η εμπειρία της παρακολούθησης των μαθημάτων της Όρασης Υπολογιστών και της Αναγνώρισης Προτύπων, που διδάσκονται στη σχολή μας από τον καθηγητή Π. Μαραγκό, υπήρξε καθοριστική όχι μόνο για την επιλογή ενός θέματος διπλωματικής εργασίας από τις παραπάνω περιοχές έρευνας αλλά και γενικά για τη διαμόρφωση των ακαδημαϊκών μου ενδιαφερόντων. Μέσα από τα εν λόγω μαθήματα μου δόθηκε η ευκαιρία να έρθω σε επαφή με το ιδιαίτερα ελκυστικό μαθηματικό υπόβαθρο καθώς και τις εντυπωσιακές σύγχρονες εφαρμογές των περιοχών αυτών. Ο κινητοποιητικός τρόπος παρουσίασης και διδασκαλίας του καθηγητή Π. Μαραγκού, η απόκτηση γοητευτικής σχετικής γνώσης, η εκπόνηση υπολογιστικών εργασιών με ιδιαίτερα σύγχρονα θέματα και η ανάδειξη από όλα τα παραπάνω ολοένα και περισσότερων ερεθισμάτων για μελλοντική εργασία και έρευνα, αποτέλεσαν τους θεμελιώδεις παράγοντες στη γέννηση ενσυνείδητου ενδιαφέροντος και ζήλου για περαιτέρω ενασχόληση και τριβή με τα παραπάνω αντικείμενα έρευνας. Ύστερα από συζήτηση με τον καθηγητή Π. Μαραγκό επιλέξαμε το θέμα της διπλωματικής εργασίας το οποίο προσανατολίστηκε σε ένα ιδιαίτερα απαιτητικό μέχρι σήμερα πρόβλημα της Όρασης Υπολογιστών.

Καθ' όλη τη διάρκεια της περίοδου εκπόνησης της εργασίας, η ενασχόληση και η καθοδήγηση του καθηγητή Πέτρου Μαραγκού ήταν συνεχής και πολύτιμη. Θέλω να τον ευχαριστήσω θερμά για τον χρόνο και το ενδιαφέρον που μου αφιέρωσε καθώς και για την ενθάρρυνση και την παροχή πρωτότυπων κατευθύνσεων από πλευράς του κατά τη διάρκεια των συναντήσεών μας.

Καθοριστική υπήρξε, κατά τη διάρκεια της εργασίας μου, η συνεργασία με τον μεταδιδακτορικό ερευνητή Γιώργο Ευαγγελόπουλο. Πριν ακόμα από την ανάληψη της εργασίας, η συζήτησή μας σχετικά με τις διάφορες πλευρές και προκλήσεις του προβλήματος της ανάλυσης και ερμηνείας των βίντεο αποτέλεσε σημαντικό κίνητρο για τον προσδιορισμό του θέματος. Η παραχώρηση του πρωτότυπου λογισμικού για την αποδιαμόρφωση εικόνων επιτάχυνε σημαντικά την ενσωμάτωση σχετικών χαρακτηριστικών στο πλαίσιο της παρούσας εργασίας. Καίρια ήταν και η συμβολή του στη διασαφήνιση του εννοιολογικού πλαισίου και στον καθορισμό των επιλογών σχεδιασμού του ανιχνευτή που αναπτύξαμε στο Κεφάλαιο 5 της εργασίας. Για όλα τα παραπάνω καθώς και για την αμείωτη διάθεσή του για ανταλλαγή ιδεών και συνεργασία θέλω να του εκφράσω τις ιδιαίτερες ευχαριστίες μου. Επιπλέον, πολύτιμη στάθηκε η βοήθεια των ερευνητών του εργαστηρίου CVSP που ευγενικά μου προσέφεραν οποιαδήποτε στιγμή τους ζητήθηκε.

Φυσικά, από μία παράγραφο ειδικών ευχαριστιών δεν είναι δυνατόν να απουσιάζει η αναφορά στα μέλη της οικογένειάς μου. Ευχαριστώ θερμά τον αδερφό μου για την αγάπη

του, την παρότρυνση και τις πολύτιμες συμβουλές του και τους γονείς μου για την αμέριστη αγάπη και υποστήριξη που μου προσφέρουν, ενθαρρύνοντας την πραγμάτωση κάθε στόχου μου.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Γενικά για την Όραση Υπολογιστών

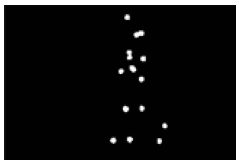
Ο θεμελιώδης στόχος της Όρασης Υπολογιστών (Computer Vision), ως διεπιστημονικής περιοχής, είναι η εύρεση μιας όσο το δυνατόν πληρέστερης συμβολικής περιγραφής των αντικειμένων του τρισδιάστατου κόσμου που ενυπάρχουν σε μια σκηνή, δοθείσας μίας ή περισσοτέρων δισδιάστατων εικόνων. Η εξαγωγή της εν λόγω συμβολικής περιγραφής προκύπτει από μια διαδικασία αναζήτησης και συλλογής πληροφορίας που σχετίζεται με την ταυτότητα και τα χαρακτηριστικά σχήματος και εμφάνισης των αντικειμένων της σκηνής, τη θέση και την κίνησή τους καθώς καί τις σχέσεις και τις ομοιότητες των αντικειμένων που εμφανίζονται. Σε αντίθεση με άλλες σχετικές επιστήμες εικόνων, όπως η επεξεργασία εικόνων και τα γραφικά υπολογιστών, οι οποίες πραγματεύονται προβλήματα μετασχηματισμού ή σύνθεσης εικόνων αντίστοιχα, η Όραση Υπολογιστών ασχολείται με την εννοιολογικά αντίστροφη διαδικασία, αυτή της εξαγωγής συμπερασμάτων για την ύπαρξη ή μη, τις ιδιότητες της επιφάνειας και το είδος των αντικειμένων της φυσικής σκηνής, έχοντας ως είσοδο αριθμητικά δεδομένα εικόνων ή εικονοσειρών.
Η ανάπτυξη της Όρασης Υπολογιστών ως επιστημονικού πεδίου εντοπίζεται χρονικά στη δεκαετία του 1960, οπότε άρχισε να σχηματίζεται και να εξελίσσεται με τη συνέργεια τριών κυρίως επιστημονικών περιοχών, της Τεχνητής Νοημοσύνης, της Επεξεργασίας Σημάτων και της Αναγνώρισης Προτύπων. Έκτοτε, υποστηριζόμενη από σχετικές επιστημονικές περιοχές όπως η νευροβιολογία, η ψυχοφυσική ή τα εφαρμοσμένα μαθηματικά, έχει διαμορφωθεί σε μία ραγδαία εξελισσόμενη και ευρεία διεπιστημονική περιοχή με σημαντικές εφαρμογές σε πολλούς ερευνητικούς τομείς. Παραδείγματα αυτών αποτελούν η επεξεργασία πληροφοριών των εικόνων, η ρομποτική και τα συστήματα αυτομάτου ελέγχου, η ανάλυση βιοϊατρικών εικόνων, η ανάλυση και ερμηνεία βίντεο, η επικοινωνία ανθρώπου και υπολογιστή.
Η εξαγωγή συμβολικών αναπαραστάσεων από εικόνες και η μηχανική οπτική αντίληψη του φυσικού κόσμου μέσω υπολογιστικών μεθόδων, όπως αυτές που μετέρχεται η Όραση Υπολογιστών, παραμένει ακόμα και σήμερα ένα ιδιαίτερα πολύπλοκο και απαιτητικό πρόβλημα. Τα κυριότερα προβλήματα που προσπαθεί να επιλύσει αποτελούν πρόκληση για περαιτέρω έρευνα και ανάπτυξη καινοτόμων προσεγγίσεων αποτελεσματικής αντιμετώπισής τους. Για περισσότερες πληροφορίες σχετικά με την Όραση Υπολογιστών

παραπέμπουμε τον αναγνώστη στην ολοκληρωμένη συγγραφική δουλειά του Π. Μαραγκού στο [4], στην οποία στηρίχτηκε η παρούσα ενότητα.

1.2 Το Πρόβλημα της Αναγνώρισης Ανθρώπινων Δράσεων σε Βίντεο. Διαφορετικές Προσεγγίσεις και Εφαρμογές[1]

Η Αναγνώριση Ανθρώπινων Δράσεων αποτελεί ένα ιδιαίτερα ενεργό πεδίο της Όρασης Υπολογιστών και έχει κερδίσει έντονο ερευνητικό ενδιαφέρον τελευταία, χάρη στις πολλαπλές θεμελιώδεις εφαρμογές του σε τομείς όπως η ρομποτική, η οπτική επιτήρηση, η αλληλεπίδραση ανθρώπου-υπολογιστή και η ανάκτηση πολυμέσων. Ως αναγνώριση δράσεων ορίζεται η διαδικασία της ονοματοδοσίας των ανθρώπινων δράσεων με τη χρήση αισθητηριακών παρατηρήσεων και συνιστά εννοιολογικά ένα πρόβλημα ταξινόμησης. Στα πλαίσια της παρούσας διπλωματικής εργασίας θα ασχοληθούμε με την αυτόματη αναγνώριση δράσεων που προκύπτει αποκλειστικά από την πληροφορία του οπτικού αισθητηριακού καναλιού, δηλαδή των οπτικών παρατηρήσεων σε ακολουθίες εικόνων.
Η αναγνώριση δράσεων εντάσσεται εγγενώς στη γενικότερη περιοχή έρευνας της αυτόματης ανάλυσης της ανθρώπινης κίνησης σε ακολουθίες εικόνων και επομένως συγγενεύει, από πλευράς εφαρμογών, με προβλήματα όπως η *ανάκτηση ανθρώπινης πόζας* (*human pose recovery*), η *ανίχνευση* και ο *εντοπισμός ανθρώπων* (*human detection, human localization*). Η τομή στην συστηματική οπτική ανάλυση της ανθρώπινης κίνησης αποδίδεται σε μεγάλο βαθμό στην εργασία του G. Johansson [6], ο οποίος έδειξε πειραματικά ότι συμπαγείς αναπαραστάσεις της κίνησης, μέσω αντιπροσωπευτικών δεικτών τοποθετημένων στους συνδέσμους του ανθρώπινου σώματος *(Moving Light Displays (MLDs))*, αρκούν να οδηγήσουν τον παρατηρητή σε επιτυχή αναγνώριση της κίνησης, του φύλου των δρώντων και της γωνίας λήψης (Σχήμα 1.1). Απο τότε η σχετική έρευνα έχει προχωρήσει σημαντικά στο σχεδιασμό προσεγγίσεων που μπορούν να αξιοποιήσουν, χωρίς την ανάγκη τέτοιων τεχνικών προεπεξεργασίας και επίβλεψης, σημαίνοντα χαρακτηριστικά ή μετρήσεις εικόνων για τα διάφορα υποπροβλήματα της οπτικής ανάλυσης κίνησης σε ακολουθίες εικόνων.
Μία ανθρώπινη δράση (human action) μπορεί να ιδωθεί ως μια αλληλουχία κινήσεων που εκτελεί ο δρών κατά την εκτέλεση μιας εργασίας. Υπό αυτόν τον ορισμό, η ανθρώπινη δράση συντίθεται από επιμέρους, σε επίπεδο των μελών του ανθρώπινου σώματος, κινήσεις και ως έννοια περιλαμβάνει τη συνολική κίνηση ολόκληρου του σώματος. Το πρόβλημα της αναγνώρισης δράσεων αναφέρεται στην απόδοση στη δράση μιας "ετικέτας"-ονόματος που δύναται, ακόμα και στη μορφή ενός απλού ρήματος, να την περιγράψει καλύτερα, ανεξάρτητα από τις μεταβολές στην εμφάνιση των δρώντων, το ρυθμό εξέλιξής της, το περιβάλλον στο οποίο εκτυλίσσεται, την οπτική γωνία λήψης ή τις συνθήκες εγγραφής. Η διαδικασία αυτή της συσχέτισης παρατήρησης και "ετικέτας"δράσης (label) μπορεί να βασιστεί είτε σε *παραγωγικές προσεγγίσεις (generative approaches)* είτε σε *διακρίνουσες προσεγγίσεις (discriminative approaches)*. Οι πρώτες

[1] Η παρούσα ενότητα βασίστηκε κατά πολύ στο [5].



Σχήμα 1.1: Μεμονωμένο frame που εικονίζει άνθρωπο που βηματίζει σε σκοτεινό παρασκήνιο έχοντας στους συνδέσμους του φωτεινά σημεία *(Moving Light Displays)* (επανεκτύπωση από τις διαφάνειες του I. Laptev[2]).

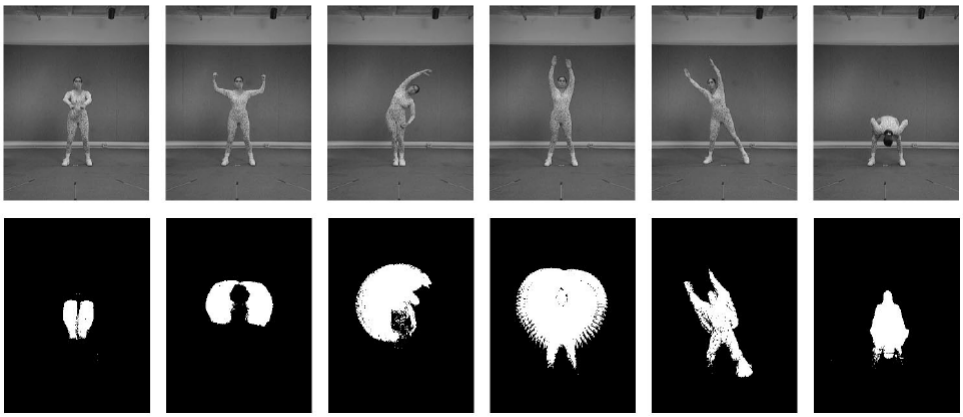
μοντελοποιούν παραμετρικά την αντιστοίχιση με φορά από την κλάση δράσης προς την εικόνα και έτσι είναι ικανές, στηριζόμενες συνήθως σε μοντέλα του ανθρώπινου σώματος, να γεννήσουν με κάποιο επαναληπτικό σχήμα την παρατήρηση, δοθείσας της ετικέτας δράσης. Οι *διακρίνουσες προσεγγίσεις* απο την άλλη ακολουθούν αντίστροφη φορά αντιστοίχισης, από την παρατήρηση προς την κλάση δράσης, χρησιμοποιούν δηλαδή δεδομένα εκμάθησης με σκοπό την εκπαίδευσή τους ώστε να διακρίνουν μεταξύ των δράσεων, δοθείσας μιας παρατήρησης.

Το πρόβλημα της αναγνώρισης ανθρώπινων δράσεων επομένως συνοψίζεται στην απόδοση ονομάτων που αντιστοιχούν σε κλάσεις δράσεων σε ακολουθίες εικόνων όπου θεωρείται δεδομένος, από προηγούμενο βήμα, ο εντοπισμός ανθρώπων και αντίστοιχων δράσεων. Η βασισμένη στην οπτική πληροφορία αναγνώριση δράσεων αποτελείται από δύο διακριτά βήματα, την εξαγωγή χαρακτηριστικών σε βίντεο και την μετέπειτα ταξινόμηση σε κλάσεις δράσεων των προκυπτουσών αναπαραστάσεων εικόνας. Τα χαρακτηριστικά εξάγονται είτε συνυπολογίζοντας την διάσταση του χρόνου στο στάδιο της αναπαράστασης των εικόνων είτε με βάση μόνο τις δύο χωρικές διαστάσεις με επεξεργασία κάθε frame της ακολουθίας ξεχωριστά. Στη δεύτερη περίπτωση, οι χρονικές μεταβολές πρέπει να ληφθούν υπόψη κατά τη διαδικασία της ταξινόμησης. Για την τελευταία υιοθετείται η μάθηση στατιστικών μοντέλων από τα χαρακτηριστικά ενός *Συνόλου Εκπαίδευσης (Training Set)* και έπειτα τα μοντέλα χρησιμοποιούνται για την ταξινόμηση νέων παρατηρήσεων. Συνήθης επιλογή για το στάδιο της ταξινόμησης είναι η χρήση ενός αλγορίθμου ταξινόμησης μηχανικής μάθησης.

Οι αναπαραστάσεις εικόνας πρέπει να είναι πλούσιες ώστε να προσδίδουν επαρκή διακριτική ικανότητα για το πρόβλημα της αναγνώρισης κάθε δράσης από τις υπόλοιπες. Ωστόσο, τα εξαγόμενα από τα βίντεο χαρακτηριστικά πρέπει ταυτόχρονα να διατηρούν ένα αξιόπιστο επίπεδο ευρωστίας απέναντι σε μεταβολές στην εμφάνιση των προσώπων, το στήσιμο του παρασκηνίου, τον τρόπο και την ταχύτητα εκτέλεσης των δράσεων, τη γωνία λήψης της κάμερας, τις συνθήκες φωτισμού ή το θόρυβο και την παρουσία οπτικών εμποδίων. Οι αναπαραστάσεις εικόνας μπορούν να διακριθούν, σύμφωνα με τον R. Poppe [5], σε δύο κατηγορίες: τις *τοπικές αναπαραστάσεις (local representations)* και τις *ολικές αναπαραστάσεις (global representations)*.

Οι *ολικές αναπαραστάσεις* βασίζονται σε πρώτο στάδιο σε τεχνικές *αφαίρεσης παρασκηνίου (background subtraction)* ή *εντοπισμού (tracking)* προκειμένου να απομονώσουν την ανθρώπινη μορφή στις εικόνες, ορίζοντας έτσι την *περιοχή ενδιαφέροντος*. Οι κωδι-

[2] http://www.di.ens.fr/willow/events/cvml2010/materials/INRIA_summer_school_2010_Ivan.pdf

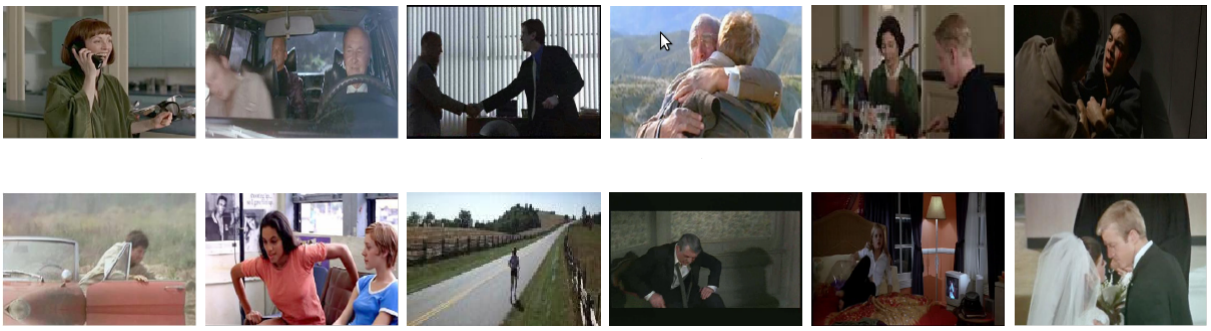


Σχήμα 1.2: Χαρακτηριστικά frames από ακολουθίες εικόνων με ασκήσεις aerobic και αντίστοιχες *Εικόνες Ενέργειας Κίνησης (Motion Energy Images)* από τους Bobick και Davis (επανεκτύπωση από το [1]).

κοποιημένες αναπαραστάσεις εικόνων υπολογίζονται επί του συνόλου της *περιοχής ενδιαφέροντος* και προκύπτουν από χαρακτηριστικά σχήματος (ανθρώπινες σιλουέτες, ακμές) ή μετρήσεις *οπτικής ροής (optical flow)*. Κάποιες προσεγγίσεις επιλέγουν μια χωρική δομή πλέγματος εισάγοντας έτσι ένα τοπικό επίπεδο κωδικοποίησης της παρατήρησης σε επιμέρους τμήματα-"κελιά" των εικόνων με στόχο να μετριάσουν την ευαισθησία της ολικής αναπαράστασης σε παραγόντες όπως ο θόρυβος, η γωνία λήψης και τα οπτικά εμπόδια. Σε κάθε περίπτωση, η συνένωση των χαρακτηριστικών σιλουέτας ή *οπτικής ροής* από όλα τα frames της ακολουθίας εικόνων οδηγεί σε έναν τρισδιάστατο όγκο που τροφοδοτεί το στάδιο της ταξινόμησης.

Χαρακτηριστικός εκπρόσωπος των *ολικών αναπαραστάσεων* που εξάγουν ανθρώπινες σιλουέτες από ακολουθίες εικόνων είναι η σχετική εργασία των Bobick και Davis [1], οι οποίοι υπολογίζουν και συσσωρεύουν τις διαφορές μεταξύ διαδοχικών frames του βίντεο. Η δυαδική εικόνα στην έξοδο που ονομάζουν *Εικόνα Ενέργειας Κίνησης (Motion Energy Image)* αποτελεί μια σύνοψη των περιοχών με παρουσία κίνησης της αρχικής ακολουθίας (Σχήμα 1.2). Επιπλέον, αποδίδοντας σε επίπεδο pixel υψηλότερες τιμές για πιο πρόσφατα εμφανιζόμενες κινήσεις της σιλουέτας στο δείγμα βίντεο παράγουν μια *Εικόνα Ιστορικού Κίνησης (Motion History Image)*. Οι Weinland et al. [7] επεκτείνουν την τελευταία στις τρεις διαστάσεις και συνδυάζουν σιλουέτες από πολλαπλές κάμερες σε ένα κοινό μοντέλο. Όταν είναι δύσκολο να επιτευχθεί η *αφαίρεση παρασκηνίου*, οι *ολικές αναπαραστάσεις* καταφεύγουν σε μετρήσεις οπτικής ροής, των προσανατολισμένων δηλαδή διαφορών μεταξύ των frames, εντός της *περιοχής ενδιαφέροντος*. Σε αυτή την κατηγορία εντάσσεται η εργασία των Efros et al. [8] οι οποίοι εφαρμόζουν ανόρθωση ημίσεως κύματος στην οριζόντια και κάθετη συνιστώσα της οπτικής ροής και λαμβάνουν τον τελικό περιγραφέα κίνησης από τη γκαουσιανή θόλωση των τεσσάρων προκυπτόντων καναλιών. Συμπερασματικά, οι *ολικές αναπαραστάσεις* συνιστούν μια πλούσια κωδικοποίηση της *περιοχής ενδιαφέροντος* με χαρακτηριστικά περιγραμμάτων των ανθρώπων ή οπτικής ροής. Ωστόσο, η εφαρμογή τους προαπαιτεί τον ακριβή εντοπισμό των ανθρώπων ή την *αφαίρεση παρασκηνίου* και κατά συνέπεια καθίστανται ευάλωτες σε παράγοντες

όπως η μερική απόκρυψη της ανθρώπινης φιγούρας, η οπτική γωνία ή ο θόρυβος. Η μελέτη *ολικών αναπαραστάσεων* για το πρόβλημα της Αναγνώρισης Ανθρώπινων Δράσεων σε Βίντεο δεν περικλείεται στους στόχους της παρούσας διπλωματικής εργασίας.
Οι *τοπικές αναπαραστάσεις* έχουν κερδίσει έντονο ερευνητικό ενδιαφέρον τα τελευταία χρόνια χάρη καί στην πρόσφατη επιτυχία τους σε άλλα προβλήματα της Οράσης Υπολογιστών όπως η αναγνώριση αντικειμένων, ανθρώπων ή υφής σε ακίνητες εικόνες. Η λογική τους βασίζεται στην περιγραφή των αρχικών ακολουθιών εικόνων μέσω μιας συλλογής ανεξάρτητων *τοπικών τεμαχίων (local patches)* περιγραφής. Η εξαγωγή χαρακτηριστικών συντίθεται στην ανίχνευση *οπτικά "σημαντικών "χωροχρονικών σημείων ενδιαφέροντος* στο δείγμα βίντεο και στην μετέπειτα περιγραφή της χωροχρονικής γειτονιάς τους με χαρακτηριστικά εμφάνισης ή κίνησης. Η τελική αναπαράσταση του αρχικού τρισδιάστατου όγκου βίντεο προκύπτει συνήθως από την εμπειρική στατιστική κατανομή σε αυτό πρότυπων τοπικών τεμαχίων που έχουν εξαχθεί από ένα *Σετ Εκπαίδευσης*, χωρίς να διατηρείται έτσι πληροφορία για τις χωροχρονικές συντεταγμένες των χαρακτηριστικών (βλ. Κεφάλαιο 4). Για τούτο, δεν απαιτείται μια εκ προοιμίου ρητή απομόνωση της *περιοχής ενδιαφέροντος* όπως συμβαίνει στις *ολικές αναπαραστάσεις*. Πρόσφατες έρευνες, ωστόσο, εξετάζουν τη συμβολή της γεωμετρικής συσχέτισης των τοπικών τεμαχίων στην απόδοση του συνολικού πλαισίου αναγνώρισης. Η παρούσα διπλωματική εργασία μελετά το πρόβλημα της αναγνώρισης δράσεων υπό το πρίσμα τέτοιων *τοπικών αναπαραστάσεων* και για αυτό δεν θα επεκταθούμε σε αυτό το σημείο στην παροχή περαιτέρω λεπτομερειών. Στοιχεία βιβλιογραφίας σχετικά με ανιχνευτές και περιγραφείς τοπικών χωροχρονικών σημείων ενδιαφέροντος σε ακολουθίες εικόνων μπορούν επομένως να αναζητηθούν στις αντίστοιχες ενότητες του Κεφάλαιου 2.
Αφού ολοκληρωθεί η εξαγωγή χαρακτηριστικών και διαμορφωθεί η τελική αναπαράσταση των ακολουθιών εικόνων, η αναγνώριση δράσεων μετασχηματίζεται πλέον σε ένα κοινό πρόβλημα ταξινόμησης. Οι δύο κατηγορίες πρακτικών που ακολουθούνται σε αυτό το στάδιο συνοψίζονται στην *απευθείας ταξινόμηση* και την *ταξινόμηση με χρονικά μοντέλα του χώρου κατάστασης (temporal state-space models)*. Συχνά πριν την ταξινόμηση εφαρμόζεται μια τεχνική μείωσης διαστάσεων στις αναπαραστάσεις, όπως η μέθοδος PCA. Η *απευθείας ταξινόμηση* δεν λαμβάνει υπόψη τη διάσταση του χρόνου για την αντιστοίχιση των δειγμάτων βίντεο στις κλάσεις δράσεων και υλοποιείται με *Ταξινομητές Κοντινότερου Γείτονα (Nearest Neighbor Classifiers)* ή *Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines)*. Στη δεύτερη κατηγορία, η ταξινόμηση βασίζεται στις πιθανότητες μεταξύ των καταστάσεων (states) και των παρατηρήσεων και σε αυτές των καταστάσεων μεταξύ τους. Μια κατάσταση μπορεί να ιδωθεί ως η σύλληψη των χαρακτηριστικών της δράσης σε μια δεδομένη χρονική στιγμή. Παραδείγματα τέτοιων προσεγγίσεων αποτελούν τα *Κρυφά Μαρκοβιανά Μοντέλα (Hidden Markov Models)* και η μέθοδος *Dynamic Time Warping*.
Η Αυτόματη Αναγνώριση Ανθρώπινων Δράσεων σε Βίντεο βρίσκει πολύ σημαντικές εφαρμογές σε μια εποχή στην οποία το βίντεο συνιστά μια πανταχού παρούσα πηγή πληροφορίας είτε σε μορφή προσωπικού αρχείου είτε σε διαθέσιμο online υλικό. Τα τελευταία χρόνια παρατηρείται μια ραγδαία αύξηση των δεδομένων βίντεο που είναι οργανωμένα σε ιστότοπους τέτοιου υλικού (YouTube, BBC Motion Gallery, Video Google). Παράλληλα, η μεγάλη διάδοση υψηλής χωρητικότητας μέσων αποθήκευσης επιτρέπει πλέον στον



Σχήμα 1.3: Ενδεικτικά frames δειγμάτων βίντεο από τη Βάση Δεδομένων *Hollywood2 Actions Dataset* που περιέχουν ανθρώπινες δράσεις (επανεκτύπωση από το [2]).

καθένα τη συλλογή μεγάλου όγκου δεδομένων ψηφιακού βίντεο. Τα παραπάνω υπαγορεύουν την ανάγκη ανάπτυξης προηγμένων μεθόδων αυτόματης ανάλυσης, ερμηνείας, περίληψης, ανάκτησης και ταξινόμησης δεδομένων βίντεο, με την αναγνώριση ανθρώπινων δράσεων να αξιολογείται ως πρωταρχικής σημασίας σε αυτό το πλαίσιο. Μια πιο "έξυπνη" αναζήτηση και περιήγηση σε βάσεις δεδομένων βίντεο διευκολύνει ταυτόχρονα εμπορικές εφαρμογές όπως είναι για παράδεγμα στη βιομηχανία της τηλεόρασης η οργάνωση και κατηγοριοποίηση του ψηφιακού αρχείου, η ανάκτηση των πιο σημαντικών γεγονότων σε βίντεο με αθλητικούς αγώνες ή της χειραψίας δύο πολιτικών αρχηγών σε μια κρίσιμη συνάντηση. Απο την άλλη, η στοχευμένη σε συγκεκριμένο περιεχόμενο αναζήτηση σε μια μεγάλη βάση δεδομένων βίντεο μπορεί να παρέχει γρήγορα σχετική πληροφορία για την εκπόνηση μιας επιστημονικής έρευνας, όπως για παράδειγμα η εξέταση της επίδρασης σκηνών καπνίσματος σε ταινίες στο κάπνισμα των εφήβων. Ακόμα, η ανίχνευση, η αναγνώριση και ο εντοπισμός ανθρώπινων κινήσεων χρησιμοποιείται ευρύτατα στη βιομηχανία του κινηματογράφου σε υλοποιήσεις που αφορούν animation και ειδικά εφέ.

Μια άλλη κατηγορία εφαρμογών στην οποία η αναγνώριση ανθρώπινων δράσεων συνδράμει αποφασιστικά, με τη μορφή της ανίχνευσης αντικανονικής ή ασυνήθιστης ανθρώπινης κίνησης ή συμπεριφοράς, είναι η οπτική επιτήρηση και παρακολούθηση ιδιωτικών ή δημόσιων χώρων. Η ανάπτυξη σχετικών αυτόνομων συστημάτων κρίνεται ουσιώδης σήμερα και εφαρμόζεται για τη φύλαξη οικιών ηλικιωμένων, χώρων μεγάλων εμπορικών κέντρων ή ακόμα και για στρατιωτικούς σκοπούς όπως η αυτόματη επιτήρηση περιοχών των συνόρων.

Τέλος, ο σχεδιασμός και η υλοποίηση διεπαφών αλληλεπίδρασης ανθρώπου-μηχανής επωφελείται ουσιαστικά από την αναγνώριση ανθρώπινων δράσεων. Η κατανόηση και ερμηνεία των ανθρώπινων κινήσεων και δράσεων συχνά αποτελούν αναπόσπαστο τμήμα της ανάπτυξης έξυπνων ρομποτικών μηχανισμών, υλικού διαδραστικών εφαρμογών για υπολογιστές ή εφαρμογών επαυξημένης πραγματικότητας (augmented reality). Επιπλέον, τεχνικές αναγνώρισης ανθρώπινων δράσεων συνεισφέρουν σημαντικά στην εξέλιξη της βιομηχανίας των παιχνιδιών για υπολογιστή.

Η Αναγνώριση Ανθρώπινων Δράσεων σε Βίντεο παραμένει μέχρι σήμερα ένα ιδιαίτερα απαιτητικό πρόβλημα της Όρασης Υπολογιστών. Αυτό οφείλεται πρωταρχικά σε παράγοντες όπως οι μεταβολές στην εμφάνιση, την έκφραση, την πόζα και την κίνηση

μεταξύ των δρώντων, ακόμα και σε δείγματα βίντεο που συγκαταλέγονται στην ίδια κατηγορία δράσης. Επιπρόσθετα σχετίζεται με την ύπαρξη διαφορετικών συνθηκών στον φωτισμό, τις κινήσεις της κάμερας και τις γωνίες λήψης και διαφορετικού επιπέδου παρουσίας θορύβου ή οπτικών εμποδίων. Στα παραπάνω δεν μπορουμε να παραλείψουμε τη σημαντική επίδραση που έχει στο πρόβλημα της αναγνώρισης το στήσιμο και η πολυπλοκότητα του παρασκηνίου μέσα στο οποίο εκτελούνται. Στόχος της σχετικής έρευνας, που επιδεικνύει έντονο ενδιαφέρον με υποσχόμενα αποτελέσματα την τελευταία δεκαετία, είναι ο σχεδιασμός εύρωστων απέναντι στους παραπάνω ανασταλτικούς παράγοντες σχημάτων που θα επιτυγχάνουν υψηλές επιδόσεις στο πρόβλημα της αναγνώρισης δράσεων επιφέρωντας παράλληλα το μικρότερο δυνατό υπολογιστικό κόστος υλοποίησης.
Η παρούσα διπλωματική εργασία μελετά το πρόβλημα της Αναγνώρισης Ανθρώπινων Δράσεων σε Βίντεο υπό το πρίσμα των τοπικών αναπαραστάσεων. Οι κυριότερες συνεισφορές της στο εν λόγω πρόβλημα συνοψίζονται στα ακόλουθα σημεία:

- Παρέχεται μια λεπτομερής επισκόπηση των τοπικών ανιχνευτών και περιγραφέων χωροχρονικών σημείων ενδιαφέροντος, με έμφαση στις κυριότερες σχετικές προσεγγίσεις.

- Τροποποιείται πρότερη υλοποίηση αποδιαμόρφωσης εικόνων και επιτυγχάνονται χαμηλότεροι χρόνοι εκτέλεσης που επιτρέπουν την εφαρμογή της σε ακολουθίες βίντεο.

- Εξετάζονται και ενσωματώνονται χαρακτηριστικά αποδιαμόρφωσης και υφής των εικόνων στο στάδιο της ανίχνευσης και περιγραφής των σημείων.

- Διεξάγονται πειράματα εκμάθησης και ταξινόμησης δράσεων στα οποία συγκρίνονται καθιερωμένα και εναλλακτικά σχήματα τοπικών χαρακτηριστικών.

- Σχεδιάζεται και υλοποιείται ένας νέος ανιχνευτής σημείων ενδιαφέροντος σε βίντεο, ο οποίος αξιολογείται για πρώτη φορά μέσω πειραμάτων αναγνώρισης δράσεων.

1.3 Διάρθρωση της Διπλωματικής Εργασίας

Στο **Κεφάλαιο 2** μελετάμε την εξαγωγή τοπικών χωροχρονικών χαρακτηριστικών σε ακολουθίες εικόνων για το πρόβλημα της αναγνώρισης ανθρώπινων δράσεων. Περιγράφουμε αρχικά τη μεθοδολογία που ακολουθείται για τη λήψη τέτοιων τοπικών αναπαραστάσεων και αναφέρουμε τα πλεονεκτήματά τους. Στη συνέχεια ασχολούμαστε με τους ανιχνευτές χωροχρονικών σημείων ενδιαφέροντος, παραθέτοντας αρχικά τους πιο σημαντικούς από αυτούς σύμφωνα με τη σχετική βιβλιογραφία. Παρουσιάζουμε διεξοδικά τον Ανιχνευτή Harris3D που χρησιμοποιείται σε πειράματα αναγνώρισης σε επόμενα κεφάλαια και τον Ανιχνευτή Cuboid που υλοποιήθηκε στα πλαίσια της παρούσας εργασίας ενώ εξετάζουμε καί τη μέθοδο της πυκνής δειγματοληψίας σε δείγματα βίντεο. Το υπόλοιπο μέρος του κεφαλαίου αφιερώνεται στη μελέτη των τοπικών περιγραφέων χωροχρονικών σημείων, ξεκινώντας με παρόμοιο τρόπο από την επισκόπηση των κυριότερων σχετικών προσεγγίσεων. Έπειτα αναφερόμαστε εκτενώς στους θεμελιώδεις περιγραφείς

προσανατολισμού εμφάνισης και κίνησης των Dalal και Triggs και τελειώνουμε το κεφάλαιο με την παρουσίαση του περιγραφέα HOG/HOF, που αποτελεί τη βασική επιλογή για το στάδιο της περιγραφής στα μαζικά πειράματα της εργασίας. Οι ανιχνευτές και περιγραφείς που αναλύονται με λεπτομέρεια συνοδεύονται από ενδεικτικά πειραματικά αποτελέσματα σε μεμονωμένα δείγματα βίντεο.

Στο **Κεφάλαιο 3** εξετάζουμε την AM-FM αποδιαμόρφωση εικόνων με τη χρήση τελεστών ενέργειας και τις προκύπτουσες αναπαραστάσεις της υφής. Αρχικά παρέχουμε το απαραίτητο μαθηματικό υπόβαθρο σχετικά με τον τελεστή ενέργειας Teager-Kaiser και τις επεκτάσεις του. Έπειτα παρουσιάζουμε τον αλγόριθμο διαχωρισμού της ενέργειας στα σήματα πλάτους και συχνότητας διαμόρφωσης των Maragos και Bovik για την αποδιαμόρφωση συνεχών και διακριτών σημάτων στενής ζώνης συχνοτήτων. Στο τρίτο μέρος του κεφαλαίου μελετάμε την αποδιαμόρφωση εικόνων που ακολουθούν το μοντέλο πολλαπλών AM-FM συνιστωσών, αναφερόμενοι αρχικά στο φιλτράρισμα σε πολλαπλές ζώνες συχνοτήτων και στην αποδιαμόρφωση στις εξόδους των επιμέρους καναλιών της συστοιχίας φίλτρων. Συνεχίζουμε με την παρουσίαση του Gabor αλγορίθμου διαχωρισμού της ενέργειας των Kokkinos et al. και κατόπιν εξετάζουμε την ανάλυση συνιστωσών καναλιού (CCA) και τις δύο μορφές της ανάλυσης κυρίαρχων συνιστωσών (ADCA, EDCA) για την αναπαράσταση εικόνων υφής. Τέλος, παρατίθενται οι λεπτομέρειες υλοποίησης των σχημάτων CCA και DCA και απεικονίζονται χαρακτηριστικά υφής που λάβαμε από την εφαρμογή τους σε frames δειγμάτων βίντεο.

Στο **Κεφάλαιο 4** αρχικά αναφερόμαστε στην προσέγγιση Bag-Of-Features, εξηγώντας έτσι τον τρόπο δημιουργίας της τελικής αναπαράστασης των βίντεο από τη συλλογή των τοπικών χαρακτηριστικών. Έχοντας παρουσιάσει συνοπτικά τις βασικές αρχές λειτουργίας των μηχανών ταξινόμησης SVM, προχωρούμε στην διεξαγωγή πειράματος αναγνώρισης έξι δράσεων της βάσης δεδομένων Hollywood2 με τη χρήση του ανιχνευτή Harris3D και του περιγραφέα HOG/HOF. Έπειτα αναφέρουμε λεπτομερώς τα διακριτά στάδια υλοποίησης του υπολογισμού της κυρίαρχης EDCA ενέργειας των εικόνων που προσαρμόσαμε για αποτελεσματική χρήση σε ακολουθίες εικόνων. Τέλος, προτείνουμε δύο εναλλακτικά σχήματα ενσωμάτωσης της EDCA ενέργειας των frames στο στάδιο της ανίχνευσης σημείων από τον Harris3D ή καί της περιγραφής τους από τον HOG/HOF και αξιολογούμε συγκριτικά την απόδοσή τους στο κοινό πλαίσιο του πειράματος αναγνώρισης των έξι δράσεων.

Στο **Κεφάλαιο 5** στηριζόμαστε σε πρόσφατες ιδέες των Maragos, Evangelopoulos και Dimitriadis για την ανάπτυξη και υλοποίηση ενός νέου ανιχνευτή χωροχρονικών σημείων ενδιαφέροντος που ονομάζουμε dca3D και ο οποίος εφαρμόζει χωροχρονικό φιλτράρισμα με συστοιχίες Gabor φίλτρων και ανάλυση κυρίαρχων συνιστωσών στον χώρο και το χρόνο. Αρχικά αναλύουμε τα διάφορα στάδια υπολογισμού της συνάρτησης απόκρισης του ανιχνευτή τόσο από πλευράς μαθηματικής θεμελίωσης όσο και από πλευράς περαιτέρω σχεδιαστικών επιλογών. Στη συνέχεια αναφέρονται οι λεπτομέρειες της υλοποίησης και εμφανίζουμε χαρακτηριστικά πειραματικά αποτελέσματα σε δείγματα βίντεο. Στο τελευταίο μέρος του κεφαλαίου παρουσιάζονται πειράματα αναγνώρισης σε υποσύνολα δράσεων της βάσης δεδομένων Hollywood2 με ανιχνευτή τον dca3D που συνοδεύονται από αντίστοιχα αποτελέσματα με βάση το σχήμα Harris3D - HOG/HOF.

Στο **Κεφάλαιο 6** γίνεται μια συμπερασματική αποτίμηση του έργου της παρούσας

διπλωματικής εργασίας και της συμβολής της στο πρόβλημα υπό εξέταση. Τέλος, προτείνονται κάποιες εφικτές προεκτάσεις για μελλοντική έρευνα.

Κεφάλαιο 2

Τοπικά Χωροχρονικά Χαρακτηριστικά για την Αναπαράσταση Ανθρώπινων Δράσεων σε Βίντεο

2.1 Γενικά

Τα τοπικά χαρακτηριστικά ή τεμάχια στατικών εικόνων ή ακολουθιών εικόνων (videos) έχουν αποδειχθεί εξαιρετικά επιτυχή για διάφορα προβλήματα αναγνώρισης της Όρασης Υπολογιστών όπως η αναγνώριση αντικειμένων ή σκηνών καθώς επίσης και για την αναγνώριση ανθρώπινων δράσεων σε βίντεο. Τέτοιες τοπικές αναπαραστάσεις περιγράφουν την παρατήρηση ως μια συλλογή τοπικών περιγραφέων (descriptors) ή τεμαχίων (patches) που υπολογίζονται στη γειτονιά ανιχνευθέντων σημείων ενδιαφέροντος ή σημείων που προέκυψαν από πυκνή δειγματοληψία, ακολουθώντας έτσι μια καθοδική (bottom-up) προσέγγιση. Τα τοπικά αυτά τεμάχια δεν έχουν καμία σύνδεση με τις συντεταγμένες της εικόνας ή του βίντεο εκτός των περιπτώσεων στις οποίες ομαδοποιούνται σε πλέγματα (grids), διατηρώντας ως ένα βαθμό την τοπική ή χρονική πληροφορία. Κατόπιν, ενσωματώνονται σε μια τελική τοπική αναπαράσταση ικανή να συλλάβει την εμπειρική στατιστική κατανομή των τοπικών χαρακτηριστικών από όλες τις περιοχές πριν οδηγηθεί στα επόμενα στάδια της αναγνώρισης.
Περνώντας στο πρόβλημα της αναγνώρισης ανθρώπινων δράσεων, τα τοπικά χαρακτηριστικά αποτελούν πλέον κοινό εργαλείο της Όρασης Υπολογιστών για αυτή την εφαρμογή κερδίζοντας έδαφος έναντι των ολιστικών προσεγγίσεων σε περιπτώσεις που οι τελευταίες βρίσκουν αντικειμενικές δυσκολίες επιτυχημένης υλοποίησης. Τα τοπικά τεμάχια είναι δισδιάστατα ή τρισδιάστατα και εξάγονται απευθείας από το βίντεο εισόδου.

Οι λόγοι για τους οποίους η χρήση τους έχει γίνει τόσο διαδεδομένη συνοψίζονται ως εξής

- Ανιχνεύουν επιτυχώς χαρακτηριστικά κίνησης και εμφάνισης.
- Δεν απαιτούν σαφή ανίχνευση και εντοπισμό του ανθρώπου, διαχωρισμό προσκηνίου - παρασκηνίου ή κατάτμηση κίνησης.
- Παρέχουν μια σχετικά ανεξάρτητη αναπαράσταση σε σχέση με χωροχρονικές μετατοπίσεις και κλίμακες.
- Εξασφαλίζουν ευρωστία απέναντι στο θόρυβο παρασκηνίου, στη μερική εμφάνιση οπτικών εμποδίων και στην ύπαρξη πολλαπλών κινήσεων στη σκηνή.
- Συνήθως εξάγονται χωρίς την ανάγκη προεπεξεργασίας των frames.

Στις ακολουθίες εικόνων, τα τοπικά χαρακτηριστικά εννοιολογικά συμπεριλαμβάνουν δύο διακριτές διεργασίες, την ανίχνευση ή πυκνή δειγματοληψία χωροχρονικών σημείων και την τοπική περιγραφή τους. Κάποιοι ανιχνευτές ή περιγραφείς αποτελούν αποτέλεσμα επέκτασης επιτυχημένων δισδιάστατων ομοίων τους στην διάσταση του χρόνου ενώ άλλοι αναπτύχθηκαν επί τούτου για εφαρμογές σε ακολουθίες εικόνων, όπως αυτές του εντοπισμού ή αναγνώρισης ανθρώπινων δράσεων.
Οι ανιχνευτές τοπικών χαρακτηριστικών αναζητούν χωροχρονικά σημεία και κλίμακες ενδιαφέροντος που αντιστοιχούν σε περιοχές μη σταθερής και σύνθετης κίνησης στο βίντεο εισόδου μεγιστοποιώντας συναρτήσεις οπτικής σημαντικότητας (saliency). Εναλλακτικά της ανίχνευσης σημείων σε διάφορες κλίμακες με βάση τα τοπικά μέγιστα μιας συνάρτησης απόκρισης μπορεί να συντελεστεί δειγματοληψία σε τακτές θέσεις και συνδυασμούς χωροχρονικών κλιμάκων. Αυτή η δειγματοληψία εκτελείται με χωρική και χρονική επικάλυψη, εξάγει πολύ μεγαλύτερο πλήθος σημείων σε σχέση με τους ανιχνευτές και έχει παρουσιάσει πρόσφατα επιτυχία στο πρόβλημα της αναγνώρισης αντικειμένων και ανθρώπινων δράσεων σε ρεαλιστικά σενάρια.
Οι τοπικοί περιγραφείς υπολογίζουν το σχήμα και την κίνηση στη γειτονιά των ήδη επιλεγμένων σημείων μέσω υπολογισμών όπως τα 2D ή 3D gradients ή της οπτικής ροής συνεχόμενων frames. Οι περιγραφείς υπολογίζουν τα επιθυμητά χαρακτηριστικά είτε στο ενιαίο χωροχρονικό τεμάχιο που περιβάλλει το επιλεγμένο σημείο είτε ξεχωριστά σε επιμέρους κελιά (cells) στα οποία διαιρείται το τεμάχιο.
Οι πιο δημοφιλείς τοπικοί ανιχνευτές και περιγραφείς παρουσιάζονται συνοπτικά στις ενότητες 2.2 και 2.3. Επιλέξαμε να παρουσιάσουμε την πυκνή δειγματοληψία στην ενότητα των ανιχνευτών χωροχρονικών σημαντικών σημείων λόγω της αξιολόγησης τους σε κοινό πλαίσιο σε πολλά άρθρα της σχετικής βιβλιογραφίας ([3]) και προς χάριν συνέχειας των νοημάτων ως προς τα διακριτά στάδια της αναγνώρισης ανθρώπινων δράσεων. Επιμέρους ανιχνευτές και περιγραφείς που αξιολογήσαμε εκτενέστερα στην παρούσα διπλωματική παρουσιάζονται στη συνέχεια σε ξεχωριστές υποενότητες του παρόντος κεφαλαίου.
Αξίζει να σημειώσουμε ότι υπάρχει μια αυξανόμενη τάση στην έρευνα για την εκμετάλλευση συσχετίσεων στο χώρο και στο χρόνο των τοπικών τεμαχίων. Αυτό διευκολύνει σημαντικά την αναγνώριση αποδίδοντας ετικέτες στα τοπικά τεμάχια ανάλογα με την περιοχή στην οποία βρίσκονται, επικεντρώνοντας έτσι την εξαγωγή χαρακτηριστικών στα

τεμάχια που εμπίπτουν στην περιοχή ενδιαφέροντος. Ωστόσο, η μελέτη και αξιολόγηση τέτοιων μεθόδων δεν συμπεριλαμβάνεται στους στόχους της παρούσας διπλωματικής.

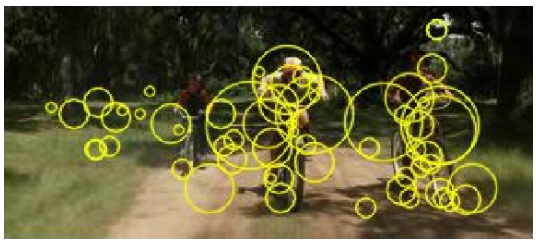
2.2 Τοπικοί Ανιχνευτές Χωροχρονικών Σημείων Ενδιαφέροντος

Οι Ανιχνευτές Χωροχρονικών Σημείων Ενδιαφέροντος (Spatiotemporal Interest Points Detectors) αναζητούν σημεία σε ακολουθίες εικόνων που χαρακτηρίζονται από απότομες μεταβολές στην κίνηση ή την εμφάνιση. Είναι συνήθως σημεία που έχουν τη δομή γωνιών (corners) ή σταγόνων (blobs) και είναι προικισμένα με τη μεγαλύτερη διακριτική ικανότητα και πληροφοριακό περιεχόμενο για το πρόβλημα της αναγνώρισης ανθρώπινων δράσεων. Οι ανιχνευτές τέτοιων σημείων συνήθως πυροδοτούν ανιχνεύσεις σε σημεία που παρουσιάζουν μια πολύπλοκη και όχι μια απλά μεταγραφική (translational) κίνηση στο χρόνο.
Πολλοί ανιχνευτές χωροχρονικών οπτικά σημαντικών (salient) σημείων έχουν επινοηθεί τα τελευταία χρόνια και στέκονται απέναντι στη μέθοδο πυκνής δειγματοληψίας αντλώντας πιο αραιά αλλά εύρωστα και ξεχωριστά σημεία στο βίντεο εισόδου. Θεμελιώνονται συνήθως με βάση μια μαθηματική συνάρτηση απόκρισης της οποίας τα τοπικά μέγιστα αντιστοιχούν σε εξέχουσες περιοχές ενδιαφέροντος στο χώρο και το χρόνο. Αρκετές φορές επιτελούν κατωφλιοποίηση στην έξοδο τους για τη ρύθμιση της πυκνότητας των εξαγόμενων σημείων ενδιαφέροντος χωρίς σε καμία περίπτωση να επιτυγχάνουν σταθερό πλήθος ανιχνεύσεων άνα frame κάτι που θα αντίβαινε στην λογική τους που βασίζεται στην άντληση σημείων ανάλογα με την υπάρχουσα μεταβλητή πληροφορία στο χώρο και το χρόνο. Αποφεύγουν ανιχνεύσεις στα σύνορα των πλαισίων των βίντεο εισόδου που αντιστοιχούν σε επισφαλή "σημαντικά"σημεία. Η υψηλή υπολογιστική πολυπλοκότητα της πλειονότητας των ανιχνευτών επιτρέπει την εφαρμογή τους σε μικρής διάρκειας ή χαμηλής ανάλυσης βίντεο. Εντούτοις, τα τελευταία χρόνια έχουν εμφανιστεί διαθέσιμες online υλοποιήσεις δημοφιλών ανιχνευτών που παρουσιάζουν υψηλές ταχύτητες επεξεργασίας (προγραμματισμένες σε C, C++ ή MATLAB).
Οι Ανιχνευτές Χωροχρονικών Σημείων ενδιαφέροντος διαφέρουν ως προς τον τύπο τους, τη μορφή της συνάρτησης οπτικής σημαντικότητας που εισάγουν, τις διαφορετικές δομές σημείων που αναζητούν, την αραιότητα του πλήθους σημείων που εξάγουν, την υπολογιστική πολυπλοκότητά τους και το αναλλοίωτο ή όχι ως προς τις χωρικές και χρονικές κλίμακες ανίχνευσης. Μιλώντας για την επιλογή των χωροχρονικών κλιμάκων στις οποίες συντελείται η ανίχνευση, έχουν αναπτυχθεί ανιχνευτές αναλλοίωτοι ως προς την κλίμακα που υποστηρίζουν την αυτόματη επιλογή κλίμακας. Άλλοι απαιτούν τον ορισμό από το χρήστη μιας μοναδικής κλίμακας ανάλογα με την εφαρμογή ενώ σε άλλες περιπτώσεις υλοποιείται ανίχνευση σε προκαθορισμένες πολλαπλές κλίμακες χώρου και χρόνου με σκοπό την εξαγωγή πιο χονδροειδών σημείων ή γεγονότων σε μεγάλες κλίμακες και πιο λεπτομερών σε μικρές κλίμακες.
Ξεκινώντας την επισκόπηση των υπαρχόντων μεθόδων ανίχνευσης σημαντικών σημείων σε ακολουθίες εικόνων, οι Laptev και Lindeberg [9] επεκτείνουν τον ανιχνευτή γωνιών Harris [10] στις τρείς διαστάσεις. Ο ανιχνευτής **Harris3D**, όπως αναφέρεται συχνά

στη βιβλιογραφία, πυροδοτεί ανιχνεύσεις σε σημεία που επιδεικνύουν υψηλή μεταβολή των τιμών της εικόνας στο χώρο και μη σταθερή κίνηση στο χρόνο. Ο I. Laptev στο [11] προτείνει έναν επαναληπτικό αλγόριθμο για την αυτόματη επιλογή χωρικής και χρονικής κλίμακας με στόχο την ανάκτηση της χωροχρονικής έκτασης κάθε ανιχνευμένου γεγονότος, ενώ η σχετική εργασία επεκτείνεται για να συμπεριλάβει και αντιστάθμιση των σχετικών κινήσεων της κάμερας στο [12]. Ο ανιχνευτής Harris3D αποτελεί σημείο αναφοράς για τη σχετική έρευνα των χωροχρονικών σημείων ενδιαφέροντος ενώ ο συνδυασμός του με περιγραφείς που θα μελετήσουμε στην επόμενη ενότητα παρέχει υψηλά ποσοστά αναγνώρισης σε γνωστές βάσεις δεδομένων ανθρώπινων δράσεων. Εξετάστηκε και εφαρμόστηκε εκτενώς σε πειράματα της παρούσας διπλωματικής και για τούτο θα παρουσιαστεί με λεπτομέρεια στην υποενότητα 2.2.1.
Οι Dollár et al. [13] ισχυρίζονται ότι απλές τρισδιάστατες επεκτάσεις των κοινών δισδιάστατων ανιχνευτών σημείων ενδιαφέροντος είναι ακατάλληλες για την εξαγωγή τοπικών χωροχρονικών χαρακτηριστικών και βασίζονται σε χρονικά Gabor φίλτρα για την ανάπτυξη ενός νέου ανιχνευτή, του ανιχνευτή **Cuboid**. Τα τοπικά μέγιστα της συνάρτησης απόκρισης στην οποία θεμελιώνουν τον ανιχνευτή αντιστοιχούν σε περιοχές με διακριτικά χαρακτηριστικά στο χώρο που περιέχουν περιοδικούς συντελεστές συχνότητας ή γενικότερα υφίστανται μια σύνθετη κίνηση. Ο ανιχνευτής Cuboid δε συλλαμβάνει περιοχές που διέπονται από απλή μεταγραφική κίνηση ούτε περιοχές που δεν εμπεριέχουν διακριτές μεταβολές στην ένταση της εικόνας και απαιτεί την επιλογή χωρικής και χρονικής κλίμακας από το χρήστη. Αποτελεί έναν από τους πιο δημοφιλείς ανιχνευτές αραιών χωροχρονικών σημείων ενδιαφέροντος σε ακολουθίες εικόνων και συχνά χρησιμοποιείται ως μηχανή παραγωγής "σημαντικών"σημείων για την αξιολόγηση τοπικών περιγραφέων λόγω της ευρωστίας του. Ασχοληθήκαμε ιδιαίτερα με τον παραπάνω ανιχνευτή καί σε πλαίσιο πειραμάτων και εκθέτουμε τις λεπτομέρειες θεμελίωσής του με μεγαλύτερη λεπτομέρεια και τα προκύπτοντα πειραματικά αποτελέσματα σε βίντεο στην υποενότητα 2.2.2.
Οι Oikonomopoulos et al. [14] επεκτείνουν την εργασία των Kadir και Brady [15] για την ανίχνευση δισδιάστατων οπτικά "σημαντικών"σημείων στις τρείς διαστάσεις. Ακολουθούν μια προσέγγιση βασισμένη στη θεωρία της πληροφορίας ενσωματώνοντας την εντροπία των κυλινδρικών χωροχρονικών γειτονιών των σημείων σε πολλαπλές κλίμακες στο μετρικό οπτικής σημαντικότητας που προτείνουν. Εκτελώντας κατωφλιοποίηση με βάση τις τιμές της συνάρτησης απόκρισης και συσταδοποίηση καταλήγουν στην εξαγωγή "σημαντικών"περιοχών ενώ οι κλίμακες επιλέγονται αυτόματα ως εκείνες που αντιστοιχούν στα μέγιστα τιμών της εντροπίας.
Οι Rapantzikos et al. [16] εφαρμόζουν διακριτούς μετασχηματισμούς κυματιδίων σε καθεμιά από τις διαστάσεις του όγκου του βίντεο εισόδου και χρησιμοποιούν ως μετρικό σημαντικότητας το άθροισμα των ενεργειών όλων των υπο-ζωνών φιλτραρίσματος στη γειτονιά κάθε στοιχειώδους όγκου. Με αυτή την προσέγγιση κυματιδιακών μετασχηματισμών ανακτούν τις χωροχρονικές συχνότητες και κατ΄ επέκταση κατευθύνσεις του σήματος που αποδεικνύονται σημαντικά πληροφοριακές για την ανίχνευση και ανάλυση δυναμικών γεγονότων στο βίντεο εισόδου. Οι ίδιοι συγγραφείς πρόσφατα [17] ενσωματώνουν χαρακτηριστικά όπως η ένταση, το χρώμα και η κατεύθυνση (κίνηση) σε μια διαδικασία ολικής ελαχιστοποίησης για την εκτίμηση της οπτικής σημαντικότητας.

(a) **Harris3D**

(b) **Cuboid**



(c) **Hessian**

Σχήμα 2.1: Ανιχνεύσεις στο ίδιο frame δείγματος βίντεο της Βάσης Δεδομένων *Hollywood2* από τρείς διαφορετικούς Ανιχνευτές (επανεκτύπωση από τις διαφάνειες του A. Kläser[1]).

Οι Wong et al. [18] αντί για την ανίχνευση χωροχρονικών σημείων στις τρείς διαστάσεις, διαμερίζουν το βίντεο εισόδου σε "συνιστώσες κίνησης" που αντιστοιχούν χονδρικά στα κινούμενα μέρη του σώματος και αναζητούν δισδιάστατα και μονοδιάστατα σημεία ενδιαφέροντος στους προκύπτοντες υποχώρους και τους συντελεστές τους αντίστοιχα. Οι Bregonzio et al. [19] αρχικά υπολογίζουν τις διαφορές μεταξύ των πλαισίων του βίντεο για την εξαγωγή των περιοχών όπου επικεντρώνεται το ενδιαφέρον και στη συνέχεια εφαρμόζουν φιλτράρισμα με δισδιάστατα Gabor φίλτρα σε διάφορες κατευθύνσεις στις επιλεγμένες περιοχές. Οι Gilbert et al. [20] εφαρμόζουν τον δισδιάστατο ανιχνευτή γωνίων Harris στους τρείς συνδυασμούς καναλιών $(x, y), (x, t), (y, t)$ και εξάγουν πολύ πυκνότερες γωνίες από τους Laptev et al. τις οποίες στη συνέχεια ομαδοποιούν ιεραρχικά σε σύνθετα χαρακτηριστικά που αποδεικνύονται εξαιρετικής διακριτικής ικανότητας για την βάση δεδομένων KTH.
Θα ολοκληρώσουμε την σύντομη επισκόπηση των Ανιχνευτών Χωροχρονικών Σημείων Ενδιαφέροντος με τον Ανιχνευτή **Hessian** που μαζί με τους ανιχνευτές Harris3D και Cuboid αποτελεί έναν από τους πιο δημοφιλείς στη βιβλιογραφία και συνοδεύεται επίσης από μια online διαθέσιμη υλοποίηση[2]. Αναπτύχθηκε από τους Willems et al. [21] ως τρισδιάστατη επέκταση του Hessian μετρικού σημαντικότητας για την ανίχνευση δομών "σταγόνας" σε εικόνες [22]. Οι ανιχνεύσεις αντιστοιχούν στα τοπικά μέγιστα της ορίζουσας της τρισδιάστατης μήτρας Hessian ενώ οι χωρικές και χρονικές κλίμακες επιλέγονται αυτόματα χωρίς τη χρήση επαναληπτικού σχήματος. Για την επιτάχυνση της υλοποίησης γίνεται χρήση των *ολοκληρωτικών βίντεο* (*integral videos*) και η ορίζουσα

[1] http://videolectures.net/bmvc09_klaser_elst/
[2] http://homes.esat.kuleuven.be/~gwillems/research/Hes-STIP/

της 3D Hessian υπολογίζεται σε διάφορες οκτάβες καθεμία από τις οποίες αποτελείται από πέντε διακριτές χωρικές ή χρονικές κλίμακες, με τις τρείς εσωτερικές να προοδεύουν με λόγο ανάμεσα στις τιμές 1.2 και 1.5. Μετά τον υπολογισμό όλων των κυβοειδών ενεργοποιείται ένας αλγόριθμος καταστολής των μη μεγίστων (*non-maximum suppression algorithm*) για την ανάκτηση των μεγίστων στον χώρο των πέντε διαστάσεων που αποτελείται από τις συντεταγμένες (x, y, t) και τις κλίμακες (σ, τ).

2.2.1 Ο Ανιχνευτής Harris3D

Στην παρούσα υποενότητα θα ασχοληθούμε εκτενέστερα με έναν από τους πιο σημαντικούς και αποτελεσματικούς Ανιχνευτές Χωροχρονικών Σημείων Ενδιαφέροντος που αποτέλεσε πολύτιμο εργαλείο για πολλά από τα πειράματα που εκπονήθηκαν στα πλαίσια της διπλωματικής αυτής εργασίας. Πρόκειται για τον Ανιχνευτή *Harris3D* που προτάθηκε από τους I. Laptev και T. Lindeberg [9] το 2003 και χρησιμοποιήθηκε εκτενώς για την παραγωγή τοπικών χαρακτηριστικών ενδιαφέροντος στο πλαίσιο του προβλήματος της αναγνώρισης ανθρώπινων δράσεων σε ακολουθίες εικόνων. Παρουσιάζεται στη βιβλιογραφία σε συνδυασμό με πολλούς διαφορετικούς τοπικούς περιγραφείς και εφαρμόζεται στις πιο γνωστές βάσεις δεδομένων ανθρωπίνων δράσεων. Χωρίς αμφιβολία συνιστά την πρώτη αναφορά όταν γίνεται λόγος σε επιστημονικά άρθρα για τη σχετική εργασία στους ανιχνευτές χωροχρονικών σημείων ενδιαφέροντος στο πλαίσιο της ανάλυσης, αναπαράστασης και αναγνώρισης ανθρώπινων δράσεων.
Ο Ανιχνευτής *Harris3D* αναζητά και εξάγει τοπικές δομές στον χώρο και το χρόνο όπου οι τιμές της εικόνας χαρακτηρίζονται από σημαντικές τοπικές μεταβολές τόσο στις χωρικές όσο και στην χρονική διάσταση. Με άλλα λόγια, σημεία της ακολουθίας εικόνων που παρουσιάζουν διακριτική εμφάνιση στο χώρο και διέπονται από μη σταθερή κίνηση στο χρόνο πυροδοτούν την ανίχνευση από τον *Harris3D*. Οι συγγραφείς διατείνονται ότι τέτοια χαρακτηριστικά αντιστοιχούν σε γεγονότα με έντονο πληροφοριακό περιεχόμενο στο βίντεο εισόδου.
Η ιδέα του τρισδιάστατου ανιχνευτή των Laptev et al. στηρίχθηκε στις ιδέες δισδιάστατων ανιχνευτών σημείων ενδιαφέροντος των Harris και Förstner ([10], [23]) και για τούτο θα παρουσιάσουμε αρχικά το μαθηματικό υπόβαθρο για την ανίχνευση σημείων στο χώρο.

Χωρικά Σημεία Ενδιαφέροντος. Ο *Harris* ανιχνευτής γωνιών ανιχνεύει σημεία που επιδεικνύουν σημαντική μεταβολή στις τιμές της εικόνας $f(x, y)$ καί στις δύο καρτεσιανές συνιστώσες. Αν θεωρήσουμε ως τοπική κλίμακα της παρατήρησης την σ_l^2, με τη χρήση γκαουσιανών παραγώγων και μιας κλίμακας ολοκλήρωσης $\sigma_i^2 = s\sigma_l^2$, όπου s μια σταθερά, κατασκευάζεται μια παραθυροποιημένη *μήτρα δεύτερων στιγμών*

$$M = g\left(\cdot; \sigma_i^2\right) * \begin{pmatrix} L_x^2 & L_x L_y \\ L_x L_y & L_y^2 \end{pmatrix} \tag{2.2.1}$$

όπου οι γκαουσιανές παράγωγοι L_x και L_y ορίζονται ως

$$\begin{aligned} L_x\left(\cdot; \sigma_l^2\right) &= \partial_x\left(g\left(\cdot; \sigma_l^2\right) * f\right) \\ L_y\left(\cdot; \sigma_l^2\right) &= \partial_y\left(g\left(\cdot; \sigma_l^2\right) * f\right) \end{aligned} \tag{2.2.2}$$

και g είναι ο δισδιάστατος γκαουσιανός πυρήνας

$$g(x, y; \sigma^2) = \frac{1}{2\pi\sigma^2} \exp(-\left(x^2 + y^2\right)/2\sigma^2) \tag{2.2.3}$$

Για την αποφυγή σύγχυσης μεταξύ των κλιμάκων σ_l και σ_i σημειώνουμε ότι η πρώτη είναι η τοπική κλίμακα (*local scale*), με την οποία υπολογίζονται οι γκαουσιανές παράγωγοι L_x και L_y, ενώ η κλίμακα σ_i (*integration scale*) προκύπτει από τη σχέση $\sigma_i^2 = s\sigma_l^2$ και είναι η κλίμακα του γκαουσιανού πυρήνα $g\left(\cdot; \sigma_i^2\right)$ που επιτελεί averaging της μήτρας M.
Αν συμβολίσουμε με λ_1, λ_2 τις ιδιοτιμές της μήτρας M όπου $\lambda_1 \leq \lambda_2$, αυτές αντιπροσωπεύουν σημαντικές μεταβολές της εικόνας f στις δύο διαστάσεις και επομένως σημαντικές τιμές τους αντιστοιχούν σε σημεία ενδιαφέροντος. Σύμφωνα με τους Harris και Stephens [10] τα χωρικά σημεία ενδιαφέροντος μπορούν να βρεθούν από τα θετικά μέγιστα της συνάρτησης γωνιών που θεμελιώνουν ως εξής

$$H = \det(M) - k\mathrm{trace}^2(M) = \lambda_1\lambda_2 - k(\lambda_1 + \lambda_2)^2 \tag{2.2.4}$$

όπου με τους συμβολισμούς *det* και *trace* αναφερόμαστε στην *ορίζουσα* και το *ίχνος* αντίστοιχα της μήτρας M.

Χωροχρονικά Σημεία Ενδιαφέροντος. Ο δισδιάστατος ανιχνευτής χωρικών οπτικά "σημαντικών"σημείων σε ακίνητες εικόνες επεκτείνεται από τους Laptev και Lindeberg [9] στο πεδίο του χωρο-χρόνου ώστε να είναι σε θέση να ανιχνεύει σημεία που έχουν εξέχουσες τιμές εικόνας καί στη χωρική και στη χρονική διάσταση. Σημεία που εντάσσονται σε αυτό το προφίλ των 3D γωνιών είναι τα χωρικά προεξέχοντα σημεία στις στιγμές που υφίστανται μια μη σταθερή κίνηση σε σχέση με τη χωροχρονική γειτονία τους. Η ακολουθία εικόνων $f : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$ μοντελοποιείται με τη γραμμική αναπαράσταση κλίμακας-χώρου $L : \mathbb{R}^2 \times \mathbb{R} \times \mathbb{R}_+^2 \mapsto \mathbb{R}$ συνελίσσοντας την $f(x, y, t)$ με τον γκαουσιανό πυρήνα με χωρική και χρονική μεταβλητότητα αντίστοιχα σ_l^2 και τ_l^2

$$L(\cdot; \sigma_l^2, \tau_l^2) = g(\cdot; \sigma_l^2, \tau_l^2) * f \tag{2.2.5}$$

Για τον χωροχρονικά διαχωρίσιμο γκαουσιανό πυρήνα ισχύει

$$g(x, y, t; \sigma_l^2, \tau_l^2) = \frac{\exp(-(x^2 + y^2)/2\sigma_l^2 - t^2/2\tau_l^2)}{\sqrt{(2\pi)^3\sigma_l^4\tau_l^2}} \tag{2.2.6}$$

Εδώ η 3×3 *μήτρα δεύτερων στιγμών* κατασκευάζεται πάλι από τις γκαουσιανές παραγώγους ως προς τις χωρικές και τη χρονική συντεταγμένη και συνελίσσεται με μια γκαουσιανή συνάρτηση βάρους με κλίμακες $\sigma_i^2 = s\sigma_l^2$ και $\tau_i^2 = s\tau_l^2$, όπου σ_l και τ_l οι τοπικές κλίμακες

$$M = g(\cdot; \sigma_i^2, \tau_i^2) * \begin{pmatrix} L_x^2 & L_xL_y & L_xL_t \\ L_xL_y & L_y^2 & L_yL_t \\ L_xL_t & L_yL_t & L_t^2 \end{pmatrix} \tag{2.2.7}$$

Οι γκαουσιανές παράγωγοι πρώτης τάξης ορίζονται ως

$$L_\xi\left(\cdot;\sigma_l^2,\tau_l^2\right)=\partial_\xi\left(g*f\right) \tag{2.2.8}$$

Όπως καί στη δισδιάστατη περίπτωση, αναζητούμε τα σημεία ενδιαφέροντος σε περιοχές όπου οι ιδιοτιμές $\lambda_1,\lambda_2,\lambda_3$ έχουν υψηλές τιμές. Οι Laptev και Lindeberg [9] ακολουθούν τη λογική των Harris και Stephens κάνοντας χρήση της ορίζουσας και του ίχνους της μήτρας M για να θεμελιώσουν τη δική τους εκδοχή της συνάρτησης H με παρόμοιο τρόπο

$$H_{3D}=\det(M)-k\mathrm{trace}^3(M)=\lambda_1\lambda_2\lambda_3-k(\lambda_1+\lambda_2+\lambda_3)^3 \tag{2.2.9}$$

Για να δείξουν ότι μεγάλες τιμές των $\lambda_1,\lambda_2,\lambda_3$ ($\lambda_1\leq\lambda_2\leq\lambda_3$) μπορούν να αναζητηθούν στα θετικά μέγιστα της συνάρτησης H_{3D} αναδιατυπώνουν τη σχέση (2.2.9) ως εξής $H_{3D}=\lambda_1^3(\alpha\beta-k(1+\alpha+\beta)^3)$ όπου $\alpha=\lambda_2/\lambda_1$ και $\beta=\lambda_3/\lambda_1$ και από την απαίτηση μη αρνητικότητας της H_{3D} καταλήγουν στη σχέση $k\leq\alpha\beta/(1+\alpha+\beta)^3$. Εύκολα προκύπτει ότι όσο το k πλησιάζει τη μέγιστη τιμή του $k=1/27$ οι λόγοι α και β τείνουν στη μονάδα. Επομένως, για μεγάλες τιμές του k τα τοπικά μέγιστα της H_{3D} αντιστοιχούν σε περιοχές με μεγάλες μεταβολές καί στα χωρικά καί στο χρονικό πεδίο. Η συνάρτηση H_{3D} αποτελεί, λοιπόν, το μετρικό σημαντικότητας (saliency measure) των Laptev και Lindeberg με την έννοια ότι τα τοπικά χωροχρονικά μέγιστα αυτής αποκαλύπτουν χωροχρονικά σημεία της ακολουθίας εικόνων f με σημαίνον περιεχόμενο. Γωνίες στο χώρο τη στιγμή αλλαγής της κατεύθυνσης κίνησής τους διεγείρουν ανιχνεύσεις από τον *Harris3D*.

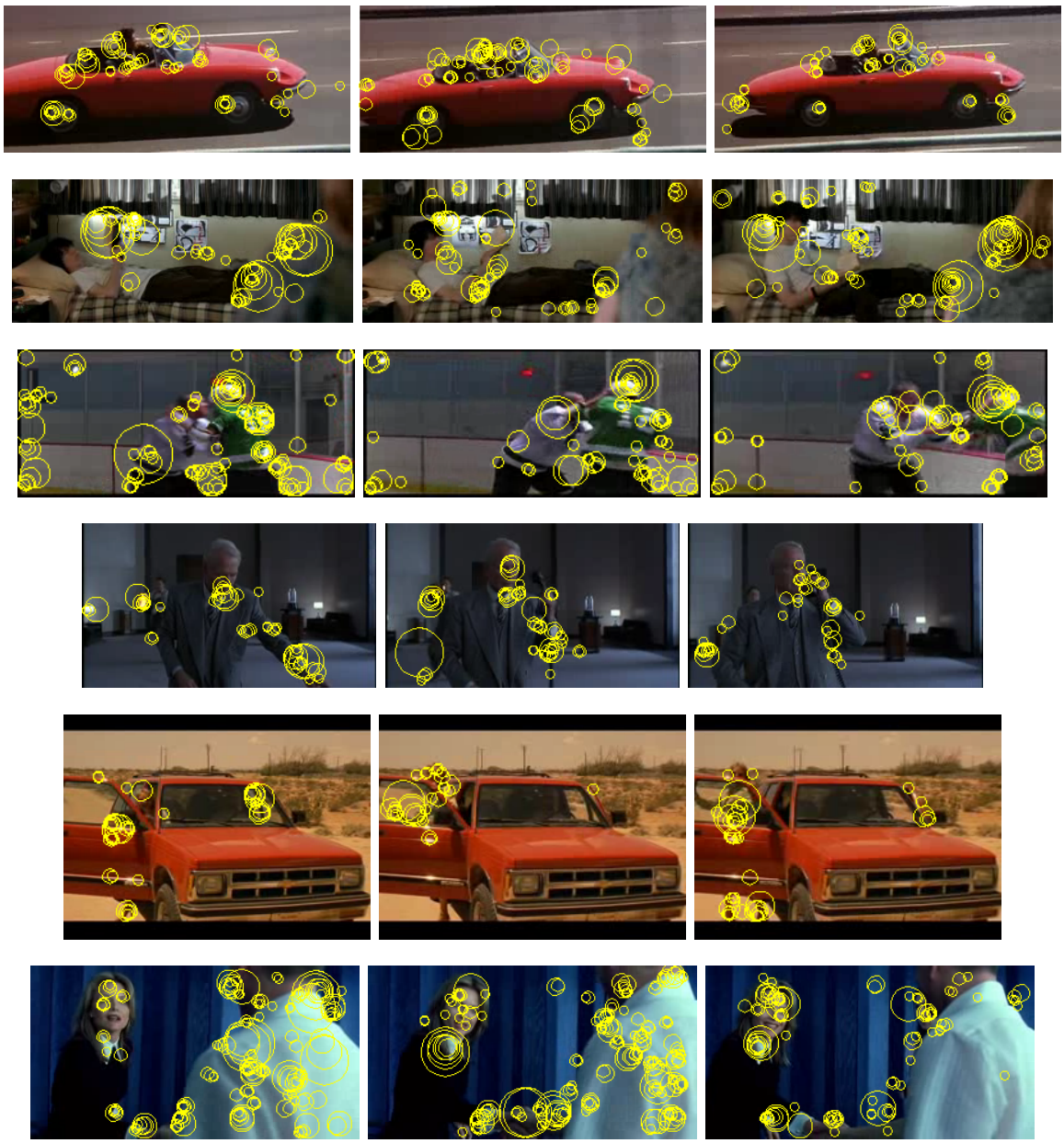
Στην πρωτότυπη εργασία τους οι συγγραφείς εισήγαγαν και μια μέθοδο ανάκτησης του χωροχρονικού περιεχομένου της κάθε ανίχνευσης, δηλαδή της *αυτόματης επιλογής της χωρικής και χρονικής τοπικής κλίμακας* σ_l και τ_l αντίστοιχα. Συγκεκριμένα στο [11] παρουσιάζεται αναλυτικά ο επαναληπτικός αλγόριθμος για την εκτίμηση των κλιμάκων ανίχνευσης που στηρίζεται στην εύρεση των τοπικών μεγίστων του κανονικοποιημένου χωροχρονικού Λαπλασιανού τελεστή και την προσαρμογή των θέσεων των ανιχνεύσεων βάσει των καινούργιων εκτιμημένων τιμών των κλιμάκων. Ωστόσο, αυτό το προτεινόμενο σχήμα επιβαρύνει τη μέθοδο με σημαντικό υπολογιστικό βάρος, συχνά προκαλεί επισφαλείς εκτιμήσεις κλίμακας ενώ σε ορισμένες περιπτώσεις ο επαναληπτικός αλγόριθμος αποκλίνει. Για τους παραπάνω λόγους και ενθαρρυμένοι από την επιτυχημένη αναγνώριση που παρέχουν μέθοδοι πυκνής δειγματοληψίας κλιμάκων, οι Laptev et al. εγκατέλειψαν την ιδέα του επαναληπτικού σχήματος και υιοθετούν πλέον την ανίχνευση σε πολλαπλά επίπεδα προεπιλεγμένων χωρικών και χρονικών κλιμάκων στα πειράματά τους [24].

Ο ανιχνευτής *Harris3D* παραμένει αναλλοίωτος κάτω από τρισδιάστατες περιστροφές της ακολουθίας εικόνων, διαδικασία που στερείται φυσικής σημασίας σε αντίθεση με περιστροφές στις δύο διαστάσεις. Από την άλλη, η χρονική διάσταση είναι ευάλωτη σε *Galilean* μετασχηματισμούς όπως στις περιπτώσεις σχετικής κίνησης μεταξύ της κάμερας και των "σημαντικών"γεγονότων του βίντεο. Για να αντιμετωπίσουν τέτοιου τύπου επιδράσεις από σχετικές κινήσεις, οι Laptev et al. [12] προτείνουν μια διορθωμένη ως προς την κίνηση έκδοση του *Harris3D* ανιχνευτή στην οποία γίνεται χρήση μιας τροποποιημένης εκδοχής της *μήτρας δεύτερων στιγμών* M (2.2.7), προσεγγιστικά αναλλοίωτη σε *Galilean* μετασχηματισμούς. Η προσαρμογή της ταχύτητας εντάσσεται στο επαναληπτικό σχήμα για την εκτίμηση των κλιμάκων και της επαναπροσαρμογής της θέσης των

ανιχνεύσεων. Ωστόσο, αυτή η ιδέα δεν αποτέλεσε αντικείμενο περαιτέρω έρευνας από τους συγγραφείς, δεν χρησιμοποιείται σε πρόσφατα πειράματά τους και δεν παρέχεται ως επιλογή στη διαθέσιμη υλοποίηση του ανιχνευτή και για τούτο την αναφέραμε μόνο επιγραμματικά.

Λεπτομέρειες Υλοποίησης - Πειραματικά Αποτελέσματα. Ο Ανιχνευτής Χωροχρονικών Σημείων Ενδιαφέροντος *Harris3D* χρησιμοποιήθηκε εκτενώς στα πειράματα της παρούσας διπλωματικής. Για την εκπόνηση των πειραμάτων έγινε χρήση της διαθέσιμης online υλοποίησης[3] που παρέχεται από την ιστοσελίδα του I. Laptev. Πρόκειται για εκτελέσιμο κώδικα που υλοποιεί ανίχνευση *Harris3D* σημείων ενδιαφέροντος, υπολογισμό περιγραφέων *HOG/HOF* (βλ. υποενότητα 2.3.2) στα τοπικά τεμάχια των επιλεγμένων σημείων και οπτικοποίηση των ανιχνεύσεων πάνω στις ακολουθίες εικόνων εισόδου. Περαιτέρω λεπτομέρειες σχετικά με την επιλογή των παραμέτρων του ανιχνευτή ή του περιγραφέα που χρησιμοποιήθηκαν θα δωθούν στα επόμενα κεφάλαια για κάθε μεμονωμένο σχετικό πείραμα. Αναφορικά με τις παραμέτρους του *Harris3D* ανιχνευτή, στα μεμονωμένα πειράματα της παρούσας υποενότητας ακολουθήσαμε τις προτεινόμενες από τους συγγραφείς επιλογές τιμών.
Συγκεκριμένα, η χωρική και χρονική κλίμακα ανίχνευσης αντιστοιχούν στις τοπικές κλίμακες σ_l και τ_l με τις οποίες υπολογίζονται οι γκαουσιανές παράγωγοι της μήτρας M (2.2.7) μέσω της σχέσης (2.2.8). Για αυτές επιλέχθηκε η προσέγγιση πολλαπλών κλιμάκων με πανομοιότυπο τρόπο όπως στο [24]. Για τις τιμές των χωρικών και χρονικών κλιμάκων (σ_i^2, τ_j^2) η επιλογή ήταν $\sigma_i = 2^{(1+i)/2}, \quad i = 1, \ldots, 6$ και $\tau_j = 2^{j/2}, \quad j = 1, 2$ καταλήγοντας έτσι σε δεκαέξι συνδυασμούς χωρικών και χρονικών κλιμάκων στις οποίες αναζητήθηκαν οι ανιχνεύσεις. Η παράμετρος k της συνάρτησης H_{3D} (2.2.9) τέθηκε στην τιμή $k = 5 \cdot 10^{-4}$ ενώ το κατώφλι για την απόρριψη "αδύναμων" ανιχνεύσεων στην τιμή 10^{-9}. Τέλος, για την αποφυγή επίπλαστων ανιχνεύσεων στα όρια των πλαισίων (frames) του βίντεο εισόδου, αγνοήθηκαν ανιχνεύσεις που βρίσκονταν σε απόσταση μέχρι και πέντε pixels από τις συντεταγμένες ορίων του κάθε πλαισίου.
Στο Σχήμα 2.2 απεικονίζονται σημεία ενδιαφέροντος από τον ανιχνευτή *Harris3D* σε χαρακτηριστικά frames δειγμάτων της Βάσης Δεδομένων *Hollywood2*. Επιλέξαμε να παρουσιάσουμε ανιχνεύσεις ανά δώδεκα frames (που αντιστοιχούν περίπου σε χρονική διάρκεια του μισού δευτερολέπτου βάσει του frame rate των δειγμάτων βίντεο) ώστε να δώσουμε μια εικόνα της μεταβολής των ανιχνεύσεων για μικρά χρονικά διαστήματα. Οι ακτίνες των κύκλων που περιβάλλουν τα σημεία ενδιαφέροντος είναι ανάλογες των χωρικών κλιμάκων στις οποίες αυτά ανιχνεύθηκαν. Μπορεί εύκολα κανείς να παρατηρήσει την καλή εστίαση των σημείων ενδιαφέροντος σε περιοχές της εικόνας όπου προεξέχοντα σημεία στο χώρο αλλάζουν την κατεύθυνση της κίνησης τους. Τέτοια σημεία είναι τα μέλη του σώματος στη δράση της χειραψίας ή της πάλης ή οι τροχοί και η πόρτα του αυτοκινήτου στις δράσεις της οδήγησης και της εξόδου από το αυτοκίνητο αντίστοιχα. Στα επόμενα κεφάλαια θα καταστεί σαφές με ποιο τρόπο αυτό το πλήθος των τοπικών σημείων ενδιαφέροντος παρέχει την ευρωστία και τη διακριτική ικανότητα για την αναγνώριση ανθρώπινων δράσεων σε βίντεο.

[3] http://www.irisa.fr/vista/Equipe/People/Laptev/download.html



Σχήμα 2.2: Ανιχνεύσεις του *Harris3D* σε έξι δείγματα βίντεο της Βάσης Δεδομένων *Hollywood2 Actions Dataset*. Για κάθε δείγμα οι ανιχνεύσεις εμφανίζονται σε τρείς χρονικές στιγμές που απέχουν δώδεκα frames με τη μικρότερη χρονική στιγμή στην αριστερή στήλη και τη μεγαλύτερη στη δεξία. Απεικονίζονται δείγματα από έξι διαφορετικές ανθρώπινες δράσεις της *Hollywood2*. Από Επάνω προς τα Κάτω: *DriveCar, SitUp, FightPerson, AnswerPhone, GetOutCar, HandShake*

2.2.2 Ο Ανιχνευτής Cuboid

Ο δεύτερος Ανιχνευτής Χωροχρονικών Σημείων Ενδιαφέροντος με τον οποίο ασχοληθήκαμε εκτενέστερα καί σε επίπεδο πειραμάτων στην παρούσα διπλωματική είναι ο Ανιχνευτής *Cuboid* που προτάθηκε από τους Dollár et al. [13]. Συνιστά έναν από τους δημοφιλέστερους ανιχνευτές μαζί με τον Ανιχνευτή *Harris3D* και τον Ανιχνευτή *Hessian*. Έπεται χρονικά του Ανιχνευτή *Harris3D* καθώς πρωτοεμφανίστηκε το 2005, όταν η σχετική έρευνα για τους χωροχρονικούς ανιχνευτές ήταν ακόμα σε πρώιμο στάδιο, πέραν της εργασίας των Laptev και Lindeberg [9]. Εξάγει τα πιο πυκνά τοπικά χαρακτηριστικά σε σύγκριση με όλους τους άλλους ανιχνευτές και πήρε το όνομά του από τους περιγραφείς που υιοθετούν οι συγγραφείς, οι οποίοι υπολογίζονται σε μια "κυβοειδή" τρισδιάστατη περιοχή γύρω από κάθε επιλεγμένο σημείο ενδιαφέροντος (βλ. ενότητα 2.3).

Ένα στοιχείο για τον ανιχνευτή *Cuboid* που παρουσιάζει ιδιαίτερο ενδιαφέρον είναι ότι αποτελεί τον μοναδικό ανιχνευτή χωροχρονικών σημείων ενδιαφέροντος που δεν έχει πανομοιότυπο "ταίρι" στις δύο διαστάσεις. Οι συγγραφείς θεωρούν εσφαλμένη τη λογική απευθείας επέκτασης των ανιχνευτών "σημαντικών" σημείων στο χώρο στις τρείς διαστάσεις. Ισχυρίζονται ότι η χωροχρονική επέκταση του ανιχνευτή *Harris* είναι αποτελεσματική μόνο σε περιπτώσεις ανθρώπινων δράσεων που χαρακτηρίζονται ικανοποιητικά από την αντιστροφή της κατεύθυνσης κίνησης των χεριών και ποδιών, όπως οι κατηγορίες δράσεων της Βάσης Δεδομένων *KTH* στην οποία ο *Harris3D* παρουσιάζει υψηλά ποσοστά αναγνώρισης.

Συνάρτηση Απόκρισης του Ανιχνευτή Cuboid. Οι Dollár et al. [13] κάνουν χρήση χρονικών Gabor φίλτρων και δισδιάστατου γκαουσιανού πυρήνα για να θεμελιώσουν την συνάρτηση απόκρισής τους. Με τη χρήση των Gabor φίλτρων δεν συλλαμβάνονται μόνο τοπικές μεταβολές στο πεδίο του χρόνου αλλά και επαναλαμβανόμενα γεγονότα σταθερής συχνότητας. Για τη συνάρτηση απόκρισης (response function) ισχύει

$$R = (I * g * h_{ev})^2 + (I * g * h_{odd})^2 \tag{2.2.10}$$

όπου $g(x, y; \sigma)$ είναι ο δισδιάστατος γκαουσιανός πυρήνας εξομάλυνσης που εφαρμόζεται μόνο στο χώρο, $I(x, y, t)$ είναι η ακολουθία εικόνων και h_{ev} και h_{odd} είναι το ζευγάρι μονοδιάστατων Gabor φίλτρων τετραγωνισμού (quadrature pair) που εφαρμόζονται στο χρόνο. Τα χρονικά Gabor φίλτρα δίνονται από τις σχέσεις

$$h_{ev}(t; \tau, \omega) = -\cos(2\pi t\omega)\, e^{-t^2/\tau^2} \tag{2.2.11}$$

$$h_{odd}(t; \tau, \omega) = -\sin(2\pi t\omega)\, e^{-t^2/\tau^2} \tag{2.2.12}$$

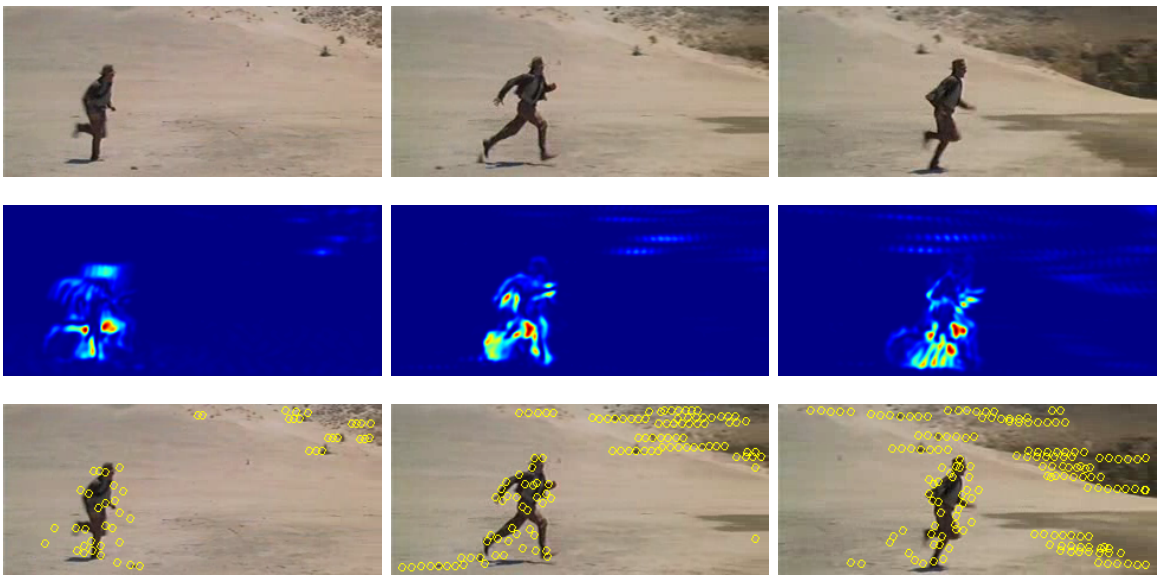
Για την κυκλική συχνότητα ω των φίλτρων Gabor οι συγγραφείς προτείνουν τη σχέση $\omega = 4/\tau$, αφήνοντας για τη συνάρτησή R δύο παραμέτρους που απαιτούν ορισμό, τις σ και τ, που προσεγγιστικά συσχετίζουν με τη χωρική και χρονική κλίμακα ανίχνευσης της μεθόδου τους.

Ο ανιχνευτής *Cuboid* δίνει προτεραιότητα στην ανίχνευση μεγάλων μεταβολών της δομής της εικόνας που χαρακτηρίζονται από περιοδική κίνηση. Η προσέγγιση αυτή αντλεί

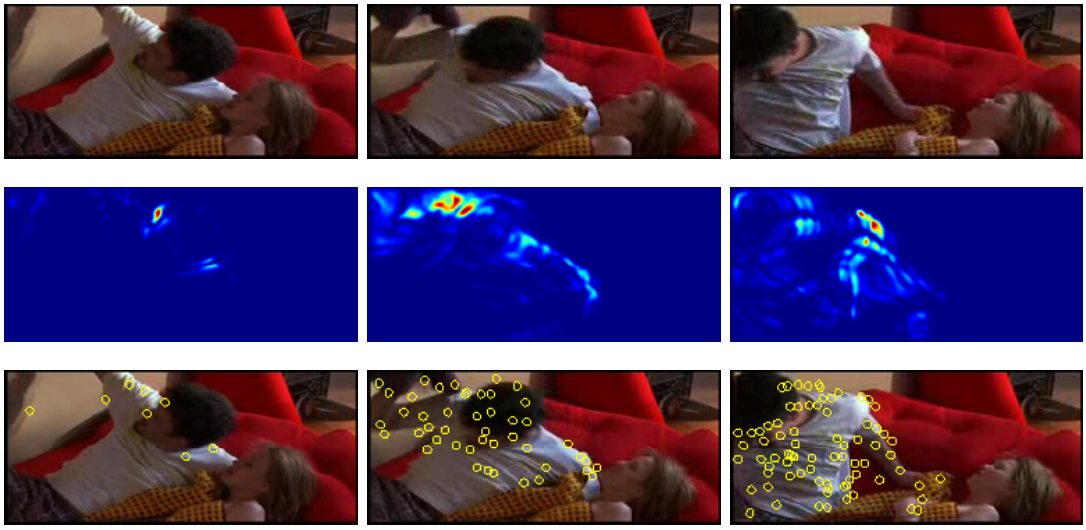
έμπνευση από ανάλυση συμπεριφορών ζώων και ανθρώπων στις οποίες οι περιοδικές κινήσεις έχουν προεξέχοντα ρόλο, όπως στο μάσημα τροφής, το περπάτημα ή το φτερούγισμα. Εντούτοις, ο ανιχνευτής δεν ανταποκρίνεται μόνο σε κινήσεις με περιοδικό περιεχόμενο αλλά και σε άλλες κινήσεις όπως οι χωροχρονικές γωνίες, με χαμηλότερη προτεραιότητα βέβαια. Γενικότερα, οι περιοχές που πυροδοτούν ανιχνεύσεις ικανοποιούν δύο κριτήρια: διακριτικά χαρακτηριστικά στο χώρο και σύνθετη κίνηση στο χρόνο.

Λεπτομέρειες Υλοποίησης - Πειραματικά Αποτελέσματα. Η υλοποίηση του Ανιχνευτή *Cuboid* έγινε στο λογισμικό MATLAB, όπου για την κατασκευή των μονοδιάστατων χρονικών φίλτρων Gabor χρησιμοποιήσαμε τη συνάρτηση `filterGabor1d.m` που περικλείεται στο *toolbox*[4] που διατίθεται online από τους συγγραφείς. Να σημειώσουμε εδώ ότι για την κεντρική κυκλική συχνότητα ω των φίλτρων υιοθετήσαμε την επιλογή $\omega = 1/\tau$ που προτείνεται στην εν λόγω ρουτίνα, όπου τ η σταθερά που καθορίζει το εύρος ζώνης (bandwidth) των φίλτρων και συμπίπτει με τη χρονική κλίμακα ανίχνευσης. Η χωρική κλίμακα ανίχνευσης σ ταυτίζεται με την τυπική απόκλιση του Gaussian πυρήνα εξομάλυνσης. Για τη χωρική και χρονική κλίμακα επιλέξαμε τις τιμές $\sigma = 2$, $\tau = 4$ που χρησιμοποιούνται στα πειράματα της εργασίας των Wang et al. [3], με αποτέλεσμα την τιμή $\omega = 0.25$ για την κεντρική συχνότητα.
Ο υπολογισμός των συνελίξεων στο πεδίο του χρόνου έγινε μέσω της συναρτησης `convn` του MATLAB. Υπολογίζοντας τις τιμές της συνάρτησης απόκρισης (2.2.10) για τα δείγματα βίντεο εισόδου, προχωρήσαμε στην εύρεση των *τοπικών μεγίστων* της στο χώρο και το χρόνο. Από το πλήθος των *τοπικών μεγίστων* αγνοήθηκαν αυτά που αντιστοιχούν σε ανιχνεύσεις στα χωρικά όρια των frames μήκους 5 pixels και στα χρονικά όρια της ακολουθίας εικόνων μήκους 5 frames, για την αποφυγή λήψης επίπλαστων ανιχνεύσεων. Τέλος εφαρμόσαμε την τεχνική *καταστολής των μη μεγίστων (non-maxima suppression)* εκτελώντας κατωφλιοποίηση στο σύνολο των *τοπικών μεγίστων*. Συγκεκριμένα, διατηρήθηκαν οι ανιχνεύσεις με τιμή της συνάρτησης απόκρισης μεγαλύτερη του 1% του ολικού μεγίστου της. Αποφύγαμε την επιβολή μεγαλύτερων τιμών κατωφλίου καθώς αυτό είχε ως αποτέλεσμα την υπερβολικά "αραιή"κατανομή ανιχνεύσεων ανά frame.
Στα Σχήματα 2.3 και 2.4 απεικονίζονται οι τιμές της συνάρτησης απόκρισης και τα ανιχνευθέντα σημεία σε χαρακτηριστικά frames δειγμάτων βίντεο της Βάσης Δεδομένων *Hollywood2 Actions Dataset* που ανήκουν στις δράσεις *Run* και *SitUp* αντίστοιχα. Τα frames έχουν χρονική απόσταση μεταξύ τους αυτή των δώδεκα frames, που μεταφράζεται χρονικά σε διάστημα περίπου μισού δευτερολέπτου. Από την έγχρωμη απεικόνιση των τιμών της συνάρτησης απόκρισης παρατηρούμε τη συγκέντρωση υψηλών τιμών σε περιοχές της εικόνας που αντιστοιχούν στα κινούμενα μέλη των δρώντων κατά την εξέλιξη της δράσης. Στα frames του Σχήματος 2.3 πυροδοτούνται αρκετές ανιχνεύσεις καί στο παρασκήνιο κάτι που ως ένα βαθμό εξηγείται από την ύπαρξη κινούμενης κάμερας στο συγκεκριμένο δείγμα. Γενικότερα, τα πειραματικά αποτελέσματα επιβεβαιώνουν την ανάδειξη σημείων με προεξέχοντα χαρακτηριστικά στο χώρο που υπεισέρχονται σύνθετες κινήσεις από τον Ανιχνευτή *Cuboid*.

[4] http://vision.ucsd.edu/~pdollar/toolbox/doc/



Σχήμα 2.3: Συνάρτηση απόκρισης και ανιχνευθέντα χωροχρονικά σημεία από τον Ανιχνευτή *Cuboid* σε τρία χαρακτηριστικά frames με βήμα δώδεκα από δείγμα βίντεο της *Hollywood2 Actions Dataset* με τη δράση *Run*. Από Επάνω προς τα Κάτω: αρχικά έγχρωμα frames, τιμές της συνάρτησης απόκρισης σε έγχρωμη απεικόνιση, ανιχνεύσεις που αντιστοιχούν στα ισχυρά τοπικά μέγιστα της συνάρτησης.



Σχήμα 2.4: Συνάρτηση απόκρισης και ανιχνευθέντα χωροχρονικά σημεία από τον Ανιχνευτή *Cuboid* σε τρία χαρακτηριστικά frames με βήμα δώδεκα από δείγμα βίντεο της *Hollywood2 Actions Dataset* με τη δράση *SitUp*. Από Επάνω προς τα Κάτω: αρχικά έγχρωμα frames, τιμές της συνάρτησης απόκρισης σε έγχρωμη απεικόνιση, ανιχνεύσεις που αντιστοιχούν στα ισχυρά τοπικά μέγιστα της συνάρτησης.

2.2.3 Πυκνή Δειγματοληψία (Dense Sampling)

Οι Wang et al. [3] το 2009 παρουσίασαν μια εμπεριστατωμένη εργασία αξιολόγησης τοπικών Ανιχνευτών και Περιγραφέων Χωροχρονικών Σημείων Ενδιαφέροντος για το πρόβλημα της Αναγνώρισης Ανθρώπινων Δράσεων σε Βίντεο σε ένα κοινό πειραματικό σκελετό σε τρείς διαφορετικές δημοφιλείς Βάσεις Δεδομένων, την *KTH Actions Dataset, UCF Sport Actions Dataset* και την *Hollywood2 Actions Dataset*. Στόχος τους ήταν η δημιουργία μιας συνεπούς και συγκρίσιμης αξιολόγησης για τέσσερεις τοπικούς ανιχνευτές και έξι τοπικούς περιγραφείς χωροχρονικών σημείων ενδιαφέροντος και να συζητηθούν τα πλεονεκτήματα και οι περιορισμοί της κάθε προσέγγισης. Θα παρουσιάσουμε ενδιαφέροντα στοιχεία από την εργασία τους καί σε επόμενα κεφάλαια της παρούσας διπλωματικής.
Πλάι στους Ανιχνευτές Αραιών Χωροχρονικών Σημείων Ενδιαφέροντος για ακολουθίες εικόνων (*Harris3D, Cuboid, Hessian*), παρουσιάζουν για πρώτη φορά για το πρόβλημα της αναγνώρισης ανθρώπινων δράσεων μια εναλλακτική προσέγγιση που συνίσταται στην *πυκνή δειγματοληψία* σε τακτές θέσεις και κλίμακες στο χώρο και στο χρόνο. Η ιδέα τους αντλεί έμπνευση από την πρόσφατη επιτυχία τεχνικών πυκνής δειγματοληψίας σε ακίνητες εικόνες για την αναγνώριση αντικειμένων.
Σύμφωνα με τα πειράματα των συγγραφέων, αυτή η μέθοδος πυκνής εξαγωγής σημείων παράγει 15-20 φορές περισσότερα τοπικά χαρακτηριστικά ανά frame κατά μέσο όρο σε σχέση με τους γνωστούς ανιχνευτές αραιών χωροχρονικών σημείων. Ανάμεσα στους τελευταίους, ο Ανιχνευτής *Cuboid* παράγει τις πυκνότερες ανιχνεύσεις ανά frame ενώ ο Ανιχνευτής *Hessian* των Willems et al. [21] τα πιο αραία χωροχρονικά τοπικά χαρακτηριστικά. Η μέση τιμή ανιχνεύσεων ανά frame για τους αραιούς ανιχνευτές και την πυκνή δειγματοληψία, σύμφωνα με το [3], παρουσιάζεται συγκεντρωτικά στον Πίνακα 2.1.
Οι Wang et al. προτείνουν *πυκνή δειγματοληψία* σε προκαθορισμένες κανονικές θέσεις και κλίμακες, συνολικά δηλαδή στο χώρο πέντε διαστάσεων (x, y, t, σ, τ) όπου (x, y, t) οι συντεταγμένες θέσης και σ και τ η χωρική και χρονική κλίμακα αντίστοιχα. Πραγματοποιούν πειράματα για διαφορετικό ελάχιστο μέγεθος για το τοπικό τεμάχιο της ανίχνευσης κάθε φορά και παρατηρούν ότι τα ποσοστά αναγνώρισης φθάνουν σε κορεσμό σε δεδομένο μέγεθος του 3D τοπικού τεμαχίου. Για τούτο, στα πειράματα με πυκνή δειγματοληψία που πραγματοποιήσαμε σε μεμονωμένα δείγματα βίντεο επιλέξαμε τις προτεινόμενες διαστάσεις $18 \times 18 \times 10$ για τις τρείς διαστάσεις του ελάχιστου τοπικού μπλόκ δειγματοληψίας.
Η πυκνή δειγματοληψία πραγματοποιείται σε διάφορες χωρικές και χρονικές κλίμακες όπως αυτές διαμορφώνονται με βάση την επιλογή των διαστάσεων του τοπικού τεμαχίου δειγματοληψίας. Η πρόοδος των κλιμάκων στα διάφορα επίπεδα γίνεται πολλαπλασι-

	Harris3D	**Hessian**	**Cuboid**	**Dense Sampling**
Ανιχνεύσεις/Frame	31	19	44	643

Πίνακας 2.1: Μέση τιμή του αριθμού παραγόμενων τοπικών χαρακτηριστικών ανά frame των αραιών ανιχνευτών και της πυκνής δειγματοληψίας σύμφωνα με το [3]

άζοντας τη χωρική και χρονική κλίμακα σ και τ αντίστοιχα με $\sqrt{2}$. Η χωρική κλίμακα αποδείχθηκε μείζονος σημασίας σε σχέση με τη χρονική γι΄ αυτό συνολικά επιλέχθηκαν οκτώ χωρικές και δύο χρονικές κλίμακες, καταλήγοντας σε ένα συνδυασμό δεκαέξι διαφορετικών περιπτώσεων στις οποίες γίνεται δειγματοληψία σε κάθε ακολουθία εικόνων. Είναι σημαντικό να τονίσουμε εδώ ότι η δειγματοληψία πραγματοποιείται με χωρική και χρονική επικάλυψη συνήθως 50% κάτι που δικαιολογεί εν μέρει και το πυκνό πλήθος δειγμάτων ανα frame. Τόσο μεγάλο μέγεθος δεδομένων είναι συχνά δύσκολα διαχειρίσιμο λόγω περιορισμών της μνήμης ταχείας προσπέλασης (RAM) και για αυτό μια πιο βιώσιμη επιλογή είναι η τιμή επικάλυψης 25% στο χώρο καί στο χρόνο στα πειράματα με πυκνή δειγματοληψία.

Για την εκπόνηση των πειραμάτων πυκνής δειγματοληψίας χρησιμοποιήθηκε η διαθέσιμη online υλοποίηση[5] του Alexander Kläser που πραγματοποιεί πυκνή δειγματοληψία και περιγραφή των χαρακτηριστικών με *HOG3D* περιγραφέα (βλ. 2.3). Από τα παραγόμενα αρχεία εξόδου της υλοποίησης διατηρήσαμε φυσικά μόνο τις θέσεις και κλίμακες της πυκνής δειγματοληψίας για την οπτικοποίηση των αποτελεσμάτων πυκνής δειγματοληψίας σε δείγματα βίντεο.

Σύμφωνα με το [3] η μέθοδος της πυκνής δειγματοληψίας σε συνδυασμό με αποτελεσματικούς περιγραφείς τοπικών χαρακτηριστικών (όπως οι *HOG/HOF* και *HOG3D*) ξεπέρασε σε απόδοση τους Ανιχνευτές Αραιών Χωροχρονικών Σημείων Ενδιαφέροντος για το πρόβλημα αναγνώρισης ανθρώπινων δράσεων σε περιπτώσεις ακολουθιών εικόνων με ρεαλιστικά σενάρια και υψηλή μεταβλητότητα όπως αυτές της *Hollywood2 Actions Dataset* και της *UCF Sport Actions Dataset*. Εντούτοις, στο απλούστερο σκηνικό των δειγμάτων βίντεο της *KTH Actions Dataset* παρουσίασε χαμηλότερα ποσοστά από όλους τους "αραιούς" ανιχνευτές.

Τα υψηλά ποσοστά αναγνώρισης της πυκνής δειγματοληψίας στο πρόβλημα αναγνώρισης ανθρώπινων δράσεων έχει αναδείξει τους περιορισμούς των Χωροχρονικών Σημείων Ενδιαφέροντος, ενθαρρύνοντας την πρόοδο της σχετικής έρευνας που αποσκοπεί στην αύξηση της αποτελεσματικότητάς τους. Υπάρχει η αυξανόμενη αντίληψη για την ανάγκη εύρεσης μιας συμβιβαστικής προσέγγισης ανάμεσα στο θόρυβο που εισάγουν ανιχνεύσεις στο παρασκήνιο των εικόνων (background clutter) και στο πληροφοριακό *περιεχόμενο συμφραζομένων* (contextual information) που τέτοιες ανιχνεύσεις παρέχουν σε βίντεο με ρεαλιστικό σκηνικό. Ένα παράδειγμα έρευνας σε αυτή την κατεύθυνση αποτελεί η εργασία των Marszalek et al. στο [25] στην οποία προτείνουν τη συνδυασμένη αναγνώριση ανθρώπινων δράσεων και σκηνών. Πολύ πρόσφατα ερευνήθηκε για πρώτη φορά από τον Tuytelaars [26] μια υβριδική προσέγγιση συνδυασμού των πλεονεκτημάτων των σημείων ενδιαφέροντος και της πυκνής δειγματοληψίας σε ένα επιτυχημένο σχήμα που αποκαλούν *Πυκνά Σημεία Ενδιαφέροντος*.

Στο Σχήμα 2.5 απεικονίζονται frames δειγμάτων της Βάσης Δεδομένων *Hollywood2* μαζί με τα αποτελέσματα της πυκνής δειγματοληψίας για κάθε frame. Είναι εύκολο να παρατηρήσει κανείς την πυκνή κάλυψη των εικόνων μετά τη διαδικασία δειγματοληψίας σε ένα σταθερό πλέγμα από σημεία σε όλες τις περιοχές της εικόνας. Ο μέσος όρος των δειγμάτων για τα frames του Σχήματος 2.5 ανέρχεται στα 937 δείγματα ανά frame.

[5] http://lear.inrialpes.fr/people/klaeser/software_3d_video_descriptor

Σχήμα 2.5: Αποτελέσματα πυκνής δειγματοληψίας με χωρική και χρονική επικάλυψη 25% σε δείγματα βίντεο της Βάσης Δεδομένων *Hollywood2 Actions Dataset*. Στην αριστέρη στήλη απεικονίζονται τα original frames ενώ στη δεξιά τα επιλεγμένα δείγματα για τα αντίστοιχα frames με κύκλους ακτίνας ανάλογης της χωρικής κλίμακας δειγματοληψίας για κάθε σημείο. Τα εικονιζόμενα δείγματα ανήκουν σε πέντε διαφορετικές δράσεις της *Hollywood2*. Από Επάνω προς τα Κάτω: *StandUp, Run, HandShake, Eat, SitDown*

2.3 Τοπικοί Περιγραφείς Χωροχρονικών Σημείων Ενδιαφέροντος

Στην προηγούμενη ενότητα μελετήσαμε Ανιχνευτές Χωροχρονικών Σημείων Ενδιαφέροντος που έχουν σχεδιαστεί και χρησιμοποιηθεί εκτενώς στο πρόβλημα της Αναγνώρισης Ανθρώπινων Δράσεων σε Βίντεο. Για την εξαγωγή τοπικών χαρακτηριστικών η δεύτερη φάση συνίσταται στην περιγραφή της περιοχής των επιλεγμένων σημείων μέσω δισδιάστατων παραθύρων για ακίνητες εικόνες ή τρισδιάστατων όγκων για ακολουθίες εικόνων, με τη χρήση τοπικών μετρήσεων των εικόνων ικανών να αναπαραστήσουν την εμφάνιση ή την κίνηση των περιοχών ενδιαφέροντος. Μια παρουσίαση των πιο σημαντικών περιγραφέων (local descriptors) που έχουν αναπτυχθεί στη βιβλιογραφία και εφαρμοστεί για την αναγνώριση ανθρώπινων δράσεων θα δωθεί σε αυτή την ενότητα. Θα μελετηθούν εκτενέστερα τοπικοί περιγραφείς που είχαν θέση στα πειράματα που εκπονήθηκαν στην παρούσα διπλωματική εργασία.

Ένας ιδανικός τοπικός περιγραφέας παρέχει μια αναπαράσταση ενός τοπικού τεμαχίου εικόνας ή βίντεο που παραμένει αναλλοίωτη στο θόρυβο παρασκηνίου, την εμφάνιση και τα οπτικά εμπόδια, και ενδεχομένως στην περιστροφή ή τις μεταβολές της κλίμακας. Είναι κοινή πρακτική ο καθορισμός της χρονικής και χωρικών διαστάσεων του τρισδιάστατου όγκου υπολογισμού του περιγραφέα από τη χρονική και τις χωρικές κλίμακες αντίστοιχα στις οποίες ανιχνεύθηκε το σημείο ενδιαφέροντος. Με άλλα λόγια, τοπικοί περιγραφείς σημείων που ανιχνεύθηκαν σε "χονδροειδείς", μεγάλες κλίμακες υπολογίζονται σε μεγαλύτερου μεγέθους τοπικά χωροχρονικά τεμάχια από τους περιγραφείς σημείων που πυροδότησαν τοπικούς ανιχνευτές σε πιο λεπτές, μικρές κλίμακες. Η πλειονότητα των τρισδιάστατων τοπικών περιγραφέων αντλεί έμπνευση από επιτυχημένους περιγραφείς εικόνων όπως ο *SIFT* [27].

Θα ξεκινήσουμε την επισκόπηση με περιπτώσεις περιγραφέων που υπολογίζονται σε ολόκληρο το τοπικό τεμάχιο χωρίς να κρατούν καμία πληροφορία για τις τοπικές συντεταγμένες. Οι Laptev και Lindeberg [9] προτείνουν χωροχρονικούς περιγραφείς (*jets*) με τη χρήση γκαουσιανών παραγώγων στο χώρο και το χρόνο μέχρι και τρίτης τάξης της μορφής $L_{x^m y^n t^k} = \sigma^{m+n}\tau^k \left(\partial_{x^m y^n t^k} g\right) * f$ όπου f είναι η ακολουθία εικόνων, $g(x, y, t; \sigma_l^2, \tau_l^2)$ ο χωροχρονικά διαχωρίσιμος πυρήνας της σχέσης (2.2.6) και (σ^2, τ^2) οι κλίμακες ανίχνευσης των σημείων από τον Ανιχνευτή *Harris3D*. Χάρη στην κανονικοποίηση ως προς τις κλίμακες των παραγώγων, ο περιγραφέας προκύπτει αναλλοίωτος σε χωρικές και χρονικές μετάβολες κλίμακας. Οι Dollár et al. [13] πειραματίζονται με διάφορους περιγραφείς, που υπολογίζονται σε "κυβοειδείς"περιοχές γύρω από τα σημεία ενδιαφέροντος, διαστάσεων κατά προσέγγιση έξι φορές την αντίστοιχη κλίμακα ανίχνευσης που χρησιμοποιήθηκε, και στη συνέχεια τοποθετούνται στη σειρά για να σχηματίσουν ένα διάνυσμα περιγραφής. Μελετούν τοπικές μετρήσεις όπως οι κανονικοποιημένες τιμές εικόνας, τα *gradients* της εικόνας ή οι μετρήσεις *οπτικής ροής* (*optical flow*). Στην περίπτωση των *gradients* αυτά υπολογίζονται ως προς καί τις τρείς διαστάσεις της εικόνας ενώ για την οπτική ροή προκύπτουν δύο κανάλια, ένα για κάθε διάσταση στο χώρο, για κάθε ζευγάρι συνεχόμενων frames του "κυβοειδούς". Για μείωση των διαστάσεων του τελικού περιγραφέα εφαρμόζουν τη μέθοδο PCA. Οι συγγραφείς καταλήγουν πειραματικά στο συμπέρασμα ότι η απλή διανυσματική αναπαράσταση του "κυβοειδούς"είναι πιο

αποτελεσματική από τον υπολογισμό ολικών ή τοπικών ιστογραμμάτων των τιμών στο τρισδιάστατο τοπικό τεμάχιο. Παρόμοια λογική ακολουθούν και οι Niebles et al. [28] που υιοθετούν την επιλογή των *gradients* για την περιγραφή των σημείων ενδιαφέροντος που εξάγουν με τον Ανιχνευτή *Cuboid*.

Μια άλλη μορφή περιγραφέων περιλαμβάνει αυτούς που βασίζουν τους υπολογισμούς τους σε μια τρισδιάστατη *δομή πλέγματος* του τοπικού τεμαχίου (*grid-based descriptors*). Ο τελικός περιγραφέας διαμορφώνεται από τη σύνοψη των παρατηρήσεων των επιμέρους "κελιών" από τα οποία απαρτίζεται το τοπικό τεμάχιο. Οι Scovanner et al. [29] προτείνουν μια επέκταση του γνωστού δισδιάστατου περιγραφέα *SIFT* [27] στις τρείς διαστάσεις. Κάθε pixel περιγράφεται από το μέτρο και την κατεύθυνση του δισδιάστατου *gradient* αλλά και από μία επιπρόσθετη γωνία που αντιπροσωπεύει την απόκλιση από τη διεύθυνση του δισδιάστατου *gradient*. Σε κάθε υποπεριοχή της τρισδιάστατης γειτονιάς του σημείου ενδιαφέροντος, υπολογίζεται ένα δισδιάστατο υπο-ιστόγραμμα των κατευθύνσεων των *gradients*. Το τελικό ιστόγραμμα του τοπικού τεμαχίου κατασκευάζεται από τη συσσώρευση των υπο-ιστογραμμάτων σε ένα διάνυσμα. Οι Laptev et al. [24] εισάγουν τους **HOG/HOF** περιγραφείς (*Histograms of Oriented Gradient/Histograms of Optic Flow*). Πρόκειται για τον υπολογισμό ιστογραμμάτων των κατευθύνσεων των *gradients* και της *οπτικής ροής* σε μια *δομή πλέγματος* στην οποία διαμερίζεται το τοπικό τεμάχιο. Με τα εν λόγω ιστογράμματα στοχεύουν στο να συλλάβουν την τοπική εμφάνιση και κίνηση αντίστοιχα, στις γειτονιές των σημείων ενδιαφέροντος. Αυτοί οι περιγραφείς χρησιμοποιήθηκαν εκτενώς σε πειράματα αναγνώρισης ανθρώπινων δράσεων στην παρούσα διπλωματική εργασία και για τούτο κρίνουμε σκόπιμο να τους αναπτύξουμε με λεπτομέρεια στην υποενότητα 2.3.2.

Οι Kläser et al. [30] προτείνουν μια διαφορετική επέκταση του *SIFT* στις τρείς διαστάσεις, τον περιγραφέα **HOG3D**, στηριζόμενοι στον υπολογισμό κατευθύνσεων χωροχρονικών *gradients*. Τα *3D gradients* υπολογίζονται αποτελεσματικά με τη χρήση των *ολοκληρωτικών βίντεο* (*integral videos*) ενώ οι κατευθύνσεις τους κβαντοποιούνται βάσει *κανονικών πολύεδρων* (στα πειράματά τους επιλέγουν το εικοσάεδρο). Υιοθετείται και εδώ η λογική του υπολογισμού ιστογραμμάτων σε "κελιά" στα οποία χωρίζεται το χωροχρονικό τεμάχιο γύρω από το σημείο ενδιαφέροντος και στη συνένωσή τους σε ένα τελικό διάνυσμα χαρακτηριστικών που κανονικοποιείται με την $\mathcal{L}_2$ νόρμα. Για τις διαστάσεις των τοπικών τεμαχίων προτείνονται οι τιμές $\Delta_x = \Delta_y = 8\sigma$ και $\Delta_t = 6\tau$ όπου (σ, τ) οι κλίμακες ανίχνευσης ενώ για τη διαμέριση σε "κελιά" το σχήμα $4 \times 4 \times 3$. Πρόκειται για έναν αποτελεσματικό περιγραφέα για το πρόβλημα της αναγνώρισης ανθρώπινων δράσεων σε βίντεο καθώς συλλαμβάνει ταυτόχρονα τοπική πληροφορία εμφάνισης και κίνησης. Οι Ke et al. [31] προτείνουν τον υπολογισμό των καναλιών οπτικής ροής στα τοπικά τεμάχια με τη χρήση των *ολοκληρωτικών βίντεο*. Οι Willems et al. [21] προτείνουν μια επέκταση του *SURF* περιγραφέα, καθιερώνοντας τον δικό τους **E-SURF** (extended SURF) περιγραφέα. Με παρόμοιο τρόπο όπως μέθοδοι που περιγράψαμε παραπάνω, τα τοπικά τεμάχια διαμερίζονται σε "κελιά". Κάθε "κελί" περιγράφεται από ένα άθροισμα με διαφορετικά βάρη $v = (\sum d_x, \sum d_y, \sum d_t)$ των αποκρίσεων των *Haar* χαρακτηριστικών (d_x, d_y, d_t) (των γνωστών από την εργασία των Viola-Jones στην ανίχνευση προσώπου) στις τρείς διαστάσεις.

Συνοψίζοντας, είδαμε επιγραμματικά μερικούς από τους σημαντικότερους Περιγραφείς

Χωροχρονικών Σημείων Ενδιαφέροντος. Οι περισσότεροι από αυτούς αναπτύχθηκαν με βάση επιτυχημένα "ζευγάρια" τους στις δύο διαστάσεις που έχουν στο παρελθόν αποδειχθεί εύρωστα για το πρόβλημα της αναγνώρισης αντικειμένων ή ατόμων σε ακίνητες εικόνες. Είναι εμφανής η προτίμηση των τοπικών μετρήσεων των *gradients* και της *οπτικής ροής* για την αναπαράσταση των τοπικών τεμαχίων στην πλειονότητα των περιγραφέων. Τα χαρακτηριστικά αυτά στοχεύουν στην αιχμαλώτιση του περιεχομένου εμφάνισης και κίνησης στις ακολουθίες εικόνων. Και τα δύο παρουσιάζουν το πλεονέκτημα ότι δεν απαιτούν την αφαίρεση του background όπως συμβαίνει με τις ολιστικές προσεγγίσεις. Ωστόσο, παρουσιάζουν ευαισθησία σε μεταβολές μεγεθών όπως είναι η υφή ή οι συνθήκες φωτισμού. Οι μετρήσεις *οπτικής ροής* παρέχουν διακριτική ικανότητα μόνο για τα κινούμενα αντικείμενα ενώ τα *gradients* προσκομίζουν πληροφορία καί για στατικά αντικείμενα, κάτι που ανά περιπτώσεις αποδεικνύεται επιβλαβές καθώς προσδίδουν σημαντικότητα σε ανιχνεύσεις του ακίνητου παρασκηνίου. Πρόσφατες τάσεις στη σχετική έρευνα επιχειρούν το συνδυασμό των δύο ειδών χαρακτηριστικών με πολύ ικανοποιητικά αποτελέσματα σε περιπτώσεις ακολουθιών εικόνων με ρεαλιστικά σενάρια (βλ. [3]).

2.3.1 Περιγραφείς Προσανατολισμού Εμφάνισης και Κίνησης

Όπως είδαμε στην εισαγωγή της παρούσας ενότητας, η πλειονότητα των Περιγραφέων Χωροχρονικών Σημείων Ενδιαφέροντος χρησιμοποιούν τοπικές μετρήσεις προσανατολισμού των *gradients* και της *οπτικής ροής* με στόχο την παραγωγή τοπικών χαρακτηριστικών που συλλαμβάνουν την εμφάνιση και την κίνηση στις επιλεγμένες περιοχές ενδιαφέροντος των πλαισίων της ακολουθίας εικόνων εισόδου. Τέτοιες προσεγγίσεις κυριαρχούν στους τοπικούς περιγραφείς για το πρόβλημα της Αναγνώρισης Ανθρώπινων Δράσεων σε Βίντεο λόγω της ευρωστίας και της αποτελεσματικότητας τους για το εν λόγω πρόβλημα.
Θεωρήσαμε σκόπιμο να μελετήσουμε και να παρουσιάσουμε τις πρώτες προσπάθειες χρήσης ιστογραμμάτων προσανατολισμού εμφάνισης και κίνησης που έγιναν από τους Dalal, Triggs και Schmid ([32],[33]) με εφαρμογή στην ανίχνευση ανθρώπων σε στατικές εικόνες ή εικονοσειρές και αποτέλεσαν τον πρόδρομο για τη μετέπειτα έρευνα σε αυτή την κατηγορία περιγραφέων στο πρόβλημα της αναγνώρισης ανθρώπινων δράσεων. Παρόμοιοι περιγραφείς χρησιμοποιήθηκαν εκτενώς σε μαζικά πειράματα της παρούσας διπλωματικής εργασίας που παρουσιάζονται σε επόμενα κεφάλαια.

2.3.1.1 Ιστογράμματα Προσανατολισμού των gradients για την Περιγραφή Εμφάνισης

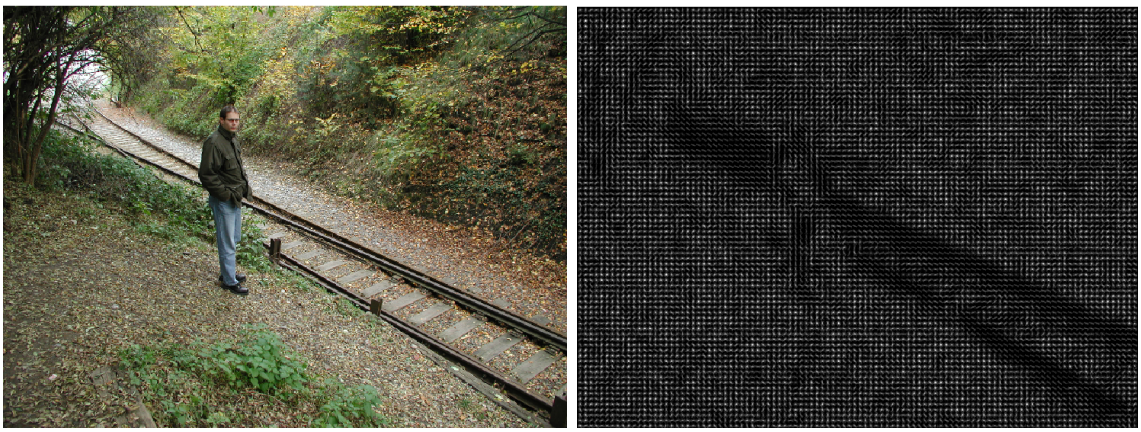
Οι Dalal και Triggs [32] το 2005, αντλώντας έμπνευση από ιστογράμματα προσανατολισμού ακμών, περιγραφείς τύπου *SIFT* [27] και εκφράσεις σχήματος, εισήγαγαν έναν νέο περιγραφέα εμφάνισης, τα *Ιστογράμματα Προσανατολισμού των Gradients* (*Histograms of Oriented Gradients - HOG*) με εφαρμογή στην ανίχνευση ανθρώπων και αντικειμένων. Πρόκειται για μία προσέγγιση καλά κανονικοποιημένων τοπικών ιστογραμμάτων

των κατευθύνσεων των gradients της εικόνας σε ένα δομή κανονικού πυκνού πλέγματος (dense grid). Σε αυτή την εργασία τους μελέτησαν εκτενώς την επιρροή κάθε μεμονωμένης παραμέτρου της διαδικασίας εξαγωγής των χαρακτηριστικών τους καταλήγοντας σε μια υλοποίηση εξαιρετικά αποτελεσματική για την ανίχνευση ανθρώπων. Η βασική ιδέα του νέου περιγραφέα συνίσταται στην παρατήρηση ότι η εμφάνιση και το σχήμα των αντικειμένων μπορεί να χαρακτηριστεί καλά από την κατανομή των τοπικών τιμών των gradients της εικόνας ή των κατευθύνσεων των ακμών, χωρίς να απαιτείται η γνώση της ακριβούς θέσης τους στο χώρο της εικόνας.

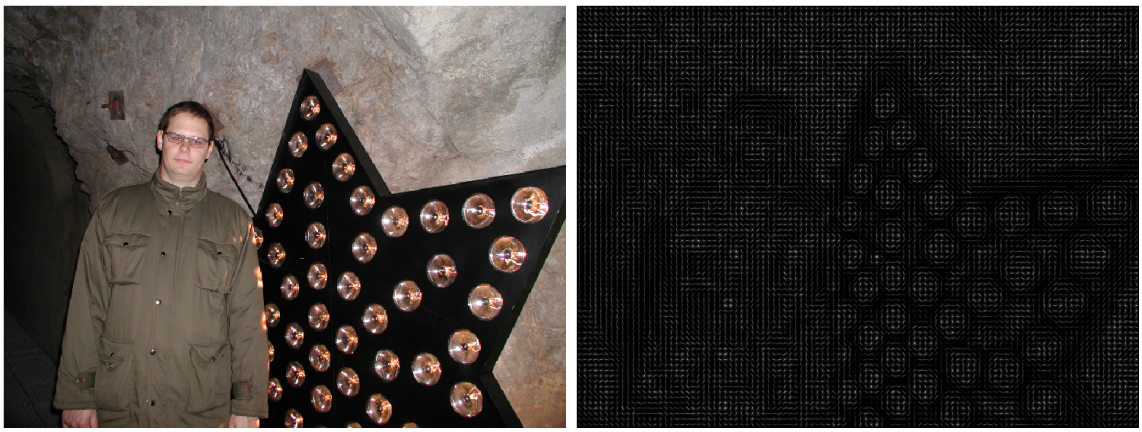
Επισκόπηση της μεθόδου. Η πρακτική υλοποίηση της μεθόδου των Dalal και Triggs [32] στηρίζεται στην διαίρεση της εικόνας σε χωρικές υπο-περιοχές "κελιά"για καθένα από τα οποία υπολογίζεται ένα τοπικό μονοδιάστατο ιστόγραμμα των κατευθύνσεων των gradients για όλα τα pixels του "κελιού". Το διάνυσμα του gradient της εικόνας υπολογίζεται για κάθε pixel και η γωνία του συνεισφέρει στην αντίστοιχη ράβδο (bin) του ιστογράμματος με ένα βάρος ανάλογο του μέτρου του. Χρησιμοποιείται *διγραμμική παρεμβολή* με τα κέντρα των γειτονικών ράβδων για να τοποθετηθεί το κάθε gradient στην κατάλληλη ράβδο κατεύθυνσης και χώρου.
Ιδιαίτερο βάρος δίνουν οι συγγραφείς στην τοπική κανονικοποίηση των αποκρίσεων κάθε υπο-περιοχής για την αντιμετώπιση ανεπιθύμητων επιδράσεων από το φωτισμό, τη σκίαση και τη διαφορά αντίθεσης μεταξύ προσκηνίου και παρασκηνίου. Αυτό επιτυγχάνεται με την ομαδοποίηση των "κελιών"σε μεγαλύτερες επικαλυπτομένες χωρικές περιοχές "μπλόκ"τα οποία κανονικοποιούνται ξεχωριστά, με τρόπο ώστε η απόκριση κάθε "κελιού"να συμμετέχει με μια σειρά διαφορετικών κανονικοποιήσεων στον τελικό περιγραφέα. Αυτή η φαινομενικά πλεονασματική πληροφορία αποδεικνύουν πειραματικά ότι είναι κρίσιμη για τη βελτίωση της απόδοσης του περιγραφέα.
Οι κανονικοποιημένες αποκρίσεις από όλα τα επικαλυπτόμενα "μπλόκ"στην εσωτερική περιοχή του παραθύρου που υιοθετούν για την ανίχνευση ανθρώπων συσσωρεύονται σε ένα τελικό ιστόγραμμα που συνιστά και τον τελικό περιγραφέα. Να σημειώσουμε ότι ο υπολογισμός των gradients γίνεται με απλές μονοδιάστατες "μάσκες"διακριτών παραγώγων $[-1, 0, 1]$ αφού η χρήση γκαουσιανών παραγώγων αποδεικνύεται ότι υποβαθμίζει τις επιδόσεις.

Λεπτομέρειες Υλοποίησης - Πειραματικά Αποτελέσματα. Όπως και στο [32] εφαρμόζουμε γκαουσιανή εξομάλυνση (Gaussian smoothing) στις εικόνες πριν τις τροφοδοτήσουμε στο παραπάνω σχήμα υπολογισμού των ιστογραμμάτων HOG δοκιμάζοντας διάφορες τιμές μεταβλητότητας σ σε κάθε εικόνα εξέτασης. Αναφορικά με τις υπόλοιπες παραμέτρους, οι κατευθύνσεις των gradients κβαντοποιήθηκαν σε εννέα ράβδους προσανατολισμού στην περιοχή $0° - 180°$ (αγνοείται το πρόσημο των gradients) ενώ πειραματικά αποδείχθηκε καταλληλότερο μέγεθος "κελιού"αυτό των 6×6. Για την τοπική κανονικοποίηση αντίθεσης (contrast-normalization) ορίστηκε τετραπλάσιο μέγεθος "μπλόκ"με αποτέλεσμα κάθε "κελί"να συμμετέχει σε τέσσερεις διαφορετικές κανονικοποίησεις των επικαλυπτόμενων "μπλόκ"στον τελικό περιγραφέα. Οι Dalal και Triggs δοκιμάζουν καί κυκλικού σχήματος "μπλόκ"στην εργασία τους με συγκρίσιμα αποτελέσματα. Στις έγχρωμες RGB εικόνες εφαρμόζουν $\gamma-$κανονικοποίηση χωρίς αξι-

όλογη αύξηση της επίδοσης και για τούτο στα πειράματα μας μετατρέπουμε τις έγχρωμες εικόνες σε γκρίζες (grayscale) για το απλούστερο της υλοποίησης.
Τα πειράματα υπολογισμού των ιστογραμμάτων HOG έγιναν με τη χρήση της αντίστοιχης ρουτίνας για το λογισμικό MATLAB του *Piotr's Image & Video Toolbox*[6]. Για τις εικόνες εξέτασης αντλήσαμε δείγματα από την Βάση Δεδομένων *INRIA Person Dataset*[7]. Χαρακτηριστικά αποτελέσματα εφαρμογής του Περιγραφέα HOG εικονίζονται στο Σχήμα 2.6 και το Σχήμα 2.7.



Σχήμα 2.6: Δείγμα εικόνων της Βάσης Δεδομένων *INRIA Person Dataset* (Αριστερά) και το αντίστοιχο *Ιστόγραμμα Προσανατολισμού των Gradients (HOG)* (Δεξιά)



Σχήμα 2.7: *Ιστόγραμμα Προσανατολισμού των Gradients (HOG)* σε εικόνα με παρουσία ανθρώπου και έντονης υφής παρασκηνίου

[6] http://vision.ucsd.edu/~pdollar/toolbox/doc
[7] http://pascal.inrialpes.fr/data/human

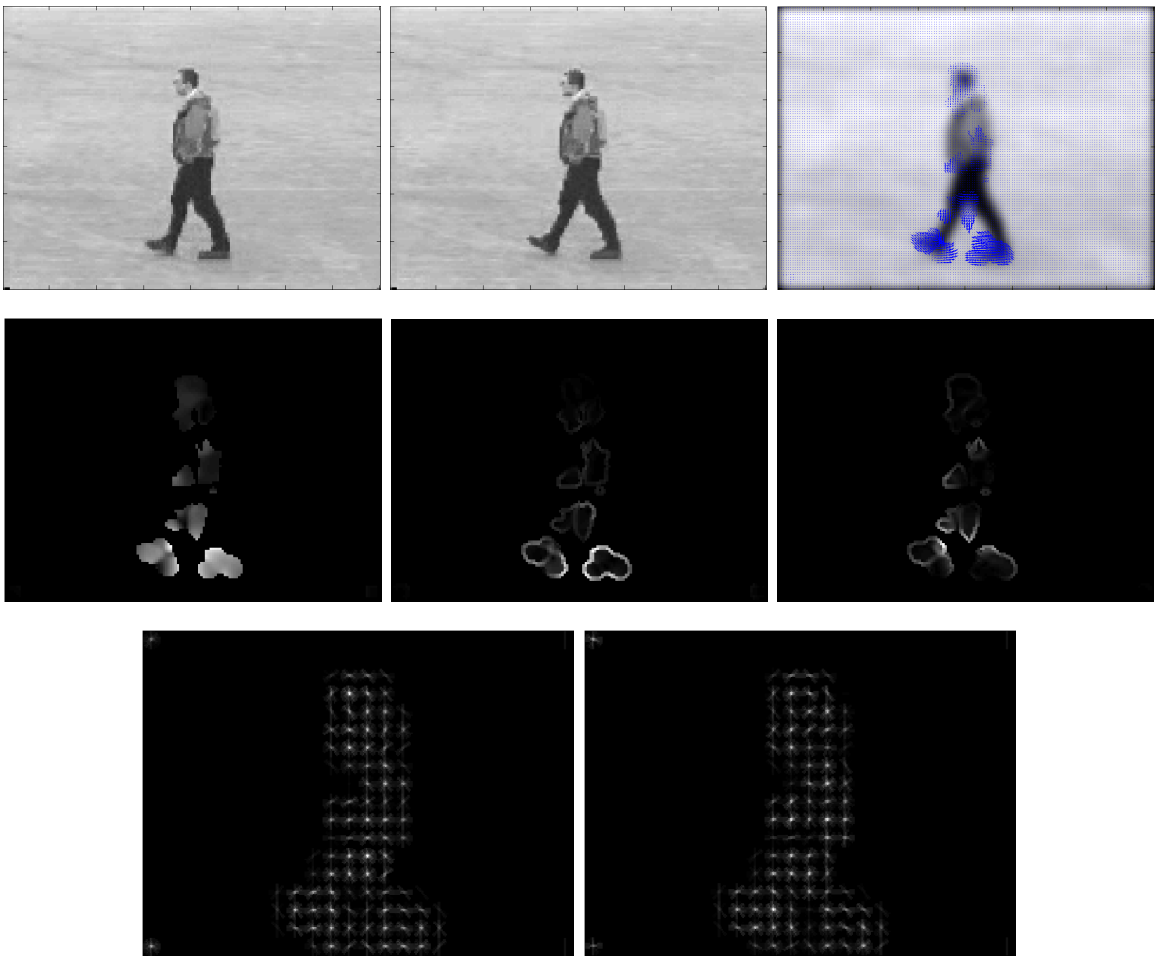
2.3.1.2 Ιστογράμματα Προσανατολισμού της Διαφορικής Οπτικής Ροής για την Περιγραφή Κίνησης

Οι Dalal et al. [33] το 2006 μελετούν εκ νέου το πρόβλημα της αναγνώρισης ανθρώπων, αυτή τη φορά καί στις ακολουθίες εικόνων (βίντεο) εισάγοντας νέα σχήματα περιγραφέων που στηρίζονται σε διαφορετικές κωδικοποιήσεις κίνησης με κοινό γνώμονα τα *Ιστογράμματα Προσανατολισμού της Διαφορικής Οπτικής Ροής* (*Histograms of Oriented Differential Optical Flow*). Στόχος τους είναι η εξαγωγή χαρακτηριστικών κίνησης που επιδεικνύουν ευρωστία σε σχετικές κινήσεις της κάμερας και σε δυναμικές δομές παρασκηνίου (dynamic backgrounds). Σύμφωνα με τους συγγραφείς, οι μετρήσεις της διαφορικής οπτικής ροής εξαλείφουν επιδράσεις περιστροφής της κάμερας και είναι ικανές να περιγράψουν τόσο σχετικές κινήσεις μεταξύ κάμερας, υποκειμένου και παρασκηνίου όσο και ανεξάρτητες κίνησεις που επικεντρώνονται κυρίως στα σύνορα της κίνησης (motion boundaries). Με λίγα λόγια, επιδεικνύουν μεγάλες αποκρίσεις στις κινήσεις του ανθρώπινου σώματος και των άκρων, παρέχοντας μια αποτελεσματική αναπαράσταση για την ανθρώπινη φιγούρα στην ακολουθία εικόνων. Τα καινούργια αυτά σχήματα συνδυάζονται με τα *Ιστογράμματα Προσανατολισμού των Gradients (HOG)* που είδαμε παραπάνω με σκοπό την σύλληψη αμφότερων της εμφάνισης και της κίνησης και την κατασκευή ενός αποτελεσματικού ανιχνευτή.
Δε θα επιχειρήσουμε την πλήρη παρουσίαση όλων των σχημάτων κωδικοποίησης κίνησης που προτείνουν οι Dalal et al., κάτι που ξεφεύγει από τους στόχους της παρούσας διπλωματικής εργασίας. Θα περιγράψουμε μόνο συνοπτικά τη νέα κωδικόποιηση που εισάγεται για τα σύνορα κίνησης και έχει στόχο την αιχμαλώτιση των τοπικών κατευθύνσεων των ακμών ανεξάρτητων κινήσεων σε ακολουθίες εικόνων. Για τις κωδικοποιήσεις εσωτερικών και σχετικών δυναμικών κίνησης παραπέμπουμε τον αναγνώστη στο [33].

Ιστογράμματα Συνόρων Κίνησης - Πειραματικά Αποτελέσματα. Προκειμένου να αναπαρασταθεί η τοπική κατανομή των κατευθύνσεων των ακμών της κίνησης, οι Dalal et al. [32] ακολουθούν τη λογική των *Ιστογράμματων Προσανατολισμού των Gradients*. Για κάθε δύο συνεχόμενα πλαίσια (frames) εξάγονται η οριζόντια και κάθετη συνιστώσα της οπτικής ροής $\mathcal{I}^x$ και $\mathcal{I}^y$ αντίστοιχα και στη συνέχεια αντιμετωπίζονται ως δύο ανεξάρτητες εικόνες. Για καθεμιά τους υπολογίζονται το μέτρο και η κατεύθυνση των gradients τα οποία ακολούθως συνεισφέρουν στις αντίστοιχες ράβδους των μονοδιάστατων ιστογραμμάτων προσανατολισμού με πανομοιότυπο τρόπο όπως στον περιγραφέα HOG για γκρίζες εικόνες που είδαμε πιο πάνω. Οι συγγραφείς προτείνουν τη ξεχωριστή χρήση των καναλιών οπτικής ροής καί όχι το συνδυασμό τους ενώ για τις χωρικές παραγώγους που απαιτούνται για τον υπολογισμό των gradients προτείνουν και πάλι την μονοδιάστατη μάσκα $[-1, 0, 1]$ χωρίς γκαουσιανή εξομάλυνση.
Τα προκύπτοντα ιστογράμματα (δύο για κάθε συνεχόμενα frames, ένα για κάθε συνιστώσα της οπτικής ροής) καλούνται *Ιστογράμματα Συνόρων Κίνησης* (*Motion Boundary Histograms*). Για τα σχετικά πειράματα σε ακολουθίες εικόνων που διενεργήθηκαν στην παρούσα διπλωματική εργασία, διατηρήθηκαν οι ίδιες επιλογές παραμέτρων για τα *Ιστογράμματα Προσανατολισμού των Gradients* ενώ ο υπολογισμός της οπτικής ροής υλοποιήθηκε βάσει του αλγορίθμου των Lucas-Kanade με την αντίστοιχη ρουτίνα του

Piotr's Image & Video Toolbox[8]. Η υλοποίηση βασίστηκε εξ' ολοκλήρου στο λογισμικό MATLAB.

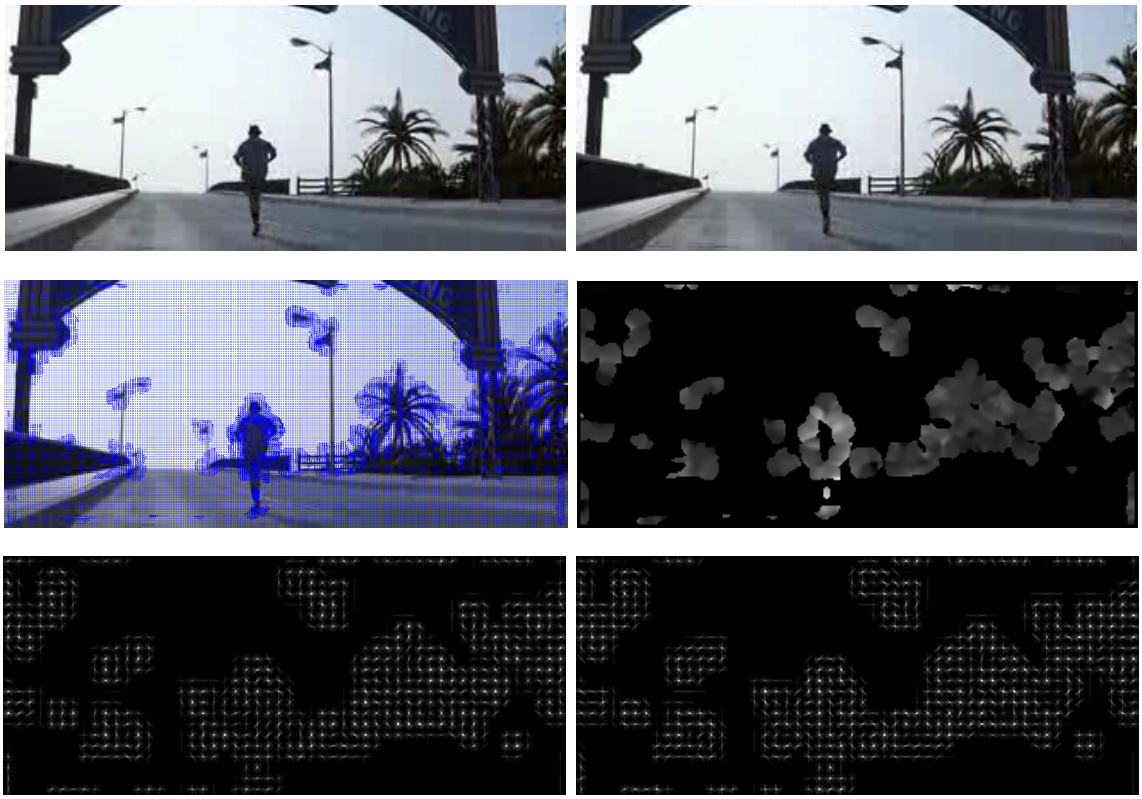
Στο Σχήμα 2.8 απεικονίζονται τα αποτελέσματα διαφόρων μετρήσεων που αφορούν τα *Ιστογράμματα Συνόρων Κίνησης* σε ένα δείγμα βίντεο που ανήκει στην κατηγορία ανθρώπινης δράσης *walking* της Βάσης Δεδομένων KTH Actions Dataset. Το εν λόγω δείγμα περιλαμβάνει τη δράση του περπατήματος σε ένα σκηνικό απλού παρασκηνίου.



Σχήμα 2.8: Ιστογράμματα Συνόρων Κίνησης (Motion Boundary Histograms (MBH)) για δύο συνεχόμενα frames δείγματος βίντεο από τη δράση *walking* της Βάσης Δεδομένων KTH Actions Dataset. Από Επάνω προς τα Κάτω και απο Αριστερά προς τα Δεξιά: Πρωτότυπο frame τη χρονική στιγμή 74, Πρωτότυπο frame τη χρονική στιγμή 75, Διανύσματα οπτικής ροής, Μέτρο της οπτικής ροής, Μέτρο των διανυσμάτων gradients της συνιστώσας οπτικής ροής $\mathcal{I}^x$, Μέτρο των διανυσμάτων gradients της συνιστώσας οπτικής ροής $\mathcal{I}^y$, Ιστόγραμμα Συνόρων Κίνησης για τη συνιστώσα $\mathcal{I}^x$, Ιστόγραμμα Συνόρων Κίνησης για τη συνιστώσα $\mathcal{I}^y$

[8] http://vision.ucsd.edu/~pdollar/toolbox/doc

Παρόμοιες μετρήσεις παρουσιάζονται και στο Σχήμα 2.9. Εδώ εξετάζουμε τα *Ιστογράμματα Συνόρων Κίνησης* σε ένα δείγμα βίντεο της δράσης *Run* της Βάσης Δεδομένων *Hollywood2 Actions Dataset*. Στα εικονιζόμενα frames παρατηρούμε ένα πιο πολύπλοκο σκηνικό παρασκηνίου και για τούτο οι μετρήσεις της διαφορικής οπτικής ροής αναδεικνύουν εκτός από την ανθρώπινη του τρεξίματος πολλαπλές σχετικές κινήσεις κάμερας-παρασκηνίου στη σκηνή όπως η σχετική κίνηση των δέντρων.



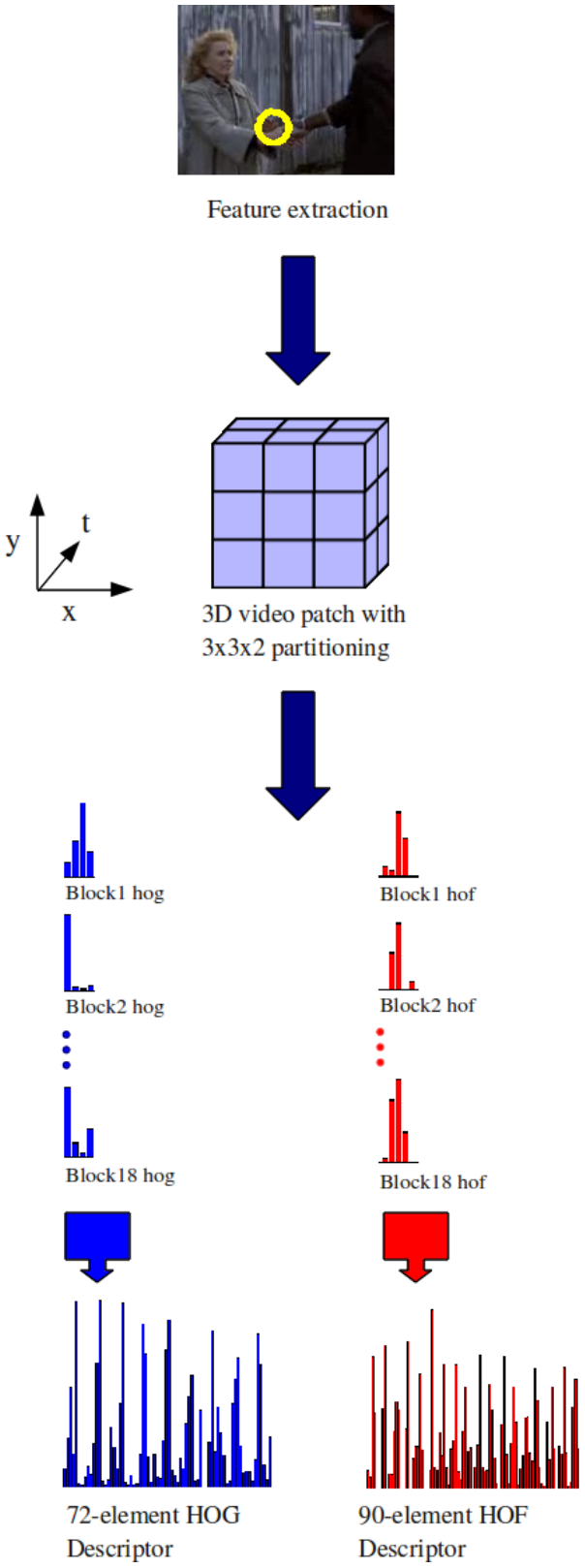
Σχήμα 2.9: Ιστογράμματα Συνόρων Κίνησης (Motion Boundary Histograms (MBH)) για δύο συνεχόμενα frames δείγματος βίντεο από τη δράση *Run* της Βάσης Δεδομένων *Hollywood2 Actions Dataset*. Από Επάνω προς τα Κάτω και απο Αριστερά προς τα Δεξιά: Πρωτότυπο frame τη χρονική στιγμή 91, Πρωτότυπο frame τη χρονική στιγμή 92, Διανύσματα οπτικής ροής, Μέτρο της οπτικής ροής, Ιστόγραμμα Συνόρων Κίνησης για την οριζόντια συνιστώσα οπτικής ροής $\mathcal{I}^x$, Ιστόγραμμα Συνόρων Κίνησης για την κατακόρυφη συνιστώσα οπτικής ροής $\mathcal{I}^y$

2.3.2 Συνδυασμένος Περιγραφέας Εμφάνισης και Κίνησης HOG/HOF

Όπως αναφέραμε παραπάνω, οι Περιγραφείς Χωροχρονικών Σημείων Ενδιαφέροντος που συνδυάζουν τοπικά χαρακτηριστικά μετρήσεων των *gradients* και της *οπτικής ροής* στα τοπικά τεμάχια που περιβάλλουν τα ανιχνευθέντα σημεία ενδιαφέροντος σε ακολουθίες εικόνων αποτελούν πολύ δημοφιλείς και αποτελεσματικές αναπαραστάσεις για το πρόβλημα της Αναγνώρισης Ανθρώπινων Δράσεων. Το 2008 οι Laptev et al. [24] παρουσίασαν ένα νέο περιγραφέα, τον Περιγραφέα *HOG/HOF* (*Histograms of Oriented Gradients and Optic Flow*) με στόχο τον συνδυασμένο χαρακτηρισμό της κίνησης και της εμφάνισης των τοπικών χαρακτηριστικών.
Ο Περιγραφέας *HOG/HOF* συνίσταται στον υπολογισμό ιστογραμμάτων χωρικών gradients και οπτικής ροής στη χωροχρονική γειτονιά των σημείων ενδιαφέροντος που έχουν προηγουμένως εξαχθεί από κάποιο ανιχνευτή στην ακολουθία εικόνων (βίντεο) εισόδου. Θεωρώντας τη χωρική και χρονική κλίμακα σ και τ αντίστοιχα στην οποία ανιχνεύθηκαν τα σημεία, οι διάστασεις του τρισδιάστατου τοπικού τεμαχίου περιγραφής $(\Delta_x, \Delta_y, \Delta_t)$ ορίζονται ως εξής

$$\Delta_x, \Delta_y = 2k_{sp}\sigma \quad \text{και} \quad \Delta_t = 2k_t\tau \tag{2.3.1}$$

όπου στην πρωτότυπη σχετική εργασία τους στο [24] είναι $k_{sp} = k_t$. Κάθε τοπικό τεμάχιο διαμερίζεται σε ένα τρισδιάστατο πλέγμα $n_x \times n_y \times n_t$ "κελιών"για καθένα από τα οποία υπολογίζονται τα ιστογράμματα προσανατολισμού των gradients και της οπτικής ροής, τα *HOG/HOF*. Να σημειώσουμε εδώ ότι σε καμία περίπτωση δεν ακολουθείται η ίδια διαδικασία υπολογισμού των εν λόγω ιστογραμμάτων όπως στους περιγραφείς των Dalal et al. που περιγράψαμε στην παραπάνω υποενότητα (2.3.1). Για την αποφυγή σύγχυσης, οι περιγραφείς των Laptev et al. θα ονομάζονται από δω και στο εξής με την συντομογραφία τους *HOG/HOF*.
Για τα ιστογράμματα *HOG*, οι προσανατολισμοί των χωρικών gradient υπολογίζονται για όλα τα pixel του "κελιού"μέσω γκαουσιανών παραγώγων της μορφής της σχέσης (2.2.2) και διακριτοποιούνται σε ιστογράμματα τεσσάρων ράβδων προσανατολισμού, συμμετέχοντας σε αυτά με βάρος ανάλογο του μέτρου του gradient [34]. Για τα ιστογράμματα οπτικής ροής *HOF*, η κβαντοποίηση γίνεται σε πέντε ράβδους, οι τέσσερεις των οποίων αντιστοιχούν σε τέσσερεις διακριτές κατευθύνσεις κίνησης και η πέμπτη στην απουσία κίνησης. Τα τοπικά ιστογράμματα όλων των "κελιών", αφού πρώτα κανονικοποιηθούν ξεχωριστά, συνενώνονται σε ένα κοινό δίανυσμα χαρακτηριστικών (feature vector) που συνιστά την τελική αναπαράσταση του περιγραφέα.
Για όλα τα σχετικά πειράματα της παρούσας διπλωματικής στην οποία χρησιμοποιήθηκε ο συνδυασμένος Περιγραφέας *HOG/HOF* επιλέχθηκαν οι τιμές παραμέτρων $n_x = 3, n_y = 3, n_t = 2$. Με τη δεδομένη επιλογή, κάθε χωροχρονικό τεμάχιο χωρίζεται σε 18 επιμέρους "κελιά"καταλήγοντας επομένως σε συνολικό μήκος περιγραφέα 72 και 90 στοιχείων για τους Περιγραφείς *HOG* και *HOF* αντίστοιχα. Ο συνδυασμένος Περιγραφέας *HOG/HOF* προκύπτει από την απλή συνένωση των δύο επιμέρους περιγραφέων.
Το σχηματικό διάγραμμα υπολογισμού του Περιγραφέα *HOG/HOF* αποτυπώνεται στο Σχήμα 2.10.



Σχήμα 2.10: Σχηματική απεικόνιση της διαδικασίας υπολογισμού του Περιγραφέα *HOG/HOF* από τη διαμέριση του τοπικού τεμαχίου μέχρι καί το σχηματισμό του τελικού διανύσματος χαρακτηριστικών.

Κεφάλαιο 3

Ενεργειακοί Τελεστές και Χαρακτηριστικά AM-FM Αποδιαμόρφωσης Εικόνων

Στο κεφάλαιο αυτό θα ασχοληθούμε με χαρακτηριστικά εικόνων που προκύπτουν από την εφαρμογή *Τελεστών Ενέργειας* (*Energy Operators*) και μοντέλων *AM-FM Αποδιαμόρφωσης (AM-FM Demodulation)* σε εικόνες. Θα μελετήσουμε τη χρήση των ενεργειακών τελεστών για την εξαγωγή χαρακτηριστικών ενέργειας και για το σκοπό της αποδιαμορφώσης εικόνων στενής ή ευρείας ζώνης συχνοτήτων που έχουν υποστεί διαμόρφωση κατά πλάτος και συχνότητα. Τέλος, θα παρουσιάσουμε τη χρησιμότητα τέτοιων χαρακτηριστικών υφής που προκύπτουν για διαμορφωμένες εικόνες σε διάφορα σχήματα αναπαράστασης που θα αναλύσουμε.

3.1 Ο Τελεστής Ενέργειας Teager-Kaiser

Στην παρούσα ενότητα θα μελετήσουμε τον Ενεργειακό Τελεστή Teager Kaiser στις διάφορες μορφές που εμφανίζεται για μονοδιάστατα, δισδιάστατα και πολυδιάστατα βαθμωτά ή διανυσματικά σήματα, συνεχούς ή διακριτού χώρου και πραγματικών ή μιγαδικών τιμών.

3.1.1 Μονοδιάστατος Ενεργειακός Τελεστής Teager-Kaiser

Ο διαφορικός ενεργειακός τελεστής Teager-Kaiser αναπτύχθηκε αρχικά από τους HM. Teager και SM. Teager [35] για το σκοπό της μη γραμμικής επεξεργασίας σημάτων φωνής. Είναι μη γραμμικός, παραμένει αναλλοίωτος σε απλές μετατοπίσεις στο χρόνο και δίνεται από τη σχέση

$$\Psi[f(t)] \triangleq [f'(t)]^2 - f(t)f''(t) \tag{3.1.1}$$

όπου $f(t)$ το συνεχές μονοδιάστατο πραγματικό σήμα. Το ανάλογο του τελεστή για διακριτά σήματα κατά τους συγγραφείς ορίζεται

$$\Psi_d[f(n)] \triangleq f^2(n) - f(n-1)f(n+1) \tag{3.1.2}$$

Ο "ενεργειακός"τελεστής βρήκε τη συστηματική καθιέρωση της ονομασίας και της χρήσης του από τον Kaiser [36], στην εργασία του οποίου αποτέλεσε εργαλείο για την εξαγωγή της ενέργειας σημάτων απλών αρμονικών ταλαντωτών. Έτσι, αν θεωρήσουμε $f(t) = A\cos(\omega_c t + \theta)$ τη μετατόπιση ενός απλού συστήματος γραμμικού ταλαντωτή μάζας-ελατηρίου με μάζα m και σταθερά ελατηρίου k, η συνολική ενέργεια (άθροισμα κινητικής και δυναμικής) του ταλαντωτή δίνεται από τη σχέση

$$TotalEnergy = \frac{1}{2}(m[f'(t)]^2 + kf^2(t)) = \frac{1}{2}mA^2{\omega_c}^2 \tag{3.1.3}$$

είναι ανάλογη δηλαδή του τετραγώνου του γινομένου πλάτους και συχνότητας. Επομένως, ο ενεργειακός τελεστής Ψ εφαρμοζόμενος στο σήμα $f(t)$ μπορεί να συλλάβει την ενέργεια ("ημι-μοναδιαίας"μάζας) της πηγής που παρήγαγε το σήμα ταλάντωσης αφού

$$\Psi[f(t)] = \Psi[A\cos(\omega_c t + \theta)] = (A\omega_c)^2 \tag{3.1.4}$$

Η ικανότητα εντοπισμού της ενέργειας ταλάντωσης από τον ενεργειακό τελεστή Teager-Kaiser δεν περιορίζεται μόνο στους ταλαντωτές σταθερού πλάτους και συχνότητας αλλά έχει ισχύ και για σήματα AM-FM διαμορφωμένου πλάτους $\alpha(t)$ και διαμορφωμένης συχνότητας $\phi(t)$ της μορφής

$$f(t) = \alpha(t)\cos(\phi(t)) \tag{3.1.5}$$

σύμφωνα με τους Maragos et al. [37] αφού όπως έδειξαν

$$\Psi[\alpha(t)\cos(\phi(t))] = \Psi\left[\alpha(t)\cos\left(\int_0^t \omega_i(\tau)d\tau + \theta\right)\right] \approx [\alpha(t)\omega_i(t)]^2 \tag{3.1.6}$$

υπό την προϋπόθεση ότι τα σήματα διαμόρφωσης $\alpha(t)$ και $\omega_i(t)$ δε μεταβάλλονται τόσο γρήγορα (ως προς το ρυθμό μεταβολής της τιμής τους) ή τόσο πολύ (ως προς το πεδίο τιμών τους) στο χρόνο σε σχέση με τη συχνότητα φέροντος ω_c (carrier frequency). Όπως παρατηρούμε, καί σε αυτή την περίπτωση στην έξοδο του τελεστή παίρνουμε προσεγγιστικά το τετράγωνο του γινομένου του πλάτους $\alpha(t)$ και της στιγμιαίας συχνότητας $\omega_i(t)$.

3.1.2 Πολυδιάστατος Ενεργειακός Τελεστής

Οι Maragos και Bovik [38] εισάγουν την επέκταση του ενεργειακού τελεστή Teager-Kaiser για πραγματικής τιμής πολυδιάστατα σήματα $f(\mathbf{x})$ συνεχούς χρόνου, όπου $\mathbf{x} \in \mathbb{R}^d, \quad d \geq 2$. Η πολυδιάστατη εκδοχή Φ του τελεστή Teager-Kaiser είναι

$$\Phi[f(\mathbf{x})] \triangleq \|\nabla f(\mathbf{x})\|^2 - f(\mathbf{x})\nabla^2 f(\mathbf{x}) \tag{3.1.7}$$

όπου ∇f το gradient του f

$$\nabla f = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_d} \right) \tag{3.1.8}$$

$\|\nabla f\|^2$ το τετράγωνο της ευκλείδιας νόρμας του gradient

$$\|\nabla f\|^2 = \left(\frac{\partial f}{\partial x_1} \right)^2 + \cdots + \left(\frac{\partial f}{\partial x_d} \right)^2 \tag{3.1.9}$$

και $\nabla^2 f$ ο Λαπλασιανός τελεστής για το f

$$\nabla^2 f = \sum_{k=1}^{d} \frac{\partial^2 f}{\partial {x_k}^2} \tag{3.1.10}$$

Από τον ορισμό του, ο πολυδιάστατος τελεστής Φ (3.1.7) μπορεί να εκφραστεί ως

$$\Phi(f) = \sum_{k=1}^{d} \left(\frac{\partial f}{\partial x_k} \right)^2 - f \left(\frac{\partial^2 f}{\partial {x_k}^2} \right) = \sum_{k=1}^{d} \Psi_k(f) \tag{3.1.11}$$

Με λίγα λόγια, η έξοδος του πολυδιάστατου τελεστή προκύπτει ισοδύναμα από το άθροισμα των εξόδων του μονοδιάστατου τελεστή για σήματα συνεχούς χρόνου Ψ όταν ο τελευταίος εφαρμοστεί χωριστά σε κάθε συνιστώσα x_k του αρχικού πολυδιάστατου σήματος $f(\mathbf{x})$.
Ας θεωρήσουμε εδώ ένα $d-$διάστατο συνημίτονο $f(\mathbf{x})$ σταθερού πλάτους A και σταθερού διανύσματος συχνοτήτων $\boldsymbol{\omega}_c = (\omega_{c,1}, \ldots, \omega_{c,d})$ της μορφής

$$f(\mathbf{x}) = A \cos(\boldsymbol{\omega}_c \mathbf{x} + \theta) \tag{3.1.12}$$

όπου θ μια σταθερά διαφοράς φάσης. Η εφαρμογή του πολυδιάστατου τελεστή Φ στο σήμα $f(\mathbf{x})$ δίνει στην έξοδο

$$\Phi[A \cos(\boldsymbol{\omega}_c \mathbf{x} + \theta)] = A^2 \left(\sum_{k=1}^{d} {\omega_{c,k}}^2 \right) = A^2 \|\boldsymbol{\omega}_c\|^2 \tag{3.1.13}$$

δηλαδή το τετράγωνο του γινομένου του πλάτους A και του μέτρου του διανύσματος συχνοτήτων $\boldsymbol{\omega}_c$, που παρόμοια με τη μονοδιάστατη περίπτωση που είδαμε πιο πάνω αντιστοιχεί στη συνολική ενέργεια της ταλάντωσης.
Περνώντας στη γενικότερη περίπτωση των πολυδιάστατων AM-FM σημάτων, θεωρούμε τώρα ένα σήμα $f(\mathbf{x})$ d διαστάσεων με μεταβαλλόμενο στο $d-$διάστατο χώρο πλάτος $\alpha(\mathbf{x})$ και μεταβαλλόμενο διάνυσμα συχνοτήτων $\boldsymbol{\omega}(\mathbf{x}) = \nabla\phi(\mathbf{x}) = [\omega_1(\mathbf{x}), \ldots, \omega_d(\mathbf{x})]$ της μορφής

$$f(\mathbf{x}) = \alpha(\mathbf{x}) \cos(\phi(\mathbf{x})) \tag{3.1.14}$$

όπου $\phi(\mathbf{x})$ η φάση του σήματος και $\omega_k(\mathbf{x})$ η $k-$οστή γωνιακή συνιστώσα του διανύσματος συχνοτήτων $\boldsymbol{\omega}(\mathbf{x})$.
Η τελευταία μπορεί να εκφρασθεί ως εξής

$$\omega_k(\mathbf{x}) = \omega_{c,k} + \omega_{m,k} q_k(\mathbf{x}) \tag{3.1.15}$$

όπου $\omega_{c,k}$ η $k-$οστή συχνότητα φέροντος, $q_k(\mathbf{x})$ το $k-$οστό σήμα διαμόρφωσης συχνότητας κανονικοποιημένο στις τιμές $[-1,1]$ και $\omega_{m,k}$ η μέγιστη απόκλιση της γωνιακής συνιστώσας $\omega_k(\mathbf{x})$ από την αντίστοιχη συχνότητα φέροντος, για την οποία ισχύει $0 \leq \omega_{m,k} \leq |\omega_{c,k}|$ ώστε η $\omega_k(\mathbf{x})$ να έχει σταθερό πρόσημο για όλα τα $\mathbf{x}$.
Θεωρούμε επιπλέον ότι το σήμα πλάτους $\alpha(\mathbf{x})$ έχει συχνοτικό φάσμα φραγμένο από μια σφαίρα συχνοτήτων ακτίνας ω_a και ότι κάθε συνιστώσα ω_k του διανύσματος συχνοτήτων $\boldsymbol{\omega}(\mathbf{x})$ είναι σήμα περιορισμένου εύρους συχνοτήτων, με εύρος ζώνης $\omega_{f,k} \leq |\omega_{c,k}|$. Υπό τις παραπάνω προϋποθέσεις, οι Maragos και Bovik [38] αποδεικνύουν ότι ισχύει

$$\Phi\left[\alpha(\mathbf{x})\cos(\phi(\mathbf{x})\right] \approx \left(\alpha(\mathbf{x})\|\boldsymbol{\omega}(\mathbf{x})\|\right)^2 \tag{3.1.16}$$

με ένα σφάλμα προσέγγισης που εκφυλίζεται σε τιμές πολύ μικρότερες της μονάδας όταν πληρούνται οι δύο παρακάτω συνθήκες

$$\begin{aligned} \omega_a &\ll \min_k |\omega_{c,k}| \\ \sum_{k=1}^{d} \omega_{m,k}\omega_{f,k} &\ll \|\boldsymbol{\omega}_c\|^2 \end{aligned} \tag{3.1.17}$$

συνθήκες που πρακτικά απαιτούν τα σήματα πλάτους και συχνότητας να μη μεταβάλλονται πολύ γρήγορα στον $d-$διάστατο χώρο ή πολύ σε τιμή σε σχέση με τα φέροντα σήματα των αντίστοιχων συνιστωσών.
Βλέπουμε λοιπόν ότι κάτω από ρεαλιστικές υποθέσεις, ο πολυδιάστατος ενεργειακός τελεστής Φ αποφέρει στην έξοδο προσεγγιστικά το ίδιο αποτέλεσμα με την (3.1.13) όταν εφαρμόζεται σε πολυδιάστατα συνημιτονοειδή μεταβαλλόμενου πλάτους και διανύσματος συχνοτήτων.

Οι Maragos και Bovik [38] προτείνουν διάφορες μεθοδολογίες μεταφοράς του πολυδιάστατου ενεργειακού τελεστή στα σήματα διακριτού χρόνου. Ακολουθώντας τη λογική διακριτοποίησης του μονοδιάστατου τελεστή Teager-Kaiser, αυτό θα μπορούσε να επιτευχθεί με την προσέγγιση των παραγώγων του τελεστή Φ (3.1.7) με διαφορές ενός δείγματος (one-sample differences).
Μια άλλη εναλλακτική προσέγγιση προκύπτει από την παρατήρηση ότι ο πολυδιάστατος ενεργειακός τελεστής μπορεί να ιδωθεί ώς άθροισμα των εξόδων των αποκρίσεων του μονοδιάστατου τελεστή Teager-Kaiser στις επιμέρους διαστάσεις του πολυδιάστατου σήματος, όπως είδαμε καί στη σχέση (3.1.11). Επομένως, ο πολυδιάστατος ενεργειακός τελεστής Φ_d για διακριτά σήματα μπορεί να κατασκευαστεί από το άθροισμα των ενεργειακών συνιστωσών των επιμέρους κατευθύνσεων από τις εξόδους του μονοδιάστατου ενεργειακού τελεστή για διακριτά σήματα Ψ_d της σχέσης (3.1.2).
Έτσι, σε ένα βαθμωτό διακριτού χώρου σήμα δύο διαστάσεων $f(m,n)$, όπως είναι μια γκρίζα (greyscale) ακίνητη εικόνα, η μορφή που θα πάρει ο δισδιάστατος διακριτού χώρου ενεργειακός τελεστής Φ_d είναι

$$\begin{aligned} \Phi_d[f(m,n)] &\triangleq \Psi_{d,1}[f(m,n)] + \Psi_{d,2}[f(m,n)] \\ &= 2f^2(m,n) - f(m-1,n)f(m+1,n) - f(m,n-1)f(m,n+1) \end{aligned} \tag{3.1.18}$$

όπου, όπως παρατηρούμε, ο μονοδιάστατος τελεστής $\Psi_{d,1}$ ενεργεί στην κάθετη κατεύθυνση (στις στήλες) της εικόνας ενώ ο $\Psi_{d,2}$ στην οριζόντια κατεύθυνση (στις γραμμές).

3.1.3 Ενεργειακός Τελεστής για Διανυσματικά Σήματα

Οι Maragos και Bovik [38] εισήγαγαν επιπλέον μια εκδοχή του πολυδιάστατου ενεργειακού τελεστή για διανυσματικά σήματα. Θεωρούμε ένα μονοδιάστατο διανυσματικό σήμα $f : \mathbb{R} \to \mathbb{R}^n$

$$\mathbf{f}(t) = [f_1(t), \ldots, f_n(t)] \tag{3.1.19}$$

με την k τάξης παράγωγό του να δίνεται από τη σχέση

$$\mathbf{f}^{(k)}(t) = [f_1^{(k)}, \ldots, f_n^{(k)}], \quad k = 1, 2, \ldots \tag{3.1.20}$$

Ο ενεργειακός τελεστής διανυσματικών σημάτων ορίζεται τότε ως εξής

$$\Theta(\mathbf{f})(t) \triangleq \|\mathbf{f}'(t)\|^2 - \mathbf{f}(t)\mathbf{f}''(t) \tag{3.1.21}$$

όπου $\mathbf{f}'$ και $\mathbf{f}''$ δίνονται από την (3.1.20) για $k = 1$ και $k = 2$ αντίστοιχα. Ο ενεργειακός τελεστής Θ μπορεί να εκφραστεί καί με τον παρακάτω τρόπο

$$\Theta(\mathbf{f}) = \sum_{l=1}^{n} \Psi(f_l) \tag{3.1.22}$$

όπου Ψ ο μονοδιάστατος ενεργειακός τελεστής Teager-Kaiser (3.1.1). Με άλλα λόγια, η έξοδος του ενεργειακού τελεστή διανυσματικών σημάτων μπορεί να ληφθεί από το άθροισμα των εξόδων του μονοδιάστατου τελεστή Teager-Kaiser όταν ο τελευταίος εφαρμοστεί χωριστά σε καθένα από τα βαθμωτά σήματα από τα οποία απαρτίζεται το $\mathbf{f}$. Παρακάτω θα δούμε μία ειδική εφαρμογή των ενεργειακών τελεστών διανυσματικών σημάτων, συγκεκριμένα στα μιγαδικά σήματα:

Μιγαδικά Σήματα Για τα μιγαδικά σήματα $\mathbf{f}(t)$ είναι δυνατό να οριστεί, σύμφωνα με τους Maragos και Bovik [38], ένας ενεργειακός τελεστής ως εξής:

$$C(\mathbf{f})(t) \triangleq \|\mathbf{f}'(t)\|^2 - \mathrm{Re}[\mathbf{f}^*(t)\mathbf{f}''(t)] \tag{3.1.23}$$

Μπορούμε όμως να εκφράσουμε την παραπάνω σχέση χρησιμοποιώντας μόνο τους προαναφερθέντες ενεργειακούς τελεστές Θ και Ψ, αν δούμε το μιγαδικό σήμα ώς ένα διανυσματικό σήμα με $n = 2$ βαθμωτές συνιστώσες $f_1 = \mathrm{Re}(\mathbf{f})$ και $f_2 = \mathrm{Im}(\mathbf{f})$ οπότε η (3.1.23) γράφεται

$$C(\mathbf{f})(t) = \Theta\{[\mathrm{Re}(\mathbf{f}), \mathrm{Im}(\mathbf{f})]\} = \Psi[\mathrm{Re}(\mathbf{f})] + \Psi[\mathrm{Im}(\mathbf{f})] \tag{3.1.24}$$

Η παραπάνω σχέση συνιστά ένα πολύ χρήσιμο αποτέλεσμα, καθώς φανερώνει ότι η εξαγωγή χαρακτηριστικών από ενεργειακούς τελεστές Teager-Kaiser σε μιγαδικά σήματα μπορεί να απλοποιηθεί στην ανάλυση ξεχωριστά του πραγματικού και του φανταστικού μέρους. Η χρησιμότητα αυτή θα γίνει εμφανής παρακάτω στην παρούσα διπλωματική εργασία όπου θα ασχοληθούμε με ενεργειακά χαρακτηριστικά μιγαδικών εικόνων που προκύπτουν από ζωνοπερατό φιλτράρισμα με Gabor φίλτρα.

3.2 Εφαρμογή του Τελεστή Ενέργειας Teager-Kaiser για την αποδιαμόρφωση AM-FM σημάτων

Οι Maragos et al. [37] έδειξαν ότι μη γραμμικοί συνδυασμοί των εξόδων AM-FM σημάτων και των παραγώγων τους από τον τελεστή ενέργειας Teager-Kaiser μπορούν αποτελεσματικά να χρησιμοποιηθούν για την εκτίμηση των συνιστωσών διαμόρφωσης πλάτους και συχνότητας. Όπως είδαμε πιο πάνω, η εφαρμογή του τελεστή ενέργειας Teager-Kaiser σε σήματα αυτής της κατηγορίας παράγει στην έξοδο, κάτω από ρεαλιστικές συνθήκες, την ενέργεια της πηγής που προκάλεσε την ταλάντωση του συνημιτοειδούς σήματος, στη μορφή του τετραγώνου του γινομένου του πλάτους και του μέτρου του διανύσματος συχνοτήτων. Για τούτο οι προτεινόμενοι αλγόριθμοι ονομάστηκαν *Αλγόριθμοι Διαχωρισμού της Ενέργειας (Energy Separation Algorithms)* ακριβώς για την ικανότητά τους να διαχωρίζουν την ενέργεια στην έξοδο του τελεστή στα επιμέρους σήματα πλάτους και συχνότητας διαμόρφωσης.
Αν και η αρχική σχετική εργασία των Maragos et al. [37] παρουσιάστηκε για μονοδιάστατα σήματα με εφαρμογή στην ανάλυση σημάτων φωνής, στην παρούσα ενότητα θα περιγράψουμε συνοπτικά τους *Αλγορίθμους Διαχωρισμού της Ενέργειας* που ανέπτυξαν μεταγενέστερα για πολυδιάστατα AM-FM σήματα στο συνεχή χώρο. Για σήματα διακριτού χώρου θα μελετήσουμε μόνο τον αντίστοιχο αλγόριθμο στην περίπτωση δισδιάστατων σημάτων, όπως είναι οι εικόνες.

3.2.1 Συνεχής Αλγόριθμος Διαχωρισμού της Ενέργειας

Αρχικά θα θεωρήσουμε την περίπτωση $d-$διάστατου συνημιτονοειδούς σήματος f σταθερού πλάτους A και σταθερού διανύσματος συχνοτήτων $\boldsymbol{\omega}_c = (\omega_{c,1}, \ldots, \omega_{c,d})$ της μορφής (3.1.12) για το οποίο η εφαρμογή του τελεστή Φ δίνει, όπως είδαμε, στην έξοδο

$$\Phi[A\cos(\boldsymbol{\omega}_c\mathbf{x} + \theta)] = A^2\|\boldsymbol{\omega}_c\|^2 \tag{3.2.1}$$

Εφαρμόζουμε τον τελεστή στις d μερικές παραγώγους του σήματος παίρνοντας τις d ακόλουθες εξισώσεις

$$\Phi\left(\frac{\partial f}{\partial x_k}\right)(\mathbf{x}) = (A\omega_{c,k})^2\|\boldsymbol{\omega}_c\|^2 \quad , \quad k = 1, \ldots, d \tag{3.2.2}$$

Με το συνδυασμό των εξισώσεων (3.2.1) και (3.2.2) μπορούμε πλέον να εκτιμήσουμε τις $d+1$ παραμέτρους $|A|, |\omega_{c,1}|, \ldots, |\omega_{c,d}|$. Η λύση του συστήματος δίνει τις ακριβείς εκτιμήσεις του μέτρου του πλάτους και των συνιστωσών του διανύσματος συχνότητας

$$|A| = \frac{\Phi(f)}{\sqrt{\sum_{k=1}^{d}\Phi\left(\frac{\partial f}{\partial x_k}\right)}} \tag{3.2.3}$$

$$|\omega_{c,k}| = \sqrt{\frac{\Phi\left(\frac{\partial f}{\partial x_k}\right)}{\Phi(f)}} \quad , \quad k = 1, \ldots, d \tag{3.2.4}$$

Η γνώση του προσήμου των $\omega_{c,k}$ είναι ουσιώδης καθώς καθορίζει τον προσανατολισμό του διανύσματος συχνοτήτων. Εντούτοις, με το παραπάνω σχήμα δεν είναι εφικτή η εκτίμηση του προσήμου των συνιστωσών του διανύσματος συχνοτήτων αλλά μόνο του μέτρου τους.

Περνάμε τώρα στην περίπτωση των πολυδιάστατων AM-FM σημάτων της μορφής (3.1.14) με μεταβαλλόμενο πλάτος $\alpha(\mathbf{x})$ και μεταβαλλόμενο διάνυσμα συχνοτήτων $\boldsymbol{\omega}(\mathbf{x}) = \nabla\phi(\mathbf{x}) = [\omega_1(\mathbf{x}), \ldots, \omega_d(\mathbf{x})]$. Όπως είδαμε αναλυτικά στην υποενότητα για τον Πολυδιάστατο Ενεργειακό Τελεστή, οι Maragos και Bovik [38] έδειξαν ότι με την εκπλήρωση των συνθηκών (3.1.17) η απόκριση του σήματος f στον ενεργειακό τελεστή δίνεται με σφάλμα προσέγγισης πολύ μικρότερο της μονάδας από τη σχέση

$$\Phi(f) = \Phi\left[\alpha(\mathbf{x})\cos(\phi(\mathbf{x})\right] \approx \left(\alpha(\mathbf{x})\|\boldsymbol{\omega}(\mathbf{x})\|\right)^2 \tag{3.2.5}$$

Οι μερικές παράγωγοι του σήματος f εδώ δίνονται από τη σχέση

$$\frac{\partial f}{\partial x_k} = \frac{\partial \alpha}{\partial x_k}\cos(\phi) - \alpha\omega_k \sin(\phi) \tag{3.2.6}$$

Εξαιτίας των συνθηκών (3.1.17), στο παραπάνω άθροισμα ο δεύτερος όρος έχει μέγιστη απόλυτη τιμή πολύ μεγαλύτερης τάξης μεγέθους σε σχέση με τον πρώτο και για τούτο η έκφραση για τις μερικές παραγώγους μπορεί να προσεγγιστεί μόνο από αυτόν. Έχοντας λοιπόν $\partial f/\partial x_k = -\alpha\omega_k \sin(\phi)$ και χρησιμοποιώντας τον προσεγγιστικό τύπο (3.2.5) παίρνουμε

$$\Phi\left(\frac{\partial f}{\partial x_k}\right) \approx \alpha^2 {\omega_k}^2 \|\boldsymbol{\omega}\|^2 \quad , \quad k = 1, \ldots, d \tag{3.2.7}$$

Από το συνδυασμό των $d+1$ σχέσεων (3.2.5) και (3.2.7) παίρνουμε τις παρακάτω εκτιμήσεις για την απόλυτη τιμή του πλάτους α και των γωνιακών συνιστωσών συχνότητας ω_k

$$|\alpha(\mathbf{x})| \approx \frac{\Phi(f)}{\sqrt{\sum_{k=1}^{d} \Phi\left(\frac{\partial f}{\partial x_k}\right)}} \tag{3.2.8}$$

$$|\omega_k(\mathbf{x})| \approx \sqrt{\frac{\Phi\left(\frac{\partial f}{\partial x_k}\right)}{\Phi(f)}} \quad , \quad k = 1, \ldots, d \tag{3.2.9}$$

Οι παραπάνω σχέσεις (3.2.8) και (3.2.9) συνιστούν τον *Συνεχή Αλγόριθμο Διαχωρισμού της Ενέργειας (Continuous Energy Separation Algorithm (CESA))*.

Να θυμίσουμε εδώ ότι για τις γωνιακές συνιστώσες ω_k ισχύει η έκφραση

$$\omega_k(\mathbf{x}) = \omega_{c,k} + \omega_{m,k} q_k(\mathbf{x}) \tag{3.2.10}$$

όπου $-1 \leq q_k \leq 1$ και $0 \leq \omega_{m,k} \leq |\omega_{c,k}|$.
Επομένως, για την περίπτωση των AM-FM σημάτων είναι δυνατός ο προσδιορισμός των

προσήμων των ω_k, τα οποία συμπίπτουν με τα πρόσημα των αντιστοίχων συχνοτήτων φέροντος $\omega_{c,k}$ αφού ο δεύτερος όρος του παραπάνω αθροίσματος δεν μπορεί να επιφέρει αλλαγή προσήμου λόγω της φραγμένης τιμής του. Όπως θα δούμε παρακάτω, τα πρόσημα των συχνοτήτων φέροντος $\omega_{c,k}$ μπορούν να ληφθούν στη δισδιάστατη περίπτωση με την εφαρμογή του αλγορίθμου στις ζωνοπερατά φιλτράρισμένες εξόδους του αρχικού σήματος όπου οι συχνότητες φέροντος προσεγγιστικά έχουν ίδια τιμή με τις κεντρικές συχνότητες των αντίστοιχων φίλτρων.

3.2.2 Διακριτός Αλγόριθμος Διαχωρισμού της Ενέργειας

Εδώ θα μελετήσουμε τη μορφή που λαμβάνει ο *Αλγόριθμος Διαχωρισμού της Ενέργειας* για την περίπτωση των δισδιάστατων σημάτων διακριτού χώρου ή ισοδύναμα για γκρίζες (greyscale) εικόνες, όπως προτάθηκε από τους Maragos και Bovik [38].
Η εφαρμογή του δισδιάστατου διακριτού χώρου ενεργειακού τελεστή Φ_d (3.1.18) σε ένα συνημιτονοειδές διακριτού χώρου σταθερού πλάτους και συχνότητας δίνει στην έξοδο

$$\Phi_d[A\cos(\Omega_1 m + \Omega_2 n + \theta)] = A^2[\sin^2(\Omega_1) + \sin^2(\Omega_2)] \tag{3.2.11}$$

όπου Ω_1 και Ω_2 η κάθετη και οριζόντια "στιγμιαία"συχνότητα αντίστοιχα. Να τονίσουμε εδώ ότι όταν αναφερόμαστε σε πολυδιάστατα σήματα AM-FM η χρήση του όρου "στιγμιαίος"για τα σήματα πλάτους και συχνότητας γίνεται για την αναφορά στις μεταβολές των σημάτων στις χωρικές διαστάσεις.
Ας προχωρήσουμε στη γενικότερη περίπτωση δισδιάστατου διακριτού AM-FM σήματος

$$f(m,n) = \alpha(m,n)\cos[\phi(m,n)] \tag{3.2.12}$$

με τις γωνιακές συχνότητες Ω_1 και Ω_2 να δίνονται από τις σχέσεις

$$\begin{aligned} \Omega_1(m,n) &= \frac{\partial \phi}{\partial m} \\ \Omega_2(m,n) &= \frac{\partial \phi}{\partial n} \end{aligned} \tag{3.2.13}$$

Υπό την πλήρωση ρεαλιστικών συνθηκών, οι Maragos και Bovik [38] απέδειξαν ότι η έξοδος του δισδιάστατου διακριτού χώρου ενεργειακού τελεστή Φ_d από το σήμα f (3.2.12) είναι με καλή προσέγγιση

$$\Phi_d[f(m,n)] \approx \alpha^2(m,n)\left(\sin^2[\Omega_1(m,n)] + \sin^2[\Omega_2(m,n)]\right) \tag{3.2.14}$$

Αντίστοιχα, η εφαρμογή του τελεστή Φ_d στις μερικές παραγώγους g_1 και g_2 του σήματος f, οι οποίες προσεγγίζονται από συμμετρικές διαφορές τριών δειγμάτων, δίνει στην έξοδο

$$\Phi_d[g_1] \approx \alpha^2 \sin^2[\Omega_1]\left(\sin^2[\Omega_1] + \sin^2[\Omega_2]\right) \tag{3.2.15}$$

$$\Phi_d[g_2] \approx \alpha^2 \sin^2[\Omega_2]\left(\sin^2[\Omega_1] + \sin^2[\Omega_2]\right) \tag{3.2.16}$$

Λύνοντας το σύστημα που διαμορφώνεται από τις σχέσεις (3.2.14), (3.2.15) και (3.2.16) ως προς την απόλυτη τιμή του σήματος διαμόρφωσης πλάτους α και των γωνιακών συχνοτήτων Ω_1 και Ω_2 παίρνουμε τις ακόλουθες σχέσεις

$$|\alpha(m,n)| \approx \frac{2\Phi_d[f(m,n]}{\sqrt{\Phi_d[f(m+1,n)-f(m-1,n)]+\Phi_d[f(m,n+1)-f(m,n-1)]}} \tag{3.2.17}$$

$$|\Omega_1(m,n)| \approx \arcsin\left(\sqrt{\frac{\Phi_d[f(m+1,n)-f(m-1,n)]}{4\Phi_d[f(m,n]}}\right) \tag{3.2.18}$$

$$|\Omega_2(m,n)| \approx \arcsin\left(\sqrt{\frac{\Phi_d[f(m,n+1)-f(m,n-1)]}{4\Phi_d[f(m,n]}}\right) \tag{3.2.19}$$

Οι παραπάνω σχέσεις συνιστούν τον *Διακριτό Αλγόριθμο Διαχωρισμού της Ενέργειας (Discrete Energy Separation Algorithm (DESA))* των Maragos και Bovik [38] και παρέχουν εκτιμήσεις για το σήμα πλάτους και τα σήματα συχνότητας ενός δισδιάστατου διακριτού σήματος AM-FM σε κάθε σημείο του χώρου, όταν πληρούνται ρεαλιστικές υποθέσεις. Για τις εκτιμήσεις των συχνοτήτων ισχύει ότι η απόλυτη τιμή τους βρίσκεται στο πρώτο τεταρτημόριο. Παρόμοια με την περίπτωση του *Συνεχή Αλγορίθμου Διαχωρισμού της Ενέργειας (CESA)* τα πρόσημα των σημάτων συχνότητας Ω_i μπορούν να βρεθούν από τα πρόσημα των αντίστοιχων συχνοτήτων φέροντος.
Ο *Διακριτός Αλγόριθμος Διαχωρισμού της Ενέργειας* βρίσκει εφαρμογή στην αποδιαμόρφωση AM-FM διαμορφωμένων εικόνων, συνθετικών ή φυσικών. Βασική προϋπόθεση για την απευθείας χρήση του είναι το δισδιάστατο σήμα να είναι ολικά ή τμηματικά *στενής ζώνης (narrowband)*. Εφαρμογές του ενεργειακού τελεστή Teager-Kaiser σε σήματα *ευρείας ζώνης (wideband)* αποφέρουν ασταθείς, θορυβώδεις εκτιμήσεις με συχνές εμφανίσεις αρνητικών τιμών. Η εφαρμογή του *DESA* σε *ευρυζωνικές* εικόνες καθίσταται δυνατή με το ζωνοπερατό φιλτράρισμα της εικόνας εισόδου και τη χρήση του αλγορίθμου ξεχωριστά στις προκύπτουσες συνιστώσες *στενής ζώνης* όταν αυτές μοντελοποιούνται επίσης από AM-FM σήματα. Η παραπάνω στρατηγική προμηθεύει επίσης το συνολικό σχήμα με την ικανότητα περιορισμού του θορυβώδους περιεχομένου και εκτίμησης των προσήμων των σημάτων συχνότητας από τις αντίστοιχες κεντρικές συχνότητες των φίλτρων.
Στην επόμενη ενότητα θα περιγράψουμε αναλυτικά την αποδιαμόρφωση AM-FM εικόνων *ευρείας ζώνης* με το ζωνοπερατό φιλτράρισμα με συστοιχίες δισδιάστατων Gabor φίλτρων και τη χρήση του αλγορίθμου διαχωρισμού της ενέργειας με στόχο την ανάκτηση των σημάτων διαμόρφωσης πλάτους και συχνοτήτων. Χαρακτηριστικά τέτοιων AM-FM μοντέλων αποκαλύπτουν τις σημαντικότερες πλευρές της υφής σε φυσικές εικόνες.

3.3 Μοντέλο Πολλαπλών AM-FM Συνιστωσών και Χαρακτηριστικά Υφής για Εικόνες

Όπως είναι γνωστό από την κλασική θεωρία Fourier, είναι εφικτή η αναπαράσταση πολυδιάστατων σημάτων μέσω μιας σύνθεσης από ημιτονοειδείς συνιστώσες σταθερού πλάτους και συχνότητας. Ωστόσο σε περιπτώσεις σημάτων που επιδεικνύουν σημαντικές τοπικές μεταβολές πλάτους και συχνότητας στο χώρο, η παραπάνω αναπαράσταση πάσχει καθώς επιφέρει την απώλεια πλούσιας πληροφορίας της υφής όταν χρησιμοποιείται για τη μοντελοποίηση εικόνων. Από την άλλη, μοντέλα AM-FM για εικόνες είναι ικανά να συλλάβουν ουσιώδη τοπική πληροφορία υφής σχετικά με τις μεταβολές της αντίθεσης της εικόνας (image contrast) μέσω του σήματος διαμόρφωσης πλάτους και με την κλίμακα και τον προσανατολισμό της ταλάντωσης (scale and orientation) μέσω του μεταβαλλόμενου στο χώρο διανύσματος συχνοτήτων.
Αν και για ορισμένες περιπτώσεις φυσικών εικόνων είναι δυνατή η μοντελοποίησή τους με μία μοναδική συνημιτονοειδή συνάρτηση AM-FM (βλ. (3.2.12)), η πλειονότητα των *ευρυζωνικών* εικόνων που παρουσιάζουν πολύπλοκες χωρικές δομές όπως ακμές, συμβολές και συνδέσμους απαιτούν την παρουσία περισσοτέρων της μίας συνιστώσας AM-FM για την άρτια αναπαράσταση του τοπικού φάσματός τους. Με τέτοιες μοντελοποιήσεις θα ασχοληθούμε στην παρούσα ενότητα, περιγράφοντας αναλυτικά τη διαδικασία εξαγωγής χαρακτηριστικών αποδιαμόρφωσης από *ευρυζωνικές* εικόνες με τη χρήση των ενεργειακών τελεστών Teager-Kaiser σε καθεμιά από τις συνιστώσες *στενής ζώνης* στις οποίες αποσυντίθεται η αρχική εικόνα ύστερα από φιλτράρισμα με κυματοειδή Gabor. Θα δούμε εναλλακτικά σχήματα *Περιγραφέων Υφής (Texture Descriptors)* που προκύπτουν από τη χρήση τέτοιων χαρακτηριστικών, ικανά να διατηρήσουν την κυρίαρχη πληροφορία της υφής αλλά και των τοπικών δομών μιας φυσικής εικόνας.
Το *Μοντέλο Πολλαπλών AM-FM Συνιστωσών (The Multicomponent AM-FM Model)* προτάθηκε και αναπτύχθηκε από τους Bovik et al. [39] και Havlicek et al. ([40],[41]). Συγκεκριμένα, μια εικόνα I μοντελοποιείται ως μια υπέρθεση K τοπικά ομαλά μεταβαλλόμενων στο χώρο και *στενοζωνικών (narrowband)* ημιτονοειδών συνιστωσών $f_k(x,y)$ και την παρουσία Λευκού Γκαουσιανού Θορύβου (White Gaussian Noise) $w(x,y)$ ως εξής

$$I(x,y) = \sum_{k=1}^{K} \underbrace{\alpha_k(x,y)\cos(\phi_k(x,y))}_{f_k(x,y)} + w(x,y) \tag{3.3.1}$$

όπου τα σήματα διαμόρφωσης πλάτους και μεταβαλλόμενου στο χώρο διανύσματος συχνοτήτων για καθεμία από τις $k = 1, \ldots, K$ συνιστώσες είναι αντίστοιχα τα $\alpha_k(x,y)$ και $\vec{\omega_k}(x,y) = \nabla\phi_k(x,y)$. Να σημειώσουμε εδώ ότι η παραπάνω αναπαράσταση δεν είναι μοναδική ούτε ώς προς την υπέρθεση των συνιστωσών ούτε ως προς τα επιμέρους K ζευγάρια σημάτων διαμόρφωσης $(\alpha_k(x,y), \vec{\omega_k}(x,y))$. Οι Havlicek et al. στο [41] αναφέρουν ότι το ζεύγος πλάτους και φάσης διαμόρφωσης για κάθε συνιστώσα γίνεται μοναδικό μόνο μετά την προσθήκη ενός φανταστικού μέρους στο σήμα και τον ορισμό της αποσύνθεσης στις συνιστώσες. Θα μιλήσουμε παρακάτω για την αποδιαμόρφωση εικόνων μιγαδικών τιμών.

Το πρόβλημα αποδιαμόρφωσης συνίσταται στην προκειμένη περίπτωση στην εκτίμηση των σημάτων $\alpha_k(x,y)$ και $\nabla\phi_k(x,y)$ για όλες τις συνιστώσες της παραπάνω αποσύνθεσης. Μάλιστα, το *Μοντέλο Πολλαπλών AM-FM Συνιστωσών* εξ' ορισμού επιτρέπει τη χρήση του δισδιάστατου ενεργειακού τελεστή Teager-Kaiser και των *Αλγορίθμων Διαχωρισμού της Ενέργειας (CESA, DESA)* στις επιμέρους συνιστώσες $f_k(x,y)$ αφού οι τελευταίες αποτελούν *στενοζωνικά* τοπικά ομαλά μεταβαλλόμενα AM-FM σήματα. Επομένως, για την αποδιαμόρφωση μιας εικόνας της μορφής (3.3.1) υπεισέρχεται η ανάγκη αποτελεσματικού φασματικού διαχωρισμού των συνιστωσών και της απομόνωσης της απόκρισης συχνότητας καθεμιάς με τρόπο ώστε να περιορίζονται στο ελάχιστο οι παρεμβολές από τις αποκρίσεις των υπολοίπων κατά τη διαδικασία αποδιαμόρφωσης. Μια κοινή λύση σε τούτο πρόβλημα που ταυτόχρονα παρέχει και μείωση του θορυβώδους περιεχομένου αποτελεί το ζωνοπερατό φιλτράρισμα της εικόνας με μια συστοιχία δισδιάστατων φίλτρων (filterbank), ικανών να καλύψουν πυκνά το πεδίο συχνοτήτων. Τα εν λόγω φίλτρα θα πρέπει αφενός να είναι καλά τοποθετημένα στο φασματικό χώρο για την εύρωστη απεμπλοκή των συνιστωσών μεταξύ τους αλλά και να έχουν καλή χωρική τοποθέτηση προκειμένου να ανιχνεύουν τοπικές μεταβολές της δομής του σήματος. Η ιδανική επιλογή για τη συστοιχία των φίλτρων είναι τα δισδιάστατα μιγαδικά ισοτροπικά Gabor φίλτρα σε διάφορες κλίμακες και προσανατολισμούς για την επαρκή κάλυψη του συχνοτικού χώρου. Η προτίμησή τους οφείλεται κυρίως στην ικανότητα τους για την ελαχιστοποίηση της *αρχής αβεβαιότητας (uncertainty principle)* ταυτόχρονα για τη χωρική και φασματική τους τοποθέτηση. Έτσι, το πρόβλημα της αποδιαμόρφωσης του αρχικού δισδιάστατου σήματος διαμελίζεται στην αποδιαμόρφωση των επιμέρους συνιστωσών για τις οποίες γνωρίζουμε *a priori* τον αριθμό και το φασματικό εύρος τους μέσω της προκαθορισμένης συστοιχίας των φίλτρων. Κατόπιν του φιλτραρίσματος, είναι δυνατή η εφαρμογή των *Αλγορίθμων Διαχωρισμού της Ενέργειας* που παρουσιάστηκαν στην προηγούμενη ενότητα για την εκτίμηση των σημάτων πλάτους και φάσης στις εξόδους των καναλιών της συστοιχίας φίλτρων που αντιστοιχούν στις *στενοζωνικές* συνιστώσες. Στην επόμενη υποενότητα παρουσιάζεται αναλυτικά η διαδικασία κατασκευής της συστοιχίας των δισδιάστατων Gabor φίλτρων που χρησιμοποιήθηκε σε όλα τα πειράματα της παρούσας διπλωματικής εργασίας.

3.3.1 Φιλτράρισμα σε Πολλαπλές Ζώνες με Συστοιχία Gabor Φίλτρων

Ο σχεδιασμός της συστοιχίας των Gabor φίλτρων είναι πανομοιότυπος με εκείνον που προτάθηκε από τους Havlicek et al. στο [40]. Χρησιμοποιούνται συνολικά σαράντα (40) δισδιάστατα μιγαδικά ισοτροπικά Gabor φίλτρα που καλύπτουν το πεδίο συχνοτήτων σε μια *διάταξη τύπου κυματοειδών (wavelet-like tesselation)* σε πέντε (5) κλίμακες (scales) που προχωρούν με γεωμετρική πρόοδο και οκτώ (8) προσανατολισμούς (orientations). Η κρουστική απόκριση ενός τέτοιου φίλτρου με μοναδιαία $\mathcal{L}^2$ νόρμα είναι

$$g_k(\mathbf{x}) = \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left[-\frac{1}{4{\sigma_k}^2}\mathbf{x}^T\mathbf{x}\right] \exp\left[j2\pi\mathbf{\Omega}_k^T\mathbf{x}\right], \quad \mathbf{x}\in\mathbb{R}^2 \tag{3.3.2}$$

όπου η ακτινική κεντρική συχνότητα είναι $r_k = |\mathbf{\Omega}_k|$ και ο προσανατολισμός $\theta_k = \arg[\mathbf{\Omega}_k]$ για $k = 1, \ldots, 40$.
Το $\eta-$peak ακτινικό εύρος ζώνης οκτάβας ($\eta-$peak radial octave bandwidth) δίνεται από τη σχέση

$$B = \log_2 \left[\frac{r_k + \frac{\sqrt{-\ln \eta}}{2\pi\sigma_k}}{r_k - \frac{\sqrt{-\ln \eta}}{2\pi\sigma_k}} \right] \tag{3.3.3}$$

Ο συμβολισμός $\eta-$peak αναφέρεται στο κλάσμα της μέγιστης απόκρισης συχνότητας στην οποία διαδοχικά φίλτρα που ανήκουν στην ίδια ακτίνα (ισοδύναμα έχουν τον ίδιο προσανατολισμό) τέμνονται και βάσει αυτών των σημείων τομής γίνεται ο υπολογισμός του B. Όταν $\eta = \frac{1}{2}$ το ακτινικό εύρος ζώνης οκτάβας ισοδυναμεί με το εύρος ζώνης 3-dB του φίλτρου.
Το $\eta-$peak εύρος ζώνης προσανατολισμού ($\eta-$peak orientation bandwidth) είναι

$$\Theta = 2 \arctan \sqrt{\gamma} \quad , \quad \text{όπου} \quad \gamma = \frac{(2^B - 1)^2}{(2^B + 1)^2} \tag{3.3.4}$$

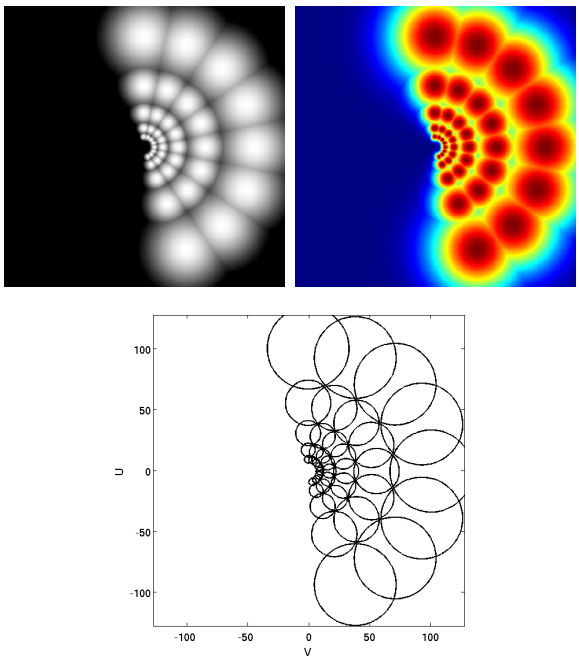
Η τοποθέτηση των φίλτρων γίνεται σε ένα πλέγμα πολικών συντεταγμένων με τρόπο ώστε τέσσερα φίλτρα που βρίσκονται σε διαδοχικούς προσανατολισμούς και διαδοχικές κλίμακες να έχουν αποκρίσεις συχνότητας που να τέμνονται στο ίδιο σημείο στο πεδίο συχνοτήτων στην τιμή ημίσειας κορυφής τους (half-peak). Οι ακτινικές κεντρικές συχνότητες r_k αυξάνονται με γεωμετρική πρόοδο λόγου R σε κάθε διαφορετική γωνία προσανατολισμού για να καλυφθούν οι πέντε (5) κλίμακες, ξεκινώντας με μικρότερη τιμή την $r_0 = 9.6$ κύκλοι ανά εικόνα. Η γωνιακή χωροθέτηση των ακτίνων διαφορετικού προσανατολισμού γίνεται με διάστημα

$$\Lambda = 2 \arcsin \left[(4R)^{-\frac{1}{2}} \sqrt{(R^2 + 1)(\gamma - 1) + 2R(\gamma + 1)} \right] \tag{3.3.5}$$

για να καλυφθεί στο χώρο συχνοτήτων μια συνολική γωνία π από οκτώ (8) προσανατολισμούς. Για τις τυπικές αποκλίσεις σ_k ισχύει

$$\sigma_k = \frac{\sqrt{-\ln(\eta)/\gamma}}{2\pi r_k} \tag{3.3.6}$$

Οι τιμές των παραμέτρων που χρησιμοποιούνται από τους Havlicek et al. στο [40] και υιοθετήθηκαν καί στα πειράματά μας είναι $\eta = \frac{1}{2}$, $R = 1.8$, $B = 1$ οκτάβα και $\gamma = \frac{1}{9}$.
Η αναπαράσταση του παραπάνω σχεδιασμού της συστοιχίας φίλτρων Gabor στο πεδίο συχνοτήτων δίνεται στο Σχήμα 3.1, συμμετρικά προσαρμοσμένη για εικόνα διαστάσεων 288×288.



Σχήμα 3.1: Σχεδιασμός Συστοιχίας Gabor Φίλτρων σε πέντε κλίμακες και οκτώ προσανατολισμούς. Επάνω Αριστερά: Γκρίζα απεικόνιση των φίλτρων στο δισδιάστατο πεδίο συχνοτήτων, Επάνω Δεξιά: Έγχρωμη απεικόνιση των φίλτρων στον δισδιάστατο πεδίο συχνοτήτων, Κάτω: Απεικόνιση των περιγραμμάτων που αντιστοιχούν στο εύρος ζώνης ημίσειας κορυφής (half-peak) της απόκρισης συχνότητας των φίλτρων.

3.3.2 Αποδιαμόρφωση στις Εξόδους των Καναλιών της Συστοιχίας Φίλτρων

Ύστερα από τη διαδικασία του φιλτραρίσματος της εικόνας εισόδου σε πολλαπλές ζώνες με τη συστοιχία των Gabor φίλτρων που περιγράψαμε παραπάνω, στην έξοδο των καναλιών έχουμε πλέον δισδιάστατα AM-FM συνεχή σήματα *στενής ζώνης* συχνοτήτων με φασματικό περιεχόμενο εντοπισμένο εντός του εύρους ζώνης των αντίστοιχων φίλτρων. Θεωρώντας ότι στην έξοδο κάθε καναλιού επικρατεί η απόκριση μιας μόνο συνιστώσας της υπέρθεσης (3.3.1) με τη συμβολή των υπολοίπων να κρίνεται αμελητέα, μπορούμε να εφαρμόσουμε τον *Συνεχή Αλγόριθμο Διαχωρισμού της Ενέργειας (CESA)* που περιγράψαμε στην ενότητα 3.2 για να εκτιμήσουμε τα σήματα διαμόρφωσης πλάτους και μεταβαλλόμενου διανύσματος συχνότητας των επιμέρους συνιστωσών του μοντέλου (3.3.1).
Εντούτοις, οι έξοδοι των καναλιών της συστοιχίας μιγαδικών φίλτρων Gabor είναι μιγαδικά σήματα και για τούτο απαιτείται η χρήση του μιγαδικού ενεργειακού τελεστή Teager-Kaiser για δισδιάστατα συνεχή μιγαδικά σήματα και αντίστοιχης τροποποίησης του *CESA* για την αποδιαμόρφωση. Αν I είναι η εικόνα εισόδου και g_k η κρουστική απόκριση του $k-$οστού φίλτρου που δίνεται από τη σχέση (3.3.2), στην έξοδο του $k-$οστού καναλιού της συστοιχίας λαμβάνουμε τη μιγαδική *στενοζωνική* συνιστώσα f_k

που θα έχει τη μορφή

$$f_k(x,y) = I(x,y) * g_k(x,y) = \alpha_k(x,y)\exp(j\phi_k(x,y)), \quad k = 1,\ldots,40 \tag{3.3.7}$$

Παρόμοια με την περίπτωση του μιγαδικού ενεργειακού τελεστή για μονοδιάστατα σήματα της σχέσης (3.1.24), εδώ ο μιγαδικός τελεστής C μπορεί να εκφραστεί ως το άθροισμα των εξόδων της δισδιάστατης μορφής του τελεστή Φ της σχέσης (3.1.7) στο πραγματικό και φανταστικό μέρος της f_k

$$C(f_k) = \Phi[\mathrm{Re}\{f_k\}] + \Phi[\mathrm{Im}\{f_k\}], \quad k = 1,\ldots,40 \tag{3.3.8}$$

οπότε με τη χρήση της προσεγγιστικής σχέσης (3.1.16) ο μιγαδικός δισδιάστατος τελεστής δίνει στην έξοδο

$$C(f_k) \approx 2{\alpha_k}^2\|\boldsymbol{\omega}_k\|^2 \tag{3.3.9}$$

Η παραπάνω σχέση σε συνδυασμό με τις εξόδους του τελεστή C από τις μερικές παραγώγους της f_k ως προς την κάθετη και οριζόντια κατεύθυνση $\partial f_k/\partial x$ και $\partial f_k/\partial y$ αντίστοιχα οδηγεί στην τροποποιημένη εκδοχή του σχήματος αποδιαμόρφωσης [42] δίνοντας πανομοιότυπες εκτιμήσεις για τις συνιστώσες του διανύσματος συχνοτήτων $\boldsymbol{\omega}_k$ με αυτές της σχέσης (3.2.9) για $d = 2$ και την ελαφρά τροποποιημένη εκτίμηση για το σήμα διαμόρφωσης πλάτους

$$|\alpha_k(x,y)| \approx \frac{C(f_k)}{\sqrt{2}\sqrt{C\left(\frac{\partial f_k}{\partial x}\right) + C\left(\frac{\partial f_k}{\partial y}\right)}} \tag{3.3.10}$$

Να υπενθυμίσουμε εδώ ότι τα πρόσημα για τις συνιστώσες του εκτιμημένου διανύσματος συχνοτήτων προσδιορίζονται από τα πρόσημα των συνιστωσών της κεντρικής συχνότητας του αντίστοιχου φίλτρου για κάθε επιμέρους σήμα f_k.
Εντούτοις, το φασματικό περιεχόμενο κάθε συνιστώσας του *Μοντέλου Πολλαπλών AM-FM συνιστωσών* της σχέσης (3.3.1) δεν είναι απαραίτητα συμμετρικά κατανεμημένο γύρω από την κεντρική συχνότητα του αντίστοιχου Gabor φίλτρου με αποτέλεσμα μια ενδεχόμενη εσφαλμένη εκτίμηση του πλάτους α_k. Για να αντισταθμίσουν την επίδραση αυτής της απόκλισης οι Kokkinos et al. [42] προτείνουν τη διαίρεση της παραπάνω εκτίμησης για το πλάτος με την τιμή της απόκρισης συχνότητας του αντίστοιχου φίλτρου στην εκτιμημένη από τον αλγόριθμο διαχωρισμού συχνότητα $\boldsymbol{\omega}_k$. Συμβολίζοντας με G_k την απόκριση συχνότητας του $k-$οστού φίλτρου η νέα εύρωστη εκτίμηση του πλάτους A_k δίνεται από τη σχέση

$$A_k = \frac{\alpha_k}{|G_k(\boldsymbol{\omega}_k)|} \tag{3.3.11}$$

Είδαμε, επομένως, τον τρόπο με τον οποίο μπορεί να εφαρμοστεί ο *Συνεχής Αλγόριθμος Διαχωρισμού της Ενέργειας* στις μιγαδικές εξόδους των καναλιών της συστοιχίας φίλτρων για την αποδιαμόρφωση των επιμέρους *στενοζωνικών* συνιστωσών. Στην περίπτωση όμως εικόνων διακριτού χώρου, επιβάλλεται η χρήση των διακριτών αναλόγων του ενεργειακού τελεστή Teager-Kaiser και αντίστοιχων διακριτών αλγορίθμων αποδιαμόρφωσης, όπως αυτούς που μελετήσαμε στην ενότητα 3.2. Υπάρχουν πολλές μεθοδολογίες

τέτοιων διακριτοποιήσεων, η πλειονότητα των οποίων βασίζεται στην προσέγγιση των μερικών παραγώγων του σήματος με απλές διακριτές διαφορές δειγμάτων, συμμετρικές ή μή. Τέτοιες προσεγγίσεις εισάγουν σφάλματα μη αμελητέα στο συνολικό σχήμα και θόρυβο ενώ συχνά οδηγούν σε αλλοπρόσαλλες εκτιμήσεις των σημάτων πλάτους και συχνότητας.

3.3.3 Gabor Αλγόριθμος Διαχωρισμού της Ενέργειας

Οι Dimitriadis και Maragos στο [43] προτείνουν δύο εναλλακτικά εύρωστα σχήματα αποδιαμόρφωσης μονοδιάστατων σημάτων φωνής, το πρώτο εκ των οποίων βασίζεται στην προσέγγιση διακριτών σημάτων με συνεχείς συναρτήσεις της μορφής *B-spline* ενώ το δεύτερο στη χρήση παραγώγων Gabor φίλτρων για την "κανονικοποίηση"του ενεργειακού τελεστή και του αλγορίθμου ESA. Τα αποτελέσματά τους υποστηρίζουν την επίλογη του δεύτερου σχήματος με το οποίο αποφεύγεται ο θόρυβος των διακριτοποιήσεων των παραγώγων και παράγονται πιο ομαλές εκτιμήσεις των παραγώγων του σήματος όταν οι τιμές SNR είναι χαμηλές, δηλαδή όταν το σήμα θορύβου είναι ισχυρό. Αυτή η "κανονικοποιημένη"εκδοχή του ESA που πρότειναν οι Dimitriadis και Maragos [43] ονομάστηκε *Gabor ESA*.
Οι Kokkinos et al. [42] επέκτειναν τον *Gabor ESA* για δισδιάστατα διακριτά σήματα πραγματικής τιμής και εισήγαγαν τον *"Κανονικοποιημένο"Αλγόριθμο Διαχωρισμού της Ενέργειας (Regularized ESA)* ή διαφορετικά *2D Gabor ESA*, τον οποίο θα μελετήσουμε στην παρούσα υποενότητα. Ωστόσο, όπως θα δούμε παρακάτω, ο εν λόγω αλγόριθμος μπορεί να εφαρμοστεί καί σε μιγαδικά σήματα με χρήση του μιγαδικού τελεστή Teager-Kaiser που περιγράψαμε στην προηγούμενη υποενότητα.
Έστω $I(x,y)$ η συνεχής εικόνα εισόδου, $g(x,y)$ η κρουστική απόκριση ενός δισδιάστατου πραγματικού Gabor φίλτρου και $f(x,y) = I(x,y) * g(x,y)$ η φιλτραρισμένη έξοδος. Χάρη στην αντιμεταθετική ιδιότητα της συνέλιξης και της παραγώγισης μεταξύ τους, η έξοδος του δισδιάστατου τελεστή Teager-Kaiser Φ από τη φιλτραρισμένη εικόνα f μπορεί να γραφεί ως εξής

$$\begin{aligned}\Phi(f) = \Phi(I*g) &= \|I*\nabla g\|^2 - (I*g)(I*\nabla^2 g) \\ &= \|I*(g_x, g_y)\|^2 - (I*g)(I*(g_{xx}+g_{yy}))\end{aligned} \tag{3.3.12}$$

Παρατηρούμε λοιπόν ότι δεν απαιτείται ο υπολογισμός των μερικών παραγώγων της αρχικής εικόνας I ή της φιλτραρισμένης εξόδου f (και επομένως η διακριτοποίησή τους για την περίπτωση διακριτών σημάτων) αλλά μόνο ο υπολογισμός των παραγώγων του φίλτρου, οι οποίες υπολογίζονται αναλυτικά σε κλειστό τύπο.
Για την παραγωγή των εκτιμήσεων για το πλάτος και τις συνιστώσες του μεταβαλλόμενου διανύσματος συχνοτήτων στις δύο κατευθύνσεις της εικόνας f χρειαζόμαστε δύο επιπλέον σχέσεις, όπως για κάθε δισδιάστατο αλγόριθμο ESA, που προκύπτουν από τις εξόδους του ενεργειακού τελεστή όταν αυτός εφαρμοστεί στις μερικές παραγώγους του σήματος ως προς x και y.

Ακολουθώντας την ίδια λογική υπολογίζουμε τα $\Phi(f_x)$ και $\Phi(f_y)$

$$\begin{aligned}\Phi(f_x) &= \|I * \nabla g_x\|^2 - (I * g_x)(I * \nabla^2 g_x) \\ &= \|I * (g_{xx}, g_{xy})\|^2 - (I * g_x)\,(I * (g_{xxx} + g_{xyy})) \qquad (3.3.13)\\ \Phi(f_y) &= \|I * \nabla g_y\|^2 - (I * g_y)(I * \nabla^2 g_y) \\ &= \|I * (g_{yx}, g_{yy})\|^2 - (I * g_y)(I * (g_{yxx} + g_{yyy})) \qquad (3.3.14)\end{aligned}$$

Η αντικατάσταση των τιμών των $\Phi(f)$, $\Phi(f_x)$ και $\Phi(f_y)$ από τις σχέσεις (3.3.12), (3.3.13) και (3.3.14) αντίστοιχα στις σχέσεις του *Συνεχούς Αλγορίθμου Διαχωρισμού της Ενεργειας (CESA)* (3.2.8) και (3.2.9) παρέχει τον διαχωρισμό της ενέργειας ταλάντωσης στα εκτιμώμενα σήματα πλάτους και συχνότητας. Το παραπάνω ενοποιημένο σχήμα φιλτραρίσματος και αποδιαμόρφωσης αποτελεί τον *Gabor Αλγόριθμο Διαχωρισμού της Ενέργειας (2D Gabor ESA)*.

Για τον υπολογισμό της ενέργειας του σήματος f και των παραγώγων του απαιτούνται ο υπολογισμός των παραγώγων g_x, g_y, g_{xx}, g_{yy}, g_{xy} και των Λαπλασιανών τελεστών $\nabla^2 g_x$ και $\nabla^2 g_y$ και οι αντίστοιχες συνελίξεις με την αρχική εικόνα εισόδου $I(x, y)$. Για να περιορίσουν το υπολογιστικό κόστος που προσθέτουν οι συνελίξεις στον αλγόριθμο, οι Kokkinos et al. [42] τις αντικαθιστούν στον *2D Gabor ESA* με πολλαπλασιασμούς στο πεδίο της συχνότητας μέσω του αποτελεσματικού αλγορίθμου *FFT (Fast Fourier Transform)* , υπολογίζοντας τους μετασχηματισμούς Fourier των παραγώγων του φίλτρου βάσει της παρακάτω σχέσης

$$\mathcal{F}\left\{\frac{\partial^{k+l} g}{\partial x^k \partial y^l}\right\} = \mathcal{F}\{g\}(j\omega_x)^k (j\omega_y)^l \qquad (3.3.15)$$

όπου $\mathcal{F}\{g\}$ ο μετασχηματισμός Fourier του φίλτρου Gabor g.

Η επιλογή χρήσης της σχέσης (3.3.15) υποδεικνύει την αντικατάσταση των πραγματικών από μιγαδικά φίλτρα Gabor όπως αυτά της σχέσης (3.3.2). Στην προκειμένη περίπτωση, ο τελεστής Φ αντικαθίσταται από τον μιγαδικό ανάλογό του C της σχέσης (3.3.8) ενώ η εκτίμηση για το πλάτος είναι η τροποποιημένη της σχέσης (3.3.10). Σε κάθε περίπτωση υιοθετείται η αντιστάθμιση για το πλάτος της σχέσης (3.3.11).

Από τα παραπάνω, γίνεται φανερό ότι ο *2D Gabor ESA* είναι απευθείας εφαρμόσιμος για την περίπτωση σημάτων διακριτού χώρου με απλή δειγματοληψία των συνεχών σημάτων αφού δεν περιέχει καμία διακριτοποίηση των παραγώγων. Τα σφάλματα που εισάγουν τέτοιες διακριτοποιήσεις στις εκτιμήσεις της αποδιαμόρφωσης από τον *DESA* εξαλείφονται στο συγκεκριμένο σχήμα, με μοναδικά σφάλματα τις προσεγγίσεις του αλγορίθμου διαχωρισμού ενέργειας και τις επιδράσεις του σήματος θορύβου.

Επιστρέφοντας στο πρόβλημα της αποδιαμόρφωσης ενός σήματος εικόνας *ευρείας ζώνης* συχνοτήτων που ακολουθεί το *Μοντέλο Πολλαπλών AM-FM Συνιστωσών* της σχέσης (3.3.1), είδαμε παραπάνω ότι αυτή επιτυγχάνεται σε δύο στάδια. Αρχικά φιλτράρεται το δισδιάστατο σήμα εισόδου σε πολλαπλές ζώνες συχνοτήτων μέσω της συστοιχίας Gabor φίλτρων που περιγράψαμε (βλ. Σχήμα 3.1) και ακολούθως εφαρμόζεται η τροποποιημένη για μιγαδικά σήματα εκδοχή του *Συνεχούς Αλγορίθμου Διαχωρισμού της Ενέργειας (CESA)* για την αποδιαμόρφωση ξεχωριστά της *στενοζωνικής* εξόδου κάθε καναλιού της συστοιχίας.

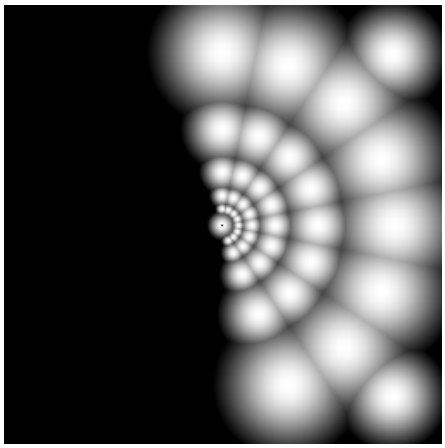
Ωστόσο, προκειμένου να εφαρμοστεί το παραπάνω σχήμα σε εικόνες διακριτού χώρου θα πρέπει σε κάθε φιλτραρισμένη έξοδο καναλιού της συστοιχίας να χρησιμοποιηθεί ο *Διακριτός Αλγόριθμος Διαχωρισμού της Ενέργειας (DESA)*, προφανώς με αντίστοιχη τροποποίηση για μιγαδικά σήματα. Αυτό πρακτικά συνεπάγεται την εισαγωγή σφαλμάτων διακριτοποίησης των παραγώγων στο σχήμα αποδιαμόρφωσης κάθε εξόδου καναλιού ή ισοδύναμα κάθε συνιστώσας της υπέρθεσης (3.3.1) με τις επιβλαβείς επιδράσεις που περιγράψαμε παραπάνω.
Για την αντιμετώπιση του εν λόγω προβλήματος, αντί της πρακτικής του φιλτραρίσματος σε πολλαπλές ζώνες και της μετέπειτα *DESA* αποδιαμόρφωσης σε κάθε κανάλι, μπορεί να γίνει πολλαπλή χρήση του σχήματος του *2D Gabor ESA* για όλα τα φίλτρα της συστοιχίας φίλτρων Gabor. Με άλλα λόγια, με την εφαρμογή του *2D Gabor ESA* κάθε φορά για διαφορετικό φίλτρο της συστοιχίας (3.3.2) επιτυγχάνεται ταυτόχρονα το φιλτράρισμα στη συγκεκριμένη ζώνη συχνοτήτων και η εύρωστη, χωρίς διακριτοποίηση παραγώγων, αποδιαμόρφωση του προκύπτοντος *στενοζωνικού* σήματος. Αυτό το συνδυασμένο σχήμα φιλτραρίσματος-αποδιαμόρφωσης χρησιμοποιούμε σε όλα τα σχετικά πειράματα της παρούσας διπλωματικής. Να σημειώσουμε ότι ο *2D Gabor ESA* αυτής της προσέγγισης αναφέρεται σε διακριτά σήματα και επιπλέον απαιτεί μιγαδικά φίλτρα Gabor. Επομένως, ισχύουν οι τροποποιήσεις που αναφέραμε παραπάνω για τη χρήση του μιγαδικού τελεστή και των αντίστοιχων εκτιμήσεων του αλγορίθμου διαχωρισμού ενώ όπως κάναμε σαφές το διακριτό ανάλογο του *2D Gabor ESA* δεν απαιτεί καμία επιπλέον τροποποίηση.

3.3.4 Αποδιαμόρφωση Εικόνων και Περιγραφείς Υφής

Τα χαρακτηριστικά υφής έχουν αποδειχθεί πολύτιμα για την ανάλυση εικόνων σε πολλά προβλήματα της Όρασης Υπολογιστών όπως η κατάτμηση εικόνων, η υπολογιστική στερέοψη βασισμένη στη φάση, η εκτίμηση του 3D σχήματος από την υφή και η ανάκτηση εικόνων. Στα επόμενα κεφάλαια θα μελετήσουμε αναλυτικά τον τρόπο με τον οποίο *Περιγραφείς Υφής (Texture Descriptors)* που προκύπτουν από τη βασισμένη στην ενέργεια αποδιαμόρφωση των εικόνων παρέχουν εύρωστα χαρακτηριστικά *οπτικής σημαντικότητας (visual saliency)* και οδηγούν σε βελτιωμένες επιδόσεις για το πρόβλημα της Αναγνώρισης Ανθρώπινων Δράσεων σε Βίντεο.

3.3.4.1 Ανάλυση Συνιστωσών Καναλιού (Channelized Component Analysis)

Η διαδικασία φιλτραρίσματος σε πολλαπλές ζώνες και αποδιαμόρφωσης ενός *ευρυζωνικού* σήματος εικόνας της μορφής (3.3.1) που περιγράψαμε παραπάνω μας προμηθεύει μια πλούσια μεγάλης διάστασης αναπαράσταση της εικόνας. Συγκεκριμένα, η αποδιαμόρφωση στην έξοδο κάθε καναλιού της συστοιχίας K φίλτρων (στην περίπτωσή μας $K = 40$) παρέχει σε κάθε pixel της εικόνας ένα διάνυσμα $(A_k, \nabla\phi_k)$ για κάθε κανάλι, που αντιστοιχεί στα σήματα AM-FM διαμόρφωσης κάθε *στενοζωνικής* συνιστώσας της υπέρθεσης (3.3.1). Συνολικά λοιπόν το προκύπτον διάνυσμα των χαρακτηριστικών υφής έχει διάσταση $3 \times K$ για κάθε pixel. Το παραπάνω σχήμα εισήχθη από τους Havlicek et al. στο [41] με την ονομασία *Ανάλυση Συνιστωσών Καναλιού (Channelized Component*



Σχήμα 3.2: Επαυξημένη Συστοιχία 43 μιγαδικών φίλτρων Gabor για την *Ανάλυση Συνιστωσών Καναλιού CCA*. Στο κέντρο του πεδίου συχνοτήτων εικονίζεται το βαθυπερατό φίλτρο χαμηλής κεντρικής συχνότητας και στις πάνω δεξιά και κάτω δεξιά γωνίες τα δύο ειδικά υψηλής κεντρικής συχνότητας φίλτρα.

Analysis (CCA)).
Η *Ανάλυση Συνιστωσών Καναλιού CCA* είναι μια ολική αποσύνθεση της εικόνας σε επίπεδο συνιστωσών που βασίζεται στην υπόθεση ότι η απόκριση κάθε καναλιού της συστοιχίας κυριαρχείται ολικά από μία μοναδική *στενοζωνική* συνιστώσα για όλα τα pixel της εικόνας. Με άλλα λόγια, θεωρούμε ότι τα κανάλια της συστοιχίας έχουν την ικανότητα να απομονώσουν σε καθολικό επίπεδο τις AM-FM συνιστώσες. Η μέθοδος *CCA* συνήθως ενσωματώνει στη συστοιχία των μιγαδικών Gabor φίλτρων του Σχήματος 3.1 ένα επιπλέον βαθυπερατό φίλτρο για τη σύλληψη δομών χαμηλής συχνότητας στην εικόνα όπως οι υψηλής κλίμακας σκιάσεις και μεταβολές έντασης της εικόνας. Επιπλέον, για την αιχμαλώτιση υψηλής συχνότητας ακμών της δομής της εικόνας η συστοιχία των Gabor φίλτρων μπορεί να συμπληρωθεί επιπλέον από δύο ειδικά φίλτρα υψηλής συχνότητας που βρίσκονται στις γωνίες του δεξιού μέρους του πεδίου συχνοτήτων. Η παραπάνω επαυξημένη συστοιχία των σαράντα τριών (43) φίλτρων εικονίζεται στο Σχήμα 3.2. Η *Ανάλυση Συνιστωσών Καναλιού* συνιστά μια αποτελεσματική αναπαράσταση της εικόνας δεδομένου ότι περιγράφει με πληρότητα τα χαρακτηριστικά διαμόρφωσης της εικόνας. Για τούτο εμφανίζει υψηλές επιδόσεις για το πρόβλημα της ανακατασκευής της εικόνας από τις επιμέρους *στενής ζώνης* συνιστώσες της. Θα παρουσιάσουμε σχετικά πειραματικά αποτελέσματα στην επόμενη ενότητα.
Η μέθοδος *CCA* που εκτελεί διαχωρισμό συνιστωσών σε καθολικό επίπεδο, ωστόσο, παρουσιάζει σημαντικά προβλήματα και περιορισμούς σε ορισμένες περιπτώσεις σύμφωνα με τους Havlicek et al. [41]. Συχνά τα σήματα διαμόρφωσης πλάτους αρκετών συνιστωσών καναλιού είναι χαμηλής τιμής σε μεγάλες περιοχές της εικόνας κάτι που αναπόφευκτα οδηγεί στην πλεονασματική αναπαράσταση χαρακτηριστικών υφής μέσω πολλαπλών συνιστωσών σε τέτοιες περιοχές. Το σχήμα *CCA* πάσχει επίσης σε περιοχές *στενής ζώνης συχνοτήτων* της εικόνας, όπου το φασματικό περιεχόμενο των αποκρίσεων πολλών συνιστωσών καναλιού εντοπίζεται πολύ μακριά από το στενό φάσμα συχνοτήτων της εικόνας στις δεδομένες περιοχές με αποτέλεσμα την εισαγωγή θορύβου που αποφέρει

λανθασμένες εκτιμήσεις αποδιαμόρφωσης. Τέλος, η μεγάλη διάσταση του διανύσματος χαρακτηριστικών που παρέχει η *CCA* το καθιστά δύσκολα διαχειρίσιμο και σε περιπτώσεις πλεονασματικό και θορυβώδες για προβλήματα της Όρασης Υπολογιστών όπως είναι η κατάτμηση υφής (texture segmentation).
Παρακάτω θα μελετήσουμε έναν εναλλακτικό, πιο συμπαγή και ομαλό *Περιγραφέα Υφής*, ικανό να συλλάβει την προεξέχουσα δομή των σημάτων υφής.

3.3.4.2 Ανάλυση Κυρίαρχων Συνιστωσών (Dominant Component Analysis)

Οι Havlicek et al. στο [41], εκτός της μεθόδου *CCA*, παρουσίασαν μια νέα τεχνική ανάλυσης της υφής που βασίζεται στις αποκρίσεις των καναλιών της συστοιχίας φίλτρων για να εξάγει τα σήματα πλάτους και συχνότητας της *στενοζωνικής* συνιστώσας που κυριαρχεί σε κάθε pixel της εικόνας. Σε αντίθεση με την *CCA* που επενεργεί σε ολικό επίπεδο για να διαχωρίσει τις επιμέρους συνιστώσες, η νέα μέθοδος στηρίζεται στην υπόθεση ότι το πολύ μία συνιστώσα του μοντέλου (3.3.1) κυριαρχεί στην απόκριση κάθε καναλιού της συστοιχίας τοπικά σε κάθε pixel. Με βάση την παραπάνω υπόθεση, σε κάθε pixel (x, y) επιλέγεται το πλησιέστερο κανάλι $i(x, y)$ της συστοιχίας στην επικρατούσα συνιστώσα και στη συνέχεια αποδιαμορφώνεται η έξοδός του με ένα σχήμα διαχωρισμού της ενέργειας ώστε να κατασκευαστεί ο περιγραφέας υφής για το δεδομένο pixel με τα σήματα AM-FM της κυρίαρχης συνιστώσας.
Η παραπάνω μέθοδος ονομάστηκε από τους Havlicek et al. [41] *Ανάλυση Κυρίαρχων Συνιστωσών (Dominant Component Analysis (DCA))*. Η επιλογή του καναλιού $i(x, y)$ που αντιστοιχεί στην κυρίαρχη συνιστώσα σε επίπεδο pixel γίνεται βάσει της μεγιστοποίησης ενός κριτηρίου $\Gamma_k(x, y)$ ως εξής

$$i(x, y) = \arg \max_{1 \le k \le K} \{\Gamma_k(x, y)\} \tag{3.3.16}$$

όπου K ο αριθμός των καναλιών της συστοιχίας.
Μετά την επιλογή του καναλιού, ο προκύπτων *Περιγραφέας Υφής* για το pixel (x, y) είναι ένα διάνυσμα τριών στοιχείων, του σήματος διαμόρφωσης πλάτους και της κάθετης και οριζόντιας συνιστώσας της συχνότητας, οι εκτιμήσεις δηλαδή από την αποδιαμόρφωση της εξόδου του επιλεγμένου καναλιού

$$A_{DCA}(x, y) = A_{i(x,y)}(x, y), \qquad \boldsymbol{\omega}_{DCA}(x, y) = \boldsymbol{\omega}_{i(x,y)}(x, y) \tag{3.3.17}$$

Θεωρούμε ως εικόνα εισόδου την $I(x, y)$ της σχέσης (3.3.1). Αν συμβολίσουμε με $y_k(x, y) = I(x, y) * g_k(x, y)$ την έξοδο του $k-$οστού καναλιού της συστοιχίας και $G_k(\boldsymbol{\Omega})$ την απόκριση συχνότητας του $k-$οστού φίλτρου Gabor, το κριτήριο $\Gamma_k(x, y)$ που χρησιμοποιήθηκε στην πρωτότυπη εργασία των Havlicek et al. [41] είναι

$$\Gamma_k(x, y) = \frac{|y_k(x, y)|}{\max_{\boldsymbol{\Omega}} |G_k(\boldsymbol{\Omega})|} \tag{3.3.18}$$

βάσει της λογικής ότι η απόκριση καναλιού y_k στο σημείο (x, y) κυριαρχείται από κάποια *στενοζωνική* συνιστώσα f_i της εικόνας ώστε να ισχύει προσεγγιστικά τοπικά στο σημείο (x, y)

$$y_k(x, y) \approx f_i(x, y) * g_k(x, y) \tag{3.3.19}$$

Λαμβάνοντας υπόψη την υπόθεση ότι το φάσμα συχνοτήτων κάθε συνιστώσας της I μετά το φιλτράρισμα είναι εντοπισμένο γύρω από την κεντρική συχνότητα του αντίστοιχου φίλτρου της συστοιχίας, στο παραπάνω κριτήριο το συχνοτικό περιεχόμενο δεν συμμετέχει στην επιλογή του καναλιού αφού κατ΄ ουσία επιλέγονται κανάλια με μεγάλες τιμές του σήματος διαμόρφωσης πλάτους. Για τούτο, ισοδύναμα το κριτήριο (3.3.18) μπορεί να εκφρασθεί καί ως εξής

$$\Gamma_k(x, y) = |\alpha_k(x, y)| \tag{3.3.20}$$

Οι Kokkinos et al. [42] έδωσαν στο παραπάνω σχήμα εξαγωγής των κυρίαρχων συνιστωσών τον όρο *Ανάλυση Κυρίαρχων Συνιστωσών βασισμένη στο Πλάτος (Amplitude-based Dominant Component Analysis (ADCA))*. Έτσι, σε κάθε pixel της εικόνας το κανάλι που επιλέγεται είναι εκείνο στου οποίου την έξοδο ο αλγόριθμος ESA έδωσε σήμα διαμόρφωσης πλάτους μεγαλύτερης τιμής σε σχέση με τα υπόλοιπα.
Όπως έχουμε δει, σε εικόνες υφής το σήμα διαμόρφωσης πλάτους στο μοντέλο AM-FM περιέχει πληροφορία για την τοπική αντίθεση (contrast) της εικόνας ενώ το μεταβαλλόμενο διάνυσμα συχνοτήτων για την κλίμακα και τον προσανατολισμό της ταλάντωσης, μέσω του μέτρου και της κατεύθυνσής του αντίστοιχα. Η μέθοδος *ADCA* επομένως ευνοεί τα κανάλια που παρουσιάζουν υψηλές μεταβολές του πλάτους και επομένως της τοπικής αντίθεσης της εικόνας ενώ αγνοεί κανάλια που αντιστοιχούν σε σημαντικά στοιχεία της δομής της εικόνας μέσω του διανύσματος συχνοτήτων.
Μια εικόνα που ακολουθεί το *Μοντέλο Πολλαπλών AM-FM Συνιστωσών* της σχέσης (3.3.1) μπορεί να ιδωθεί ως το αποτέλεσμα πηγών ταλάντωσης σε διάφορες κλίμακες και προσανατολισμούς και όπως είδαμε στην ενότητα 3.1 ο ενεργειακός τελεστής Teager-Kaiser έχει την ικανότητα να συλλαμβάνει την ενέργεια των πηγών που προκάλεσαν την ταλάντωση. Βασισμένοι στις παραπάνω ιδέες, οι Kokkinos et al. [42] πρότειναν η επιλογή καναλιού να γίνεται βάσει της μέγιστης απόκρισης του τελεστή Teager-Kaiser από τις εξόδους της συστοιχίας, με αποτέλεσμα την ανάκτηση σε κάθε pixel των διαμορφώσεων υφής που αντιστοιχούν στην πηγή με τη μεγαλύτερη ενέργεια ταλάντωσης. Επομένως, το νέο κριτήριο $\Gamma_k(x, y)$ που εισάγουν έχει την έκφραση

$$\Gamma_k(x, y) = \Phi\left[I * g_k\right](x, y) \tag{3.3.21}$$

όπου Φ η δισδιάστατη μορφή του τελεστή Teager-Kaiser και g_k η κρουστική απόκριση του $k-$οστού πραγματικού φίλτρου Gabor ($1 \leq k \leq K$). Να σημειωθεί εδώ ότι για τα μιγαδικά φίλτρα g_k της μορφής (3.3.2) της συστοιχίας του Σχήματος 3.1 ο τελεστής Φ αντικαθίσταται από τον μιγαδικό ανάλογό του C της σχέσης (3.3.8).
Θυμίζοντας ότι η έξοδος του τελεστή Φ είναι προσεγγιστικά το τετράγωνο του γινομένου του πλάτους και του μέτρου του διανύσματος συχνοτήτων (βλ.(3.1.16)), το κριτήριο (3.3.21) συλλαμβάνει ταυτόχρονα πληροφορία για το πλάτος και τη συχνότητα, καθιστώντας δυνατή την επιλογή καναλιών με χαμηλές τιμές πλάτους αλλά υψηλό μέτρο του διανύσματος συχνοτήτων. Κανάλια με τέτοιες ιδιότητες θα απορρίπτονταν εκ

προοιμίου από το σχήμα *ADCA*.
Το παραπάνω εναλλακτικό σχήμα εξαγωγής των κυρίαρχων συνιστωσών των Kokkinos et al. [42] με βάση το κριτήριο επιλογής καναλιού (3.3.21) ονομάζεται *Ανάλυση Κυρίαρχων Συνιστωσών βασισμένη στην Ενέργεια (Energy-based Dominant Component Analysis (EDCA))*. Οι συγγραφείς με πειράματά τους παρατηρούν ότι η μέθοδος *EDCA* παράγει καλύτερα τοποθετημένες και πιο έντονες μετρήσεις στα σύνορα των αντικειμένων που εμφανίζονται στην εικόνα. Αντίθετα, το σχήμα *ADCA* ευνοεί πιο μεγάλης κλίμακας μεταβολές της αντίθεσης της εικόνας (contrast), που σε ορισμένες περιπτώσεις δεν αντιστοιχούν σε διαμορφώσεις υφής. Ακόμα, η υπεροχή της *EDCA* προσέγγισης αποτυπώνεται καί στο γεγονός ότι παράγει καλύτερες ανακατασκευές της εικόνας, διατηρώντας μεγαλύτερη λεπτομέρεια της δομής της.

Συνοψίζοντας, η *Ανάλυση Κυρίαρχων Συνιστωσών DCA* παρέχει ένα χαμηλής διάστασης ομαλά μεταβαλλόμενο διάνυσμα χαρακτηριστικών (3.3.17) που συνιστά έναν αποτελεσματικό *Περιγραφέα Υφής (Texture Descriptor)*, με την έννοια ότι κωδικοποιεί σε επίπεδο pixel τα προεξέχοντα χαρακτηριστικά υφής μιας εικόνας. Η αντίθεση της εικόνας, η κλίμακα και ο προσανατολισμός σε συνδυασμό με τις τιμές της αρχικής εικόνας αποτελούν μια εύρωστη αναπαράσταση για εφαρμογές όπως η κατάτμηση εικόνων [42]. Ειδικότερα, το σχήμα *EDCA* ανιχνεύει τις κυρίαρχες διαμορφώσεις υφής χωρίς την απώλεια λεπτομερειών της δομής των εικόνων όπως είναι οι ακμές και τα περιγράμματα των εικονιζόμενων αντικειμένων και ανθρώπων. Για τούτο, σε όλα τα σχετικά πειράματα της παρούσας διπλωματικής προτιμούμε το σχήμα *EDCA* για την εξαγωγή είτε των κυρίαρχων ενεργειών είτε του συνολικού *Περιγραφέα EDCA* με τα χαρακτηριστικά αποδιαμόρφωσης.
Αξίζει να σημειωθεί ότι τη μέθοδο *DCA* δεν ενδιαφέρει η DC συνιστώσα του φάσματος Fourier και για αυτό το λόγο παραλείπεται από τη συστοιχία Gabor φίλτρων το βαθυπερατό φίλτρο που υπάρχει στη συστοιχία του σχήματος 3.2. Επιπλέον, αμφότερες οι αναλύσεις *CCA* και *DCA* είναι, κατά τους Havlicek et al. [41], ουσιαστικά ανεξάρτητες από το σχεδιασμό της συστοιχίας των φίλτρων που χρησιμοποιείται. Ωστόσο, για την *DCA* προτιμάται η συστοιχία του Σχήματος 3.1 ενώ για την *CCA* η επαυξημένη συστοιχία του Σχήματος 3.2 που αποφέρει καλύτερες ανακατασκευές της εικόνας.

Φιλτράρισμα Ομαλοποίησης των Εκτιμήσεων Πλάτους και Συχνότητας. Σε αυτή την παράγραφο θα αναφερθούμε στη μετέπειτα επεξεργασία που προτείνεται στη σχετική βιβλιογραφία για τις εκτιμήσεις του πλάτους και της συχνότητας από αλγορίθμους διαχωρισμού της ενέργειας στις εξόδους ζωνοπερατών φίλτρων. Μη ομαλές εκτιμήσεις οφείλονται κυρίως σε σφάλματα προσέγγισης των αλγορίθμων ESA όταν εφαρμόζονται σε σήματα που δεν είναι τοπικά *στενοζωνικά* και εμφανίζουν τοπικές απότομες διαταραχές. Όπως αναφέρουν οι Havlicek et al. στο [41], αυτά τα σφάλματα αντισταθμίζονται συνήθως με *βαθυπερατό μετά-φιλτράρισμα ομαλοποίησης (low-pass smoothing post-filtering)* των εκτιμήσεων με βαθυπερατά φίλτρα σταθεράς χώρου πολλαπλάσιας της σταθεράς χώρου των αντίστοιχων φίλτρων καναλιού. Οι Maragos et al. [37] εξαλείφουν ανεπιθύμητες απότομες κορυφές στις εκτιμήσεις της συχνότητας AM-FM σημάτων φωνής με χρήση φίλτρων median.

Μια άλλη πηγή προβλημάτων για τις εκτιμήσεις των σημάτων διαμόρφωσης είναι οι ασυνέχειες της φάσης της εικόνας που οφείλονται σε διάφορους παράγοντες όπως οι ασυνέχειες και αντανακλάσεις επιφανειών, τα οπτικά εμπόδια, οι σκιάσεις ή ο θόρυβος, χαρακτηριστικά που εμφανίζονται συχνά σε φυσικές εικόνες. Τέτοιες ασυνέχειες ενδέχεται να δώσουν υπερβολικά μεγάλης τιμής τοπικά μέγιστα στην εκτίμηση του πλάτους, διευρύνοντας κατά πολύ το πεδίο τιμών με συνέπεια να αλλοιώνεται σημαντικά η ανακατασκευή των γκρίζων εικόνων μέσω *CCA*. Για την αντιμετώπιση του εν λόγω προβλήματος οι Havlicek et al. [41] προτείνουν αντί του φιλτραρίσματος ομαλοποίησης του ίδιου του πλάτους, την εφαρμογή βαθυπερατού γκαουσιανού φίλτρου, τυπικής απόκλισης πολλαπλάσιας αυτής των αντίστοιχων Gabor φίλτρων καναλιού, στις εκτιμήσεις της συχνότητας. Χρησιμοποιώντας τις ομαλοποιημένες εκτιμήσεις συχνότητας, το πλάτος επανεκτιμάται μέσω του δικού τους σχήματος αποδιαμόρφωσης και έπειται διέρχεται από βαθυπερατό φιλτράρισμα.
Συμπερασματικά, οι δύο κοινές επιλογές για *μετά-φιλτράρισμα* των εκτιμήσεων πλάτους και συχνότητας είναι τα βαθυπερατά φίλτρα εξομάλυνσης, με σταθερά χώρου πολλαπλάσια των αντίστοιχων φίλτρων καναλιού, και τα φίλτρα τύπου median. Αν και προσθέτει υπολογιστικό κόστος στο συνολικό σχήμα, ειδικά όταν πρόκειται για εικόνες μεγάλων διαστάσεων, αυτή η εκ των υστέρων επεξεργασία είναι ουσιώδης για την αποφυγή μη ομαλών και αλλοπρόσαλλων εκτιμήσεων των AM-FM σημάτων.
Στην παρακάτω ενότητα θα παρουσιάσουμε πειράματα αποδιαμόρφωσης σε στατικές εικόνες ή frames δειγμάτων βίντεο με τη χρήση των σχημάτων *DCA* και *CCA* και του *Gabor Αλγορίθμου Διαχωρισμού της Ενέργειας*.

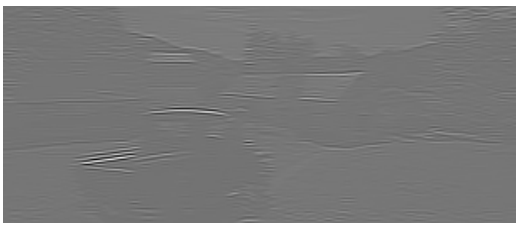
3.4 Λεπτομέρειες Υλοποίησης - Πειραματικά Αποτελέσματα

Στην παρούσα ενότητα θα παρουσιάσουμε πειραματικά αποτελέσματα από το φιλτράρισμα σε πολλαπλές ζώνες *ευρυζωνικών* εικόνων και την AM-FM αποδιαμόρφωση των επιμέρους συνιστωσών τους. Για τις εικόνες που εξετάσαμε, θα εκθέσουμε σε γκρίζα απεικόνιση τις εξόδους του ενεργειακού τελεστή Teager-Kaiser και τα χαρακτηριστικά υφής που προκύπτουν από τα σχήματα *Ανάλυσης Συνιστωσών Καναλιού (CCA)* και *Ανάλυσης Κυρίαρχων Συνιστωσών (DCA)* στις δύο εναλλακτικές της μορφές *EDCA* και *ADCA*.
Για τα εν λόγω πειράματα χρησιμοποιήθηκαν χαρακτηριστικά frames με έντονο περιεχόμενο υφής από δείγματα βίντεο της Βάσης Δεδομένων *Hollywood2 Actions Dataset*. Όλα τα frames αναλύθηκαν στη μισή ανάλυση (half resolution) και οι έγχρωμες εικόνες μετατράπηκαν αρχικά σε γκρίζες (greyscale) πριν οδηγηθούν στα σχήματα φιλτραρίσματος και αποδιαμόρφωσης. Επομένως, χρησιμοποιήθηκαν τελεστές Teager-Kaiser για βαθμωτά και όχι διανυσματικά (όπως π.χ. οι RGB εικόνες) σήματα.
Η υλοποίηση του φιλτραρίσματος πολλαπλών ζωνών και της *2D Gabor ESA* αποδιαμόρφωσης έγινε στο λογισμικό MATLAB με χρήση του *dca2D package* του G. Evangelopoulos. Για το φίλτραρισμα χρησιμοποιήθηκαν οι συστοιχίες μιγαδικών Gabor φίλτρων των Σχημάτων 3.1 και 3.2 για την ανάλυση *DCA* και *CCA* αντίστοιχα. Παρατηρήσαμε

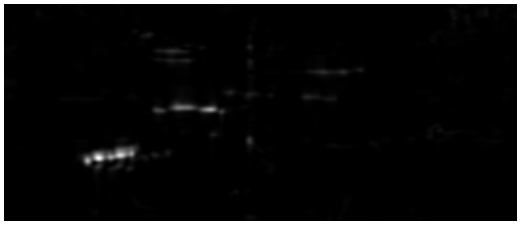
πειραματικά ότι η ακριβής προσαρμογή της σχεδίασης της συστοιχίας για το μέγεθος της εικόνας εισόδου δεν επιφέρει σημαντική βελτίωση των αποτελεσμάτων και για τούτο χρησιμοποιήσαμε την καθιερωμένη επιλογή της τοποθέτησης των φίλτρων σε πέντε (5) κλίμακες και οκτώ (8) προσανατολισμούς στο πεδίο συχνοτήτων για εικόνες μεγέθους 256×256. Το κέρδος σε υπολογιστικό χρόνο χρήσης μιας συστοιχίας για τα frames όλων των δειγμάτων βίντεο μιας μεγάλης βάσης δεδομένων είναι σημαντικό όταν πρόκειται για την εφαρμογή της ανάλυσης *EDCA* σε βίντεο, όπως στα πειράματα που θα παρουσιαστούν στο επόμενο κεφάλαιο. Επιπλέον, η χρήση των μιγαδικών φίλτρων υπαγόρευσε την αντικατάσταση του δισδιάστατου τελεστή Teager-Kaiser Φ από τον μιγαδικό ανάλογό του C της σχέσης (3.3.8) σε όλες τις σχέσεις υπολογισμού της ενέργειας ή των σημάτων αποδιαμόρφωσης. Για την αποδιαμόρφωση κάθε καναλιού έγινε χρήση του αλγορίθμου των Kokkinos et al. [42] *2D Gabor ESA* στη μορφή του για σήματα διακριτού χώρου. Για το σήμα διαμόρφωσης πλάτους χρησιμοποιήθηκε σε όλα τα πειράματα η αντιστάθμιση (3.3.11) και στη συνέχεια φίλτρο 5×5 median για την ομαλοποίηση του σήματος. Τέλος, στα σημεία ταυτόχρονου μηδενισμού των εξόδων του ενεργειακού τελεστή από τις παραγώγους του σήματος, το πλάτος τέθηκε στην τιμή μηδέν για την αποφυγή απροσδιόριστων και άπειρων τιμών στην εκτίμηση ενώ στα σημεία μηδενικής ενέργειας χρησιμοποιήθηκε παρεμβολή με φίλτρο 3×3 median για τα σήματα συχνότητας στις αντίστοιχες τιμές.

Αξίζει να παρατηρήσουμε την αποτελεσματική σύλληψη των κυρίαρχων διαμορφώσεων υφής με ταυτόχρονη διατήρηση των σημαντικότερων ακμών της εικόνας στην έξοδο του ενεργειακού τελεστή στα κυρίαρχα κανάλια από το σχήμα *EDCA*. Αυτή η συμπαγής χαμηλής διάστασης αναπαράσταση θα αποτελέσει ερέθισμα για τα πειράματα που παρουσιάζονται στο επόμενο κεφάλαιο.

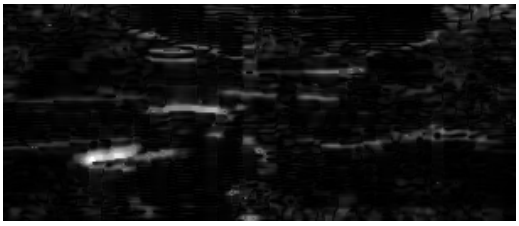
(a) Αρχική Εικόνα



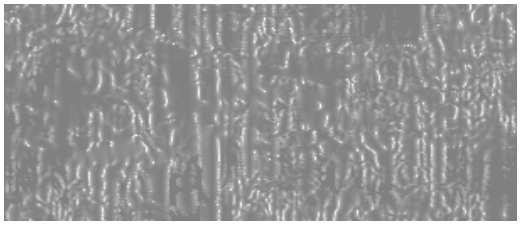
(b) Στενοζωνική Φιλτραρισμένη Εικόνα



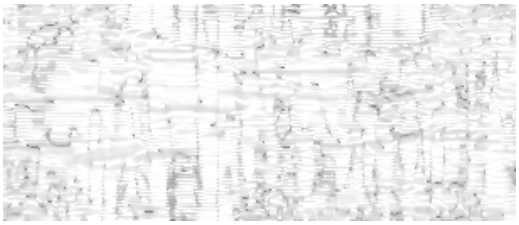
(c) Teager-Kaiser Ενέργεια Στενοζωνικής Εικόνας



(d) Σήμα Διαμόρφωσης Πλάτους

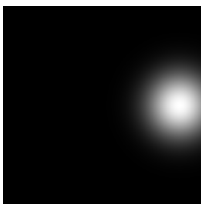


(e) Μεταβαλλόμενη Κάθετη Συχνότητα

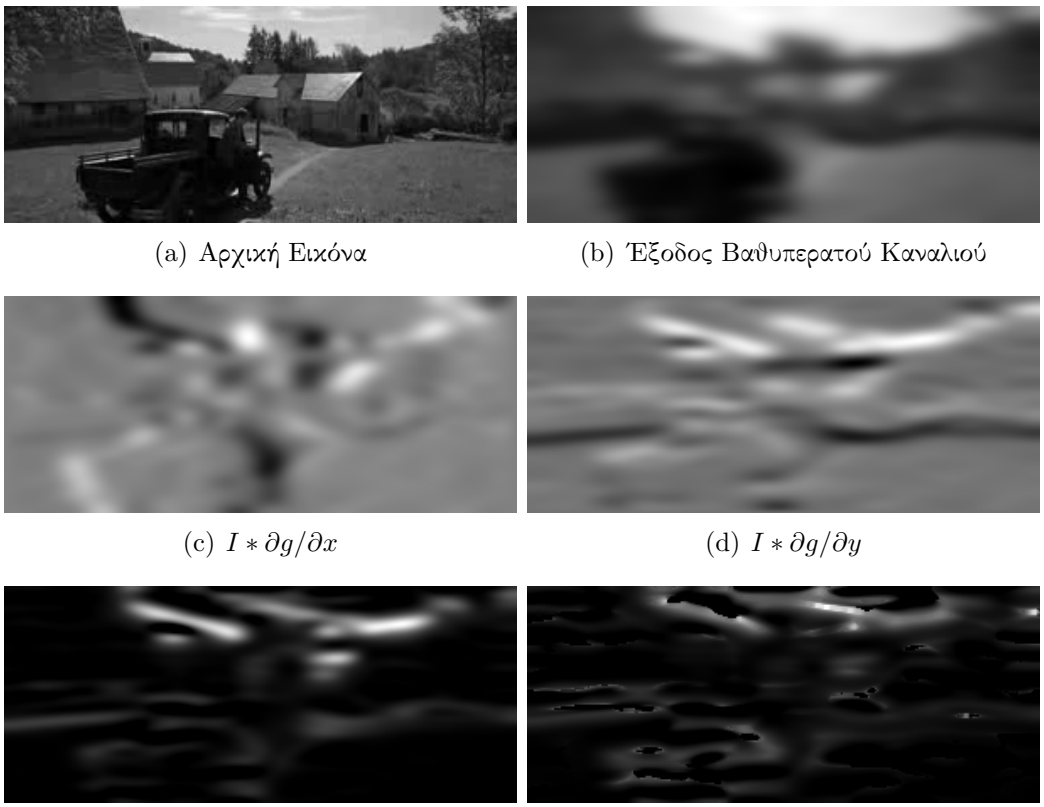


(f) Μεταβαλλόμενη Οριζόντια Συχνότητα

Σχήμα 3.3: Αποτελέσματα αποδιαμόρφωσης στην έξοδο του ζωνοπερατού καναλιού 25 της συστοιχίας του Σχήματος 3.2 ύστερα από *Ανάλυση Συνιστωσών Καναλιού CCA* σε frame δείγματος βίντεο της Βάσης Δεδομένων *Hollywood2 Actions Dataset*. Το φίλτρο 25 της συστοιχίας έχει κάθετη κεντρική συχνότητα 0 και οριζόντια κεντρική συχνότητα 100.78 κύκλοι ανά εικόνα (Σχήμα 3.4). (a) Αρχικό έγχρωμο frame, (b) Στενοζωνική φιλτραρισμένη έξοδος της αντίστοιχης γκρίζας εικόνας από το κανάλι 25, (c) Ενέργεια της στενοζωνικής συνιστώσας από την έξοδο του τελεστή Teager-Kaiser, (d) Σήμα διαμόρφωσης πλάτους της στενοζωνικής συνιστώσας εκτιμημένο από τον *2D Gabor ESA* και ομαλοποιημένο με φίλτρο median 5×5, (e) Κάθετη συνιστώσα ω_1 του διανύσματος συχνοτήτων της στενοζωνικής εικόνας εκτιμημένη από τον *2D Gabor ESA*, (f) Οριζόντια συνιστώσα ω_2 του διανύσματος συχνοτήτων εκτιμημένη από τον *2D Gabor ESA*.



Σχήμα 3.4: Απόκριση συχνότητας (αριστερά) και κρουστική απόκριση (δεξιά) του φίλτρου 25 της συστοιχίας του Σχήματος 3.2, 5ης κλίμακας και 5ου προσανατολισμού.



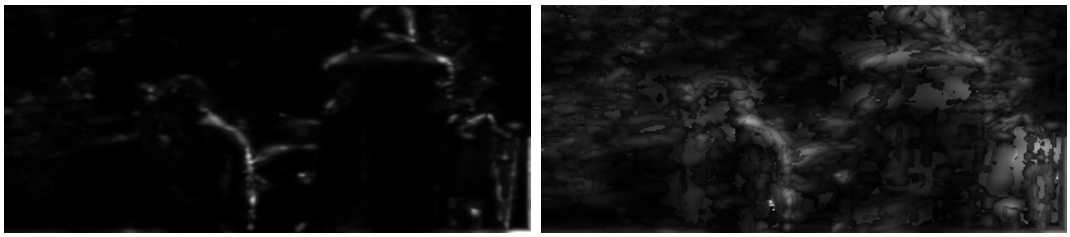
(a) Αρχική Εικόνα (b) Έξοδος Βαθυπερατού Καναλιού

(c) $I * \partial g/\partial x$ (d) $I * \partial g/\partial y$

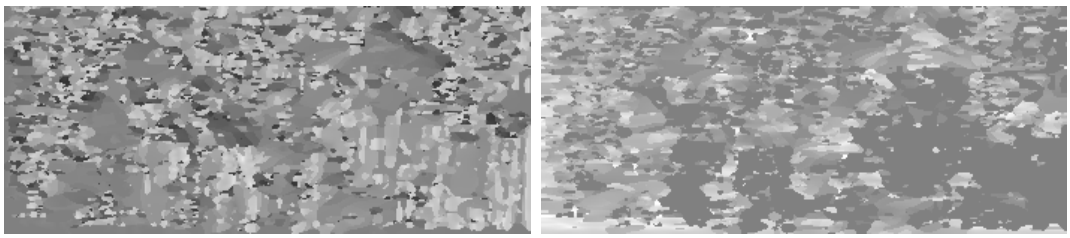
(e) Teager-Kaiser Ενέργεια (f) Σήμα Διαμόρφωσης Πλάτους

Σχήμα 3.5: Αποτελέσματα φιλτραρίσματος, ενέργεια Teager-Kaiser και σήμα διαμόρφωσης πλάτους από την έξοδο του ειδικού βαθυπερατού καναλιού της συστοιχίας του Σχήματος 3.2 ύστερα από *Ανάλυση Συνιστωσών Καναλιού CCA* στην αντίστοιχη γκρίζα εικόνα του ίδιου αρχικού έγχρωμου frame του Σχήματος 3.3. Το βαθυπερατό φίλτρο της επαυξημένης συστοιχίας των 43 φίλτρων έχει διάνυσμα κεντρικής συχνότητας $(\Omega_1, \Omega_2) = (0, 0)$.
(a) Αρχική γκρίζα εικόνα, (b) Φιλτραρισμένη έξοδος της γκρίζας εικόνας από το βαθυπερατό κανάλι, (c) Έξοδος φιλτραρίσματος με την παράγωγο του βαθυπερατού φίλτρου ως προς την κάθετη κατεύθυνση, (d) Έξοδος φιλτραρίσματος με την παράγωγο του βαθυπερατού φίλτρου ως προς την οριζόντια κατεύθυνση, (e) Ενέργεια της βαθυπερατής συνιστώσας από την έξοδο του τελεστή Teager-Kaiser, (f) Σήμα διαμόρφωσης πλάτους της βαθυπερατής συνιστώσας εκτιμημένο από τον *2D Gabor ESA* και ομαλοποιημένο με φίλτρο median 5×5.

(a) Αρχική Εικόνα (b) EDCA Σύνθεση της Γκρίζας Εικόνας

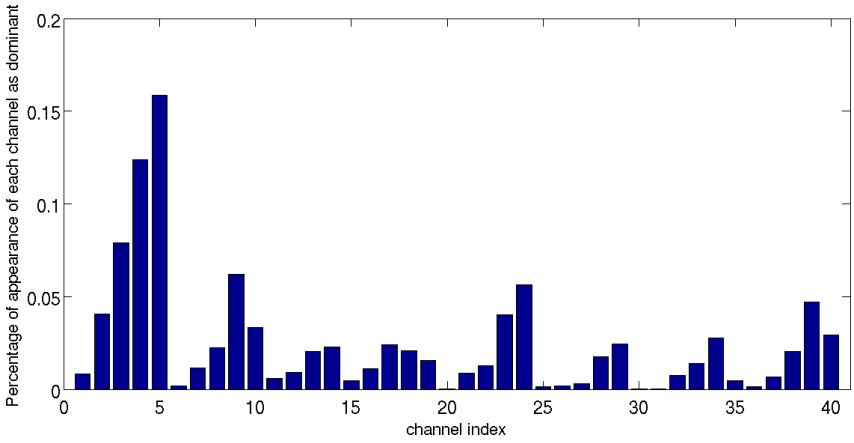


(c) EDCA Ενέργεια (d) EDCA Πλάτος

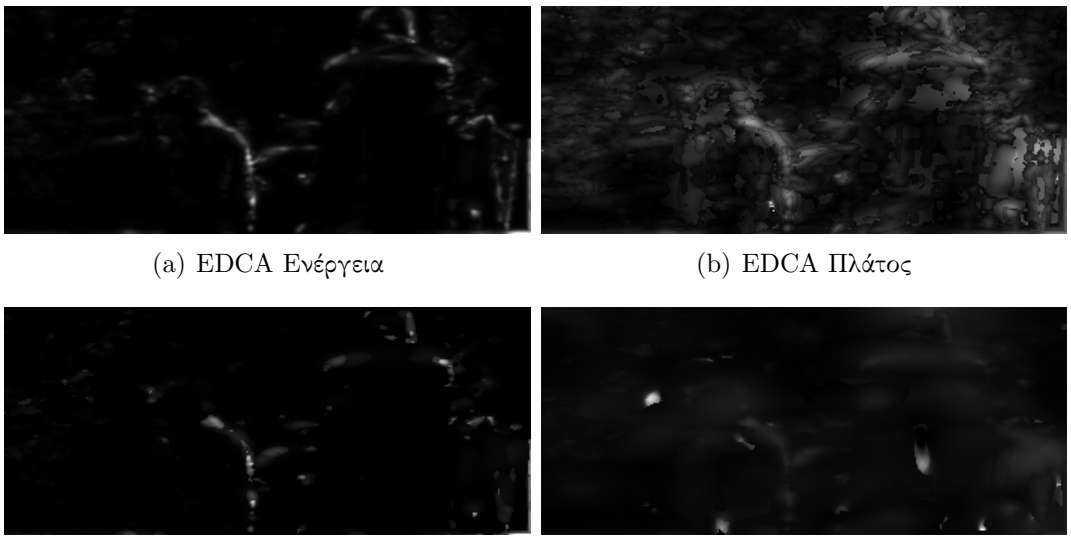


(e) EDCA Κάθετη Συχνότητα (f) EDCA Οριζόντια Συχνότητα

Σχήμα 3.6: Κυρίαρχα χαρακτηριστικά αποδιαμόρφωσης ύστερα από *Ανάλυση Κυρίαρχων Συνιστωσών βασισμένη στην Ενέργεια (EDCA)* στην αντίστοιχη γκρίζα εικόνα έγχρωμου frame δείγματος βίντεο της Βάσης Δεδομένων *Hollywood2 Actions Dataset* με χρήση της συστοιχίας του Σχήματος 3.1.
(a) Αρχική έγχρωμη εικόνα, (b) Σύνθεση της αντίστοιχης γκρίζας εικόνας από τις φιλτραρισμένες εξόδους των κυρίαρχων καναλιών σε κάθε pixel, (c) EDCA ενέργεια από τις εξόδους του τελεστή Teager-Kaiser στα κυρίαρχα κανάλια, (d) EDCA πλάτος εκτιμημένο από τον *2D Gabor ESA* και ομαλοποιημένο με φίλτρο median 5×5, (e) Κάθετη συνιστώσα της EDCA συχνότητας εκτιμημένη από τον *2D Gabor ESA*, (f) Οριζόντια συνιστώσα της EDCA συχνότητας εκτιμημένη από τον *2D Gabor ESA*.



Σχήμα 3.7: Συνολικό ποσοστό εμφάνισης κάθε καναλιού της συστοιχίας του Σχήματος 3.1 ως dominant στην Ανάλυση EDCA της εικόνας του Σχήματος 3.6.



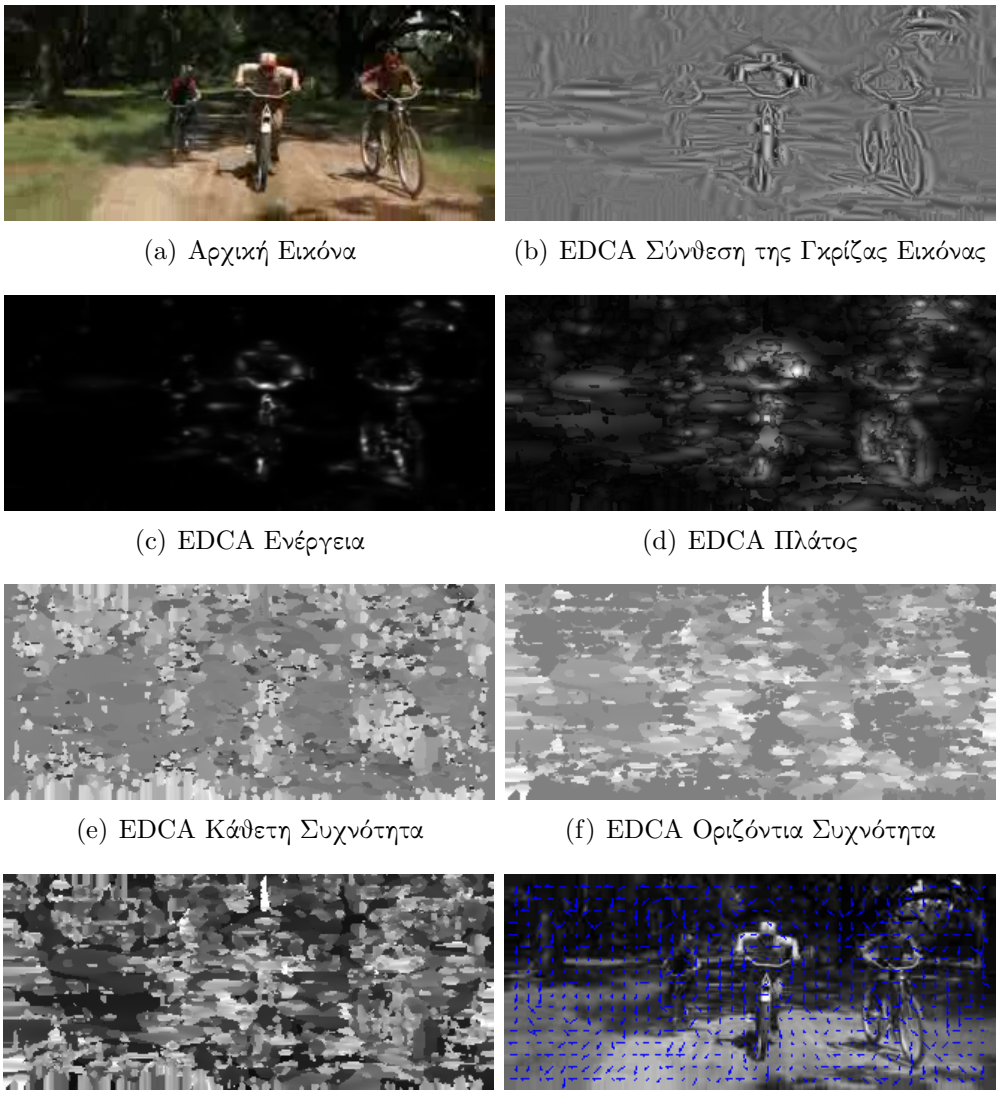
(a) EDCA Ενέργεια

(b) EDCA Πλάτος

(c) ADCA Ενέργεια

(d) ADCA Πλάτος

Σχήμα 3.8: Συγκριτική απεικόνιση των αποτελεσμάτων για την κυρίαρχη ενέργεια (αριστερά) και το κυρίαρχο πλάτος (δεξιά) από τα δύο διαφορετικά σχήματα *Ανάλυσης Κυρίαρχων Συνιστωσών* EDCA (επάνω) και ADCA (κάτω) στην αντίστοιχη γκρίζα εικόνα του ίδιου αρχικού έγχρωμου frame του Σχήματος 3.6. Για την ενέργεια παρατηρούμε την ελαφρώς καλύτερη διατήρηση των ακμών της εικόνας από το σχήμα EDCA στα περιγράμματα των ανθρώπων ενώ για το πλάτος την καλύτερη ανάδειξη των περιοχών έντονου περιεχομένου υφής στο παρασκήνιο πάλι από το σχήμα EDCA.



(a) Αρχική Εικόνα (b) EDCA Σύνθεση της Γκρίζας Εικόνας

(c) EDCA Ενέργεια (d) EDCA Πλάτος

(e) EDCA Κάθετη Συχνότητα (f) EDCA Οριζόντια Συχνότητα

(g) Μέτρο EDCA Διανύσματος Συχνοτήτων (h) Διανύσματα EDCA Μεταβαλλόμενης Συχνότητας

Σχήμα 3.9: Κυρίαρχα χαρακτηριστικά αποδιαμόρφωσης ύστερα από *Ανάλυση Κυρίαρχων Συνιστωσών βασισμένη στην Ενέργεια (EDCA)* στην αντίστοιχη γκρίζα εικόνα έγχρωμου frame δείγματος βίντεο της Βάσης Δεδομένων *Hollywood2 Actions Dataset* με χρήση της συστοιχίας του Σχήματος 3.1.
(a) Αρχική έγχρωμη εικόνα, (b) Σύνθεση της αντίστοιχης γκρίζας εικόνας από τις φιλτραρισμένες εξόδους των κυρίαρχων καναλιών σε κάθε pixel, (c) EDCA ενέργεια από τις εξόδους του τελεστή Teager-Kaiser στα κυρίαρχα κανάλια, (d) EDCA πλάτος εκτιμημένο από τον *2D Gabor ESA* και ομαλοποιημένο με φίλτρο median 5×5, (e) Κάθετη συνιστώσα της EDCA συχνότητας εκτιμημένη από τον *2D Gabor ESA*, (f) Οριζόντια συνιστώσα της EDCA συχνότητας εκτιμημένη από τον *2D Gabor ESA*, (g) Μέτρο του EDCA μεταβαλλόμενου διανύσματος συχνοτήτων $\boldsymbol{\omega}_{EDCA}(x, y)$, (h) Απεικόνιση των EDCA διανυσμάτων συχνότητας επί της αντίστοιχης γκρίζας εικόνας, με δειγματοληψία ανά 8 δείγματα σε κάθε κατεύθυνση και με κλίμακα 0.8 επί του μέτρου τους.

Κεφάλαιο 4

Αναγνώριση Ανθρώπινων Δράσεων σε Βίντεο με Ενσωμάτωση Χαρακτηριστικών Κυρίαρχης Ενέργειας των Εικόνων

Στο Κεφάλαιο 2 μελετήσαμε αναλυτικά την προσέγγιση τοπικών χαρακτηριστικών για τις ακολουθίες εικόνων, που προκύπτουν από την ανίχνευση χωροχρονικών σημείων ενδιαφέροντος και την μετέπειτα περιγραφή τους μέσω χαρακτηριστικών εμφάνισης και κίνησης στη χωροχρονική γειτονιά τους. Στο παρόν κεφάλαιο θα καταστεί σαφές με ποιό τρόπο αυτή η ακανόνιστη συλλογή τοπικών τεμαχίων (local patches) οδηγεί σε μια τελική αναπαράσταση του δείγματος βίντεο για να προωθηθεί ακολούθως στη γνωστή μέθοδο ταξινόμησης των Support Vector Machines (SVM) με σκοπό την αναγνώριση της υπάρχουσας σε αυτό ανθρώπινης δράσης. Θα παρουσιάσουμε πειράματα αναγνώρισης δράσεων στη γνωστή Βάση Δεδομένων *Hollywood2 Actions Dataset* χρησιμοποιώντας ένα επιτυχημένο σχήμα συνδυασμού Ανιχνευτή-Περιγραφέα και θα δείξουμε ότι η συνέργειά του με χαρακτηριστικά *Ανάλυσης Κυρίαρχων Συνιστωσών (Dominant Component Analysis)* (βλ. 3.3.4.2) στα επιμέρους frames του βίντεο οδηγεί σε βελτιωμένες επιδόσεις στο εν λόγω πρόβλημα.

4.1 Αναπαράσταση Βίντεο με την Προσέγγιση Bag-Of-Features

Η πρόσφατη επιτυχία των τοπικών χαρακτηριστικών (local features) σε προβλήματα της Όρασης Υπολογιστών όπως η αναγνώριση αντικειμένων και η ταξινόμηση υφής κατέστησε αναγκαία την εύρεση μιας απλής, χαμηλής υπολογιστικής πολυπλοκότητας και αποτελεσματικής αναπαράστασης των εικόνων μέσω της κατηγοριοποίησης του πλήθους των χαρακτηριστικών, ώστε η εκμάθηση των διαφορετικών κλάσεων να είναι εύρωστη

απέναντι σε προβλήματα όπως η μεγάλη μεταβλητότητα στην εμφάνιση, τις συνθήκες φωτισμού, την ύπαρξη οπτικών εμποδίων και η εντός κλάσης οπτική διαφοροποίηση αντικειμένων και προτύπων υφής. Οι Willamowski et al. [44] προσέγγισαν το πρόβλημα της *οπτικής κατηγοριοποίησης (visual categorization)* με την τεχνική *bag-of-keypoints* βάσει της οποίας μια εικόνα αναπαρίσταται από ένα ιστόγραμμα του αριθμού εμφανίσεων οπτικών προτύπων σε αυτή. Η ιδέα τους αυτή αντλεί έμπνευση από προγενέστερες ανάλογες τεχνικές *bag-of-words* που εφαρμόστηκαν με επιτυχία στην κατηγοριοποίηση κειμένου. Πρόκειται, επομένως, για μια αναπαράσταση που ομαδοποιεί τους περιγραφείς τοπικών τεμαχίων που έχουν εξαχθεί για την εικόνα σε ένα διακριτό προκαθορισμένο σύνολο "οπτικών λέξεων"("visual words") χωρίς να διατηρεί καμία πληροφορία για τη γεωμετρική δομή των θέσεων των χαρακτηριστικών. Η στατιστική κατανομή των περιγραφέων της εικόνας στις διάφορες οπτικές κατηγορίες αποτελεί το τελικό διάνυσμα χαρακτηριστικών.
Όπως αναφέρουν οι συγγραφείς, η δημιουργία του διανύσματος χαρακτηριστικών μιας εικόνας εξέτασης, για το πρόβλημα της αναγνώρισης αντικειμένων, θα μπορούσε στην ακραία μορφή της να βασιστεί στη σύγκριση καθενός από τους περιγραφείς της εικόνας με όλους τους περιγραφείς που έχουν εξαχθεί από το σετ εκπαίδευσης (training set). Κάτι τέτοιο θα οδηγούσε σε τεράστιο όγκο συγκρίσεων όταν το training set αντιστοιχεί σε μια μεγάλη βάση δεδομένων με ρεαλιστικές φυσικές εικόνες και είναι προφανώς υπολογιστικά απαιτητικό και ασύμφορο. Από την άλλη, ούτε το υπερβολικά μικρό πλήθος μεγάλων οπτικών κλάσεων για κάθε μία κατηγορία αντικειμένου έχει αποδειχθεί αποτελεσματικό. Μια πρακτική λύση στο παραπάνω πρόβλημα, που επιτελεί ένα συμβιβασμό μεταξύ υπολογιστικής πολυπλοκότητας και ακρίβειας στην αναγνώριση, είναι η κατασκευή ενός *οπτικού λεξιλογίου (visual vocabulary)* μεσαίου μεγέθους χρησιμοποιώντας κάποιον γνωστό αλγόριθμο συσταδοποίησης (clustering) ή κβαντοποίησης διανυσμάτων (vector quantization) σε ένα υποσύνολο των περιγραφέων του training set. Η προτεινόμενη επιλογή είναι ο επαναληπτικός αλγόριθμος k-means που βασίζεται σε μια τετραγωνικού-σφάλματος διαμέριση των δεδομένων, απαιτώντας βέβαια τον εκ προοιμίου καθορισμό του αριθμού των clusters. Μετά την κατασκευή του λεξιλογίου, κάθε περιγραφέας αποδίδεται στην πλησιέστερη κλάση, με βάση τις τιμές κάποιου μετρικού απόστασης από τα κέντρα των κλάσεων, και κάθε εικόνα εξέτασης αντιστοιχεί στο ιστόγραμμα της συχνότητας εμφάνισης των "οπτικών λέξεων"του λεξιλογίου σε αυτή, όπως περιγράψαμε παραπάνω.
Αυτή η στατιστική και απαλλαγμένη από τη γεωμετρία αναπαράσταση των τοπικών τεμαχίων, που συναντάται με τον γενικό όρο *bag-of-features (BOF)*, υιοθετήθηκε πρόσφατα με επιτυχία καί στο πρόβλημα της Αναγνώρισης Ανθρώπινων Δράσεων σε Βίντεο ([13],[24],[21],[30]). Αξίζει να σημειώσουμε ότι ένα χαμηλό επίπεδο γεωμετρικής πληροφορίας διατηρείται χάρη στη διαμέριση του 3D τοπικού τεμαχίου του περιγραφέα σε μια δομή πλέγματος (όπως, για παράδειγμα, η $3 \times 3 \times 2$ διαμέριση που χρησιμοποιείται από τον Περιγραφέα HOG/HOF). Αντίστοιχα με την περίπτωση των εικόνων, μια ακολουθία εικόνων εδώ αναπαρίσταται από μια συλλογή (έναν "σάκο") τοπικών χωροχρονικών χαρακτηριστικών που περιγράφονται με κάποιον από τους τοπικούς περιγραφείς που αναλύσαμε στην ενότητα 2.3. Αρχικά, κατασκευάζεται το *οπτικό λεξιλόγιο (visual vocabulary)* χρησιμοποιώντας τον αλγόριθμο k-means. Οι οπτικές πρωτότυπες

λέξεις που προκύπτουν μπορούν διαισθητικά να ιδωθούν ως το τρισδιάστατο ανάλογο των επιμέρους τμημάτων των αντικειμένων στην προσέγγιση BOF για την αναγνώριση αντικειμένων. Η τιμή $V = 4000$ για το πλήθος των clusters έχει εμπειρικά αποδειχθεί βέλτιστη για πολλές βάσεις δεδομένων σύμφωνα με το [3] και για τούτο επιλέγεται καί στα πειράματα της παρούσας διπλωματικής εργασίας. Ακολουθώντας τη σχετική εργασία των Laptev et al. [24], ένα υποσύνολο 100000 τυχαία επιλεγμένων χαρακτηριστικών από το training set χρησιμοποιείται για τη συσταδοποίηση. Η εκτέλεση του k-means για διαφορετικές αρχικοποιήσεις και η διατήρηση του αποτελέσματος με το χαμηλότερο σφάλμα είναι επιθυμητή για την επίτευξη μεγαλύτερης ακρίβειας. Η κάθε λέξη του προκύπτοντος λεξιλογίου έχει διάσταση ίση με το πλήθος στοιχείων του διανύσματος χαρακτηριστικών του τοπικου περιγραφέα που χρησιμοποιείται. Ακολουθεί η "ετικετοποίηση"(labeling) των περιγραφέων, δηλαδή η αντιστοίχισή τους στην πλησιέστερη οπτική λέξη με κριτήριο την Ευκλείδεια απόσταση. Παρόμοια, η τελική αναπαράσταση του βίντεο είναι το ιστόγραμμα της συχνότητας εμφάνισης των οπτικών λέξεων του λεξιλογίου (frequency histogram of visual word occurences), το οποίο κανονικοποιείται ως προς την $\mathcal{L}_1$ ή την $\mathcal{L}_2$ νόρμα.
Η αναπαράσταση που παρέχει η προσέγγιση Bag-Of-Features είναι απλή στην υλοποίηση και αποτελεσματική ως προς την ικανότητα διάκρισης διαφορετικών πρωτότυπων οπτικών γεγονότων σε ακολουθίες εικόνων. Καταλήγει σε ίσου μεγέθους αναπαραστάσεις για κάθε βίντεο και, όπως θα δούμε παρακάτω, συνδυάζεται με επιτυχία με την SVM τεχνική ταξινόμησης.

4.2 Ταξινόμηση με Support Vector Machines

Οι *Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines)* είναι ταξινομητές μηχανικής μάθησης που επιτυγχάνουν μεγάλα περιθώρια διαχωρισμού και έχουν πρόσφατη επιτυχία σε προβλήματα αναγνώρισης οπτικών προτύπων [45]. Είναι ακόμα η πιο συνήθης τεχνική ταξινόμησης που χρησιμοποιείται σε συνδυασμό με την προσέγγιση *Bag-of-Features* για την Αναγνώριση Ανθρώπινων Δράσεων σε Βίντεο. Θα δώσουμε μια γενική ιδέα της λειτουργίας τους μελετώντας αρχικά την περίπτωση γραμμικά διαχωρίσιμων προτύπων.
Ας θεωρήσουμε το πρόβλημα του διαχωρισμού ενός συνόλου δεδομένων εκπαίδευσης (training data) $(\mathbf{x_1}, d_1), (\mathbf{x_2}, d_2), \ldots, (\mathbf{x_m}, d_m)$ όπου $\mathbf{x}_i \in \mathcal{R}^N$ είναι ένα διάνυσμα χαρακτηριστικών και $d_i \in \{-1, +1\}$ η ετικέτα (label) της κλάσης στην οποία ανήκει. Θεωρώντας ότι τα δεδομένα, για τα οποία δεν έχουμε γνώση της κατανομής τους, μπορούν να διαχωριστούν από ένα υπερεπίπεδο με επιφάνεια απόφασης $\mathbf{w} \cdot \mathbf{x} + b = 0$, τότε το βέλτιστο υπερεπίπεδο είναι αυτό που μεγιστοποιεί το περιθώριο διαχωρισμού μεταξύ των δεδομένων με θετική και αρνητική ετικέτα. Αποδεικνύεται ότι η εύρεση των βέλτιστων $\mathbf{w}$ και b είναι λύση του προβλήματος ελαχιστοποίησης με περιορισμούς

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{υπό τους περιορισμούς} \quad & d_i(\mathbf{w} \cdot \mathbf{x_i} + b) \geq 1, \quad \forall i = 1, \ldots, m \end{aligned} \tag{4.2.1}$$

Το παραπάνω πρόβλημα επιλύεται με τη χρήση πολλαπλασιαστών Lagrange $\alpha_i (i = 1, \ldots, m)$ και έτσι προκύπτει η ακόλουθη συνάρτηση απόφασης

$$f(x) = \text{sgn}\left(\sum_{i=1}^{m} \alpha_i d_i \mathbf{w} \cdot \mathbf{x} + b\right) \tag{4.2.2}$$

Τα διανύσματα χαρακτηριστικών $\mathbf{x}_i$ που αντιστοιχούν σε μη μηδενικούς πολλαπλασιαστές Lagrange α_i ονομάζονται *διανύσματα υποστήριξης (support vectors)* και εννοιολογικά είναι αυτά που βρίσκονται πλησιέστερα στην επιφάνεια διαχωρισμού και επομένως διαδραματίζουν σπουδαιότερο ρόλο στον καθορισμό της.
Στην περίπτωση μη γραμμικά διαχωρίσιμων προτύπων, η μόνη διαφορά συνίσταται στο ότι ο περιορισμός $\alpha_i \geq 0$ για τους πολλαπλασιαστές Lagrange αντικαθίσταται από τον $0 \leq \alpha_i \leq C$, όπου C είναι μια θετική παράμετρος που ορίζει ο χρήστης και εκφράζει την ποινή για τις λανθασμένες ταξινομήσεις. Συνίσταται μεγάλη τιμή της σταθεράς C σε περιπτώσεις όπου το δείγμα εκπαίδευσης θεωρείται ποιοτικό και μη θορυβώδες σύμφωνα με τον S. Haykin [46].
Μια μη γραμμική μηχανή SVM μπορεί να κατασκευαστεί με μια αντιστοίχιση του διανύσματος εισόδου σε ένα χώρο χαρακτηριστικών $\mathcal{H}$ μεγαλύτερης διάστασης $\mathbf{x} \to \Phi(\mathbf{x})$, που δεν είναι ορατός από την είσοδο και την έξοδο, στον οποίο τα δεδομένα είναι γραμμικά διαχωρίσιμα. Με την εύρεση μιας συνάρτησης πυρήνα K για την οποία ισχύει $K(\mathbf{x}, \mathbf{x}_i) = \Phi(\mathbf{x}) \cdot \Phi(\mathbf{x}_i)$, η αντικατάσταση του εσωτερικού γινομένου της σχέσης 4.2.2 από την τιμή της συνάρτησης πυρήνα δίνει τη συνάρτηση απόφασης του μη γραμμικού SVM

$$f(x) = \text{sgn}\left(\sum_{i=1}^{m} \alpha_i d_i K(\mathbf{x}_i, \mathbf{x}) + b\right) \tag{4.2.3}$$

Με τον τρόπο αυτό κατασκευάζεται ένα διαχωριστικό υπερεπίπεδο στον χώρο των χαρακτηριστικών $\mathcal{H}$. Η παραπάνω τεχνική είναι γνωστή ως το "τέχνασμα του πυρήνα"αφού με αυτόν τον τρόπο αποφεύγεται ο ρητός υπολογισμός του διανύσματος βαρών $\mathbf{w}$. Για μια συνάρτηση πυρήνα K, ο πίνακας $\mathbf{K} = \{K(\mathbf{x_i}, \mathbf{x_j})\}_{(i,j)}^{m}$ είναι συμμετρικός θετικά ημιορισμένος και ονομάζεται *μήτρα Gram*.
Υπάρχουν πολλά είδη συναρτήσεων πυρήνα όπως η γραμμική, η πολυωνυμική ή η ακτινικής βάσης συνάρτηση πυρήνα (Radial Basis Function (RBF)). Η επιλογή του κατάλληλου πυρήνα εξαρτάται από το είδος των δεδομένων στα οποία πρόκειται να εφαρμόσουμε ταξινόμηση με μη γραμμικό SVM και το είδος της εφαρμογής. Σύμφωνα με τη σχετική βιβλιογραφία για το πρόβλημα της Αναγνώρισης Ανθρώπινων Δράσεων σε Βίντεο, όταν τα διανύσματα χαρακτηριστικών είναι ιστογράμματα, όπως στην προσέγγιση BOF που είδαμε παραπάνω, είναι αποτελεσματική η χρήση γκαουσιανών πυρήνων που εντάσσονται στην κατηγορία των RBF πυρήνων και έχουν ως συνάρτηση την

$$K(\mathbf{x}, \mathbf{y}) = \exp(-\frac{1}{A}\|\mathbf{x} - \mathbf{y}\|^2) \tag{4.2.4}$$

Έχοντας $H_i = \{h_{in}\}$ και $H_j = \{h_{jn}\}$ τα ιστογράμματα συχνότητας εμφάνισης οπτικών λέξεων και V το μέγεθος του οπτικού λεξιλογίου η παραπάνω σχέση γράφεται

$$K(H_i, H_j) = \exp\left(-\frac{1}{A}\sum_{n=1}^{V}(h_{in} - h_{jn})^2\right) \tag{4.2.5}$$

Πρόσφατα [47] χρησιμοποιήθηκαν, σε πειράματα ταξινόμησης υφής και κατηγοριών αντικειμένων σε συνδυασμό με την Bag-Of-Features προσέγγιση, γενικευμένοι γκαουσιανοί πυρήνες όπου η ευκλείδεια απόσταση στη συνάρτηση του πυρήνα αντικαθίσταται από την χ^2 απόσταση, με τη συνάρτηση να παίρνει τη μορφή

$$K(H_i, H_j) = \exp\left(-\frac{1}{2A}\sum_{n=1}^{V}\frac{(h_{in} - h_{jn})^2}{h_{in} + h_{jn}}\right) \tag{4.2.6}$$

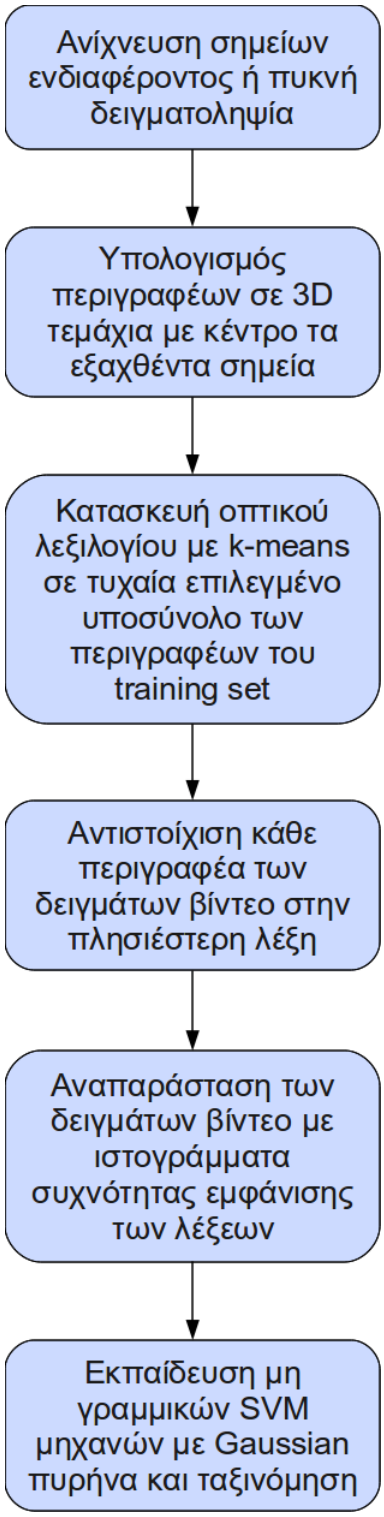
Αυτή την επιλογή για τη συνάρτηση πυρήνα υιοθετούν καί οι Laptev et al. στο [24]. Οι Zhang et al. [47] υποστήριξαν με πειράματα ότι η παράμετρος A μπορεί να τεθεί ως η μέση απόσταση (ευκλείδεια ή χ^2 αντίστοιχα) μεταξύ όλων των δειγμάτων εκπαίδευσης (training set), δίνοντας συγκρίσιμα αποτελέσματα με τη διαδικασία *διασταυρωμένης επικύρωσης (cross-validation)* και με πολύ μικρότερο υπολογιστικό κόστος.

Για την περίπτωση της ταξινόμησης των δεδομένων *σε πολλαπλές κλάσεις (multi-class classification)*, η απόφαση προκύπτει από τις επιμέρους αποφάσεις δυαδικών SVM ταξινομητών (binary SVM classifiers) με δύο διαφορετικούς τρόπους (έστω N ο αριθμός των κλάσεων):

- *One-against-one (Ένας-εναντίον-ενός)* : Εκπαιδεύεται ένας δυαδικός ταξινομητής για καθέναν από τους $\binom{N}{2}$ δυνατούς συνδυασμούς ανά δύο των κλάσεων και κάθε νέο δείγμα εξέτασης αποδίδεται στην κλάση που διαλέχθηκε από την πλειοψηφία των ταξινομητών.

- *One-against-rest (Ένας-εναντίον-των υπολοίπων)* : Εκπαιδεύονται N δυαδικοί ταξινομητές, ένας για κάθε κλάση, και κάθε νέο δείγμα εξέτασης αποδίδεται στην κλάση της οποίας η συνάρτηση απόφασης έδωσε τη μεγαλύτερη τιμή.

Οι Zhang et al. [47] αναφέρουν ότι στα πειράματά τους οι δύο διαφορετικές προσεγγίσεις παρουσίασαν σχεδόν πανομοιότυπα αποτελέσματα ταξινόμησης. Στα πειράματα ταξινόμησης σε πολλαπλές κλάσεις της παρούσας διπλωματικής εργασίας χρησιμοποιούμε την επιλογή *one-against-rest*.
Στο Σχήμα 4.1 απεικονίζονται τα επιμέρους στάδια που μεσολαβούν από την ανίχνευση χωροχρονικών σημείων ενδιαφέροντος ή την πυκνή δειγματοληψία σε ακολουθίες εικόνων μέσω της προσέγγισης BOF μέχρι να καταλήξουμε στην τελική αναπαράσταση που οδηγείται στους ταξινομητές SVM για την αναγνώριση των ανθρώπινων δράσεων.



Σχήμα 4.1: Σχηματικό διάγραμμα της αναπαράστασης Bag-Of-Features σε συνδυασμό με την ταξινόμηση SVM για το πρόβλημα της Αναγνώρισης Ανθρώπινων Δράσεων σε Βίντεο.

4.3 Πείραμα Αναγνώρισης Ανθρώπινων Δράσεων με Ανιχνευτή τον Harris3D και Περιγραφέα τον HOG/HOF

Σε αυτή την ενότητα θα παρουσιάσουμε το πείραμα αναγνώρισης ανθρώπινων δράσεων που διεξήχθη στα πλαίσια της παρούσας διπλωματικής εργασίας για ένα υποσύνολο δράσεων της Βάσης Δεδομένων *Hollywood2 Actions Dataset* χρησιμοποιώντας τον Ανιχνευτή Harris3D και τον Περιγραφέα HOG/HOF που περιγράφηκαν αναλυτικά στις υποενότητες 2.2.1 και 2.3.2 αντίστοιχα. Η αναπαράσταση των ακολουθιών εικόνων και η αναγνώριση έγιναν με την καθιερωμένη προσέγγιση Bag-Of-Features και SVM αντίστοιχα, τακτική που ακολουθήθηκε σε όλα τα σχετικά πειράματα με σκοπό την παραγωγή άμεσα συγκρίσιμων αποτελεσμάτων.

Ο Ανιχνευτής Harris3D αποτελεί έναν από τους πιο επιτυχημένους και αποτελεσματικούς ανιχνευτές αραιών χωροχρονικών σημείων ενδιαφέροντος και συνιστά σημείο αναφοράς όταν πρόκειται για τη συγκριτική απόδοση νέων προσεγγίσεων ανιχνευτών που εμφανίζονται στην βιβλιογραφία. Σε συνδυασμό με τον Περιγραφέα Κίνησης HOF, πάντοτε στο πλαίσιο της BOF προσέγγισης, εξασφαλίζει τη μεγαλύτερη επίδοση *μέσης ακρίβειας ταξινόμησης (mean classification accuracy)* στη Βάση Δεδομένων KTH Actions Dataset[1], της τάξης του 92.1% [3]. Εντούτοις, όπως έχουμε αναφέρει, η πυκνή δειγματοληψία σε συνδυασμό με τον Περιγραφέα HOG/HOF παρέχει τα καλύτερα αποτελέσματα στην *Hollywood2* από όλους τους "αραιούς" ανιχνευτές.

Ο Περιγραφέας HOG/HOF, σύμφωνα με τα πειραματικά αποτελέσματα των Wang et al. στο [3], υπερτερεί σε επιδόσεις στη Βάση Δεδομένων *Hollywood2*, σε συνδυασμό είτε με πυκνή δειγματοληψία είτε με τους γνωστούς ανιχνευτές, ενώ εξασφαλίζει τη δεύτερη καλύτερη μέση ακρίβεια μετά τον Περιγραφέα HOG3D για τη Βάση *UCF Sport Actions Dataset*. Η αποτελεσματικότητά του μπορεί να αποδωθεί στην ικανότητά του να συλλαμβάνει συνδυασμένη τοπική πληροφορία εμφάνισης και κίνησης οδηγώντας έτσι σε καλή απόδοση όταν πρόκειται για ρεαλιστικά σκηνικά ακολουθιών εικόνων.

Τα παραπάνω μας οδήγησαν στην επιλογή του Ανιχνευτή Harris3D και του Περιγραφέα HOG/HOF για την πραγματοποίηση ενός μαζικού πειράματος αναγνώρισης σε ένα υποσύνολο δράσεων της απαιτητικής βάσης *Hollywood2*. Ο κυριότερος λόγος που υπαγόρευσε την εκτέλεση του εν λόγω πειράματος ήταν η παραγωγή επιδόσεων αναγνώρισης από ένα αποτελεσματικό ζεύγος ανιχνευτή-περιγραφέα προκειμένου για την αντικειμενική σύγκριση με τις αντίστοιχες επιδόσεις των εναλλακτικών σχημάτων που θα προτείνουμε στη συνέχεια.

Η *Hollywood2 Actions Dataset*[2] αποτελεί μια συλλογή δειγμάτων βίντεο με ανθρώπινες δράσεις, που αποτελούν αποσπάσματα από ταινίες του Hollywood. Οι δράσεις εκτυλίσσονται σε φυσικές ρεαλιστικές σκηνές και υπάρχει μεγάλη μεταβλητότητα στα δείγματα που ανήκουν στην ίδια κατηγορία δράσης ως προς το στήσιμο του σκηνικού, τις κινήσεις της κάμερας, τις γωνίες λήψης, την ταχύτητα εναλλαγής των σκηνών, τον φωτισμό, την ύπαρξη οπτικών εμποδίων και θορύβου ή την εμφάνιση των εικονιζόμενων προσώπων.

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

[2] http://www.irisa.fr/vista/Equipe/People/Laptev/download.html

Αναμφίβολα πρόκειται για την πιο απαιτητική Βάση Δεδομένων για το πρόβλημα της αναγνώρισης ανθρώπινων δράσεων και για αυτό προτιμήθηκε στα πειράματα της παρούσας διπλωματικής εργασίας. Παρακάτω δίνουμε συνοπτικά τις βασικές πληροφορίες που αφορούν την συγκεκριμένη Βάση.

Η Βάση Δεδομένων *Hollywood2 Actions Dataset* Αποτελεί επέκταση της παλαιότερης Βάσης *Hollywood Human Actions Dataset*. Στη Βάση Δεδομένων περιέχονται δείγματα βίντεο από 69 διαφορετικές χολυγουντιανές ταινίες. Υπάρχουν δώδεκα ανθρώπινες δράσεις: AnswerPhone, DriveCar, Eat, FightPerson, GetOutCar, HandShake, HugPerson, Kiss, Run, SitDown, SitUp, StandUp. Εκτός από το *Σετ Εκπαίδευσης (Training Set)* και το *Σετ Εξέτασης (Test Set)*, οι δημιουργοί της παρέχουν και ένα αυτόματα "ετικετοποιημένο"και θορυβώδες σετ εκπαίδευσης. Συνολικά η Βάση περιλαμβάνει 1707 δείγματα βίντεο, 823 εκ των οποίων αποτελούν το Training Set και 884 το Test Set, με τα βίντεο των δύο σετ να αντλούνται από διαφορετικές ταινίες. Τα πρώτα αποτελέσματα αναγνώρισης δράσεων σε αυτή τη Βάση αναφέρονται από τους Marszalek et al. στο [25]. Οι συγγραφείς προτείνουν ως μετρικό αναγνώρισης τον υπολογισμό της *μέσης ακρίβειας (average precision)* για κάθε κατηγορία δράσης ξεχωριστά και την αναφορά της μέσης τιμής της σε όλες τις δράσεις (*mean average precision*). Το μέγεθος των frames των δειγμάτων ποικίλει με μια μέση τιμή πλάτους περί τα 600-800 pixels και ύψους περί τα 400 pixels. Είναι ακόμα διαθέσιμα σε αρχεία κειμένου τα όρια των λήψεων (shot boundaries), ώστε να υπάρχει η επιλογή της απόρριψης τεχνητών επισφαλών ανιχνεύσεων σε εκείνα τα frames. Τέλος, τα δείγματα βίντεο της Βάσης ανήκουν σε 10 διαφορετικές κατηγορίες σκηνών και για αυτό το λόγο η πλήρης ονομασία της είναι *Hollywood2 Human Actions and Scenes Dataset* (σχετικά πειράματα αναγνώρισης σκηνών μπορούν να βρεθούν στο [25]).

Λόγω του μεγάλου όγκου σε δείγματα βίντεο της Βάσης Δεδομένων, επιλέξαμε να πραγματοποιήσουμε το πείραμα της αναγνώρισης για τις έξι από τις δώδεκα δράσεις, συμπεριλαμβάνοντας για αυτές όλα τα αντίστοιχα βίντεο του Training και Test Set. Οι δράσεις που επιλέχθηκαν είναι οι: AnswerPhone, DriveCar, FightPerson, GetOutCar, HandShake και SitUp. Αξίζει να σημειώσουμε ότι οι δράσεις AnswerPhone, GetOutCar, HandShake και SitUp, σύμφωνα με τα αναφερόμενα μέχρι σήμερα αποτελέσματα στη σχετική βιβλιογραφία, αποτελούν παραδοσιακά απαιτητικές δράσεις της *Hollywood2* ως προς την αναγνώρισή τους στο πλαίσιο BOF. Με τη δεδομένη επιλογή δράσεων, καταλήξαμε σε ένα Training Set μεγέθους 310 δειγμάτων βίντεο και ένα Test Set μεγέθους 371 δειγμάτων βίντεο. Στα παραπάνω αθροίσματα τα δείγματα βίντεο που βρέθηκαν να ανήκουν σε περισσότερες από μία δράσεις υπολογίζονται μία φορά (είχαμε δύο τέτοιες περιπτώσεις στο Training Set και τέσσερεις στο Test Set). Ο αριθμός δειγμάτων ανά δράση στα προκύπτοντα Training και Test Set αναφέρεται στον Πίνακα 4.1. Για το εν λόγω πείραμα όλα τα δείγματα βίντεο προωθήθηκαν στον Ανιχνευτή Harris3D και τον Περιγραφέα HOG/HOF στην ημίσεια χωρική ανάλυση (half spatial resolution). Η τεχνική αυτή αποτελεί ένα συμβιβασμό προκειμένου να μειωθεί ο υπολογιστικός χρόνος και χρησιμοποιείται συχνά στη βιβλιογραφία για σχετικά πειράματα (βλ. [3]). Για την

	Training Set	Test Set
AnswerPhone	66	64
DriveCar	85	102
FightPerson	54	70
GetOutCar	51	57
HandShake	32	45
SitUp	24	37

Πίνακας 4.1: Αριθμός δειγμάτων βίντεο για καθεμιά από τις έξι δράσεις στο Training και Test Set.

εκπόνηση του πειράματος έγινε χρήση της διαθέσιμης online υλοποίησης[3] που παρέχεται από την ιστοσελίδα του I. Laptev. Οι μόνες εξαρτήσεις του παραπάνω εκτελέσιμου κώδικα είναι η βιβλιοθήκη *OpenCV Library* και το λογισμικό *ffmpeg* τα οποία εγκαταστήσαμε σε περιβάλλον *Linux* για τις ανάγκες του πειράματος. Ο κώδικας υπολογίζει χωροχρονικά σημεία ενδιαφέροντος με τον Ανιχνευτή Harris3D και τον Περιγραφέα HOG/HOF σε τρισδιάστατα τοπικά τεμάχια με κέντρο τα επιλεγμένα σημεία.
Οι επιλογές των παραμέτρων που κάναμε για την ανίχνευση και την περιγραφή των σημείων περιγράφονται παρακάτω.

Επιλογές Παραμέτρων για τον Ανιχνευτή Harris3D: Οι χωρικές και χρονικές κλίμακες ανίχνευσης είναι οι ίδιες με εκείνες που αναφέραμε στην υποενότητα 2.2.1, δηλαδή $\sigma^2 \in \{4, 8, 16, 32, 64, 128\}$ και $\tau^2 \in \{2, 4\}$. Η επεξεργασία των βίντεο σε ημίσεια χωρική κλίμακα, με το ύψος των frames να κυμαίνεται περί τα 200 pixels και το πλάτος τους περί τα 300-400 pixels, θα καθιστούσε παράλογη την προσθήκη μεγαλύτερων χωρικών κλιμάκων. Η παράμετρος της συνάρτησης H_{3D} (2.2.9) τέθηκε στην προτεινόμενη από την υλοποίηση τιμή $k = 5 \cdot 10^{-4}$. Το κατώφλι για την απόρριψη των "αδύναμων" ανιχνεύσεων, που παρέχεται από τον κώδικα ως επιλογή για τη ρύθμιση της πυκνότητας των ανιχνεύσεων, διατηρήθηκε στην default τιμή του 10^{-9}. Τέλος, χρησιμοποιήσαμε το όριο των πέντε pixels για την αποφυγή πυροδότησης ανιχνεύσεων στα σύνορα των frames.

Επιλογές Παραμέτρων για τον Περιγραφέα HOG/HOF: Με τη διαμέριση του 3D τοπικού τεμαχίου σε $3 \times 3 \times 2$ "κελιά" το συνολικό διάνυσμα χαρακτηριστικών του περιγραφέα έχει 162 στοιχεία (βλ. υποενότητα 2.3.2). Υπενθυμίζουμε εδώ ότι το μέγεθος του τοπικού τεμαχίου (patch) σε κάθε διάσταση ορίζεται από τις αντίστοιχες κλίμακες ανίχνευσης ως εξής $\Delta_x, \Delta_y = 2k_{sp}\sigma$ και $\Delta_t = 2k_t\tau$ για τις χωρικές και τη χρονική διάσταση αντίστοιχα. Για τις σταθερές χωρικής και χρονικής υποστήριξης k_{sp} και k_t του περιγραφέα επιλέξαμε την προτεινόμενη τιμή 5. Σύμφωνα με αυτή την τιμή της σταθεράς και τις τιμές κλιμάκων για τον ανιχνευτή, το μικρότερο 3D τεμάχιο στο οποίο μπορεί να υπολογιστεί ο HOG/HOF έχει μέγεθος $20 \times 20 \times 15$, επιλογή που είναι λογική αν λάβουμε υπόψη μας το μέγεθος των frames και ότι το frame rate των

[3] http://www.irisa.fr/vista/Equipe/People/Laptev/download.html

δειγμάτων παίρνει τιμές περί τα 23-25 frames το δευτερόλεπτο.

Ο εκτελέσιμος κώδικας *stipdet* παράγει στην έξοδο αρχεία κειμένου από τα οποία μπορούμε να αντλήσουμε τις θέσεις και κλίμακες των ανιχνεύσεων και τους αντίστοιχους HOG/HOF περιγραφείς. Στο σύνολο των 681 δειγμάτων βίντεο του Training και Test Set, ο συνολικός αριθμός των frames ανέρχεται στα 236736. Ο συνολικός αριθμός των ανιχνευθέντων από τον Harris3D σημείων στο σύνολο των frames υπολογίστηκε στα 3390048 σημεία. Έτσι, ο Harris3D για το εν λόγω πείραμα έδωσε ένα μέσο όρο ανιχνεύσεων περί τα 14 σημεία ανά frame.
Οι απαραίτητες συναρτήσεις για τη διαχείριση των αρχείων κειμένου και τις διαδικασίες που απαιτούνται από την Bag-Of-Features προσέγγιση (βλ. 4.1) προγραμματίστηκαν στο λογισμικό MATLAB. Για την κατασκευή του λεξιλογίου, εφαρμόστηκε ο αλγόριθμος k-means με τυχαία αρχικοποίηση για τη συσταδοποίηση 100000 τυχαία επιλεγμένων περιγραφέων του Training Set σε 4000 οπτικές λέξεις. Ο μέγιστος αριθμός επαναλήψεων για τη σύγκλιση του αλγορίθμου τέθηκε στις 12 και ο χρόνος εκτέλεσης για αυτές τις επαναλήψεις, με χρήση της mex υλοποίησης του M. Everingham[4], υπολογίστηκε περί τα 30 λεπτά. Να σημειώσουμε εδώ ότι δεν έγινε χρήση της αντίστοιχης συνάρτησης του MATLAB καθώς ο μεγάλος όγκος των δεδομένων οδηγούσε σε εξάντληση της μνήμης RAM. Το συνολικό άθροισμα των εντός κλάσης αποστάσεων των δεδομένων από τα κεντροειδή για όλες τις κλάσεις ανήλθε στο 340475. Για την αντιστοίχιση των περιγραφέων στις λέξεις του οπτικού λεξιλογίου χρησιμοποιήθηκε το κριτήριο της Ευκλείδειας απόστασης. Τέλος, τα ιστογράμματα της συχνότητας εμφάνισης των λέξεων για κάθε δείγμα βίντεο κανονικοποιήθηκαν με την $\mathcal{L}_2$ νόρμα.
Για την αναγνώριση των ανθρώπινων δράσεων στα δείγματα βίντεο του Test Set χρησιμοποιήσαμε την ταξινόμηση με *Support Vector Machines (SVM)* που αναλύσαμε στην ενότητα 4.2. Στο εν λόγω πείραμα έχουμε το πρόβλημα ταξινόμησης σε πολλαπλές κλάσεις για το οποίο ακολουθήσαμε την *one-against-rest* λογική, εκπαιδεύοντας έναν δυαδικό ταξινομητή για κάθε κατηγορία δράσης. Το δείγμα βίντεο που εξετάζεται αποδίδεται τελικά στην κλάση που έδωσε τη μεγαλύτερη τιμή πρόβλεψης. Όπως είδαμε, ως μέτρο απόδοσης της αναγνώρισης χρησιμοποιούμε τη μέση τιμή της *μέσης ακρίβειας* επί όλων των κλάσεων (*mean average precision - mAP*). Θεωρώντας ένα δυαδικό πρόβλημα ταξινόμησης με ετικέτες $d_i \in \{-1, +1\}$ τα μεγέθη *precision* και *recall* ορίζονται ως

$$\begin{aligned} recall &= \frac{\#TruePositives}{\#TruePositives + \#FalseNegatives} \\ precision &= \frac{\#TruePositives}{\#TruePositives + \#FalsePositives} \end{aligned} \tag{4.3.1}$$

Παρακάτω εκθέτουμε τις επιλογές που έγιναν για τους ταξινομητές SVM στο εν λόγω πείραμα.

[4] http://server.cs.ucf.edu/~vision/source.html

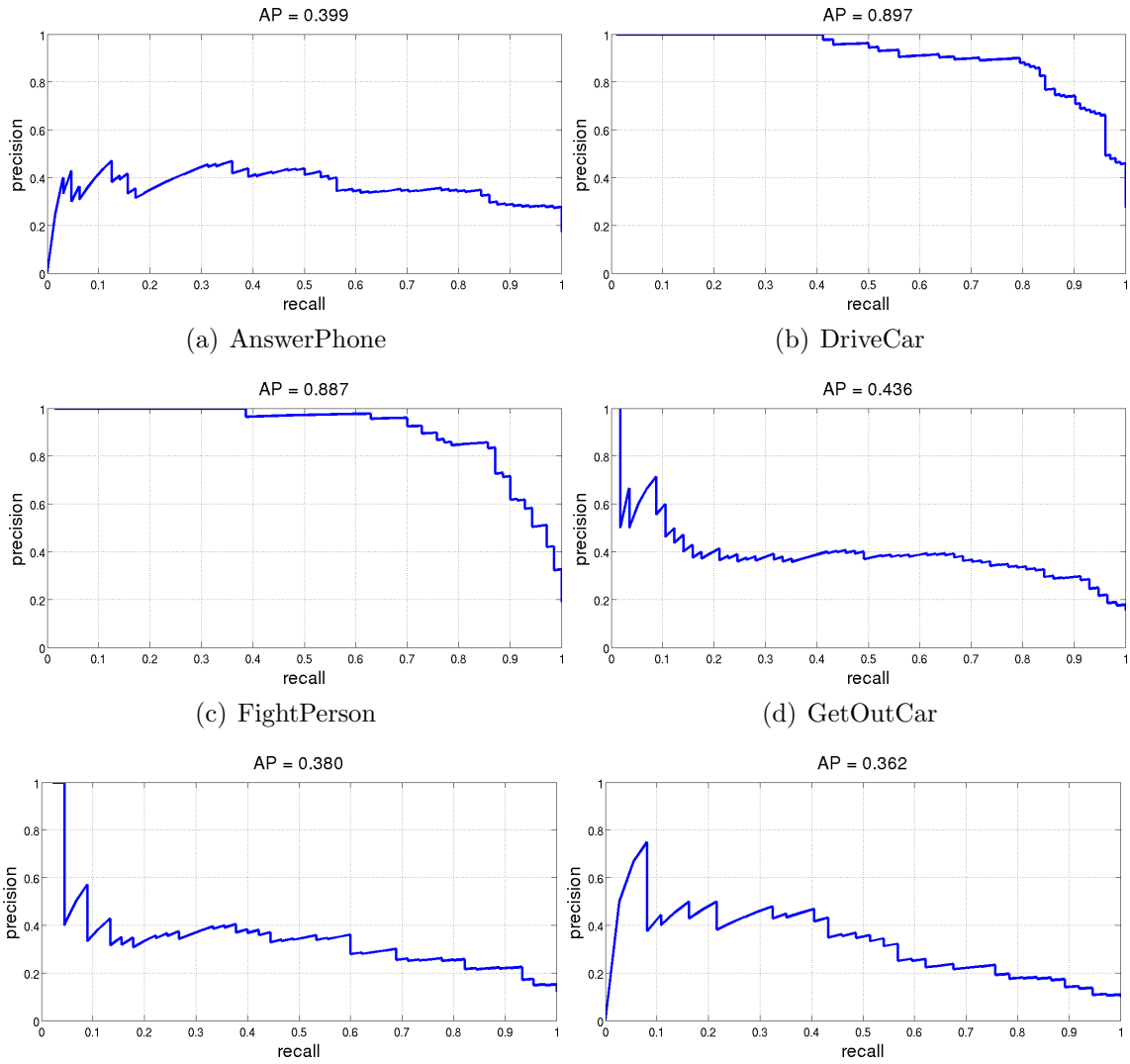
Επιλογές ταξινομητών SVM και των παραμέτρων τους: Για τους ταξινομητές SVM χρησιμοποιήσαμε την υλοποίηση *SVMLight*[5] για το λογισμικό MATLAB. Δεδομένου ότι δεν παρέχεται στις βιβλιοθήκες η επεκτεταμένη μορφή του Gaussian πυρήνα με απόσταση χ^2, επιλέξαμε τον καθιερωμένο RBF Gaussian πυρήνα με τη συνάρτησή του να δίνεται από τη σχέση (4.2.5). Η παράμετρος gamma του Gaussian πυρήνα τέθηκε στην τιμή $1/A$ όπου A η ευκλείδεια απόσταση μεταξύ όλων των ιστογραμμάτων του Training Set. Η παράμετρος c της υλοποίησης είναι η λεγόμενη παράμετρος "ομαλοποίησης"και ισούται με το αντίστροφο της παραμέτρου ποινής για λανθασμένες ταξινομήσεις C που είδαμε στην ενότητα 4.2. Σύμφωνα με τον Haykin [46], εκφράζει εννοιολογικά το συμβιβασμό που πρέπει να γίνει μεταξύ της πιστότητας σε σχέση με τα δεδομένα και της ομαλότητας της λύσης για το υπερεπίπεδο διαχωρισμού στο χώρο των χαρακτηριστικών. Ύστερα από πολλές δοκιμές, για τις οποίες παρατηρήσαμε αξιόλογες μεταβολές της τιμής *mean average precision (mAP)*, θέσαμε την παράμετρο στην τιμή $c = 0.01$ με την οποία λάβαμε το υψηλοτερο ποσοστό mAP.

Στον Πίνακα 4.2 παραθέτονται συγκεντρωτικά τα αποτελέσματα για τη μέση ακρίβεια ταξινόμησης-αναγνώρισης κάθε δράσης καθώς καί τη μέση τιμή της επί όλων των δράσεων (mean average precision). Τα αποτελέσματα επιβεβαιώνουν τη δυσκολία αναγνώρισης των τεσσάρων δράσεων AnswerPhone, GetOutCar, HandShake, SitUp ενώ για τις εναπομείνασες δύο δράσεις η απόδοση είναι αρκετά υψηλή. Τέτοιες μεγάλες μεταβολές στην απόδοση σε επιμέρους δράσεις είναι σε συμφωνία με αναφερόμενα στη βιβλιογραφία αποτελέσματα σχετικών πειραμάτων, χωρίς να καθίσταται εφικτή μια άμεση σύγκριση, καθώς το παρόν πείραμα αναφέρεται στο υποσύνολο των έξι δράσεων.
Στο Σχήμα 4.2 απεικονίζονται οι καμπύλες *precision-recall* με βάση τις προβλέψεις καθενός από τους έξι δυαδικούς ταξινομητές SVM του πειράματος.

	Harris3D + HOG/HOF
AnswerPhone	39.9%
DriveCar	89.7%
FightPerson	88.7%
GetOutCar	43.6%
HandShake	38.0%
SitUp	36.2%
mAP	**56.0%**

Πίνακας 4.2: Αποτελέσματα Average Precision για καθεμιά από τις έξι δράσεις του πειράματος και mean Average Precision επί όλων των δράσεων.

[5] http://svmlight.joachims.org/



(a) AnswerPhone (b) DriveCar

(c) FightPerson (d) GetOutCar

(e) HandShake (f) SitUp

Σχήμα 4.2: Καμπύλες *Precision-Recall* για καθεμιά από τις έξι δράσεις του πειράματος.

4.4 Πειράματα Αναγνώρισης με Εναλλακτικά Σχήματα Υπολογισμού των Ανιχνεύσεων και των Περιγραφέων στις Εικόνες Κυρίαρχης Ενέργειας

Στο προηγούμενο κεφάλαιο μελετήσαμε αναλυτικά τη χρήση του ενεργειακού τελεστή Teager-Kaiser για την αποδιαμόρφωση εικόνων που ακολουθούν το *Μοντέλο Πολλαπλών AM-FM συνιστωσών* της σχέσης (3.3.1) καθώς και σχήματα περιγραφής της υφής μέσω χαμηλής διάστασης ομαλά μεταβαλλόμενων διανυσμάτων χαρακτηριστικών. Πιο συγκεκριμένα, στην υποενότητα 3.3.4.2 είδαμε ότι με την *Ανάλυση Κυρίαρχων Συνιστωσών βασισμένη στην Ενέργεια (Energy-based Dominant Component Analysis (EDCA))* συλλαμβάνονται αποτελεσματικά οι κυρίαρχες διαμορφώσεις υφής μέσω της εξόδου του ενεργειακού τελεστή από τα κυρίαρχα κανάλια χωρίς την απώλεια της ουσιώδους γεωμετρικής δομής της εικόνας. Παράλληλα με τις περιοχές προεξόχοντος πλάτους διαμόρφωσης, με το κριτήριο της ενέργειας ταλάντωσης που χρησιμοποιείται από την *EDCA* για την επιλογή των κυρίαρχων καναλιών σε κάθε pixel, αναδεικνύονται καί περιοχές που, αν και δεν εμφανίζουν υψηλές τιμές του σήματος πλάτους, εντούτοις χαρακτηρίζονται από υψηλό μέτρο του μεταβαλλόμενου διανύσματος συχνοτήτων.
Οι εικόνες κυρίαρχης χωρικής ενέργειας που προκύπτουν από το σχήμα *EDCA* συμπυκνώνουν την κυρίαρχη πληροφορία ενέργειας ταλάντωσης της αρχικής *ευρυζωνικής* εικόνας (βλ. Σχήμα 3.9). Καταστέλλονται περιοχές που επιδεικνύουν χαμηλό γινόμενο διαμόρφωσης ενώ αναδεικνύονται άλλες με υψηλής τιμής έξοδο του ενεργειακού τελεστή. Παρατηρήσαμε ότι σε πολλές περιπτώσεις αυτό μεταφράζεται στην παραγωγή χαμηλών τιμών *EDCA* ενέργειας σε περιοχές του παρασκηνίου της εικόνας ή του εσωτερικού των εικονιζόμενων μορφών και αντικειμένων που εμφανίζουν μάλλον αδιάφορο περιεχόμενο ταλάντωσης και εν γένει οπτικής σημαντικότητας (visual saliency). Από την άλλη, οι προεξέχουσες ακμές της εικόνας, όπως αυτές στα περιγράμματα των ανθρώπων ή των αντικειμένων, διατηρούνται σε πολύ ικανοποιητικό βαθμό λεπτομέρειας, κάτι που είναι εμφανές όταν οι "ενεργειακές"εικόνες απεικονίζονται ως γκρίζες (greyscale). Υψηλές τιμές *EDCA* ενέργειας δεν εντοπίζονται μόνο στο εξωτερικό περίγραμμα (silhouette) των ανθρώπινων μορφών αλλά καί στα σύνορα των μελών και άκρων τους, πληροφορία που είναι ουσιώδης για το πρόβλημα της αναγνώρισης των ανθρώπινων δράσεων σε ρεαλιστικά σκηνικά με ύπαρξη οπτικών εμποδίων.
Παρατηρούμε λοιπόν ότι οι εικόνες *EDCA* ενέργειας παρέχουν μια λιγότερο λεπτομέρη αλλά και λιγότερο θορυβώδη εκδοχή των αρχικών frames, παρέχοντας υψηλές τιμές απόκρισης σε περιοχές έντονου περιεχομένου υφής, υποβαθμίζοντας μη σημαντικές σε όρους ενέργειας ταλάντωσης περιοχές και αναδεικνύοντας τις επικρατούσες ακμές της εικόνας. Με βάση τα παραπάνω, ενισχύθηκε η πεποίθησή μας ότι η προώθηση των τιμών κυρίαρχης ενέργειας για κάθε frame με βάση το *EDCA* σχήμα αντί των αρχικών πλαισίων στην είσοδο των σταδίων ανίχνευσης ή περιγραφής, θα μπορούσε να αποφέρει ποιοτικότερες ανιχνεύσεις ή πιο διακριτικούς περιγραφείς αντίστοιχα. Δεδομένου ότι η έξοδος του ενεργειακού τελεστή από τα κυρίαρχα κανάλια συνδυάζει πληροφορία πλάτους και

συχνότητας διαμόρφωσης (όπως είδαμε αναλυτικά στο προηγούμενο κεφάλαιο), αποφασίσαμε να μην εξάγουμε μέσω του Αλγορίθμου *2D Gabor ESA* (βλ. 3.3.3) τα σήματα πλάτους και συχνότητας, διαδικασία που μάλιστα θα επιβάρυνε σημαντικά την υπολογιστική πολυπλοκότητα στην επεξεργασία κάθε frame.
Πριν προχωρήσουμε στην περιγραφή των πειραμάτων στα οποία χρησιμοποιήσαμε τις εικόνες *EDCA* ενέργειας των frames ώς είσοδο στον Ανιχνευτή Harris3D ή στον Περιγραφέα HOG/HOF, θα δώσουμε στην ενότητα που ακολουθεί τις βασικές λεπτομέρειες υλοποίησης του σχήματος εξαγωγής τους, αναφέροντας καί τους χρόνους εκτέλεσης των πιο απαιτητικών υπολογιστικά διαδικασιών.

4.4.1 Εξαγωγή των Εικόνων EDCA Ενέργειας για δείγματα Βίντεο - Ανάπτυξη Ταχύτερης Υλοποίησης

Το σύστημα υπολογισμού της *EDCA* Ενέργειας για κάθε frame του δείγματος βίντεο εισόδου υλοποιήθηκε στο λογισμικό MATLAB με τη βοήθεια του *dca2D package* του G. Evangelopoulos που αναπτύχθηκε στο CVSP Lab[6]. Ακολουθούν σε οργάνωση κατά παραγράφους οι λεπτομέρειες κάθε βήματος του αλγορίθμου που χρησιμοποιήσαμε από την ανάγνωση του δείγματος βίντεο εώς την παραγωγή της *EDCA* Ενέργειας όλων των frames στην έξοδο.

Ανάγνωση των frames και μετατροπή τους σε γκρίζες εικόνες ημίσειας ανάλυσης. Τα frames του δείγματος εισόδου (από τη Βάση *Hollywood2*) διαβάζονται στην αρχική RGB μορφή τους, μετατρέπονται σε γκρίζες (greyscale) εικόνες και αποθηκεύονται στην ημίσεια χωρική ανάλυση (δηλαδή σε υποτετραπλάσιο μέγεθος) σε δομή πίνακα *array*. Για την αποφυγή προβλημάτων *Out Of Memory* η ανάγνωση γίνεται με βήμα 50 frames, διαδικασία που προσθέτει μηδαμινή χρονική επιβάρυνση στην υλοποίηση.

Κατασκευή συστοιχίας 40 μιγαδικών Gabor φίλτρων. Ακολουθούνται οι επιλογές σχεδίασης της υποενότητας 3.3.1 για την κατασκευή της συστοιχίας δισδιάστατων μιγαδικών Gabor φίλτρων που καλύπτουν το πεδίο συχνοτήτων, διατεταγμένα σε 5 κλίμακες και 8 προσανατολισμούς. Η σχεδίαση είναι συμμετρική και προσαρμοσμένη για εικόνα διαστάσεων 288×288 που πειραματικά αποδείχθηκε λογική επιλογή δεδομένου του μεγέθους των frames στη μισή τους ανάλυση. Η συστοιχία παραμένει σταθερή στην εφαρμογή του αλγορίθμου για όλα τα βίντεο του Training και Test Set της προηγούμενης ενότητας.

Υπολογισμός των αποκρίσεων συχνότητας των φίλτρων. Οι παράμετροι εύρους ζώνης (bandwidth) σε κάθε κατεύθυνση και οι συνιστώσες της κεντρικής συχνότητας των φίλτρων κανονικοποιούνται με βάση το μέγεθος του frame στις δύο διαστάσεις και ακολούθως υπολογίζονται οι αποκρίσεις συχνότητας των 40 φίλτρων. Η διαδικασία αυτή εκτελείται μόνο μία φορά για κάθε δείγμα βίντεο.

[6] http://cvsp.cs.ntua.gr/

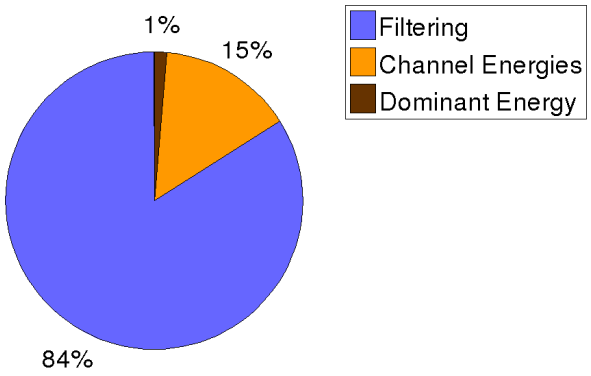
Υπολογισμός των φιλτραρισμένων εξόδων κάθε frame από τα φίλτρα και τις παραγώγους τους. Ακολουθούμε τη λογική του Αλγορίθμου *2D Gabor ESA* που περιγράψαμε στην υποενότητα 3.3.3. Για τον υπολογισμό της Teager-Kaiser ενέργειας στην έξοδο κάθε καναλιού της συστοιχίας απαιτούνται οι συνελίξεις του σήματος εικόνας με το φίλτρο g και με τις τέσσερεις μερικές παραγώγους του g_x, g_y, g_{xx}, g_{yy} (βλ. (3.3.12)). Σύμφωνα με την εναλλακτική πρόταση των Kokkinos et al. [42], υπολογίζουμε τον πολλαπλασιασμό του FFT μετασχηματισμού της εικόνας με την απόκριση συχνότητας κάθε φίλτρου και των παραγώγων του, παίρνοντας τον FFT μετασχηματισμό των παραγώγων από τη σχέση (3.3.15). Ο υπολογισμός των εξόδων φιλτραρίσματος στο πεδίο του διακριτού χώρου από το χώρο συχνοτήτων γίνεται με χρήση της συνάρτησης `ifft2` του MATLAB. Το στάδιο αυτό είναι το πιο χρονοβόρο της υλοποίησης καθώς απαιτούνται, εκτός των πολλαπλασιασμών πινάκων και μιας κλήσης της `fft2`, 5 κλήσεις της `ifft2` για κάθε φίλτρο, συνολικά δηλαδή 200 κλήσεις της `ifft2` για κάθε frame. Όταν το μέγεθος frame είναι μεγαλό, αυτό μπορεί να οδηγήσει σε αρκετά αυξημένους χρόνους εκτέλεσης αυτού του σταδίου, εντούτοις το προτιμούμε από την επιλογή των συνελίξεων που θα επέφερε μεγαλύτερο υπολογιστικό κόστος. Δεδομένου ότι τα φίλτρα είναι μιγαδικά, οι έξοδοι των καναλιών είναι επίσης μιγαδικές. Οι φιλτραρισμένες έξοδοι αποθηκεύονται σε δομή *cell* η οποία προσπελαύνεται πιο γρήγορα από το MATLAB σε σχέση με τη δομή *array*.

Υπολογισμός της Teager-Kaiser ενέργειας στην έξοδο κάθε καναλιού. Σε αυτό το στάδιο ανακτώνται τα πραγματικά και μιγαδικά μέρη των φιλτραρισμένων εξόδων καναλιού του προηγούμενου σταδίου και ακολούθως υπολογίζεται η Teager-Kaiser ενέργεια στην έξοδο κάθε καναλιού βάσει του μιγαδικού τελεστή της σχέσης (3.3.8). Για την αποφυγή αρνητικών τιμών, στα σημεία όπου η τιμή της ενέργειας είναι χαμηλότερη του γινομένου 10^{-5} επί της μέγιστης τιμής, η ενέργεια τίθεται στην τιμή μηδέν. Οι ενέργειες όλων των καναλιών αποθηκεύονται σε δομή *cell*. Το στάδιο αυτό βασίζεται κυρίως σε πράξεις πινάκων και έτσι εμφανίζει χαμηλούς χρόνους εκτέλεσης.

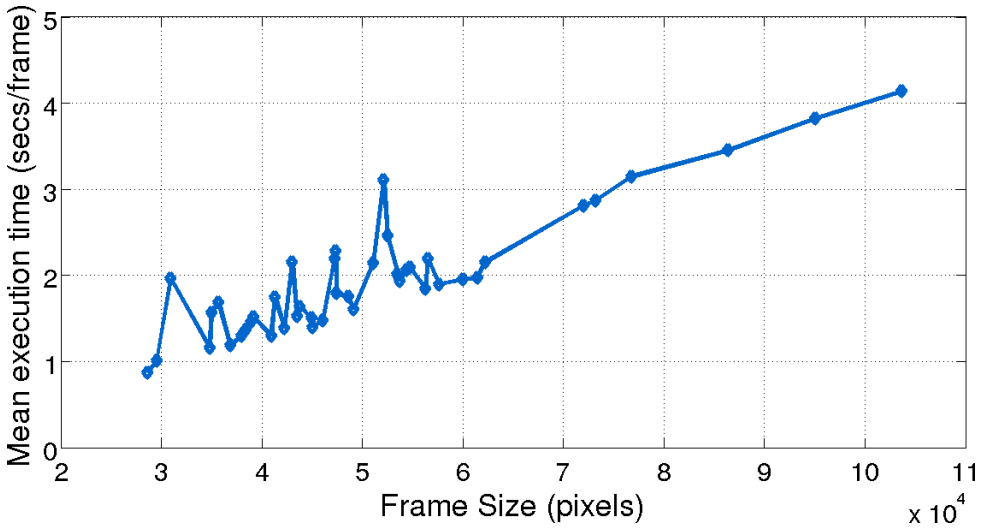
Υπολογισμός της κυρίαρχης (dominant) κατά EDCA ενέργειας. Αναζητούμε σε κάθε pixel το κανάλι στου οποίου την έξοδο ο ενεργειακός τελεστής έδωσε την υψηλότερη τιμή ενέργειας μεταξύ των 40 καναλιών. Στο *dca2D package* ακολουθείται η λογική της διανυσματοποίησης (vectorization) κάθε ενέργειας καναλιού και της συνένωσης όλων σε μια δομή *array* με μία γραμμή για κάθε κανάλι και αριθμό στηλών όσες τα pixels του frame. Ακολούθως χρησιμοποιείται η συνάρτηση `sort` για την εξαγωγή των δεικτών (indices) των κυρίαρχων καναλιών, είτε σε ολόκληρο το *array* με τις ενέργειες καναλιών είτε σε τμήματά του στην περίπτωση μεγάλου μεγέθους πίνακα. Τέλος, οι δείκτες των κυρίαρχων ενεργειών αναδιατάσσονται στο αρχικό μέγεθος εικόνας προκειμένου να υπολογιστεί ο πίνακας της κυρίαρχης ενέργειας. Αυτή η διαδικασία απαιτεί, εκτός των άλλων, την κλήση της συνάρτησης `reshape` του MATLAB 41 φορές (40 για τους πίνακες ενεργειών και 1 για τον πίνακα δεικτών) καθώς και ένα βρόχο 40 επαναλήψεων (`for` loop) για την εξαγωγή του πίνακα κυρίαρχης ενέργειας βάσει των δεικτών των κυρίαρχων καναλιών.
Στην παρούσα διπλωματική εργασία προτείνουμε μια πολύ απλούστερη και ταχύτερη

υλοποίηση αυτού του σταδίου με χρήση δύο build-in συναρτήσεων του MATLAB, δεδομένου ότι δεν έχουμε να αντιμετωπίσουμε εν προκειμένω πρόβλημα πλήρωσης της μνήμης RAM λόγω της επεξεργασίας των frames στην ημίσεια ανάλυση. Έτσι, τα στοιχεία της δομής *cell* στα οποία έχουν αποθηκευθεί οι ενέργειες καναλιού συνενώνονται με χρήση της συνάρτησης `cat` σε έναν τρισδιάστατο πίνακα, του οποίου οι δύο πρώτες διαστάσεις ταυτίζονται με αυτές της εικόνας και η τρίτη έχει μέγεθος 40, όσα και τα φίλτρα. Ο υπολογισμός της εικόνας με τις τιμές της κυρίαρχης ενέργειας ανα pixel γίνεται απλά με την κλήση της συνάρτησης `max` ως προς την τρίτη διάσταση. *Η τροποποίηση αυτή επιτρέπει στο στάδιο αυτό να εκτελείται σε περίπου 10 φορές μικρότερο χρόνο σε σχέση με την υλοποίηση που υιοθετείται στο dca2D package (ενδεικτικές τιμές χρόνων για μέγεθος frame* 152×288 *είναι 0.02 secs/frame και 0.2 secs/frame αντίστοιχα)* .

Τα τρία παραπάνω στάδια εκτελούνται για κάθε frame του δείγματος βίντεο εισόδου. Σε αρχεία κειμένου αποθηκεύεται ο μέσος χρόνος εκτέλεσης του κάθε σταδίου αλλά και ο συνολικός μέσος χρόνος για την εξαγωγή του πίνακα της *EDCA* ενέργειας σε δευτερόλεπτα ανά frame για κάθε δείγμα.
Η τρισδιάστατη δομή *array* που περιέχει τις *EDCA* ενέργειες για όλα τα frames του δείγματος βίντεο πρέπει να μετατραπεί σε δείγμα βίντεο επέκτασης *.avi* ώστε να είναι εφικτή η προώθηση του νέου βίντεο με τις "ενεργειακές"εικόνες στο στάδιο ανίχνευσης ή περιγραφής του εκτελέσιμου κώδικα *stipdet*. Με αυτόν τον τρόπο οι τιμές της κυρίαρχης ενέργειας για κάθε frame αντιστοιχίζονται στο πεδίο τιμών μιας εικόνας RGB. Η δημιουργία βίντεο με τις *EDCA* ενέργειες στο frame rate του αρχικού δείγματος υλοποιείται στο MATLAB με χρήση των συναρτήσεων `avifile`, `addframe` και `getframe`. Να σημειώσουμε εδώ ότι ιδιαίτερη προσοχή χρειάζεται στη χρήση της συνάρτησης `getframe`, καθώς το MATLAB όταν συλλαμβάνει το περιεχόμενο της εικόνας που εμφανίζεται, πολλές φορές λανθασμένα προσθέτει μία ή περισσότερες γραμμές ή στήλες από το περιθώριο στα δεδομένα, με αποτέλεσμα τη δημιουργία ενός βίντεο στην έξοδο με *στρεβλωμένα* frames. Το βίντεο εξόδου, που είναι κωδικοποιημένο σε rgb24 μορφή, το μετατρέπουμε με τη βοήθεια του λογισμικού *ffmpeg* στη συμπιεσμένη μορφή κωδικοποίησης DivX MPEG-4 για εξοικονόμηση μνήμης.
Η παραπάνω υλοποίηση υπολογισμού της *EDCA* κυρίαρχης ενέργειας των frames εφαρμόστηκε στο σύνολο των 681 δειγμάτων βίντεο του Training και Test Set του πειράματος αναγνώρισης δράσεων της προηγούμενης ενότητας. Το πείραμα διεξήχθη σε υπολογιστή με επεξεργαστή Intel Core 2 Duo 2GHz και μνήμη 3GB RAM.
Στο Σχήμα 4.3 απεικονίζεται η κατανομή του μέσου χρόνου εκτέλεσης ανά frame για τον υπολογισμό της *EDCA* ενέργειας στα τρία επιμέρους στάδια του φιλτραρίσματος, του υπολογισμού των ενεργειών καναλιού και της εξαγωγής της κυρίαρχης ενέργειας. Στο Σχήμα 4.4 δίνεται το γράφημα του μέσου χρόνου υπολογισμού της *EDCA* ενέργειας με το μέγεθος frame. Στον Πίνακα 4.3 παρατίθενται οι μέσοι χρόνοι εκτέλεσης των τριών σταδίων και ο μέσος συνολικός χρόνος σε δευτερόλεπτα ανά frame.



Σχήμα 4.3: Κατανομή του μέσου χρόνου υπολογισμού της *EDCA* Ενέργειας στα επιμέρους στάδια του φιλτραρίσματος με τα Gabor φίλτρα και τις παραγώγους τους, του υπολογισμού της ενέργειας Teager-Kaiser στις εξόδους των καναλιών και της εξαγωγής της κυρίαρχης ενέργειας με βάση το σχήμα *EDCA*.



Σχήμα 4.4: Μέσος χρόνος υπολογισμού της *EDCA* Ενέργειας για τα διάφορα μεγέθη frame των δειγμάτων του πειράματος.

	Μέσος Χρόνος Εκτέλεσης (secs/frame)
Φιλτράρισμα	1.565
Ενέργειες Καναλιών	0.276
Κυρίαρχη EDCA Ενέργεια	0.026
Σύνολο	**1.867**

Πίνακας 4.3: Μέσοι χρόνοι εκτέλεσης σε secs/frame των τριών σταδίων υπολογισμού της *EDCA* Ενέργειας και μέσος συνολικός χρόνος, υπολογισμένοι επί του συνόλου των frames των 681 δειγμάτων του πειράματος, με μέσο μέγεθος frame τα 50126 pixels.

4.4.2 Εικόνες Κυρίαρχης Ενέργειας ως είσοδος στο σχήμα Ανιχνευτή Harris3D - Περιγραφέα HOG/HOF

Στην παρούσα υποενότητα θα προωθήσουμε τα βίντεο με τις αντίστοιχες εικόνες κυρίαρχης *EDCA* Ενέργειας που έχουμε εξάγει για τα 681 δείγματα του Training και Test Set του πειράματος αναγνώρισης των έξι δράσεων της ενότητας 4.3 στο σχήμα υπολογισμού των ανιχνεύσεων από τον Harris3D και των περιγραφέων από τον HOG/HOF. Με άλλα λόγια, το ενοποιημένο σχήμα Ανιχνευτή-Περιγραφέα θα επενεργεί εδώ επί των εικόνων κυρίαρχης ενέργειας αντί των αρχικών frames των δειγμάτων για την εξαγωγή των χωροχρονικών σημείων ενδιαφέροντος και τον υπολογισμό των ιστογραμμάτων εμφάνισης και κίνησης στα αντίστοιχα τρισδιάστατα τοπικά τεμάχια. Οι λόγοι που οδήγησαν στην διεξαγωγή του εν λόγω πειράματος περιγράφηκαν στην εισαγωγή του παρόντος κεφαλαίου. Με αυτό το εναλλακτικό σχήμα επιχειρούμε να εξετάσουμε τον βαθμό στον οποίο οι εικόνες κυρίαρχης ενέργειας, που συλλαμβάνουν αποτελεσματικά το επικρατόν περιεχόμενο υφής με ταυτόχρονη ανάδειξη των κυριότερων ακμών της εικόνας, μπορούν να "καθοδηγήσουν"τον ανιχνευτή Harris3D στη σύλληψη ποιοτικότερων ανιχνεύσεων αλλά καί τον περιγραφέα HOG/HOF στον υπολογισμό πιο διακριτικών (discriminative) τοπικών τεμαχίων περιγραφής. Η μετέπειτα διαδικασία που ακολουθείται είναι πανομοιότυπη με εκείνη του πειράματος της ενότητας 4.3, με την Bag-Of-Features προσέγγιση και την ταξινόμηση SVM.

Η εξαγωγή των τοπικών χαρακτηριστικών υλοποιήθηκε με χρήση της διαθέσιμης online υλοποίησης[7] με επιλογή των ίδιων παραμέτρων για τον ανιχνευτή και τον περιγραφέα. Ωστόσο, με πειράματα σε μεμονωμένα δείγματα του Set παρατηρήσαμε ότι η χρήση του ίδιου κατωφλιού απόρριψης των "αδύναμων"ανιχνεύσεων στην τιμή 10^{-9} οδηγούσε στην παραγωγή πολύ αραιότερων ανιχνεύσεων στις εικόνες κυρίαρχης *EDCA* ενέργειας σε σχέση με τις αντίστοιχες στα αρχικά frames. Αυτό εξηγείται εύκολα από το γεγονός ότι οι "ενεργειακές"εικόνες παρουσιάζουν μια ωμή δομή πολύ μικρότερης λεπτομέρειας σε σχέση με τα αντίστοιχα frames. Όπως έχουμε αναφέρει, στο ρεαλιστικό σκηνικό των δειγμάτων της Βάσης *Hollywood2 Actions Dataset* η παραγωγή υπερβολικά αραιών ανιχνεύσεων οδηγεί σε χαμηλές αποδόσεις αναγνώρισης [3]. Για τούτο στο εν λόγω πείραμα θέσαμε το κατώφλι στην τιμή 10^{-16} για την αύξηση του αριθμού των ανιχνεύσεων στην έξοδο. Πράγματι στο σύνολο των 236736 frames των δειγμάτων του Set των έξι δράσεων, λάβαμε 3353439 χωροχρονικά σημεία ενδιαφέροντος από τον Harris3D, που αποτελούν το 98.92% του όγκου ανιχνεύσεων που έδωσε το καθιερωμένο σχήμα της ενότητας 4.3. Ο μέσος όρος ανιχνεύσεων ανέρχεται καί εδώ στα 14 σημεία ανά frame. Μάλιστα, η παραγωγή σχεδόν ταυτόσημου συνολικού αριθμού ανιχνεύσεων από τα δύο σχήματα θα καταστήσει απόλυτα "δίκαιη"τη σύγκριση της απόδοσής τους.

Στο Σχήμα 4.5 απεικονίζονται ανιχνευθέντα από τον Harris3D σημεία σε χαρακτηριστικά frames δειγμάτων του πειράματος, που έχουν προκύψει είτε από το καθιερωμένο σχήμα ανίχνευσης στα αρχικά frames είτε από το εναλλακτικό σχήμα ανίχνευσης στις εικόνες *EDCA* ενέργειας. Καί οι δύο περιπτώσεις ανιχνεύσεων εμφανίζονται επί των αρχικών πλαισίων, με κίτρινο και πράσινο χρώμα αντίστοιχα. Παρατηρώντας τις εικόνες *EDCA* ενέργειας στην τρίτη γραμμή, είναι εμφανής η διατήρηση της κυρίαρχης δομής

[7] http://www.irisa.fr/vista/Equipe/People/Laptev/download.html

Σχήμα 4.5: Ανιχνεύσεις από τον Harris3D με βάση το καθιερωμένο και το εναλλακτικό σχήμα σε χαρακτηριστικά frames δειγμάτων βίντεο που ανήκουν στις δράσεις *Handshake, AnswerPhone, GetOutCar* και *SitUp* (από Αριστερά προς τα Δεξιά). Από Επάνω προς τα Κάτω: αρχικά έγχρωμα frames, ανιχνευθέντα σημεία με είσοδο τα αρχικά frames, αντίστοιχες εικόνες *EDCA* Κυρίαρχης Ενέργειας, ανιχνευθέντα σημεία με είσοδο τις εικόνες *EDCA* Κυρίαρχης Ενέργειας εικονιζόμενα επί των αρχικών frames.

των εικόνων και των επικρατουσών ακμών τους. Στις προκύπτουσες από αυτές ανιχνεύσεις στην τέταρτη γραμμή μπορούμε να διακρίνουμε την ύπαρξη σημείων σε πολλαπλές κλίμακες σε χωρικά δεσπόζουσες περιοχές της εικόνας με μη σταθερή κίνηση όπως τα χέρια και οι αγκώνες στα δύο πρώτα δείγματα, τα σύνορα της πόρτας του αυτοκινήτου στο τρίτο ή οι ώμοι στο τέταρτο.

Ύστερα από την εξαγωγή των τοπικών χαρακτηριστικών βάσει του εναλλακτικού σχήματος, ακολουθήθηκε η διαδικασία Bag-Of-Features και SVM ταξινόμησης με τις ίδιες επιλογές όπως στο πείραμα της ενότητας 4.3. Να σημειώσουμε ότι για το στάδιο της κατασκευής του οπτικού λεξιλογίου με τον αλγόριθμο k-means το συνολικό άθροισμα των εντός κλάσης αποστάσεων των δεδομένων από τα κεντροειδή για όλες τις κλάσεις ανήλθε στο 272243 έναντι 340475 του καθιερωμένου σχήματος.

Στον Πίνακα 4.4 παρουσιάζονται τα αποτελέσματα αναγνώρισης των δράσεων στο εν λόγω πείραμα πλάι στα αντίστοιχα αποτελέσματα για τη *μέση ακρίβεια* ταξινόμησης κάθε δράσης και τη μέση τιμή της (*mean average precision*) επί όλων των δράσεων που έδωσε το πείραμα της ενότητας 4.3. Παρατηρούμε υποβάθμιση της συνολικής απόδοσης του

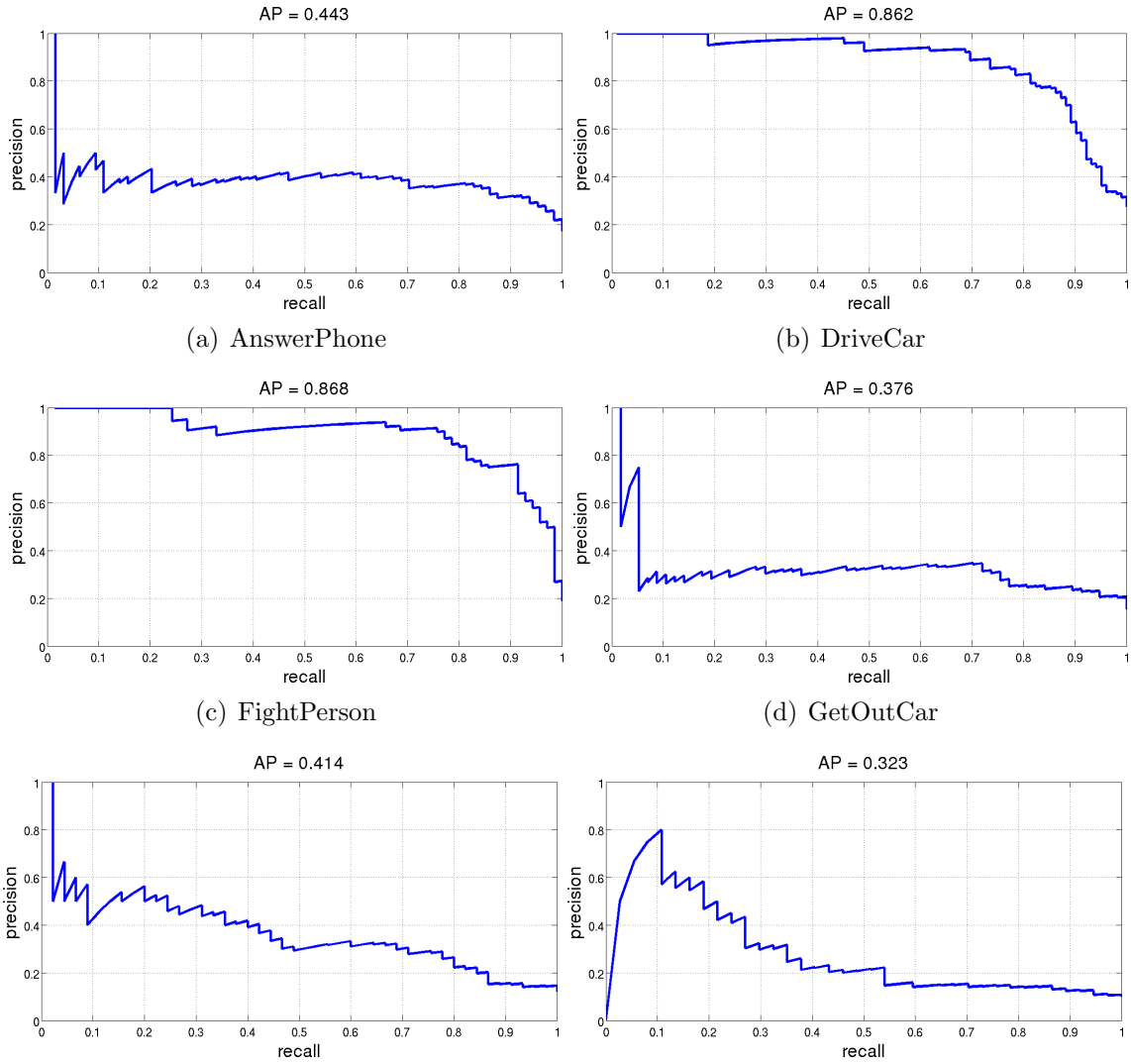
	Harris3D + HOG/HOF	**Harris3D on EDCA Energy + HOG/HOF on EDCA Energy**
AnswerPhone	39.9%	**44.3%**
DriveCar	**89.7%**	86.2%
FightPerson	**88.7%**	86.8%
GetOutCar	**43.6%**	37.6%
HandShake	38.0%	**41.4%**
SitUp	**36.2%**	32.3%
mAP	**56.0%**	54.8%

Πίνακας 4.4: Αποτελέσματα Average Precision για καθεμιά από τις έξι δράσεις του πειράματος και mean Average Precision επί όλων των δράσεων για το καθιερωμένο σχήμα εξαγωγής χαρακτηριστικών (Αριστερά) και το εναλλακτικό σχήμα υπολογισμού των ανιχνεύσεων και περιγραφέων στις εικόνες *EDCA* ενέργειας (Δεξιά).

πλαισίου αναγνώρισης κατά 1.2%, με τη δράση *GetOutCar* να εμφανίζει τη μεγαλύτερη πτώση συγκριτικά με το προηγούμενο πείραμα, της τάξης του 6%. Σημαντική μείωση στην απόδοση υπάρχει καί για τις δράσεις *SitUp* και *DriveCar* ενώ μικρή υποβάθμιση είχαμε για τη δράση *FightPerson*. Εντούτοις, το εναλλακτικό σχήμα αποφέρει σημαντική αύξηση της απόδοσης στις δύο απαιτητικές δράσεις *AnswerPhone* και *HandShake*, κατά 4.4% και 3.4% αντίστοιχα. Οι βελτιωμένες επιδόσεις σε αυτές τις δύο δράσεις έδωσαν την ώθηση για τη διεξαγωγή του επιπλέον πειράματος που θα παρουσιαστεί στην επόμενη υποενότητα και αποτελεί μια τροποποιημένη εκδοχή του εναλλακτικού σχήματος που εφαρμόσαμε εδώ.

Επιχειρώντας μια ερμηνεία της χαμηλής απόδοσης στις δράσεις *GetOutCar* και *SitUp*, πιθανολογούμε ότι αυτή εντοπίζεται κυρίως στο στάδιο της περιγραφής των σημείων από τον περιγραφέα HOG/HOF στις "ενεργειακές"εικόνες. Οι εν λόγω δράσεις είναι στενά συνυφασμένες με το σκηνικό στο οποίο συμβαίνουν. Η *GetOutCar* απαιτεί σκηνικό εξωτερικού χώρου με την παρουσία οχήματος ενώ η δράση *SitUp* εμφανίζεται στη Βάση *Hollywood2* στην πλειονότητα των περιπτώσεων σε εσωτερικό χώρο όπου είναι εμφανής η παρουσία κρεβατιού. Με πρόσφατες έρευνες στο πρόβλημα της αναγνώρισης ανθρώπινων δράσεων, ενισχύεται η πεποίθηση ότι στοιχεία εμφάνισης του παρασκηνίου, δηλωτικά του σκηνικού στο οποίο εκτυλίσσονται οι δράσεις, είναι απαραίτητα για αποτελεσματική αναγνώριση [25]. Ενδεχομένως λοιπόν, ο υπολογισμός των περιγραφέων HOG με βάση τις τιμές κυρίαρχης ενέργειας και όχι τις εντάσεις (intensities) των αρχικών frames να οδήγησε σε απώλεια σημαντικών χαρακτηριστικών εμφάνισης του περιβάλλοντος χώρου για αυτές τις δράσεις και κατά συνέπεια στη μείωση της διακριτικής ικανότητας. Η καλή απόδοση, από την άλλη, των δράσεων *AnswerPhone* και *HandShake* μπορεί να αποδοθεί στην εξαγωγή μεγάλου όγκου ανιχνεύσεων σε πολλαπλές κλίμακες στις κινήσεις των χεριών από τον Harris3D στις εικόνες κυρίαρχης ενέργειας που αναδεικνύουν συνήθως με προεξέχουσες τιμές αυτές τις περιοχές.

Στο Σχήμα 4.6 απεικονίζονται οι καμπύλες *precision-recall* για τις έξι δράσεις στο πείραμα της παρούσας υποενότητας.



(a) AnswerPhone (b) DriveCar

(c) FightPerson (d) GetOutCar

(e) HandShake (f) SitUp

Σχήμα 4.6: Καμπύλες *Precision-Recall* για καθεμιά από τις έξι δράσεις του πειράματος με το εναλλακτικό σχήμα ανίχνευσης και περιγραφής στις εικόνες *EDCA* κυρίαρχης ενέργειας.

4.4.3 Υπολογισμός Ανιχνεύσεων από τον Harris3D στις Εικόνες Κυρίαρχης Ενέργειας και Εξαγωγή Περιγραφέων HOG/HOF στα αρχικά frames

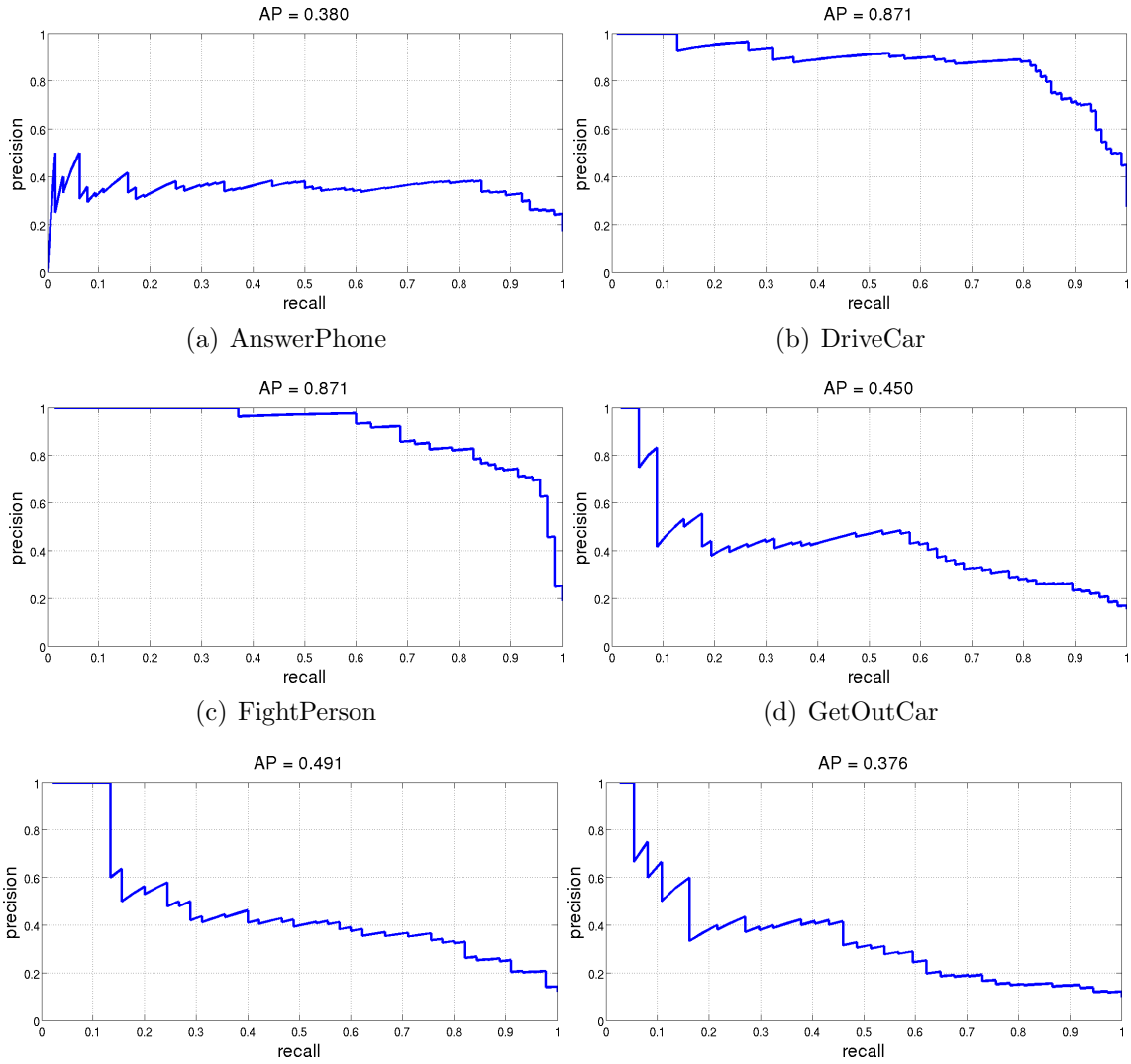
Στην προηγούμενη υποενότητα χρησιμοποιήσαμε τις εικόνες κυρίαρχης *EDCA* ενέργειας ως είσοδο καί για τα δύο στάδια εξαγωγής τοπικών χαρακτηριστικών, αυτά της ανίχνευσης χωροχρονικών σημείων ενδιαφέροντος από τον Ανιχνευτή Harris3D και της περιγραφής τους σε τρισδιάστατα τοπικά τεμάχια από τον συνδυασμένο Περιγραφέα εμφάνισης και κίνησης HOG/HOF. Στην ερμηνεία των αποτελεσμάτων αναγνώρισης των δράσεων που λάβαμε από το πείραμα, αποδώσαμε τις χαμηλές επιδόσεις σε μεμονωμένες δράσεις στην αδυναμία σύλληψης σημαντικών χαρακτηριστικών εμφάνισης του σκηνικού των δράσεων από το εναλλακτικό σχήμα υπολογισμού του Περιγραφέα HOG βάσει των τιμών κυρίαρχης ενέργειας αντί των τιμών έντασης των αρχικών frames.
Η ποιότητα των ανιχνεύσεων που παράγει το εναλλακτικό σχήμα ανίχνευσης Harris3D στις εικόνες *EDCA* ενέργειας και οι υψηλότερες τιμές *μέσης ακρίβειας* σε δύο δράσεις, συγκριτικά με το πείραμα της ενότητας 4.3, μας οδήγησε στην πραγματοποίηση ενός επιπλέον πειράματος αναγνώρισης στο ίδιο Set έξι δράσεων. Στο παρόν πείραμα, εξετάζουμε τη συμβολή της συνέργειας των χαρακτηριστικών κυρίαρχης ενέργειας με το κλασικό σχήμα ανίχνευσης του Harris3D στην απόδοση του συνολικού πλαισίου αναγνώρισης. Συγκεκριμένα, τα τρισδιάστατα τοπικά σημεία ενδιαφέροντος υπολογίζονται από τον Harris3D επί των εικόνων *EDCA* κυρίαρχης ενέργειας που έχουμε εξάγει ενώ η περιγραφή των σημείων από τον HOG/HOF γίνεται με τον καθιερωμένο τρόπο από τα αρχικά frames στις αντίστοιχες θέσεις. Το νέο εναλλακτικό σχήμα μπορεί να ιδωθεί ως ένα στάδιο προ-επεξεργασίας των αρχικών frames για την καλύτερη καθοδήγηση του πλαισίου ανίχνευσης στην παραγωγή σημείων σε προεξέχουσες, με όρους ενέργειας ταλάντωσης και υφής, περιοχές των εικόνων που εμφανίζουν μη σταθερή κίνηση.
Οι διαθέσιμες από το προηγούμενο πείραμα θέσεις και κλίμακες των ανιχνευθέντων από τον Harris3D σημείων επί των εικόνων ενέργειας, προωθήθηκαν εδώ για τον υπολογισμό των Περιγραφέων HOG/HOF αυτή τη φορά επί των αρχικών frames. Η εξαγωγή των Περιγραφέων HOG/HOF σε προϋπολογισμένες, παρεχόμενες από τον χρήστη, θέσεις καθίσταται δυνατή μόνο στη νεότερη εκδόση της online υλοποίησης *stip-2.0-linux*[8], που περιλαμβάνει εκτελέσιμο κώδικα μόνο για 64-bit Linux λειτουργικό σύστημα. Αυτή η υλοποίηση χρησιμοποιήθηκε στο παρόν πείραμα με πανομοιότυπες επιλογές για τον περιγραφέα με αυτές των προηγούμενων πειραμάτων. Η μετέπειτα διαδικασία αναπαράστασης των δειγμάτων μέσω της προσέγγισης Bag-Of-Features και ταξινόμησης των δράσεων με μη γραμμικούς SVM ταξινομητές ακολουθείται καί σε αυτό το πείραμα. Να σημειώσουμε ότι κατά την κατασκευή του λεξιλογίου οπτικών λέξεων με τον αλγόριθμο k-means το συνολικό άθροισμα των εντός κλάσης αποστάσεων των δεδομένων από τα κεντροειδή για όλες τις κλάσεις υπολογίστηκε στην τιμή 317437.
Στον Πίνακα 4.5 παρουσιάζονται τα αποτελέσματα του παρόντος πειράματος και του πειράματος της ενότητας 4.3 για τη *μέση ακρίβεια* κάθε ταξινομητή και τη μέση τιμή της επί όλων των δράσεων (*mean average precision*). Το νέο εναλλακτικό σχήμα απέφερε

[8] http://www.irisa.fr/vista/Equipe/People/Laptev/download.html

	Harris3D + HOG/HOF	**Harris3D on EDCA Energy + HOG/HOF**
AnswerPhone	**39.9%**	38.0%
DriveCar	**89.7%**	87.1%
FightPerson	**88.7%**	87.1%
GetOutCar	43.6%	**45.0%**
HandShake	38.0%	**49.1%**
SitUp	36.2%	**37.6%**
mAP	56.0%	**57.32%**

Πίνακας 4.5: Αποτελέσματα Average Precision για καθεμιά από τις έξι δράσεις του πειράματος και mean Average Precision επί όλων των δράσεων για το καθιερωμένο σχήμα εξαγωγής χαρακτηριστικών (Αριστερά) και το εναλλακτικό σχήμα υπολογισμού των ανιχνεύσεων από τον Harris3D στις εικόνες *EDCA* ενέργειας και των περιγραφέων HOG/HOF στις αντίστοιχες θέσεις των αρχικών frames (Δεξιά).

αύξηση της απόδοσης κατά 1.32%, παρέχοντας υψηλότερη *μέση ακρίβεια* για τις τρείς από τις έξι δράσεις του πειράματος. Είναι εντυπωσιακό το κέρδος απόδοσης για τη δράση *HandShake* που ανέρχεται στο 11.1%. Για τις απαιτητικές δράσεις *SitUp* και *GetOutCar*, που στο προηγούμενο εναλλακτικό σχήμα παρουσίασαν τη μεγαλύτερη υποβάθμιση στο ποσοστό *μέσης ακρίβειας*, παρατηρούμε εδώ ισόποση αύξηση της απόδοσης συγκριτικά με το καθιερωμένο σχήμα κατά 1.4%. Παράλληλα, στις εναπομείνασες τρείς δράσεις το νέο εναλλακτικό σχήμα δεν οδήγησε σε μεγάλη μείωση της απόδοσης. Στο Σχήμα 4.7 απεικονίζονται οι καμπύλες *precision-recall* για τις έξι δράσεις στο παρόν πείραμα.
Συμπερασματικά, με βάση τα αποτελέσματα του παρόντος πειράματος, κρίνεται ευεργετική για την απόδοση της αναγνώρισης των ανθρώπινων δράσεων η ενσωμάτωση των χαρακτηριστικών κυρίαρχης ενέργειας στο στάδιο της ανίχνευσης χωροχρονικών σημείων ενδιαφέροντος. Δεδομένου ότι ο Ανιχνευτής Harris3D πυροδοτεί ανιχνεύσεις σε σημεία που χαρακτηρίζονται από υψηλές τιμές της εικόνας σε σχέση με τα γειτονικά τους στο χώρο και μη σταθερή κίνηση, η καταστολή οπτικά μη "σημαντικών" περιοχών της εικόνας, με κριτήριο την ένταση των διαμορφώσεων υφής και της ενέργειας ταλάντωσης, και η απόδοση εξεχουσών τιμών στις κυρίαρχες ακμές από τις *EDCA* εικόνες ενέργειας, δρομολογεί την ανίχνευση σε σημαίνουσας οπτικής πληροφορίας περιοχές.
Απο την άλλη, οι τιμές κυρίαρχης ενέργειας οδηγούν σε απώλεια διακριτικής ικανότητας για το στάδιο της περιγραφής, που συλλαμβάνει καλύτερα τις μετρήσεις των ιστογραμμάτων των gradients και της οπτικής ροής όταν επενεργεί επί των τοπικών τεμαχίων στον αρχικό όγκο βίντεο. Ενδεχομένως, η ενσωμάτωση χαρακτηριστικών αποδιαμόρφωσης των εικόνων (πλάτους και συχνότητας) με κάποιον αλγόριθμο *ESA* στην περιγραφή των τεμαχίων παράλληλα με τον Περιγραφέα HOG/HOF να οδηγήσει σε αυξημένες αποδόσεις αναγνώρισης, κάτι που δεν εντάσσεται στους στόχους της παρούσας διπλωματικής εργασίας.



(a) AnswerPhone (b) DriveCar

(c) FightPerson (d) GetOutCar

(e) HandShake (f) SitUp

Σχήμα 4.7: Καμπύλες *Precision-Recall* για καθεμιά από τις έξι δράσεις του πειράματος με το εναλλακτικό σχήμα ανίχνευσης στις εικόνες *EDCA* κυρίαρχης ενέργειας και περιγραφής στις αντίστοιχες θέσεις των αρχικών frames.

Κεφάλαιο 5

Ανιχνευτής Βασισμένος σε Χαρακτηριστικά Κυρίαρχης Χωροχρονικής Ενεργείας και Πειράματα Αναγνώρισης Ανθρώπινων Δράσεων

Στο Κεφάλαιο 4 διεξάγαμε πειράματα Αναγνώρισης Ανθρώπινων Δράσεων με Ανιχνευτή τον Harris3D και Περιγραφέα τον HOG/HOF στο πλαίσιο της Bag-Of-Features προσέγγισης και της SVM ταξινόμησης. Εξετάσαμε εναλλακτικά σχήματα ενσωμάτωσης χαρακτηριστικών κυρίαρχης *EDCA* χωρικής ενέργειας των frames των δειγμάτων βίντεο κατά το στάδιο της ανίχνευσης ή καί της περιγραφής των σημείων ενδιαφέροντος, με το ένα από αυτά να οδηγεί σε κέρδος απόδοσης του συνολικού πλαισίου αναγνώρισης. Στο παρόν κεφάλαιο θα χρησιμοποιήσουμε σε αντίστοιχα πειράματα έναν νέο Ανιχνευτή Χωροχρονικών Σημείων Ενδιαφέροντος που μελετάται για πρώτη φορά για το πρόβλημα της Αναγνώρισης Ανθρώπινων Δράσεων σε Βίντεο και στηρίζει την εξαγωγή ανιχνεύσεων σε χαρακτηριστικά κυρίαρχης Teager-Kaiser ενέργειας στο χώρο και το χρόνο του συνολικού τρισδιάστατου όγκου του δείγματος βίντεο που έχει υποστεί χωροχρονικό φιλτράρισμα με συστοιχίες Gabor φίλτρων. Σε συνδυασμό με τον αποτελεσματικό Περιγραφέα HOG/HOF και πανομοιότυπες επιλογές για τα υπόλοιπα στάδια του πλαισίου αναγνώρισης, θα δούμε ότι ο νέος Ανιχνευτής αποφέρει παραπλήσιες ή καλύτερες επιδόσεις από αυτές που λάβαμε με τη χρήση του Harris3D.

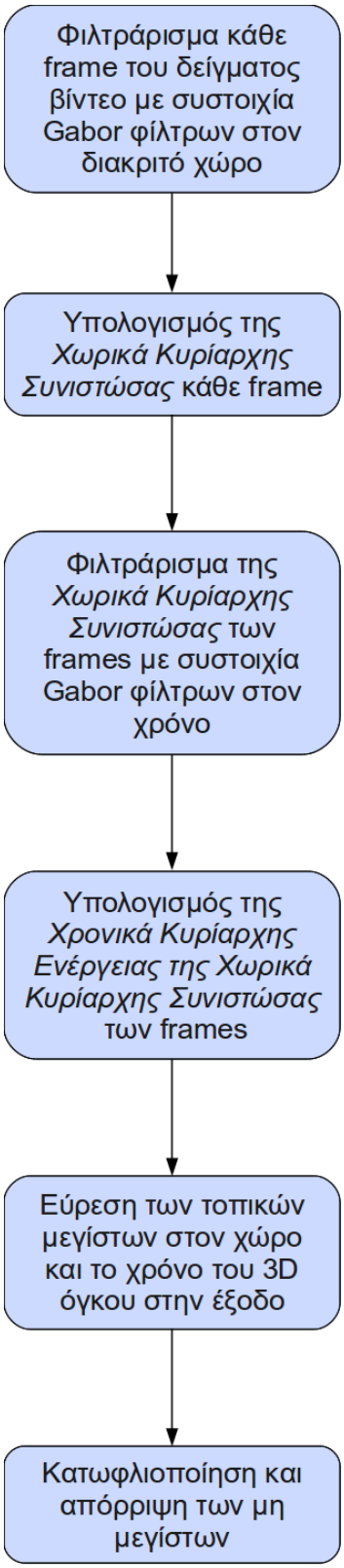
5.1 Ο Ανιχνευτής dca3D

Ο Ανιχνευτής dca3D αναπτύχθηκε με βάση πρόσφατες ιδέες των P.Maragos, D.Dimitriadis και G.Evangelopoulos, οι οποίοι ανέπτυξαν μια νέα μέθοδο χωροχρονικού φιλτραρίσματος και εξαγωγής χαρακτηριστικών ενέργειας για τη σύλληψη και τον εντοπισμό κίνησης (motion tracking) σε ακολουθίες εικόνων για το πρόβλημα της Αυτόματης Αναγνώρι-

σης Νοηματικής Γλώσσας, στο πλαίσιο του Dicta-Sign Project (Sign Language Recognition, Generation and Modelling with application in Deaf Communication). Η μεθοδολογία τους αυτή παρουσιάστηκε για πρώτη φορά στο συνέδριο του Dicta-Sign που πραγματοποιήθηκε τον Σεπτέμβριο του 2009 στο Πανεπιστήμιο του Surrey. Στα πλαίσια της παρούσας διπλωματικής εργασίας, παρουσιάζονται για πρώτη φορά πειράματα με χρήση των παραπάνω ιδεών ως σχήμα για την Ανίχνευση Χωροχρονικών Σημείων Ενδιαφέροντος στο πρόβλημα της Αναγνώρισης Ανθρώπινων Δράσεων σε Βίντεο. Πολύτιμη υπήρξε η συμβολή του G. Evangelopoulos στη διασαφήνιση των ιδεών που είχαν υιοθετηθεί στην πρωτότυπη εργασία καθώς και στις προτάσεις για τη θεμελίωση των διαφόρων σταδίων του νέου Ανιχνευτή που ονομάσαμε dca3D, και τον ευχαριστούμε ιδιαίτερα για αυτό.

Ο Ανιχνευτής dca3D πήρε το όνομά του εξαιτίας της εξαγωγής χαρακτηριστικών ενέργειας ταλάντωσης στο χώρο και το χρόνο με βάση την *Ανάλυση Κυρίαρχων Συνιστωσών (Dominant Component Analysis)*, της οποίας τη μορφή για δισδιάστατα σήματα διακριτού χώρου μελετήσαμε αναλυτικά στην υποενότητα 3.3.4.2. Η συνάρτηση απόκρισης (response function) του ανιχνευτή κατασκευάζεται ως αποτέλεσμα δύο διακριτών σταδίων. Σε πρώτο στάδιο, εφαρμόζεται φιλτραρίσμα κάθε frame του δείγματος βίντεο εισόδου σε πολλαπλές ζώνες συχνοτήτων στο διακριτό χώρο βάσει της συστοιχίας δισδιάστατων μιγαδικών φίλτρων Gabor που αναλύσαμε στην υποενότητα 3.3.1. Ακολούθως, με χρήση του κριτηρίου *EDCA* της σχέσης (3.3.21), σε κάθε frame λαμβάνουμε σε επίπεδο pixel τις φιλτραρισμένες εξόδους της αρχικής εικόνας από τα κυρίαρχα κανάλια, από αυτά δηλαδή που απέδωσαν την υψηλότερη τιμή Teager-Kaiser χωρικής ενέργειας σε κάθε pixel. Η προκύπτουσα για κάθε frame εικόνα, που μπορεί να ιδωθεί καί ως ανακατασκευή της αρχικής *ευρυζωνικής* εικόνας από τις φιλτραρισμένες κυρίαρχες "ενεργειακά"συνιστώσες, είναι αυτή που ονομάζουμε από εδώ και στο εξής *Χωρικά Κυρίαρχη Συνιστώσα (Spatial Dominant Component (SDC))*. Το πραγματικό μέρος της *Χωρικά Κυρίαρχης Συνιστώσας* όλων των frames του δείγματος βίντεο αποτελεί τον τρισδιάστατο όγκο με τον οποίο τροφοδοτούμε το δεύτερο στάδιο του ανιχνευτή, όπου αρχικά η είσοδος *SDC* φιλτράρεται στο χρόνο με μια συστοιχία μονοδιάστατων χρονικών Gabor φίλτρων, κατάλληλα διατεταγμένων στον χώρο συχνοτήτων. Αφού εξάγουμε την ενέργεια Teager-Kaiser κάθε χωροχρονικού σημείου του όγκου εισόδου από τις εξόδους όλων των καναλιών της συστοιχίας των χρονικών φίλτρων, εφαρμόζεται και πάλι το κριτήριο *EDCA* με το οποίο υπολογίζεται η *Χρονικά Κυρίαρχη Ενέργεια (Temporal Dominant Energy (TDE))* για κάθε σημείο. Οι τιμές της *Χρονικά Κυρίαρχης Ενέργειας της Χωρικά Κυρίαρχης Συνιστώσας* των frames συνιστούν τις τιμές της συνάρτησης απόκρισης του ανιχνευτή για το δείγμα βίντεο εισόδου. Να σημειώσουμε ότι όποτε αναφερόμαστε στη *Χωρικά Κυρίαρχη Συνιστώσα* εννοούμε το πραγματικό της μέρος, καθώς τα φίλτρα Gabor στο χρόνο που επιλέξαμε να χρησιμοποιούμε είναι πραγματικά και όχι μιγαδικά.

Η διαδικασία που ακολουθείται για τη λήψη των ανίχνευσεων από τον dca3D είναι η εύρεση των *τοπικών μεγίστων στο χώρο και το χρόνο* της συνάρτησης απόκρισής του. Για την απόρριψη "αδύναμων"τοπικών μεγίστων υιοθετούμε τη λογική της *καταστολής των μη μεγίστων (non-maxima suppression)*, διατηρώντας μόνο εκείνα που έχουν τιμή μεγαλύτερη ενός ποσοστού της μέγιστης τιμής επί όλων των τοπικών μεγίστων.



Σχήμα 5.1: Σχηματικό διάγραμμα των διακριτών σταδίων του Ανιχνευτή dca3D.

Ο Ανιχνευτής dca3D αρχικά επενεργεί σε κάθε frame ξεχωριστά, επιτελώντας ζωνοπερατό φιλτράρισμα της εικόνας και αποδίδοντας σε κάθε pixel την έξοδό της από το κανάλι που παρουσίασε την υψηλότερη τιμή Teager-Kaiser ενέργειας. Με αυτό τον τρόπο λαμβάνεται για κάθε frame η φιλτραρισμένη έξοδος της εικόνας από τα πιο "ενεργά", αναφορικά με την ενέργεια των πηγών που προκάλεσαν την ταλάντωση στο διακριτό χώρο, κανάλια σε επίπεδο pixel οδηγώντας σε μία μη θορυβώδη και επικρατούσα με όρους υφής αναπαράσταση της εικόνας. Έπειτα, το φιλτράρισμα με ζωνοπερατά φίλτρα Gabor στο χρόνο καθιστά δυνατή την εκτίμηση της Teager-Kaiser χρονικής ενέργειας στα επιμέρους κανάλια της συστοιχίας, με καθεμιά από αυτές να μπορεί να ιδωθεί ως ταυτόχρονη σύλληψη του πλάτους και της συχνότητας ταλάντωσης διαφορετικής συχνότητας διαμορφώσεων στο χρόνο και επομένως κινήσεων που λαμβάνουν χώρα στο δείγμα βίντεο εισόδου. Ο υπολογισμός της *Χρονικά Κυρίαρχης Ενέργειας* αναδεικνύει σε κάθε σημείο (x, y, t) της ακολουθίας εικόνων την επικρατούσα ενέργεια ταλάντωσης στο χρόνο, καταστέλλοντας τη συμμετοχή στη συνάρτηση απόκρισης της ενέργειας που προκύπτει από ζώνες συχνοτήτων μη "ενεργές"σε εκείνο το σημείο.
Παρακάτω θα μελετήσουμε σε ξεχωριστές παραγράφους και σε μεγαλύτερο επίπεδο λεπτομέρειας τα δύο διακριτά στάδια που ακολουθούνται για την εξαγωγή της συνάρτησης απόκρισης του Ανιχνευτή dca3D.

Υπολογισμός της *Χωρικά Κυρίαρχης Συνιστώσας* των frames. Για το ζωνοπερατό φιλτράρισμα των frames στον διακριτό χώρο χρησιμοποιούμε τη συστοιχία των 40 μιγαδικών ισοτροπικών δισδιάστατων φίλτρων Gabor που αναλύσαμε στην υποενότητα 3.3.1. Ακολουθεί η *Ανάλυση Κυρίαρχων Συνιστωσών βασισμένη στην Ενέργεια* με τον ίδιο τρόπο όπως την εφαρμόσαμε στην υποενότητα 4.4.1 για την εξαγωγή των εικόνων κυρίαρχης ενέργειας, με τη διαφορά ότι εδώ στο τελευταίο βήμα υπολογίζουμε μόνο τους δείκτες $i(x, y)$ των κυρίαρχων καναλιών με βάση πάλι το κριτήριο *EDCA*

$$\begin{aligned} i(x,y) &= \arg\max_{1\le k\le K}\{\Gamma_k(x,y)\} \\ \Gamma_k(x,y) &= \Phi\left[I * g_k\right](x,y) \end{aligned} \tag{5.1.1}$$

όπου $K = 40$ ο αριθμός των φίλτρων της συστοιχίας, g_k το $k-$οστό φίλτρο, I το παρόν frame και Φ ο δισδιάστατος ενεργειακός τελεστής Teager-Kaiser.
Η *Χωρικά Κυρίαρχη Συνιστώσα (Spatial Dominant Component (SDC))* για το frame $I(x, y)$ προκύπτει λαμβάνοντας στο pixel (x, y) την έξοδο της εικόνας I από το κανάλι $i(x, y)$

$$SDC(x,y) = \left(I * g_{i(x,y)}\right)(x,y) \tag{5.1.2}$$

Η χωρική κλίμακα ανίχνευσης (spatial scale) του Ανιχνευτή dca3D ταυτίζεται με την τυπική απόκλιση σ της Gaussian περιβάλλουσας του κυρίαρχου φίλτρου Gabor $i(x, y)$ σε κάθε pixel (x, y), πολλαπλασιασμένης επί του πλάτους του παρόντος frame που φιλτράρεται

$$\sigma(x,y) = \frac{\sqrt{\frac{\ln 2}{\gamma}}}{2\pi r_{i(x,y)}} \cdot \text{frame width} \tag{5.1.3}$$

όπου $\gamma = \frac{1}{9}$ και r_k η ακτινική κεντρική συχνότητα του $k-$οστού φίλτρου (βλ. (3.3.2)). Το πλάτος frame μπορεί να ιδωθεί εδώ ως συχνότητα δειγματοληψίας και ο πολλαπλασιασμός επί αυτού γίνεται για να λάβουμε από την κανονικοποιημένη κλίμακα την αντίστοιχη διακριτή τιμή της σε pixels, προσαρμοσμένης για το παρόν frame. Οι κανονικοποιημένες κλίμακες, επομένως, του Ανιχνευτή dca3D είναι πέντε, όσες και οι κλίμακες της συστοιχίας φίλτρων που χρησιμοποιείται.
Η συνέλιξη κάθε frame με τα μιγαδικά φίλτρα g_k παράγει στην έξοδο επίσης μιγαδικές τιμές για τη *Χωρικά Κυρίαρχη Συνιστώσα*. Ωστόσο, λόγω της επιλογής μας να χρησιμοποιήσουμε συστοιχία πραγματικών Gabor φίλτρων στο χρόνο για το δεύτερο στάδιο του ανιχνευτή, θεωρούμε ως *Χωρικά Κυρίαρχη Συνιστώσα* το πραγματικό της μέρος. Η παραπάνω διαδικασία που περιγράψαμε επαναλαμβάνεται για όλα τα frames ξεχωριστά της αρχικής ακολουθίας εικόνων f και στις παραπάνω σχέσεις αντικαθίσταται η εικόνα I από το τρέχον frame $f(x, y, t)$. Ο τρισδιάστατος όγκος $SDC(x, y, t)$, με τις *Χωρικά Κυρίαρχες Συνιστώσες* όλων των frames αποτελεί την είσοδο για το δεύτερο στάδιο του Ανιχνευτή dca3D.

Υπολογισμός της *Χρονικά Κυρίαρχης Ενέργειας* της *Χωρικά Κυρίαρχης Συνιστώσας* των frames. Πριν αναφέρουμε τις επιλογές που έγιναν για τη συστοιχία των πραγματικών Gabor φίλτρων στο χρόνο, θα μελετήσουμε την περίπτωση του υπολογισμού της Teager-Kaiser ενέργειας της φιλτραρισμένης *στενοζωνικής* συνιστώσας $s(t) = x(t) * g(t)$ ενός *ευρυζωνικού* μονοδιάστατου σήματος $x(t)$ που έχει υποστεί ζωνοπερατό φιλτράρισμα με ένα φίλτρο Gabor $g(t)$. Όπως έχουμε αναφέρει στο Κεφάλαιο 3, οι διάφοροι αλγόριθμοι αποδιαμόρφωσης *ESA* δεν μπορούν, λόγω εγγενών περιορισμών, να διαχειριστούν την περίπτωση *ευρυζωνικών* σημάτων. Η εκτίμηση της Teager-Kaiser ενέργειας προαπαιτεί επομένως το ζωνοπέρατο φιλτράρισμα του αρχικού σήματος όπου προτιμούνται τα φίλτρα Gabor για λόγους που έχουμε εξηγήσει, όπως η ταυτόχρονη βέλτιστη τοποθέτησή τους στα πεδία χρόνου και συχνότητας. Η λογική που ακολουθούμε σε αυτό το δεύτερο στάδιο του Ανιχνευτή dca3D ως προς το φιλτράρισμα και την εκτίμηση των ενεργειών καναλιού βασίζεται στην εργασία των D. Dimitriadis και P. Maragos [43] για την αποδιαμόρφωση σημάτων φωνής από τον αλγόριθμο *Gabor ESA*.
Το μονοδιάστατο πραγματικό φίλτρο Gabor $g(t)$ έχει κρουστική απόκριση

$$g(t) = \exp(-\beta^2 t^2)\cos(\omega_c t) \tag{5.1.4}$$

όπου β είναι η *παράμετρος εύρους ζώνης (bandwidth parameter)* και $\omega_c = 2\pi f_c$ η *γωνιακή κεντρική συχνότητα* του φίλτρου.
Η πρώτη και η δεύτερη παράγωγός του υπολογίζονται σε κλειστό τύπο ως εξής

$$\frac{\mathrm{d}g(t)}{\mathrm{d}t} = \big(-2\beta^2 t\cos(\omega_c t) - \omega_c \sin(\omega_c t)\big)\exp(-\beta^2 t^2) \tag{5.1.5}$$

$$\frac{\mathrm{d}^2 g(t)}{\mathrm{d}t^2} = \big(4\beta^2\omega_c t\sin(\omega_c t) + (4\beta^4 t^2 - 2\beta^2 - \omega_c^2)\cos(\omega_c t)\big)\exp(-\beta^2 t^2) \tag{5.1.6}$$

Επομένως, λόγω της αντιμεταθετικής ιδιότητας των τελεστών συνέλιξης και παραγώγισης μεταξύ τους, η έξοδος της φιλτραρισμένης συνιστώσας $s(t) = x(t) * g(t)$ από τον

μονοδιάστατο ενεργειακό τελεστή Teager-Kaiser δίνεται ως εξής

$$\Psi[s(t)] = \left(x(t) * \frac{\mathrm{d}g(t)}{\mathrm{d}t}\right)^2 - (x(t) * g(t)) \left(x(t) * \frac{\mathrm{d}^2 g(t)}{\mathrm{d}t^2}\right) \tag{5.1.7}$$

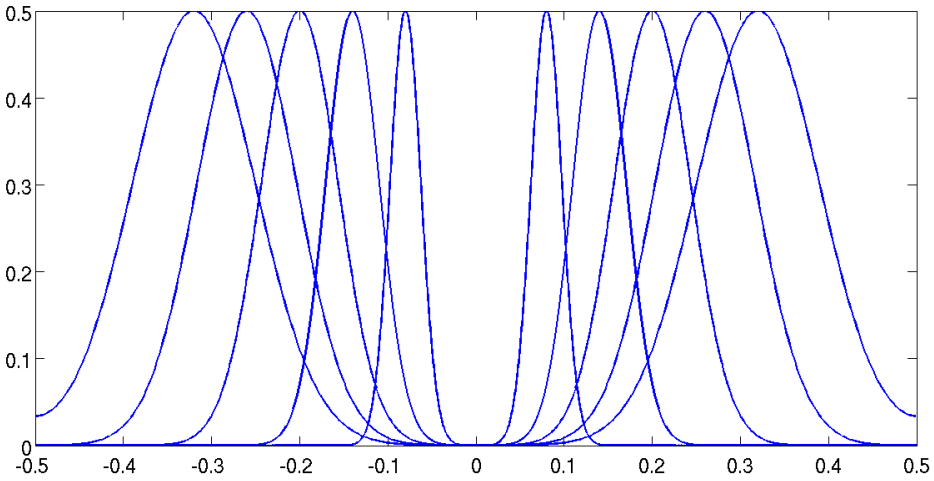
Στην περίπτωση διακριτών σημάτων η συνέλιξη πραγματοποιείται μεταξύ του διακριτού σήματος $x[n]$ και των διακριτών μορφών της συνάρτησης Gabor και των παραγώγων της

$$g^{(m)}[n] = \frac{\mathrm{d}^m}{\mathrm{d}t^m} g(t)|_{t=nT} \tag{5.1.8}$$

όπου T η περίοδος δειγματοληψίας, ενώ η έξοδος της σχέσης (5.1.7) δειγματοληπτείται επίσης σε χρονικές στιγμές $t = nT$.
Αυτή την προσέγγιση του *Gabor Τελεστή Ενέργειας Teager-Kaiser* των D. Dimitriadis και P. Maragos [43] χρησιμοποιούμε στο δεύτερο στάδιο του Ανιχνευτή dca3D. Εν προκειμένω, αντί του σήματος $x[n]$ στην είσοδο έχουμε τη μήτρα της *Χωρικά Κυρίαρχης Συνιστώσας* των frames $SDC(x, y, t)$ να φιλτράρεται στο χρόνο από φίλτρα Gabor σε πολλαπλές ζώνες και τη συχνότητα δειγματοληψίας $f_s = \frac{1}{T}$ να ισούται εδώ με το frame rate του αρχικού δείγματος βίντεο εισόδου.
Σύμφωνα με τους Dimitriadis και Maragos [43], οι κεντρικές συχνότητες $f_{c,l}$ των Gabor φίλτρων της συστοιχίας επιλέγονται ώστε να εμπίπτουν στο πεδίο συχνοτήτων $[0, f_s/2]$. Πολλές διαφορετικές επιλογές μπορούν να υιοθετηθούν για τον σχεδιασμό της συστοιχίας των πραγματικών Gabor φίλτρων στο χρόνο ενώ πολλές σχετικές ιδέες μπορούν να αντληθούν από την εργασία των Maragos et al. [37]. Οι κυριότερες αποφάσεις σχετίζονται με τον καθορισμό του τρόπου με τον οποίο προοδεύουν οι κεντρικές συχνότητες των φίλτρων και με την επιλογή των τιμών του εύρους ζώνης τους. Ύστερα από πολλούς πειραματισμούς, καταλήξαμε στο σχεδιασμό μιας συστοιχίας πέντε φίλτρων Gabor με σταθερό *εύρος ζώνης οκτάβας (octave bandwidth)* $B = 0.75$ οκτάβες, για τις ανάγκες των πειραμάτων της παρούσας διπλωματικής εργασίας. Επιλέξαμε οι κεντρικές συχνότητες των φίλτρων να έχουν σταθερό βήμα 1.5 Hz, ξεκινώντας από την τιμή 2 Hz, έχοντας υπόψη μας ότι το frame rate των δειγμάτων βίντεο της *Hollywood2 Actions Dataset* πάνω στην οποία πραγματοποιήσαμε πειράματα κυμαίνεται μεταξύ των τιμών 24 και 25. Οι κέντρικες συχνότητες $f_{c,l}$ δίνονται επομένως από τη σχέση

$$f_{c,l} = 2 + (l - 1) \cdot 1.5 \quad , \qquad 1 \leq l \leq 5 \tag{5.1.9}$$

Η προκύπτουσα τοποθέτηση των φίλτρων στον χώρο συχνοτήτων απεικονίζεται στο Σχήμα 5.2.



Σχήμα 5.2: Απόκριση συχνότητας των πραγματικών μονοδιάστατων Gabor φίλτρων της συστοιχίας, εικονιζόμενη για μοναδιαία τιμή της συχνότητας δειγματοληψίας.

Οι παράμετροι β_l των φίλτρων προκύπτουν από τη σχέση

$$\beta_l = K \frac{\pi f_{c,l}}{\sqrt{\ln 2}} \tag{5.1.10}$$

όπου $f_{c,l}$ οι κεντρικές συχνότητες που δίνονται από τη σχέση (5.1.9) και K μια σταθερά που εξαρτάται από το *εύρος ζώνης οκτάβας* B των φίλτρων ως εξής

$$K = \frac{2^B - 1}{2^B + 1} \tag{5.1.11}$$

Έχοντας πλέον ορίσει τη συστοιχία Gabor φίλτρων στο χρόνο που χρησιμοποιούμε, η διαδικασία που ακολουθείται βασίζεται στην ίδια λογική με το πρώτο στάδιο του Ανιχνευτή dca3D. Αν συμβολίσουμε με $SDC(x, y, t)$ τον τρισδιάστατο όγκο της *Χωρικά Κυρίαρχης Συνιστώσας* των frames του αρχικού δείγματος βίντεο, αυτός φιλτράρεται στη διάσταση του χρόνου με καθένα από τα φίλτρα g_l της σχέσης (5.1.4) και στην έξοδο των καναλιών υπολογίζουμε τις ενέργειες καναλιού $\Psi[SDC(x, y, t) * g_l(t)]$ βάσει του *Gabor Τελεστή Ενέργειας Teager-Kaiser* της σχέσης (5.1.7). Για την αποφυγή σύγχυσης σχετικά με τον τρόπο που εφαρμόζεται το φιλτράρισμα στο χρόνο του τρισδιάστατου όγκου με τα μονοδιάστατα φίλτρα Gabor, μπορούμε να φανταστούμε την εξέλιξη των τιμών καθενός pixel στα διάφορα frames ως ένα μονοδιάστατο σήμα που φιλτράρεται ξεχωριστά στο χρόνο. Οι δείκτες των κυρίαρχων καναλιών της συστοιχίας επιλέγονται πάλι με βάση την *Ανάλυση Κυρίαρχων Συνιστωσών βασισμένη στην Ενέργεια (EDCA)*, με κριτήριο δηλαδή την υψηλότερη τιμή Teager-Kaiser χρονικής ενέργειας μεταξύ των πέντε καναλιών της συστοιχίας

$$\begin{aligned} i(x, y, t) &= \arg \max_{1 \le l \le L} \{\Gamma_l(x, y, t)\} \\ \Gamma_l(x, y, t) &= \Psi\left(SDC(x, y, t) * g_l(t)\right) \end{aligned} \tag{5.1.12}$$

όπου $L = 5$ ο αριθμός των φίλτρων της συστοιχίας, g_l το $l-$οστό φίλτρο της μορφής (5.1.4) *παραμέτρου εύρους ζώνης* β_l και κεντρικής συχνότητας $f_{c,l}$, $SDC(x, y, t)$ η *Χωρικά Κυρίαρχη Συνιστώσα* του frame t και Ψ ο *Gabor Τελεστής Ενέργειας Teager-Kaiser* της σχέσης (5.1.7).

Η τελική συνάρτηση απόκρισης του Ανιχνευτή dca3D προκύπτει από τη *Χρονικά Κυρίαρχη Ενέργεια της Χωρικά Κυρίαρχης Συνιστώσας (Temporal Dominant Energy of Spatial Dominant Component (SDCTDE))* όπου σε κάθε σημείο (x, y, t) της αρχικής ακολουθίας εικόνων $f(x, y, t)$ αντιστοιχίζεται η τιμή της Teager-Kaiser ενέργειας από την έξοδο του κυρίαρχου καναλιού $i(x, y, t)$ της σχέσης (5.1.12)

$$R(x, y, t) = \Psi \left[SDC(x, y, t) * g_{i(x,y,t)}(t) \right] (x, y, t) \tag{5.1.13}$$

Η χρονική κλίμακα ανίχνευσης (temporal scale) του Ανιχνευτή dca3D ορίζεται με παρόμοιο τρόπο όπως καί για τη χωρική κλίμακα, από την τυπική απόκλιση δηλαδή της Gaussian περιβάλλουσας του κυρίαρχου από τη σχέση (5.1.12) φίλτρου $i(x, y, t)$, πολλαπλασιασμένης επί τη συχνότητα δειγματοληψίας f_s

$$\tau(x, y, t) = \text{round} \left[\frac{1}{\sqrt{2}\beta_{i(x,y,t)}} \cdot f_s \right] \tag{5.1.14}$$

όπου β_l η *παράμετρος εύρους ζώνης* του $l-$οστού φίλτρου που δίνεται από τη σχέση (5.1.10) και f_s το frame rate του δείγματος βίντεο εισόδου. Ο πολλαπλασιασμός επί τη συχνότητα δειγματοληψίας γίνεται πάλι εδώ ώστε να λάβουμε την προσαρμοσμένη τιμή χρονικής κλίμακας σε frames για το τρέχον δείγμα βίντεο ενώ η στρογγυλοποίηση εφαρμόζεται για τη λήψη ακεραίων τιμών χρονικής κλίμακας, διακριτών μεταξύ τους. Οι διακριτές τιμές των χρονικών κλιμάκων τ είναι πέντε, όσα καί τα φίλτρα της συστοιχίας των μονοδιάστατων φίλτρων Gabor στο χρόνο.
Τα Χωροχρονικά Σημεία Ενδιαφέροντος που εξάγει ο Ανιχνευτής dca3D αναζητούνται στα τοπικά μέγιστα της συνάρτησης απόκρισης (5.1.13). Αυτά υπολογίζονται στον διακριτό τρισδιάστατο όγκο των τιμών της συνάρτησης R από τις διαφορές τιμών κάθε σημείου (x, y, t) από τα 26 γειτονικά του σημεία στον χώρο και χρόνο. Για την αποφυγή επίπλαστων τεχνητών ανιχνεύσεων στα χωρικά σύνορα των εικόνων, απορρίπτουμε τις ανιχνεύσεις που απέχουν μέχρι 5 pixels από τα σύνορα των frames.
Ο όγκος των τοπικών μεγίστων της συνάρτησης R είναι αρκετά μεγάλος για κάθε δείγμα βίντεο εισόδου, περιλαμβάνοντας καί ανιχνεύσεις που αντιστοιχούν σε χαμηλές τιμές της συνάρτησης απόκρισης. Για τούτο ακολουθούμε τη λογική της *καταστολής των μη μεγίστων (non-maxima suppression)* εφαρμόζοντας ένα ολικό κατώφλι στις τιμές των τοπικών μεγίστων. Το τελικό σύνολο ανιχνεύσεων του Ανιχνευτή dca3D αποτελείται από τα τοπικά μέγιστα με τιμές της συνάρτησης απόκρισης που υπερβαίνουν το ολικό κατώφλι.
Συμπερασματικά, ο Ανιχνευτής dca3D εκτελεί χωροχρονικό φιλτράρισμα με συστοιχίες Gabor φίλτρων και ακολουθεί την *Ανάλυση Κυρίαρχων Συνιστωσών βασισμένη στην Ενέργεια (Energy-based Dominant Component Analysis)* τόσο στον διακριτό χώρο όσο και στον διακριτό χρόνο. Στην περίπτωση του διακριτού χώρου εξάγει σε κάθε frame

μια φιλτραρισμένη σε επίπεδο pixel εκδοχή της αρχικής εικόνας από τις εξόδους των κυρίαρχων καναλιών που αντιστοιχούν στις κυρίαρχες *στενοζωνικές* συνιστώσες της εικόνας, με κριτήριο την υψηλότερη τιμή ενέργειας ταλάντωσης. Σε κάθε χωροχρονικό σημείο του προκύπτοντα τρισδιάστατου όγκου αντιστοιχίζεται μια τιμή κυρίαρχης ενέργειας ταλάντωσης αυτή τη φορά στον χρόνο, συλλαμβάνοντας έτσι σε κάθε σημείο το τετράγωνο του γινομένου πλάτους και συχνότητας της πιο ενεργής διαμόρφωσης στον χρόνο, εννοιολογικά δηλαδή την ενέργεια των κυρίαρχων πηγών διαφορετικής έντασης και συχνότητας κινήσεων που εμφανίζονται στο δείγμα βίντεο.
Αξίζει να παρατηρήσουμε ότι η συνάρτηση απόκρισης του Ανιχνευτή *Cuboid* που μελετήσαμε στην υποενότητα 2.2.2 περιλαμβάνει επίσης φιλτράρισμα με χρονικά φίλτρα Gabor, αφού πρώτα εφαρμοστεί Gaussian εξομάλυνση των frames στον χώρο. Ωστόσο, ο τρόπος που θεμελιώνεται η συνάρτηση απόκρισης φανερώνει την αναζήτηση υψηλού πλάτους περιοδικών κινήσεων, με τον Ανιχνευτή *Cuboid* να αδυνατεί να συλλάβει αρμονικές διαμορφώσεις χαμηλού πλάτους ταλάντωσης αλλά υψηλού μέτρου συχνότητας, περιπτώσεις οι οποίες ανιχνεύονται αποτελεσματικά από τον Ανιχνευτή dca3D.
Στην επόμενη ενότητα θα καταστούν σαφείς περαιτέρω λεπτομέρειες υλοποίησης του Ανιχνευτή dca3D, θα παρουσιαστούν στοιχεία για τους χρόνους εκτέλεσης καθώς και χαρακτηριστικά αποτελέσματα αναφορικά με τη συνάρτηση απόκρισης και τα εξαγόμενα σημεία ενδιαφέροντος από τον Ανιχνευτή dca3D σε μεμονωμένα δείγματα βίντεο στα οποία πραγματοποιήσαμε σχετικά πειράματα.

5.2 Λεπτομέρειες Υλοποίησης - Πειραματικά Αποτελέσματα σε Δείγματα Βίντεο

Το σχήμα του Ανιχνευτή dca3D, από την ανάγνωση του δείγματος βίντεο εισόδου ως την παραγωγή των θέσεων και κλιμάκων των τελικών ανιχνεύσεων στην έξοδο, υλοποιήθηκε εξ'ολοκλήρου στο λογισμικό MATLAB. Θα παραθέσουμε σε ξεχωριστές παραγράφους τις λεπτομέρειες υλοποίησης για τα στάδια του υπολογισμού της *Χωρικά Κυρίαρχης Συνιστώσας*, της *Χρονικά Κυρίαρχης Ενέργειας* αυτής και της εξαγωγής των τοπικών μεγίστων της συνάρτησης απόκρισης του Ανιχνευτή dca3D.

Υπολογισμός της *Χωρικά Κυρίαρχης Συνιστώσας* των frames. Για τον υπολογισμό της *Χωρικά Κυρίαρχης Συνιστώσας (Spatial Dominant Component)* των frames του αρχικού δείγματος βίντεο εισόδου, η υλοποίηση που εφαρμόζεται είναι πανομοιότυπη μέχρι ενός σημείου με αυτή που περιγράψαμε στην υποενότητα 4.4.1 για την εξαγωγή των εικόνων *EDCA* Κυρίαρχης Ενέργειας. Γι'αυτό δε θα την αναλύσουμε διεξοδικά εδώ, παρά μόνο το τελευταίο βήμα όπου εισάγεται η διαφοροποίηση στις δύο υλοποιήσεις. Φθάνοντας στο σημείο όπου έχουν υπολογιστεί οι φιλτραρισμένες έξοδοι του τρέχοντος frame από τα 40 δισδιάστατα μιγαδικά φίλτρα και τις παραγώγους τους καθώς και οι τιμές Teager-Kaiser Ενέργειας στην έξοδο κάθε καναλιού της συστοιχίας, στο τελευταίο βήμα αντί να υπολογίσουμε την *EDCA* Κυρίαρχη Ενέργεια διατηρούμε μόνο τους δείκτες των κυρίαρχων καναλιών σε κάθε pixel. Οι τελευταίοι θα χρησιμεύσουν στον υπολογισμό της *Χωρικά Κυρίαρχης Συνιστώσας* καθώς καί της χωρικής

κλίμακας ανίχνευσης από τη σχέση (5.1.3) για κάθε επιλεγμένο από τον ανιχνευτή σημείο. Ωστόσο, το τρέχον βήμα υλοποιείται επίσης με τη συνάρτηση `max`, με την οποία αυτή τη φορά επιστρέφονται οι δείκτες και όχι οι τιμές ενέργειας των κυρίαρχων καναλιών. Για τούτο, ο χρόνος εκτέλεσης ανά frame του συγκεκριμένου βήματος παραμένει ουσιαστικά αμετάβλητος. Σε τελευταίο στάδιο, υπολογίζουμε τη *Χωρικά Κυρίαρχη Συνιστώσα* αποδίδοντας σε κάθε pixel τη φιλτραρισμένη έξοδο του τρέχοντος frame από το κυρίαρχο κατά *EDCA* κανάλι της συστοιχίας. Υπολογιστικά αυτό απαιτεί έναν επαναληπτικό βρόχο (`for` loop) 40 επαναλήψεων, όσα και τα κανάλια της συστοιχίας που χρησιμοποιούμε, και δεν επιβαρύνει σημαντικά τον συνολικό χρόνο υπολογισμού της *Χωρικά Κυρίαρχης Συνιστώσας* για κάθε frame, όπως θα δούμε παρακάτω.
Η παραπάνω διαδικασία επαναλαμβάνεται για όλα τα frames της αρχικής ακολουθίας εικόνων, με το *Spatial Dominant Component* του συνολικού τρισδιάστατου όγκου να αποθηκεύεται σε δομή *array* για να προωθηθεί έπειτα ως είσοδος στο δεύτερο στάδιο του ανιχνευτή. Για την αποφυγή προβλημάτων πλήρωσης της μνήμης RAM, η επεξεργασία του βίντεο γίνεται σε 4-5 τμήματα μικρότερης διάρκειας και τα ενδιάμεσα αποτελέσματα καθενός από αυτά απομακρύνονται προσωρινά από τη μνήμη προκειμένου να εκτελέσουμε τον αλγόριθμο στο επόμενο τμήμα. Ακόμα, όσο προχωρά η διαδικασία σε επόμενα τμήματα βίντεο, εκείνα που έχουν υποστεί επεξεργασία προηγουμένως διαγράφονται σταδιακά από τη μνήμη RAM. Αφού εξαχθούν αποτελέσματα καί για το τελευταίο τμήμα, οι πίνακες με τη *Χωρικά Κυρίαρχη Συνιστώσα* των επιμέρους τμημάτων επαναφέρονται στη μνήμη και συνενώνονται σε έναν πίνακα με μέγεθος όσο το αρχικό δείγμα βίντεο.

Υπολογισμός της *Χρονικά Κυρίαρχης Ενέργειας* της *Χωρικά Κυρίαρχης Συνιστώσας* των frames. Για την υλοποίηση του δεύτερου σταδίου του Ανιχνευτή dca3D αρχικά κατασκευάζεται η συστοιχία μονοδιάστατων πραγματικών φίλτρων Gabor που είναι όμοια με αυτή που απεικονίζεται στο πεδίο συχνότητας στο Σχήμα 5.2. Υπολογίζουμε για όλα τα φίλτρα και τις παραγώγους τους τις κρουστικές αποκρίσεις από τις σχέσεις (5.1.4), (5.1.5) και (5.1.6) αντίστοιχα, όπου οι *παράμετροι εύρους ζώνης* β_l και οι κεντρικές συχνότητες $f_{c,l}$ δίνονται από τις σχέσεις (5.1.10) και (5.1.9), διαιρεμένες με τη συχνότητα δειγματοληψίας f_s που ταυτίζεται με το frame rate του δείγματος βίντεο εισόδου. Οι Gaussian περιβάλλουσες $\exp(-\beta^2 t^2)$ των φίλτρων και των παραγώγων τους κανονικοποιούνται με βάση την $\mathcal{L}_1$ νόρμα και όλες οι αποκρίσεις είναι μηδενικής μέσης τιμής. Κάθε φίλτρο της συστοιχίας δειγματοληπτείται στον χρόνο και η απόκρισή του υπολογίζεται στο διακριτό διάστημα $(-N_l, N_l)$ όπου $N_l = \text{ceil}(4/b_{l,norm} + 1)$ όπου $b_{l,norm} = b_l/f_s$. Ακολουθούν οι συνελίξεις στον διακριτό χρόνο του τρισδιάστατου όγκου $SDC(x, y, t)$ της *Χωρικά Κυρίαρχης Συνιστώσας* με καθένα από τα φίλτρα και τις παραγώγους τους και ο υπολογισμός της ενέργειας στην έξοδο κάθε καναλιού από τη σχέση του *Gabor Τελεστή Ενέργειας Teager-Kaiser*. Παρόμοια με το πρώτο στάδιο, το φιλτράρισμα και ο υπολογισμός των ενεργειών καναλιού γίνεται σε περίπου 10 χωρικά τμήματα μικρότερου μεγέθους σε pixels, καθώς η απευθείας επεξεργασία του συνολικού όγκου SDC επιφέρει μεγάλη κατανάλωση μνήμης RAM. Για κάθε τμήμα οι ενέργειες καναλιού συνενώνονται σε έναν ενιαίο πίνακα και έπειτα, με χρήση της συνάρτησης `max`, λαμβάνουμε τους δείκτες και τις τιμές ενέργειας των κυρίαρχων καναλιών. Οι δείκτες κυρίαρχων καναλιών και οι τιμές της *Χρονικά Κυρίαρχης Ενέργειας* των επιμέρους τμη-

μάτων επαναφέρονται στη μνήμη και συγκεντρώνονται σε δύο ενιαίους πίνακες, οι οποίοι αναδιατάσσονται στο μέγεθος του αρχικού δείγματος βίντεο. Οι τιμές των δεικτών θα χρησιμεύσουν για τον προσδιορισμό της χρονικής κλίμακας του ανιχνευτή σε κάθε σημείο από τη σχέση (5.1.14) ενώ οι τιμές της *Χρονικά Κυρίαρχης Ενέργειας* της *Χωρικά Κυρίαρχης Συνιστώσας* συνιστούν τις τιμές της συνάρτησης απόκρισης (5.1.13) του Ανιχνευτή dca3D.

Εύρεση των τοπικών μεγίστων της Συνάρτησης Απόκρισης του Ανιχνευτή dca3D. Έχοντας εξάγει τον τρισδιάστατο όγκο των τιμών της Συνάρτησης Απόκρισης του Ανιχνευτή dca3D, αναζητούμε τα τοπικά μέγιστα αυτής στον χώρο και το χρόνο. Ως τοπικό μέγιστο στον τρισδιάστατο διακριτό χωροχρόνο ορίζεται ένα σημείο (x, y, t) που αντιστοιχίζεται σε τιμή μεγαλύτερη των τιμών των 26 γειτονικών του, 8 εκ των οποίων βρίσκονται στο ίδιο frame t, 9 στο frame $t-1$ και 9 στο frame $t+1$. Απορρίπτουμε τις ανιχνεύσεις σημείων που εντοπίζονται εντός των χωρικών συνόρων μήκους 5 pixels των frames.
Για τον αποκλεισμό "αδύναμων"ανιχνεύσεων, εφαρμόζουμε ένα ολικό κατώφλι (global threshold) στις τιμές των τοπικών μεγίστων. Αν συμβολίσουμε με M το σύνολο των εναπομεινάντων τοπικών μεγίστων, μετά την απόρριψη ανιχνεύσεων στα χωρικά σύνορα, το σύνολο $\mathcal{D} \subseteq M$ των τελικών ανιχνεύσεων του dca3D διαμορφώνεται από την παρακάτω σχέση

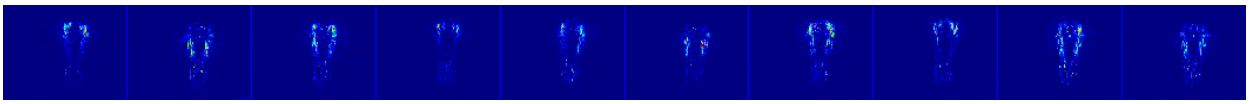
$$\mathcal{D} = \Big\{(x_i, y_i, t_i, R(x_i, y_i, t_i), \sigma_i, \tau_i) : R(x_i, y_i, t_i) > T \cdot \max_{M}(R)\Big\} \tag{5.2.1}$$

Η χωρική και χρονική κλίμακα ανίχνευσης σ_i και τ_i υπολογίζονται από τις σχέσεις (5.1.3) και (5.1.14) αντίστοιχα.

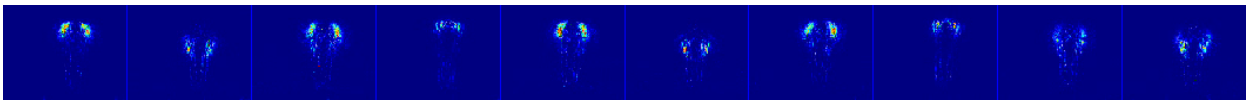
Στο Σχήμα 5.3 απεικονίζονται οι τιμές της *Χωρικά Κυρίαρχης Συνιστώσας*, των Teager-Kaiser Ενεργειών στις εξόδους των καναλιών της συστοιχίας των πέντε μονοδιάστατων φίλτρων Gabor καθώς καί της συνάρτησης απόκρισης για δέκα frames ενός δείγματος βίντεο από τη δράση *handwaving* της Βάσης Δεδομένων *KTH Actions Dataset*, που απέχουν χρονικά περί το μισό δευτερόλεπτο. Σε αυτό το απλό σκηνικό στο οποίο εξελίσσεται η δράση, είναι απόλυτα εμφανής η ικανότητα του Ανιχνευτή dca3D στη σύλληψη και τον εντοπισμό της περιοδικής κίνησης των χεριών που εκτελείται.

(a) Αρχικά frames

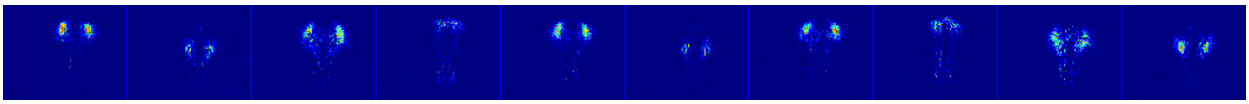
(b) Χωρικά Κυρίαρχη Συνιστώσα



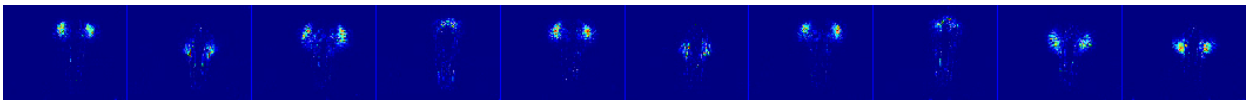
(c) ΤΚ Χρονική Ενέργεια της Χωρικά Κυρίαρχης Συνιστώσας στην έξοδο του 1ου χρονικού καναλιού



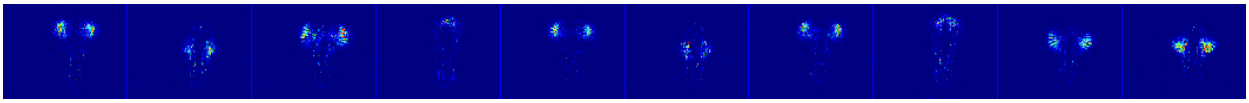
(d) ΤΚ Χρονική Ενέργεια της Χωρικά Κυρίαρχης Συνιστώσας στην έξοδο του 2ου χρονικού καναλιού



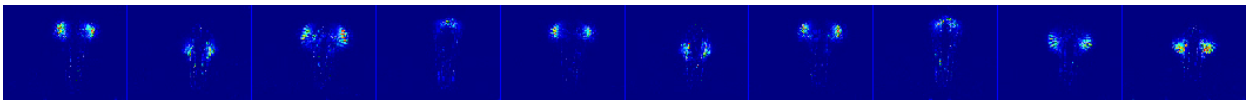
(e) ΤΚ Χρονική Ενέργεια της Χωρικά Κυρίαρχης Συνιστώσας στην έξοδο του 3ου χρονικού καναλιού



(f) ΤΚ Χρονική Ενέργεια της Χωρικά Κυρίαρχης Συνιστώσας στην έξοδο του 4ου χρονικού καναλιού



(g) ΤΚ Χρονική Ενέργεια της Χωρικά Κυρίαρχης Συνιστώσας στην έξοδο του 5ου χρονικού καναλιού

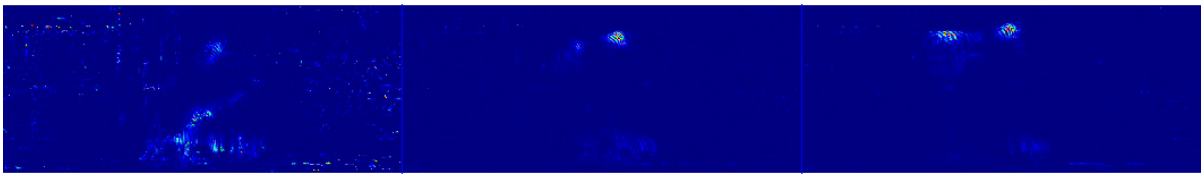


(h) Χρονικά Κυρίαρχη ΤΚ Ενέργεια της Χωρικά Κυρίαρχης Συνιστώσας

Σχήμα 5.3: Απεικόνιση των τιμών της Χωρικά Κυρίαρχης Συνιστώσας, των Ενεργειών αυτής στις εξόδους των πέντε καναλιών της συστοιχίας των χρονικών φίλτρων Gabor και της Χρονικά Κυρίαρχης Ενέργειας αυτής σε δέκα χαρακτηριστικά frames δείγματος βίντεο της *KTH Actions Dataset* από τη δράση *handwaving*. (a) Αρχικά frames που προοδεύουν χρονικά με βήμα 12 frames, (b) Χωρικά Κυρίαρχη Συνιστώσα, (c)-(g) Τιμές της Teager-Kaiser Ενέργειας της Χωρικά Κυρίαρχης Συνιστώσας στις εξόδους των πέντε καναλιών, ξεκινώντας από το κανάλι χαμηλότερης κεντρικής συχνότητας, (h) Τιμές της Συνάρτησης Απόκρισης του Ανιχνευτή dca3D.

(a) Αρχικά Έγχρωμα frames

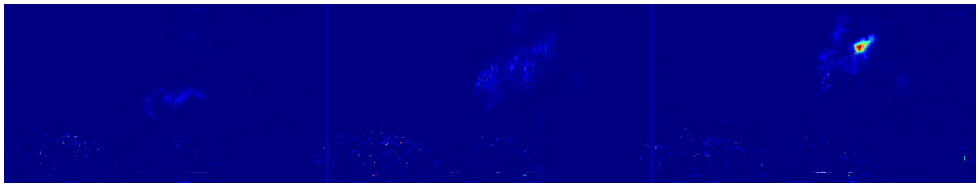
(b) Χωρικά Κυρίαρχη Συνιστώσα



(c) Χρονικά Κυρίαρχη Ενέργεια της Χωρικά Κυρίαρχης Συνιστώσας

(d) Αρχικά Έγχρωμα frames

(e) Χωρικά Κυρίαρχη Συνιστώσα



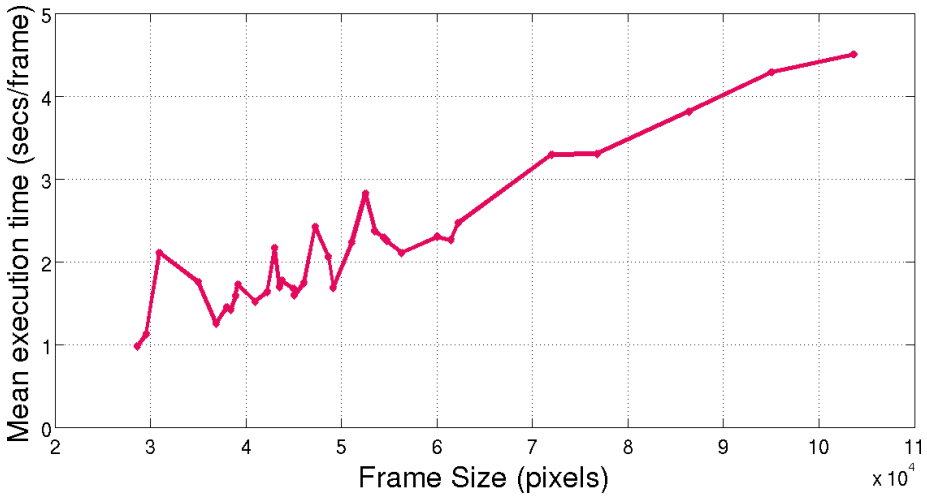
(f) Χρονικά Κυρίαρχη Ενέργεια της Χωρικά Κυρίαρχης Συνιστώσας

Σχήμα 5.4: Απεικόνιση των τιμών της Χωρικά Κυρίαρχης Συνιστώσας και της Χρονικά Κυρίαρχης Ενέργειας αυτής σε τρία χαρακτηριστικά frames δειγμάτων βίντεο της *Hollywood2 Actions Dataset* από τις δράσεις *FightPerson* (1η-3η γραμμή) και *AnswerPhone* (4η-6η γραμμή) αντίστοιχα. (a),(d) Αρχικά έγχρωμα frames των δειγμάτων που προοδεύουν χρονικά ανά 12 frames, (b),(e) Τιμές της Χωρικά Κυρίαρχης Συνιστώσας σε γκρίζα απεικόνιση, (c),(f) Τιμές της Συνάρτησης Απόκρισης του Ανιχνευτή dca3D σε έγχρωμη απεικόνιση.

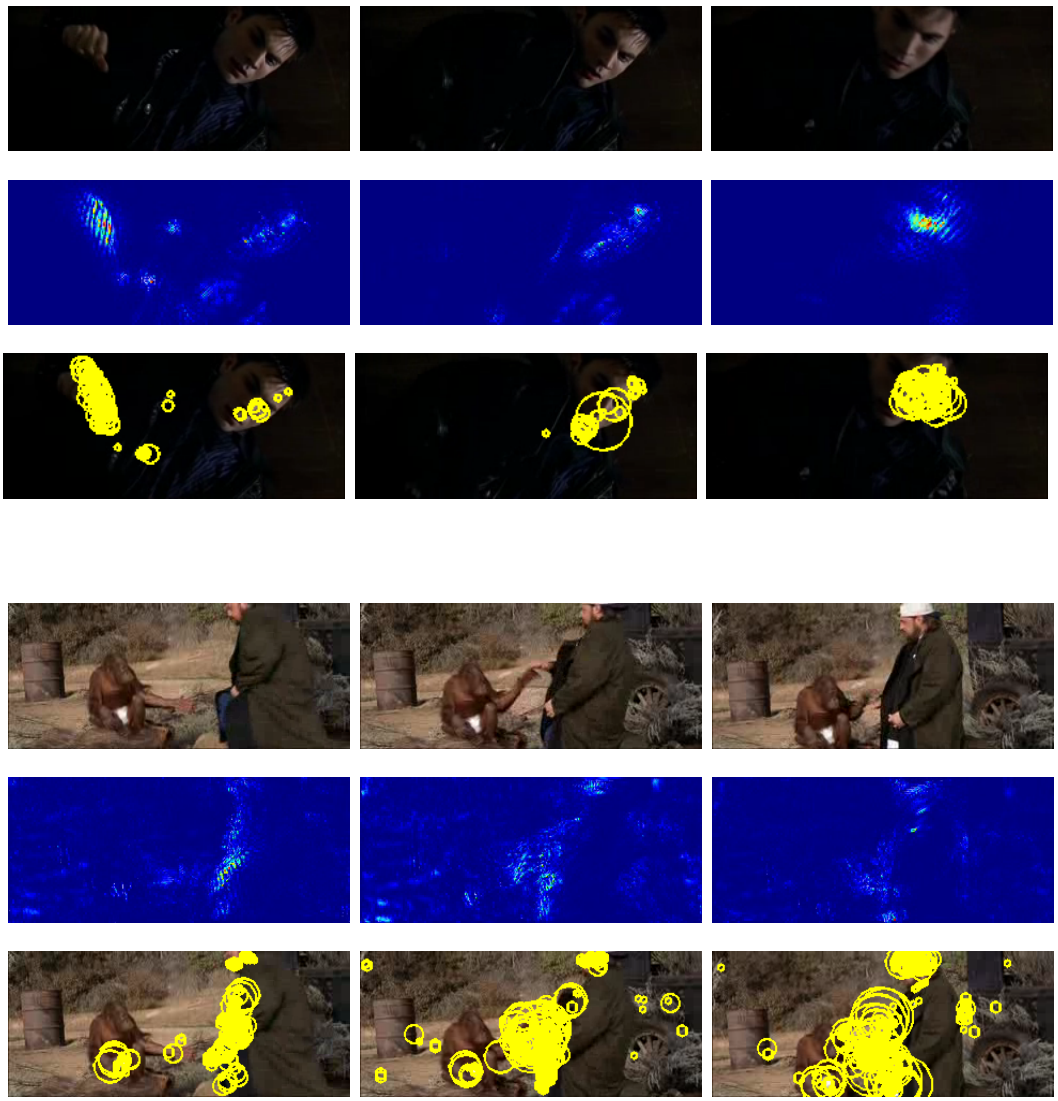
Είναι ενδιαφέρον να παρατηρήσει κανείς τις διαφορές που παρουσιάζονται στις τιμές ενέργειας των ίδιων σημείων, όταν αυτές λαμβάνονται στις εξόδους των καναλιών διαφορετικής κεντρικής συχνότητας της συστοιχίας. Στο Σχήμα 5.4 εμφανίζονται σε γκρίζα και έγχρωμη απεικόνιση αντίστοιχα οι μετρήσεις της *Χωρικά Κυρίαρχης Συνιστώσας* και της συνάρτησης απόκρισης, σε χαρακτηριστικά frames δύο δειγμάτων βίντεο από τις δράσεις *FightPerson* και *AnswerPhone* της Βάσης Δεδομένων *Hollywood2 Actions Dataset*. Παρόλο που οι δράσεις διαδραματίζονται σε ρεαλιστικό σκηνικό με αρκετά πολύπλοκο στήσιμο παρασκηνίου, η συνάρτηση απόκρισης του ανιχνευτή κατορθώνει να συλλάβει αποτελεσματικά τις επικρατούσες κινήσεις των μελών των δρώντων κατά την εξέλιξη των δράσεων.

Στο Σχήμα 5.6 εκθέτουμε χαρακτηριστικές ανιχνεύσεις του dca3D σε τρία χαρακτηριστικά frames δύο δειγμάτων βίντεο από τη Βάση Δεδομένων *Hollywood2*, που περιλαμβάνουν τις δράσεις *SitUp* και *HandShake* αντίστοιχα. Απεικονίζονται παράλληλα και οι τιμές της Συνάρτησης Απόκρισης (5.1.13) για τα αντίστοιχα frames. Στο δείγμα από τη δράση *SitUp* μπορούμε να δούμε ότι η παραγωγή ανιχνεύσεων εντοπίζεται κυρίως στις περιοχές των χεριών ή του κεφαλιού που συμμετέχουν πιο ενεργά για την πραγματοποίηση της δράσης. Στο δείγμα από τη δράση *HandShake* παρατηρούμε υψηλή συγκέντρωση ανιχνεύσεων στην περιοχή των εικόνων όπου εκτυλίσσονται οι κινήσεις των χεριών κατά την εξέλιξη της χειραψίας.

Ο μέσος χρόνος υπολογισμού της Συνάρτησης Απόκρισης του Ανιχνευτή dca3D ανά frame υπολογίστηκε στην τιμή 2.033 secs/frame, σε ένα σύνολο 138 δειγμάτων βίντεο με μέσο μέγεθος frame τα 48663 pixels. Τα εν λόγω δείγματα βίντεο, βάσει των οποίων εξήχθησαν οι χρόνοι εκτέλεσης, αντλήθηκαν από τη Βάση Δεδομένων *Hollywood2* για τις ανάγκες ενός πειράματος Αναγνώρισης Ανθρώπινων Δράσεων που θα παρουσιάσουμε στην επόμενη ενότητα. Το γράφημα του μέσου χρόνου εκτέλεσης ανά frame με το μέγεθος frame, με βάση όλα τα διακριτά μεγέθη frame που εμφανίζονται στο σύνολο των 138 δειγμάτων, απεικονίζεται στο Σχήμα 5.5.



Σχήμα 5.5: Μέσος χρόνος υπολογισμού της Συνάρτησης Απόκρισης του Ανιχνευτή dca3D ανά frame για τα διακριτά μεγέθη frame στο σύνολο 138 δειγμάτων της *Hollywood2 Actions Dataset*.



Σχήμα 5.6: Απεικόνιση των τιμών της Συνάρτησης Απόκρισης και των Χωροχρονικών Σημείων Ενδιαφέροντος του Ανιχνευτή dca3D σε τρία χαρακτηριστικά frames δειγμάτων βίντεο της *Hollywood2 Actions Dataset* από τις δράσεις *SitUp* και *HandShake*. Πρώτη και Τέταρτη Γραμμή: Αρχικά έγχρωμα frames των δειγμάτων, που προοδεύουν χρονικά με βήμα 12 frames, Δεύτερη και Πέμπτη Γραμμή: Τιμές της Συνάρτησης Απόκρισης σε έγχρωμη απεικόνιση για τα αντίστοιχα frames, Τρίτη και Έκτη Γραμμή: Ανιχνευθέντα από τον dca3D σημεία, εικονιζόμενα επί των αρχικών frames.

Στάδιο dca2D	**Μέσος Χρόνος Εκτέλεσης** (secs/frame)
Φιλτράρισμα	1.510
Ενέργειες Καναλιών	0.267
Δείκτες Κυρίαρχων EDCA Καναλιών	0.027
Χωρικά Κυρίαρχη Συνιστώσα	0.019
Σύνολο	**1.823**

Πίνακας 5.1: Μέσοι χρόνοι εκτέλεσης σε secs/frame των τεσσάρων βημάτων υπολογισμού της *Χωρικά Κυρίαρχης Συνιστώσας* και μέσος συνολικός χρόνος για το στάδιο dca2D, υπολογισμένοι επί του συνόλου των frames 138 δειγμάτων βίντεο της *Hollywood2 Actions Dataset*, με μέσο μέγεθος frame τα 48663 pixels.

Συνάρτηση Απόκρισης dca3D	**Μέσος Χρόνος Εκτέλεσης** (secs/frame)
Χωρικά Κυρίαρχη Συνιστώσα	1.823
Χρονικά Κυρίαρχη Ενέργεια της Χωρικά Κυρίαρχης Συνιστώσας	0.210
Σύνολο	**2.033**

Πίνακας 5.2: Μέσοι χρόνοι εκτέλεσης σε secs/frame των δύο σταδίων υπολογισμού της Συνάρτησης Απόκρισης του Ανιχνευτή dca3D και μέσος συνολικός χρόνος, υπολογισμένοι επί του συνόλου των frames 138 δειγμάτων βίντεο της *Hollywood2 Actions Dataset*, με μέσο μέγεθος frame τα 48663 pixels.

Στον Πίνακα 5.1 παραθέτουμε τους μέσους χρόνους εκτέλεσης ανά frame των τεσσάρων βημάτων που ακολουθούνται για το στάδιο της εξαγωγής της *Χωρικά Κυρίαρχης Συνιστώσας* (στάδιο dca2D) καθώς και τον μέσο συνολικό χρόνο ανά frame. Συγκρίνοντας τις τιμές με αυτές του Πίνακα 4.3, που περιέχει αντίστοιχες μετρήσεις για τα βήματα εξαγωγής της *EDCA* Κυρίαρχης Ενέργειας, διαπιστώνουμε ασήμαντες διαφορές στους χρόνους των τριών πρώτων βημάτων. Το τέταρτο βήμα του σταδίου dca2D, στο οποίο υπολογίζεται η *Χωρικά Κυρίαρχη Συνιστώσα* βάσει των δεικτών των Κυρίαρχων Καναλιών, δεν είναι υπολογιστικά χρονοβόρο και απασχολεί μόλις το 1% του συνολικού χρόνου του σταδίου dca2D.
Στον Πίνακα 5.2 παρουσιάζονται οι μέσοι χρόνοι εκτέλεσης ανά frame των δύο διακριτών σταδίων υπολογισμού της Συνάρτησης Απόκρισης του Ανιχνευτή dca3D καθώς και ο μέσος συνολικός χρόνος ανά frame. Όπως παρατηρούμε, το στάδιο dca2D επιφέρει πολύ μεγαλύτερη κατανάλωση χρόνου σε σχέση με το δεύτερο στάδιο, καλύπτοντας το 90% του μέσου συνολικού χρόνου εκτέλεσης.

5.3 Πειράματα Αναγνώρισης Ανθρώπινων Δράσεων σε Βίντεο με χρήση του Ανιχνευτή dca3D

Στην παρούσα ενότητα θα παρουσιάσουμε αποτελέσματα πειραμάτων Αναγνώρισης Ανθρώπινων Δράσεων σε Βίντεο που διεξάγαμε στην παρούσα διπλωματική εργασία, χρησιμοποιώντας στο στάδιο ανίχνευσης Τοπικών Χωροχρονικών Σημείων Ενδιαφέροντος τον Ανιχνευτή dca3D. Στο στάδιο της περιγραφής των εξαχθέντων σημείων υιοθετούμε τον επιτυχημένο Περιγραφέα HOG/HOF, όπως και στα πειράματα του Κεφαλαίου 4. Στα πειράματα της παρούσας ενότητας ακολουθούμε την ίδια λογική αναπαράστασης των βίντεο με ιστογράμματα της συχνότητας εμφάνισης των οπτικών λέξεων του κατασκευασμένου λεξιλογίου (Bag-Of-Features) ενώ για την αναγνώριση των δράσεων εφαρμόζουμε την τεχνική ταξινόμησης με μηχανές SVM (βλ. ενότητες 4.1 και 4.2 αντίστοιχα).

Πείραμα Ταξινόμησης των δράσεων *HandShake* και *SitUp*.

Αρχικά επιλέξαμε να εκτελέσουμε πείραμα αναγνώρισης σε δύο δράσεις της *Hollywood2 Actions Dataset*, τις *HandShake* και *SitUp*, δύο απαιτητικές για το πρόβλημα της αναγνώρισης δράσεις όπως είδαμε στο προηγούμενο κεφάλαιο. Συμπεριλαμβάνοντας όλα τα βίντεο των παραπάνω δράσεων στο υποσύνολο δειγμάτων του παρόντος πειράματος, προέκυψαν Training Set και Test Set μεγέθους 56 και 82 δειγμάτων αντίστοιχα, χωρίς να υπολογίζονται δύο φορές δείγματα που περιλαμβάνουν καί τις δύο δράσεις. Ο αριθμός δειγμάτων κάθε δράσης στα δύο Σετ αναφέρεται στον Πίνακα 4.1.
Για την παραγωγή επιδόσεων αναγνώρισης που θα χρησιμεύσουν για τη σύγκριση με τις αντίστοιχες επιδόσεις του πειράματος με ανιχνευτή τον dca3D, επιλέχθηκε ο αποτελεσματικός συνδυασμός Ανιχνευτή Harris3D και Περιγραφέα HOG/HOF για τη διεξαγωγή του πρώτου πειράματος αναγνώρισης. Οι επιλογές παραμέτρων για τον ανιχνευτή, τον περιγραφέα, την κατασκευή του οπτικού λεξιλογίου και τους ταξινομητές SVM είναι πανομοιότυπες με εκείνες του πειράματος της ενότητας 4.3. Αξίζει να αναφέρουμε την μέτρηση για τον μέσο αριθμό ανιχνεύσεων ανά frame από τον Harris3D που λάβαμε στο σύνολο των frames των 138 δειγμάτων του Training και Test Set και προέκυψε τιμής των περίπου 12 ανιχνεύσεων ανά frame. Στον Πίνακα 5.4 παραθέτονται τα αποτελέσματα για τη *μέση ακρίβεια* ταξινόμησης κάθε δράσης καθώς καί τη μέση τιμή της επί των δύο δράσεων (mean average precision).
Για το πείραμα αναγνώρισης επί του ίδιου υποσυνόλου της *Hollywood2* με Ανιχνευτή τον dca3D, εφαρμόσαμε την υλοποίηση του ανιχνευτή σε MATLAB που περιγράψαμε

	Harris3D	**dca3D**
Ανιχνεύσεις/Frame	12	18

Πίνακας 5.3: Μέση τιμή του αριθμού ανιχνευθέντων χωροχρονικών σημείων ενδιαφέροντος ανά frame από τους Ανιχνευτές Harris3D και dca3D, υπολογισμένης επί του συνόλου των frames 138 δειγμάτων βίντεο της *Hollywood2 Actions Dataset*, με μέσο μέγεθος frame τα 48663 pixels.

	Harris3D + HOG/HOF	**dca3D + HOG/HOF**
HandShake	**82.82%**	82.53%
SitUp	72.28%	**75.24%**
mAP	77.55%	**78.88%**

Πίνακας 5.4: Αποτελέσματα Average Precision για τις δύο δράσεις του πειράματος και mean Average Precision επί των δύο δράσεων για το σχήμα Ανιχνευτή Harris3D και Περιγραφέα HOG/HOF (Αριστερά) και το σχήμα Ανιχνευτή dca3D και Περιγραφέα HOG/HOF (Δεξιά).

αναλυτικά στην προηγούμενη ενότητα. Οι θέσεις και κλίμακες των ανιχνεύσεων και οι αντίστοιχες τιμές της συνάρτησης απόκρισης του ανιχνευτή προωθήθηκαν σε μορφή αρχείων κειμένου για τον υπολογισμό των Περιγραφέων HOG/HOF από την online διαθέσιμη υλοποίηση *stip-2.0-linux*[1]. Το μέγεθος σε κάθε διάσταση του τρισδιάστατου τοπικού τεμαχίου στο οποίο υπολογίζονται οι περιγραφείς ορίζεται, όπως έχουμε δει, από τις σχέσεις $\Delta_x, \Delta_y = 2k_{sp}\sigma$ και $\Delta_t = 2k_t\tau$. Να σημειώσουμε εδώ ότι η εν λόγω υλοποίηση κβαντοποιεί τις κλίμακες ανίχνευσης για εξωτερικά παρεχόμενες ανιχνεύσεις, όπως αυτές που παρείχαμε με τις τιμές κλιμάκων να δίνονται από τις σχέσεις (5.1.3) και (5.1.14), στις τιμές $\sigma \in \{4, 8, 16, 32, 64, 128 \ldots\}$ και $\tau \in \{2, 4\}$ αντίστοιχα. Για τις σταθερές χωρικής και χρονικής υποστήριξης των τεμαχίων επιλέξαμε τις τιμές $k_{sp} = 5$ και $k_t = 4$ αντίστοιχα.

Ο μέσος αριθμός ανιχνεύσεων ανά frame για το παρόν πείραμα από τον Ανιχνευτή dca3D προέκυψε ίσος με περί τις 18 ανιχνεύσεις ανά frame, περίπου 1.5 φορές περισσότερες από εκείνες του Ανιχνευτή Harris3D. Προφανώς, η πυκνότητα των ανιχνεύσεων σχετίζεται άμεσα με τον ορισμό της τιμής του ολικού κατωφλίου για την *καταστολή των μη μεγίστων*. Οι τελικές ανιχνεύσεις για το παρόν πείραμα προέκυψαν θέτοντας το ολικό κατώφλι T της σχέσης (5.2.1) στην τιμή $T = 0.1$. Παράλληλα, απορρίφθηκαν οι ανιχνεύσεις που εντοπίστηκαν στα χωρικά σύνορα μήκους 5 pixels των εικόνων ή στα χρονικά σύνορα μήκους 5 frames της ακολουθίας. Η απόρριψη ανιχνεύσεων στα χρονικά σύνορα της ακολουθίας εικόνων επιλέχθηκε για την αποφυγή ενδεχομένως τεχνητών ανιχνεύσεων σε frames που αποτελούν τα άκρα της εξόδου διακριτών συνελίξεων στον χρόνο.

Τα αποτελέσματα για τη *μέση ακρίβεια* κάθε δράσης και τη μέση τιμή αυτής επί των δύο δράσεων παρουσιάζονται στον Πίνακα 5.4, πλάι στα αντιστοιχα αποτελέσματα του σχήματος Ανιχνευτή Harris3D και Περιγραφέα HOG/HOF. Η χρήση του νέου Ανιχνευτή dca3D οδήγησε σε αύξηση της απόδοσης κατά 1.33%. Για τη δράση *HandShake* το ποσοστό *μέσης ακρίβειας* προέκυψε ελαφρά χαμηλότερο σε σχέση με την αντίστοιχη τιμή του πειράματος με Ανιχνευτή τον Harris3D, ενώ για τη δράση *SitUp* είχαμε σημαντική αύξηση της επίδοσης στη *μέση ακρίβεια* αναγνώρισης.

[1] http://www.irisa.fr/vista/Equipe/People/Laptev/download.html

Πείραμα Ταξινόμησης των δράσεων *FightPerson*, *HandShake* και *SitUp*.

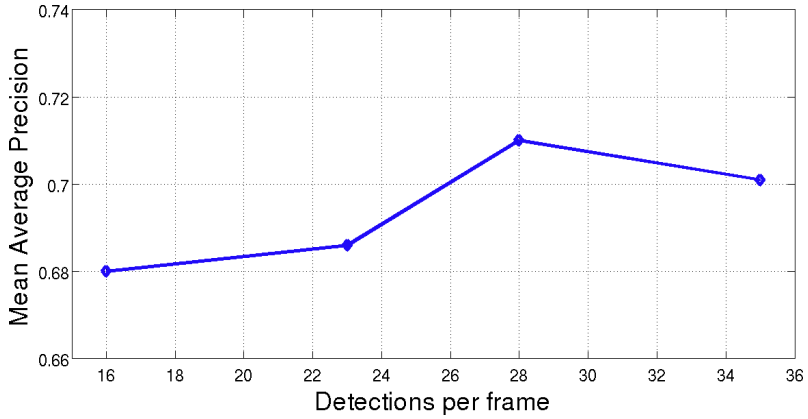
Προκειμένου να αξιολογήσουμε τη δυναμική του Ανιχνευτή dca3D ως προς τις επιδόσεις ταξινόμησης περισσοτέρων δράσεων προχωρήσαμε στην ένταξη μιας επιπλέον δράσης σε σχέση με το πείραμα που παρουσιάσαμε παραπάνω. Η δράση που επιλέχθηκε ήταν η *FightPerson*, τα δείγματα της οποίας εκτυλίσσονται συνήθως σε σκηνικά με γρήγορες εναλλαγές σκηνών και πολύπλοκες κινήσεις της κάμερας. Το πείραμα ταξινόμησης των δειγμάτων καί αυτής της δράσης θα μας δώσει επομένως την ευκαιρία να εξετάσουμε την ευρωστία του ανιχνευτή απέναντι στους εν λόγω παράγοντες, σε μεγαλύτερο βαθμό από ότι στις δύο δράσεις που πρότερα χρησιμοποιήθηκαν. Για το παρόν πείραμα προέκυψαν Training και Test Set μεγέθους 110 και 151 δειγμάτων βίντεο αντίστοιχα.
Το πρώτο πείραμα διεξήχθη, με την ίδια λογική όπως και στα προηγούμενα, με τη χρήση του Ανιχνευτή Harris3D και του Περιγραφέα HOG/HOF, τηρώντας πανομοιότυπες επιλογές για τις παραμέτρους. Τα αποτελέσματα για τη *μέση ακρίβεια* ταξινόμησης των δράσεων καθώς και τη μέση τιμή της επί των τριών δράσεων παρατίθενται στον Πίνακα 5.5.
Η χρήση του ανιχνευτή Harris3D οδήγησε σε μέσο όρο ανιχνεύσεων ανά frame, επί του συνόλου των 261 δειγμάτων, τις 19 ανιχνεύσεις ανά frame. Παρατηρώντας τις αντίστοιχες μετρήσεις για τον Harris3D και dca3D για το προηγούμενο πείραμα στον Πίνακα 5.3, γίνεται προφανές ότι ο Harris3D οδηγεί σε κατά πολύ πυκνότερες ανιχνεύσεις επί των δειγμάτων της δράσης FightPerson.
Παρά τις παραπάνω ενδείξεις, αποφασίσαμε αρχικά να διατηρήσουμε το ολικό κατώφλι *καταστολής των μη μεγίστων* της σχέσης (5.2.1) στην τιμή $T = 0.1$, που είχε τεθεί στο προηγούμενο πείραμα, για ένα πρωτο πείραμα ταξινόμησης των τριών δράσεων με ανιχνευτή τον dca3D. Ο αριθμός ανιχνεύσεων ανά frame προέκυψε τιμής 16, με άλλα λόγια η ένταξη της δράσης FightPerson οδήγησε συνολικά στην παραγωγή "αραιότερων"ανιχνεύσεων του dca3D σε σχέση με εκείνες του Harris3D, σε αντίθεση με το πείραμα των δύο δράσεων. Η τιμή της *mean average precision* για το πείραμα ταξινόμησης των τριών δράσεων με ανιχνευτή τον dca3D και τιμή κατωφλίου $T = 0.1$ ανήλθε μόλις στο 68.00%, μια απόδοση κατά πολύ υποβαθμισμένη σε σχέση με το αποτέλεσμα 76.77% που απέφερε ο Harris3D.

	Harris3D + HOG/HOF
FightPerson	94.7%
HandShake	75.2%
SitUp	60.4%
mAP	**76.77%**

Πίνακας 5.5: Αποτελέσματα Average Precision για τις τρεις δράσεις του πειράματος και mean Average Precision επί των τριών δράσεων για το σχήμα Ανιχνευτή Harris3D και Περιγραφέα HOG/HOF.

Η παράμετρος ολικού κατωφλίου καταστολής των μη μεγίστων.

Η σχετική "αραιότητα" των ανιχνεύσεων του dca3D στο πείραμα των τριών δράσεων μας οδήγησε να επαναλάβουμε το πείραμα ταξινόμησης για διάφορες τιμές του ολικού κατωφλίου. Το πείραμα διεξήχθη συνολικά για τις τιμές κατωφλίου $T \in \{0.06, 0.07, 0.08, 0.10\}$. Ο μέσος όρος ανιχνεύσεων ανά frame και οι τιμές *mean average precision* για τα πειράματα με τις διάφορες τιμές κατωφλίου φαίνονται συγκεντρωτικά στον Πίνακα 5.6.
Η υψηλότερη τιμή *mean average precision* ελήφθη για την τιμή κατωφλίου $T = 0.07$ με την οποία παρήχθησαν περί τις 28 ανιχνεύσεις ανά frame. Παρατηρούμε ότι η μείωση της τιμής κατωφλίου, ξεκινωντας από την τιμή 0.10 και φθάνοντας στην τιμή 0.07, και η κατά συνέπεια αύξηση του μέσου όρου ανιχνεύσεων οδήγησε σε βελτίωση της μέσης επίδοσης της ακρίβειας στην αναγνώριση. Εντούτοις, η περαιτέρω μείωση της τιμής στο 0.06, με αποτέλεσμα ακόμα πυκνότερες ανιχνεύσεις, υποβάθμισε την απόδοση της αναγνώρισης. Τα πειραματικά αυτά αποτελέσματα επιβεβαιώνουν την πεποίθηση ότι μεγάλυτερη μέση τιμή ανιχνεύσεων, για τοπικούς ανιχνευτές χωροχρονικών σημείων ενδιαφέροντος, δε συνεπάγεται απαραίτητα κέρδος στην απόδοση της αναγνώρισης ανθρώπινων δράσεων. Για τούτο, κρίνεται ουσιώδης η βελτιστοποίηση της παραμέτρου του ολικού κατωφλίου για τον Ανιχνευτή dca3D, όπως καί για κάθε ανιχνευτή. Το διάγραμμα της τιμής *mean average precision* με τη μέση τιμή ανιχνεύσεων ανά frame του Ανιχνευτή dca3D για το πείραμα των τριών δράσεων απεικονίζεται στο Σχήμα 5.7.



Σχήμα 5.7: Διάγραμμα της *mean Average Precision* με τη μέση τιμή των ανιχνεύσεων ανά frame στο πείραμα της ταξινόμησης των τριών δράσεων με ανιχνευτή τον dca3D.

	thresh 0.10	**thresh 0.08**	**thresh 0.07**	**thresh 0.06**
Ανιχνεύσεις/Frame	16	23	**28**	35
mean Average Precision	68.00%	68.60%	**71.00%**	70.10%

Πίνακας 5.6: Μέσος όρος ανιχνεύσεων ανά frame και τιμές mean Average Precision για το πείραμα ταξινόμησης των τριών δράσεων με χρήση του Ανιχνευτή dca3D και του Περιγραφέα HOG/HOF.

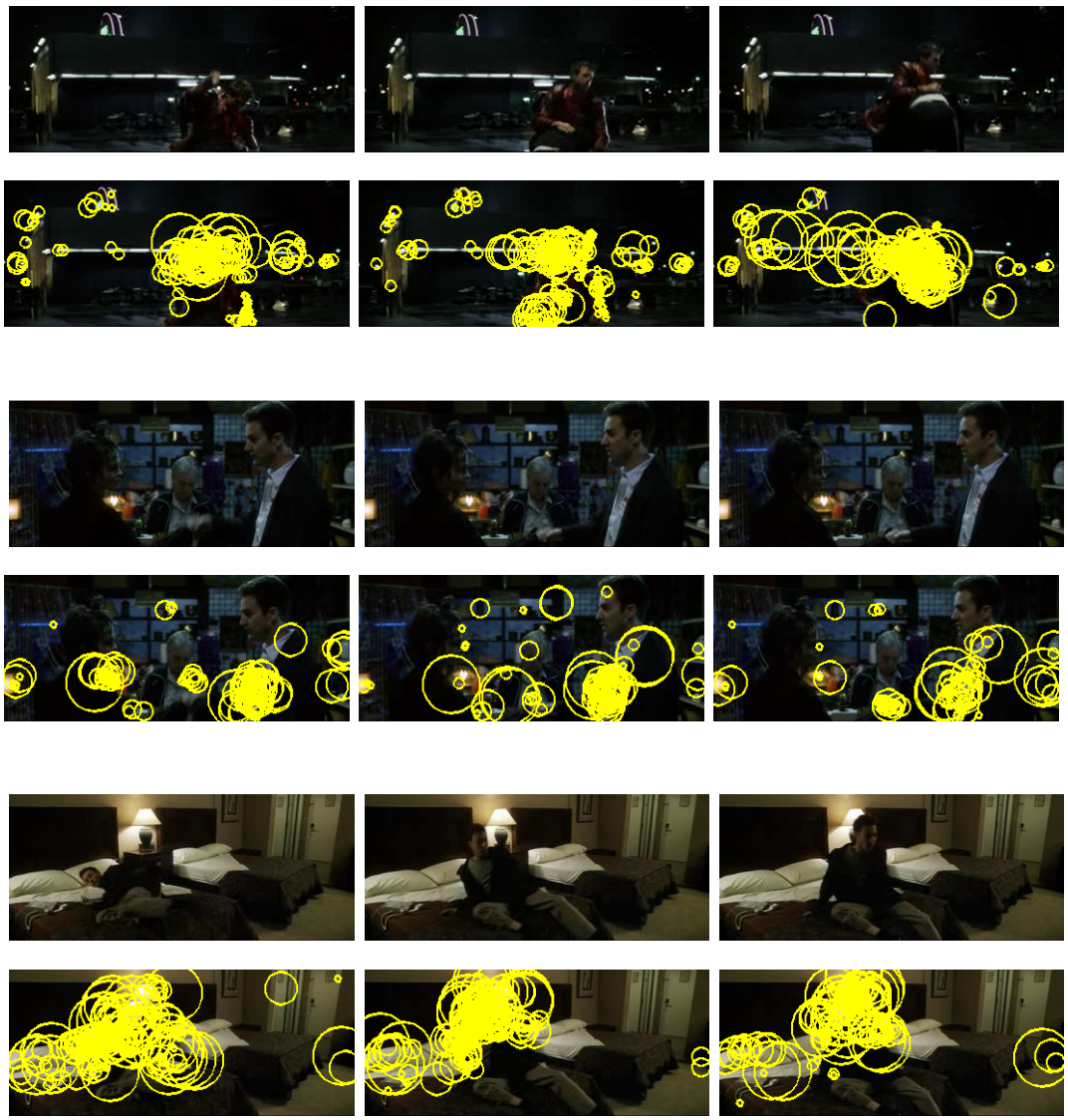
	dca3D + HOG/HOF
FightPerson	91.7%
HandShake	70.6%
SitUp	50.6%
mAP	**71.00%**

Πίνακας 5.7: Αποτελέσματα Average Precision για τις τρεις δράσεις του πειράματος και mean Average Precision επί των τριών δράσεων για το σχήμα Ανιχνευτή Harris3D και Περιγραφέα HOG/HOF. Το ολικό κατώφλι *καταστολής των μη μεγίστων* για τα άνωθι αποτελέσματα είναι $T = 0.07$.

Στο προηγούμενο πείραμα αναγνώρισης των δύο δράσεων *HandShake* και *SitUp* ακολουθήσαμε τη λογική απόρριψης των ανιχνεύσεων που βρίσκονται στα πρώτα ή τα τελευταία πέντε frames της ακολουθίας εικόνων. Για να αξιολογήσουμε την επίδραση αυτής της επιλογής στην απόδοση της αναγνώρισης εκτελέσαμε το πείραμα αναγνώρισης των τριών δράσεων με τον dca3D καί για τα δύο σενάρια της περίληψης ή απομάκρυνσης των εν λόγω ανιχνεύσεων σε κάθε δείγμα. Χρησιμοποιώντας το βέλτιστο, σύμφωνα με τα πειραματικά αποτελέσματα, ολικό κατώφλι 0.07 είδαμε ότι η απόρριψη τέτοιων ανιχνεύσεων στα χρονικά σύνορα της ακολουθίας οδήγησε σε μείωση της *mean average precision* κατά 1.4%. Ενδεχομένως αυτές οι ανιχνεύσεις συνεισφέρουν στην αιχμαλώτιση χρήσιμης πληροφορίας του παρασκηνίου στο οποίο αρχίζει ή τελειώνει η δράση. Για τούτο, όλα τα αποτελέσματα του Πίνακα 5.6 αναφέρονται σε πειράματα στα οποία δεν εκτελέσαμε απόρριψη των ανιχνεύσεων στα χρονικά σύνορα αλλά μόνο στα χωρικά σύνορα μήκους 5 frames των εικόνων.

Τα αποτελέσματα για τη *μέση ακρίβεια* ταξινόμησης κάθε δράσης αλλά και τη μέση τιμή της (*mean average precision*) επί των τριών δράσεων του πειράματος ταξινόμησης με ανιχνευτή τον dca3D και τιμή ολικού κατωφλίου $T = 0.07$ παρατίθενται στον Πίνακα 5.7. Συγκρίνοντάς τα με τα αντίστοιχα αποτελέσματα που λάβαμε με τη χρήση του Ανιχνευτή Harris3D και φαίνονται στον Πίνακα 5.5, η απόδοση του Ανιχνευτή dca3D κινείται αρκετά χαμηλότερα καί στις επιμέρους δράσεις καί στην τιμή της *mean average precision*. Η δράση *FightPerson* παρουσιάζει κοντινή επίδοση στη *μέση ακρίβεια* ταξινόμησης, η *HandShake* σημαντική σχετική υποβάθμιση ενώ το αποτέλεσμα για τη δράση *SitUp* είναι κατά 9.8% χαμηλότερο. Παρατηρούμε λοιπόν ότι η εμπειρική βελτιστοποίηση της παραμέτρου του ολικού κατωφλίου του Ανιχνευτη dca3D για το εν λόγω πείραμα δε στάθηκε δυνατή να οδηγήσει από μόνη της στην παροχή αποτελεσμάτων κοντά στα αντίστοιχα που ελήφθησαν με τον Harris3D.

Χαρακτηριστικές ανιχνεύσεις του dca3D σε δείγματα βίντεο από τις τρεις δράσεις του πειράματος απεικονίζονται στο Σχήμα 5.8. Αυτές προέκυψαν χρησιμοποιώντας την τιμή ολικού κατωφλίου $T = 0.07$, η οποία είναι υπεύθυνη για την εξαγωγή των καλύτερων αποτελεσμάτων αναγνώρισης με τον dca3D. Στα εικονιζόμενα frames μπορεί να παρατηρήσει κανείς την ποιοτική ανίχνευση οπτικά "σημαντικών"σημείων και κυρίως τη σύλληψη των κυρίαρχων κινήσεων των δρώντων. Η επισκόπηση τέτοιων ανιχνευθέντων σημείων σε μεμονωμένα δείγματα βίντεο ενίσχυσε την πεποίθησή μας ότι η εξαγωγή των τοπικών μεγίστων της συνάρτησης απόκρισης του ανιχνευτή παράγει πληροφοριακές και



Σχήμα 5.8: Απεικόνιση των Χωροχρονικών Σημείων Ενδιαφέροντος του Ανιχνευτή dca3D σε τρία χαρακτηριστικά frames δειγμάτων βίντεο της *Hollywood2 Actions Dataset* από τις δράσεις *FightPerson*, *HandShake* και *SitUp* (σημ.: τα δείγματα προέρχονται από την ταινία "Fight Club"). Πρώτη, τρίτη και πέμπτη γραμμή: Αρχικά έγχρωμα frames των δειγμάτων, που προοδεύουν χρονικά με βήμα 12 frames. Δεύτερη, τέταρτη και έκτη γραμμή: Ανιχνευθέντα από τον dca3D σημεία με τιμή ολικού κατωφλίου $T = 0.07$, εικονιζόμενα επί των αρχικών frames.

διακρίνουσες ανιχνεύσεις. Για τούτο θεωρήσαμε πιο υποσχόμενη την περαιτέρω μελέτη του σταδίου κατά το οποίο διαμορφώνονται, με βάση τις κλίμακες ανίχνευσης, οι διαστάσεις των τρισδιάστατων τοπικών τεμαχίων περιγραφής για τον υπολογισμό του Περιγραφέα HOG/HOF.

Οι κλίμακες ανίχνευσης και οι διαστάσεις των τοπικών τεμαχίων.

Όπως αναφέραμε και παραπάνω, η online διαθέσιμη υλοποίηση του Περιγραφέα HOG/HOF για εξωτερικά παρεχόμενα σημεία απαιτεί τη συνοδεία των θέσεων των σημείων από τις κλίμακες στις οποίες ανιχνεύθηκαν. Αυτά χρησιμοποιούνται για τον ορισμό της χωρικής διάστασης Δ_x και της χρονικής Δ_y του 3D τοπικού τεμαχίου από τις σχέσεις $\Delta_x, \Delta_y = 2k_{sp}\sigma$ και $\Delta_t = 2k_t\tau$ αντίστοιχα, όπου σ και τ η χωρική και χρονική κλίμακα ανίχνευσης. Σε όλα τα παραπάνω πειράματα με τον Ανιχνευτή dca3D η χωρική και χρονική κλίμακα ανίχνευσης των εξαχθέντων σημείων ορίζονται από τις σχέσεις (5.1.3) και (5.1.14) αντίστοιχα. Επίσης, για τις παραμέτρους χωρικής και χρονικής υποστήριξης θέσαμε τις τιμές $k_{sp} = 5$ και $k_t = 4$.
Ωστόσο, η εν λόγω υλοποίηση για τον Περιγραφέα HOG/HOF δεν επιτρέπει την παροχή κλιμάκων διαφορετικών από εκείνες που ανήκουν σε δύο σύνολα προκαθορισμένων τιμών για τη χωρική και χρονική κλίμακα ανίχνευσης. Κατά συνέπεια, οποιεσδήποτε διαφορετικές τιμές προωθούνται για τις κλίμακες κβαντοποιούνται με βάση αυτό το διακριτό σύνολο τιμών που υποστηρίζει η υλοποίηση. Οι εν λόγω τιμές (σ_i, τ_j) επιβάλλεται επομένως, σύμφωνα με την υλοποίηση, να δίνονται από τις σχέσεις $\sigma_i = 2^{(1+i)/2}, (i = 1, \ldots, 8)$ και $\tau_j = 2^{j/2}, (j = 1, 2)$ αντίστοιχα. Προφανώς, λαμβάνοντας υπόψη ότι επεξεργαζόμαστε τα δείγματα της Βάσης Δεδομένων *Hollywood2* στην ημίσεια χωρική ανάλυση, οι χωρικές κλίμακες για $i \in \{7, 8\}$ είναι σχεδόν πάντα αδύνατο να χρησιμοποιηθούν, για το λόγο ότι δίνουν τρισδιάστατα τεμάχια μεγαλύτερης διάστασης στο χώρο από εκείνη του αρχικού frame.
Η χωρική και χρονική κλίμακα ανίχνευσης του dca3D ανιχνευτή εξαρτώνται για κάθε δείγμα βίντεο από το πλάτος του frame και το frame rate αντίστοιχα. Αυτό έχει ως αποτέλεσμα τη μεταβολή των τιμών των κλιμάκων για διαφορετικά δείγματα βίντεο, η οποία είναι σχεδόν ασήμαντη για τη χρονική κλίμακα δεδομένου ότι το frame rate των δειγμάτων της *Hollywood2* κυμαίνεται μεταξύ 24 και 25 frames/sec. Για τη χωρική κλίμακα, ωστόσο, δεν ισχύει το ίδιο καθώς παρατηρούνται αρκετά διαφορετικά μεγέθη frame στην Βάση. Με βάση τα παραπάνω, γίνεται σαφές ότι η κβαντοποίηση της χωρικής κλίμακας στις απαιτούμενες από την υλοποίηση του HOG/HOF τιμές καταλήγει σε διαφορετικά υποσύνολα προκαθορισμένων κλιμάκων για διαφορετικά δείγματα. Για παράδειγμα, η τρίτη χωρική κλίμακα του dca3D για κάποια δείγματα θα κβαντοποιηθεί στην τιμή $2\sqrt{2}$ και για άλλα στην τιμή 4, με αποτέλεσμα στην πρώτη περίπτωση το τοπικό τεμάχιο να έχει διάσταση στον χώρο 29×29 ενώ στη δεύτερη 40×40. Ακόμα, η πρώτη κλίμακα του dca3D, που αντιστοιχεί στα φίλτρα χαμηλής κεντρικής συχνότητας, συχνά θα κβαντοποιηθεί στην τιμή 16 η οποία υπαγορεύει τον υπολογισμό του τοπικού τεμαχίου στο χώρο σε διάσταση 160×160, κάτι που σχεδόν πάντα καταλήγει στην αδυναμία εξαγωγής του περιγραφέα και την απώλεια του χαρακτηριστικού.
Οι παραπάνω προβληματισμοί μας οδήγησαν στην εξαγωγή χρήσιμων μετρήσεων σχετι-

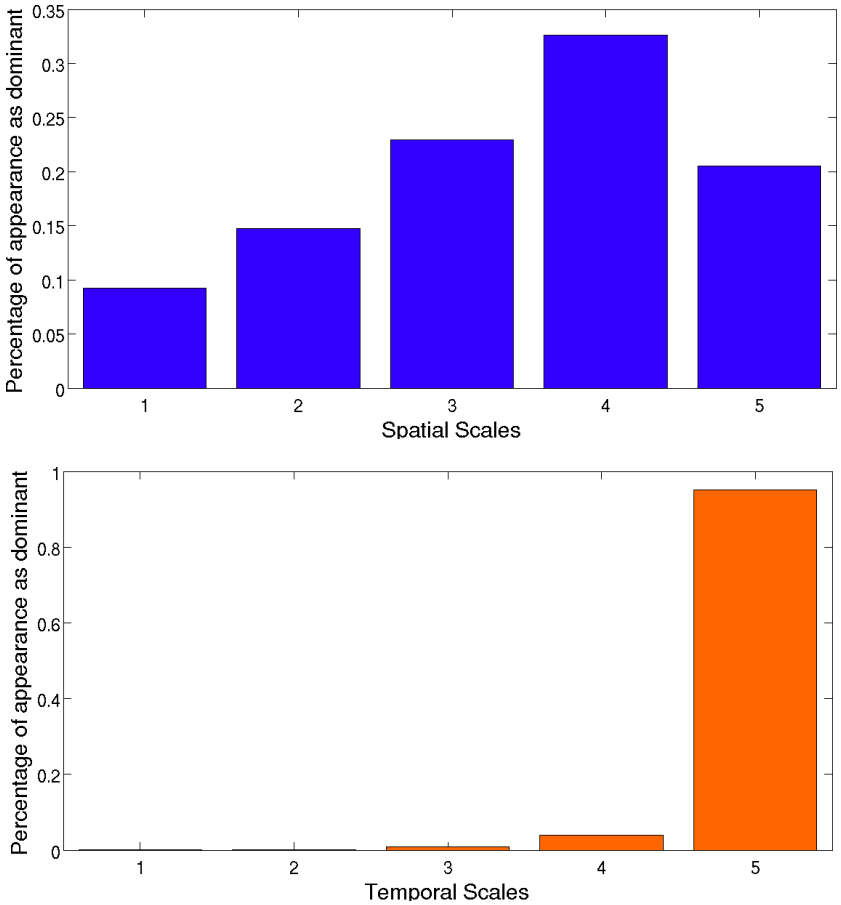
κά με τη μέση τιμή των παρατηρούμενων, στα δείγματα των τριών δράσεων, κλιμάκων καθώς και με την κατανομή των κλιμάκων που αντιστοιχούν στα κυρίαρχα κανάλια που πυροδότησαν τις ανιχνεύσεις. Η μέση τιμή, υπολογισμένη για τα 261 δείγματα του πειράματος, των κλιμάκων ανίχνευσης που αντιστοιχούν στις διακριτές κλίμακες των συστοιχίων φίλτρων Gabor στον χώρο και τον χρόνο, παρατίθεται στον Πίνακα 5.8. Μια πρώτη παρατήρηση αφορά την πρώτη χωρική κλίμακα, η οποία βάσει της μέσης τιμής της είναι πιο πιθανόν να κβαντοποιηθεί στην τιμή $8\sqrt{2}$ για την υλοποίηση του HOG/HOF. Αυτό οδηγεί σε ένα αρκετά μεγάλο χωρικό μέγεθος του τεμαχίου, με διαστάσεις στον χώρο 114×114. Επιπλέον, βλέπουμε ότι οι χωρικές κλίμακες 4 και 5 παρουσιάζουν πολύ κοντινή μέση τιμή.

Στο Σχήμα 5.9 φαίνεται η κατανομή των ποσοστών εμφάνισης κάθε χωρικής και χρονικής κλίμακας ανίχνευσης, υπολογισμένη στο σύνολο των ανιχνεύσεων που εξήγαγε ο dca3D στα δείγματα των τριών δράσεων. Καταρχήν, μπορούμε να διακρίνουμε τη σαφή υπεροχή της πέμπτης χρονικής κλίμακας, που αντιστοιχεί στο πιο υψίσυχνο κανάλι, έναντι των υπολοίπων. Η μορφή της κατανομής των χρονικών κλιμάκων καταδεικνύει το γεγονός ότι δεν είναι άτοπη η κβαντοποίηση τους από τον HOG/HOF σε δύο διακριτές τιμές ώστε να διακρίνεται το πέμπτο κανάλι από τα υπόλοιπα. Περνώντας στις χωρικές κλίμακες, εδώ η κατανομή είναι αρκετά διαφορετική, υπό την έννοια ότι συμμετέχουν με σημαντικά ποσοστά όλες οι κλίμακες στην παραγωγή ανιχνεύσεων, με μία ελαφριά επικράτηση της τέταρτης κλίμακας. Ακόμα και οι κλίμακες 1 και 2, που αντιστοιχούν σε κανάλια χαμηλού μέτρου κεντρικής συχνότητας, ευθύνονται από κοινού για περίπου το 24% των ανιχνεύσεων. Με άλλα λόγια, έχοντας πάντα κατά νου την κβαντοποίηση των κλιμάκων από την υλοποίηση του HOG/HOF, περίπου μία στις τέσσερεις ανιχνεύσεις επιφέρουν υπολογισμό του περιγραφέα σε τεμάχια με χωρικό μέγεθος μεγαλύτερο ή ίσο του 80×80.

Λαμβάνοντας υπόψη τα παραπάνω και πιστεύοντας ότι ο υψηλός αριθμός τεμαχίων μεγάλου μεγέθους στον χώρο ενδεχομένως να εισάγει πλεονάζουσα ή θορυβώδη πληροφορία πλήττοντας τη διακριτική ικανότητα των χαρακτηριστικών, προχωρήσαμε σε δύο επιπλέον πειράματα αναγνώρισης των τριών δράσεων. Για την αποφυγή των προβλημάτων κβαντοποίησης που αναφέραμε παραπάνω καθώς και για τη δυνατότητα αποφυγής μεγάλου μεγέθους τεμαχίων από τις χωρικές κλίμακες 1 και 2 αποφασίσαμε να χρησιμοποιήσουμε ως κλίμακες ανίχνευσης στον χώρο τις προκαθορισμένες τιμές της υλοποίησης του HOG/HOF. Για τις χρονικές κλίμακες ανίχνευσης διατηρήσαμε τη σχέση (5.1.14), δεδομένου ότι οι προκύπτουσες διαστάσεις των τεμαχίων στον χρόνο αντιστοιχούν σε λογικές επιλογές. Για το πρώτο πείραμα οι χωρικές κλίμακες 1 εώς 5 αντιστοιχίστηκαν στις τιμές $\{8, 4\sqrt{2}, 4, 2\sqrt{2}, 2\}$ ενώ για το δεύτερο η πρώτη κλίμακα τιμής 8 αντικαταστάθηκε με την τιμή $4\sqrt{2}$ καταλήγοντας στην αντιστοίχιση $\{4\sqrt{2}, 4\sqrt{2}, 4, 2\sqrt{2}, 2\}$. Στα εν λόγω πειράματα αποφύγαμε κλίμακες μεγαλύτερες της τιμής 8 καταλήγοντας σε μέγιστο μέγεθος τρισδιάστατου τεμαχίου $80 \times 80 \times 16$ για το πρώτο πείραμα και $57 \times 57 \times 16$ για το δεύτερο. Για το δεύτερο πείραμα αντικαταστήσαμε τη μεγαλύτερη τιμή κλίμακας 8 με μια μικρότερη από το προκαθορισμένο σετ κλιμάκων προκειμένου να αξιολογήσουμε την επίδραση της χρήσης μικρότερων τεμαχίων στην πιο "τραχειά"χωρική κλίμακα για την απόδοση του συνολικού πλαισίου ταξινόμησης. Τα αποτελέσματα των δύο πειραμάτων παρουσιάζονται συγκεντρωτικά στον Πίνακα 5.9.

	κλ. 1	**κλ. 2**	**κλ. 3**	**κλ. 4**	**κλ. 5**
Μέση Τιμή Χωρικής Κλίμακας	12.9	7.1	4.0	2.2	1.2
Μέση Τιμή Χρονικής Κλίμακας	9	5	4	3	2

Πίνακας 5.8: Μέση τιμή της χωρικής κλίμακας (σε pixels) και της χρονικής κλίμακας (σε frames) για τις πέντε κλίμακες της συστοιχίας των δισδιάστατων και μονοδιάστατων φίλτρων αντίστοιχα. Οι κλίμακες αριθμούνται από τα κανάλια χαμηλής συχνότητας προς τα κανάλια υψηλής συχνότητας στον διακριτό χώρο και χρόνο αντίστοιχα. Οι μετρήσεις αναφέρονται επί του συνόλου των 261 δειγμάτων των τριών δράσεων. (σημ.: Οι τιμές των χωρικών κλιμάκων δεν παρουσιάζονται στρογγυλοποιημένες για τη διάκριση των κλιμάκων 4 και 5).



Σχήμα 5.9: Κατανομή των χωρικών (επάνω) και χρονικών (κάτω) κλιμάκων ανίχνευσης, υπολογισμένες επί του συνόλου των τελικών ανιχνεύσεων του dca3D στα 261 δείγματα βίντεο του πειράματος.

	dca3D + HOG/HOF	
	spatial scales $\{8, 4\sqrt{2}, 4, 2\sqrt{2}, 2\}$	spatial scales $\{4\sqrt{2}, 4\sqrt{2}, 4, 2\sqrt{2}, 2\}$
FightPerson	92.3%	92.3%
HandShake	70.9%	73.0%
SitUp	45.9%	45.8%
mAP	69.70%	**70.37%**

Πίνακας 5.9: Αποτελέσματα Average Precision για τις τρεις δράσεις του πειράματος ταξινόμησης και mean Average Precision επί των τριών δράσεων για δύο εναλλακτικά σχήματα επιλογής των χωρικών κλιμάκων του Ανιχνευτή dca3D.

Παρόλο που τα παραπάνω πειράματα δεν υπερέβησαν το αποτέλεσμα 71.00% που είχαμε λάβει πιο πάνω, εντούτοις το δεύτερο πείραμα με τις τέσσερεις επιλογές χωρικών κλιμάκων ανίχνευσης απέφερε πολύ κοντινή τιμή για τη *mean average precision*. Από τη σύγκριση των αποτελεσμάτων για τις δύο διαφορετικές επιλογές των τιμών χωρικής κλίμακας του Πίνακα 5.9, παρατηρούμε αύξηση της επίδοσης *μέσης ακρίβειας* ταξινόμησης για τη δράση *HandShake* στο δεύτερο σενάριο και ταυτόσημες επιδόσεις για τις άλλες δύο δράσεις. Προκύπτει, επομένως, ότι εξαιτίας της συχνής εμφάνισης της πρώτης χωρικής κλίμακας στις ανιχνεύσεις από τον dca3D, είναι ευεργετικό για την απόδοση αναγνώρισης να τίθεται σε μια μικρότερη τιμή ώστε να αποφεύγεται ο υπολογισμός μεγάλου πλήθους περιγραφέων σε μεγάλου μεγέθους τοπικά τεμάχια.
Προσανατολιζόμενοι προς τη χρήση μικρότερων τιμών χωρικών κλιμάκων, προχωρήσαμε στη διεξαγωγή ενός τελευταίου πειράματος όπου χρησιμοποιήθηκαν πάλι τιμές κλιμάκων από τις απαιτούμενες της υλοποίησης του HOG/HOF. Παρατηρώντας τον Πίνακα 5.8 βλέπουμε ότι οι χωρικές κλίμακες ανίχνευσης 4 και 5 του dca3D έχουν παραπλήσια μέση τιμή, λαμβάνοντας υπόψη ότι αναφέρονται σε αριθμό pixel. Επιπλέον, οι παραπάνω κλίμακες αντιστοιχούν σε μεγάλη μερίδα των ανιχνεύσεων του dca3D (βλ. Σχήμα 5.9). Επομένως, υιοθετήσαμε τη λογική απόδοσης μιας ενιαίας μικρής τιμής χωρικής κλίμακας σε αυτές τις δύο κλίμακες. Συγκεκριμένα, αποδώσαμε στις αντίστοιχες ανιχνεύσεις την τιμή χωρικής κλίμακας 2. Οι χωρικές κλίμακες για το εν λόγω πείραμα αντιστοιχίζονται πλέον στις τιμές $\{4\sqrt{2}, 4, 2\sqrt{2}, 2, 2\}$. Με αυτόν τον τρόπο κινούμαστε συνολικά σε χαμηλότερες τιμές κλιμάκων και, κατά συνέπεια, χωρικού μεγέθους των τοπικών τεμαχίων. Τα αποτελέσματα του πειράματος ταξινόμησης των τριών δράσεων *FightPerson*, *HandShake* και *SitUp* με χρήση του Ανιχνευτή dca3D και των παραπάνω τιμών χωρικής κλίμακας παρατίθενται στον Πίνακα 5.10. Το νέο πείραμα απέφερε τιμή *mean average precision* μεγαλύτερη του αποτελέσματος 71.00% πρότερου πειράματος. Συγκεκριμένα, υπήρξε βελτίωση της απόδοσης σε σχέση με τα αποτελέσματα του Πίνακα 5.7 στη *μέση ακρίβεια* αναγνώρισης κάθε επιμέρους δράσης. Επιβεβαιώνεται έτσι η πεποίθησή μας ότι μικρότερες τιμές χωρικών κλιμάκων παρουσιάζονται καταλληλότερες για τον προσδιορισμό των τοπικών τεμαχίων περιγραφής.

	dca3D + HOG/HOF
	spatial scales $\{4\sqrt{2}, 4, 2\sqrt{2}, 2, 2\}$
FightPerson	92.0%
HandShake	71.1%
SitUp	51.8%
mAP	**71.63%**

Πίνακας 5.10: Αποτελέσματα Average Precision για τις τρεις δράσεις του πειράματος ταξινόμησης με ανιχνευτή τον dca3D και mean Average Precision επί των τριών δράσεων για τιμές χωρικών κλιμάκων ανίχνευσης $\{4\sqrt{2}, 4, 2\sqrt{2}, 2, 2\}$.

Συμπερασματικά, ο Ανιχνευτής dca3D παρουσιάζεται υποσχόμενος στο πλαίσιο της Αναγνώρισης Ανθρώπινων Δράσεων σε Βίντεο. Τα πειράματα που παρουσιάστηκαν σε αυτή τη διπλωματική εργασία καταδεικνύουν το γεγονός ότι μπορεί να οδηγήσει σε υψηλές αποδόσεις αναγνώρισης δράσεων σε ρεαλιστικά σκηνικά όπως αυτά της *Hollywood2 Actions Dataset*. Η εκμάθηση και ταξινόμηση ορισμένων δράσεων με χρήση ανιχνεύσεων από τον dca3D παρήγαγε αποτελέσματα που πλησιάζουν τις επιδόσεις ενός αποτελεσματικού ανιχνευτή, του Harris3D.
Ωστόσο, περαιτέρω έρευνα θα πρέπει να στοχεύσει στη βελτιστοποίηση των παραμέτρων του Ανιχνευτή dca3D, όπως το κατώφλι της καταστολής των μη μεγίστων και οι κλίμακες ανίχνευσης, κάτι που δεν είχαμε τον χρόνο να εξετάσουμε στην παρούσα διπλωματική εργασία. Ο προσδιορισμός των χωρικών κλιμάκων ανίχνευσης και του τρόπου εξαγωγής του μεγέθους που θα έχουν οι τοπικοί περιγραφείς φαίνονται εξαιρετικής σημασίας για τη βελτίωση της ευρωστίας του ανιχνευτή.
Τέλος, θα μπορούσαν να αναζητηθούν καί συνδυασμοί του Ανιχνευτή dca3D με άλλους περιγραφείς χωροχρονικών σημείων ενδιαφέροντος, όπως είναι ο επιτυχημένος Περιγραφέας HOG/HOF.

Κεφάλαιο 6

Συμπεράσματα

6.1 Συμβολή της διπλωματικής εργασίας

Η παρούσα διπλωματική εργασία ασχολήθηκε με την αναγνώριση ανθρώπινων δράσεων σε βίντεο από τη σκοπιά των αναπαραστάσεων βίντεο που μετέρχονται τοπικά χωροχρονικά χαρακτηριστικά. Βασικό μέλημα στην εργασία υπήρξε η μελέτη επιτυχημένων ανιχνευτών και περιγραφέων χωροχρονικών σημείων ενδιαφέροντος. Στα πλαίσια της εργασίας, υλοποιήσαμε και αξιολογήσαμε ποιοτικά σε μεμονωμένα δείγματα βίντεο τον αλγόριθμο για τον Ανιχνευτή Cuboid με έναν αποτελεσματικό από πλευράς χρόνων εκτέλεσης και διαχείρισης μνήμης κώδικα, που παρέχει παράλληλα τη δυνατότητα πλήρους ελέγχου των παραμέτρων στον χρήστη. Επιπλέον, υλοποιήσαμε έναν ολοκληρωμένο κώδικα για την εξαγωγή των περιγραφέων προσανατολισμού εμφάνισης και κίνησης των Dalal και Triggs, το οποίο δέχεται στην είσοδο ένα δείγμα βίντεο και υπολογίζει όλα τα εναλλακτικά σχήματα των εν λόγω περιγραφέων.
Μία από τις κύριες συνεισφορές της εργασίας μας ήταν η ενσωμάτωση χαρακτηριστικών αποδιαμόρφωσης εικόνων και ανάλυσης υφής κατά το στάδιο της ανίχνευσης και περιγραφής χωροχρονικών σημείων ενδιαφέροντος για το πρόβλημα της αναγνώρισης δράσεων. Πραγματοποιώντας πειράματα εκμάθησης και ταξινόμησης απαιτητικών ανθρώπινων δράσεων σε δείγματα ταινιών διαπιστώθηκε η ευεργετική επίδραση της ένταξης τέτοιων χαρακτηριστικών στην απόδοση αναγνώρισης του συνολικού συστήματος. Η αξιόλογηση της παραπάνω συνέργειας χαρακτηριστικών επιχειρείται, στην καλύτερη γνώση μας, για πρώτη φορά στο πλαίσιο της αναγνώρισης δράσεων. Επιπρόσθετα, για τις ανάγκες μείωσης του υπολογιστικού φόρτου που επιφέρει η εξαγωγή σε κάθε frame τέτοιων χαρακτηριστικών αποδιαμόρφωσης, βελτιστοποιήθηκαν τμήματα μιας πρότερης υλοποίησης αντικαθιστώντας χρονοβόρες για το λογισμικό MATLAB διαδικασίες με περισσότερο αποτελεσματικά σχήματα που στηρίζονται σε build-in συναρτήσεις. Προγραμματίσαμε επίσης όλα τα επιμέρους στάδια που σχετίζονται με την τεχνική Bag-Of-Features για την παραγωγή των τελικών αναπαραστάσεων βίντεο.
Καίρια συμβολή της διπλωματικής εργασίας για το πρόβλημα που μελετάται συνιστά ο σχεδιασμός και η υλοποίηση ενός νέου ανιχνευτή, του dca3D. Αναλύθηκε διεξοδικά το εννοιολογικό και μαθηματικό του υπόβαθρο ενώ διασαφηνίστηκαν και δικαιολογήθηκαν οι σχεδιαστικές επιλογές που έγιναν. Από πλευράς εφαρμογής, ο αντίστοιχος αλγόριθ-

μος προγραμματίστηκε πλήρως σε λογισμικό MATLAB μέσω του οποίου εξάγονται όλα τα ενδιάμεσα και τελικά χαρακτηριστικά σε βιώσιμους χρόνους εκτέλεσης. Ο εν λόγω κώδικας δίνει ιδιαίτερη έμφαση στη φιλοσοφία επεξεργασίας κατά τμήματα του αρχικού τρισδιάστατου όγκου με αποτέλεσμα να μην παρουσίαζονται προβλήματα πλήρωσης της μνήμης RAM κατά τα διάφορα στάδια εκτέλεσης ούτε για την περίπτωση μεγάλου μεγέθους δειγμάτων βίντεο ανθρώπινων δράσεων. Ο ανιχνευτής αξιολογήθηκε σε πειράματα ταξινόμησης δράσεων επί της απαιτητικής Βάσης Δεδομένων *Hollywood2 Actions Dataset*, τα οποία απέφεραν αποτελέσματα απόδοσης της αναγνώρισης δράσεων που στέκονται αρκετά κοντά σε αντίστοιχα αποτελέσματα που λάβαμε με τη χρήση του γνωστού ανιχνευτή Harris3D.
Τέλος, η εκτέλεση μαζικών πειραμάτων αναγνώρισης δράσεων που απαιτούσαν την εξαγωγή χαρακτηριστικών από μεγάλο αριθμό δειγμάτων βίντεο επέβαλε τη δημιουργία πλήθους συναρτήσεων για τη διαχείριση και την επεξεργασία μεγάλου μεγέθους δειγμάτων βίντεο και αρχείων κειμένου.

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

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

- Λεπτομερής ανασκόπηση της σχετικής εργασίας για ανιχνευτές και περιγραφείς τοπικών χωροχρονικών σημείων ενδιαφέροντος. Παρουσίαση του εννοιολογικού και μαθηματικού υποβάθρου τους καθώς και της φιλοσοφίας σχεδιασμού τους.

- Υλοποίηση του ανιχνευτή Cuboid και των περιγραφέων προσανατολισμού εμφάνισης και κίνησης των Dalal et al., συνοδευόμενες από σχετικά πειραματικά αποτελέσματα.

- Παρουσίαση της μεθόδου Bag-of-Features για την τελική αναπαράσταση των δειγμάτων βίντεο και των βασικών αρχών λειτουργίας των μηχανών ταξινόμησης SVM.

- Διεξοδική μελέτη των αλγορίθμων αποδιαμόρφωσης εικόνας μέσω χρήσης τελεστών ενέργειας και των μοντέλων αναπαράστασης υφής CCA και DCA.

- Τροποποίηση πρότερης υλοποίησης εξαγωγής των παραπάνω χαρακτηριστικών υφής για εικόνες και μείωση των χρόνων εκτέλεσης κατά την εφαρμογή της σε δείγματα βίντεο.

- Ενσωμάτωση χαρακτηριστικών αποδιαμόρφωσης εικόνας κατά την ανίχνευση και περιγραφή χωροχρονικών σημείων ενδιαφέροντος. Ποσοτική αξιολόγηση της επίδρασής τους στην απόδοση αναγνώρισης μεσω πειραμάτων ταξινόμησης δράσεων. Διεξαγωγή αντίστοιχων πειραμάτων με χρήση κλασικών σχημάτων για σκοπούς σύγκρισης.

- Αναλυτική παρουσίαση της θεμελίωσης και του σχεδιασμού ενός νέου ανιχνευτή, του dca3D, με βάση πρόσφατες ιδέες χωροχρονικού φιλτραρίσματος και εντοπι-

σμού κίνησης. Πλήρης διασαφήνιση των σχεδιαστικών επιλογών και των διακριτών σταδίων της υλοποίησης που αναπτύχθηκε.

- Αξιολόγηση της απόδοσης του ανιχνευτή dca3D μέσω πειραμάτων εκμάθησης και ταξινόμησης απαιτητικών ανθρώπινων δράσεων. Διεξαγωγή αντίστοιχων πειραμάτων με χρήση του ανιχνευτή Harris3D για συγκριτική αξιολόγηση των αποτελεσμάτων.

6.2 Μελλοντικές Κατευθύνσεις

Έχουμε την πεποίθηση ότι πολλαπλές πιθανές προεκτάσεις για μελλοντική έρευνα αναδύονται από τις ιδέες που πραγματεύτηκε η παρούσα διπλωματική εργασία και τα αντίστοιχα αποτελέσματα που προέκυψαν. Η παρουσίαση και η αξιολόγηση, έστω σε ποιοτικό επίπεδο για την πλειονότητά τους, των ανιχνευτών και περιγραφέων χωροχρονικών σημείων ενδιαφέροντος μπορεί να τροφοδοτήσει την εξαγωγή συμπερασμάτων, σκέψεων και ιδεών αναφορικά με τα χαρακτηριστικά που καθιστούν έναν τέτοιον ανιχνευτή ή περιγραφέα εύρωστο και αποτελεσματικό για την αναγνώριση δράσεων σε ρεαλιστικά σκηνικά βίντεο.
Τα εναλλακτικά χαρακτηριστικά που "δανείστηκε" η παρούσα εργασία από προσεγγίσεις σχετικών με το θέμα περιοχών της Όρασης Υπολογιστών όπως η αποδιαμόρφωση σημάτων στον χώρο και το χρόνο και η ανάλυση υφής των εικόνων, μπορούν να αποτελέσουν κίνητρο για περαιτέρω διερεύνηση της χρησιμότητάς τους στο πρόβλημα της αναγνώρισης ανθρώπινων δράσεων. Συγκεκριμένα, είδαμε ότι η εξαγωγή της κυρίαρχης σε επίπεδο pixel ενέργειας διαμόρφωσης στον χώρο για κάθε frame της ακολουθίας εικόνων παρέχει μια ταυτόχρονα πληροφοριακή και διακρίνουσα, από πλευράς οπτικής σημαντικότητας, δομή των εικόνων που δύναται να καθοδηγήσει αποτελεσματικά τη διαδικασία ανάδειξης σημαντικών σημείων στον χώρο. Την ευεργετική επίδραση τέτοιων χαρακτηριστικών κυρίαρχης χωρικής ενέργειας στο στάδιο ανίχνευσης την μελετήσαμε μόνο μέσω της χρήσης τους ως βήμα προεπεξεργασίας των εικόνων, που στη συνέχεια προωθήθηκαν σε γνωστό σχήμα ανιχνευτή. Κινητοποιημένοι από την πρόσφατη τάση στη σχετική έρευνα που επιδιώκει τον συνδυασμό διαφόρων ειδών χαρακτηριστικών για τη διαδικασία εκμάθησης και ταξινόμησης, πιστεύουμε ότι είναι χρήσιμο και ελπιδοφόρο να εξεταστεί καί η απευθείας συγχώνευση τέτοιων χαρακτηριστικών με εκείνα που προκύπτουν από κλασικούς ανιχνευτές με τη χρήση μηχανών SVM με *πολυκαναλικό πυρήνα (SVM with multichannel kernel)*.
Πολύς χώρος για μελλοντική τριβή και έρευνα υπάρχει στη δυνατότητα χρήσης χαρακτηριστικών αποδιαμόρφωσης στον χώρο και στον χρόνο σε σχήματα τοπικών περιγραφέων. Μελετήσαμε αναλυτικά τον τρόπο με τον οποίο μέσω των διαφόρων μορφών του αλγορίθμου *ESA*, υπό ρεαλιστικούς περιορισμούς, λαμβάνουμε την εκτίμηση των σημάτων πλάτους και συχνότητας διαμόρφωσης AM-FM σημάτων. Για την περίπτωση δισδιάστατων εικόνων, είδαμε ότι χαμηλής διάστασης περιγραφείς υφής, που συντίθενται από την ένταση της εικόνας, το πλάτος και τη συχνότητα διαμόρφωσης, αξιοποιούνται με επιτυχία για την κατάτμηση εικόνων. Με την ίδια λογική, θα μπορούσαν να αναζητηθούν προσεγγίσεις που, έπειτα από *Ανάλυση Κυρίαρχων Συνιστωσών* τόσο στον χώρο όσο

και στον χρόνο επί του αρχικού όγκου βίντεο, θα χρησιμοποιούν τις μετρήσεις των προκυπτόντων σημάτων πλάτους και διανύσματος συχνότητας σε τρισδιάστατα τοπικά τεμάχια γύρω από τα ανιχνευθέντα σημεία για την εξαγωγή τοπικών αναπαραστάσεων. Ουσιαστικό προϊόν της παρούσας διπλωματικής εργασίας υπήρξε ο σχεδιασμός και η αξιολόγηση για πρώτη φορά στο πλαίσιο της αναγνώρισης δράσεων ενός νέου ανιχνευτή χωροχρονικών σημείων ενδιαφέροντος, του dca3D. Τα προκαταρκτικά αποτελέσματα που λάβαμε από την εφαρμογή του είναι ενθαρρυντικά και υποσχόμενα καθώς πλησιάζουν τις επιδόσεις γνωστών αποτελεσματικών ανιχνευτών, όπως του Harris3D. Ωστόσο, περαιτέρω εργασία μπορεί να προσανατολιστεί στην μηχανική μάθηση παραμέτρων του ανιχνευτή όπως το ολικό κατώφλι της καταστολής των μη μεγίστων, οι κλίμακες ανίχνευσης και το προκύπτον μέγεθος των τοπικών τεμαχίων περιγραφής. Αυτό είναι δυνατό να επιτευχθεί μέσω κάποιας επαναληπτικής μεθόδου βελτιστοποίησης όπως αυτή του *gradient descent*. Ακόμα, σχετικά με τις συστοιχίες φίλτρων που χρησιμοποιούνται για το φιλτράρισμα στον χώρο και τον χρόνο, πλήθος διαφορετικών επιλογών τοποθέτησης των φίλτρων στο πεδίο συχνοτήτων μπορούν να μελετηθούν και να αξιολογηθούν.
Τέλος, η *Ανάλυση Κυρίαρχων Συνιστωσών (DCA)* παρέχει εύρωστα και πληροφοριακά χαρακτηριστικά με το κόστος, ωστόσο, σοβαρού υπολογιστικού φόρτου, ειδικότερα όταν εφαρμόζεται στον χώρο, όπως είδαμε. Εφικτή τροποποίηση, χωρίς σημαντική επίδραση στην ποιότητα αποδιαμόρφωσης, για την *DCA* στον χώρο θα μπορούσε να βρεθεί στη χρήση συστοιχίας μικρότερου αριθμού Gabor φίλτρων στον χώρο. Σε κάθε περίπτωση, το λογισμικό MATLAB παραμένει υπολογιστικά "αργό"για τις διαδικασίες φιλτραρίσματος και για τούτο απώτερος στόχος πρέπει να είναι ο προγραμματισμός του ανιχνευτή dca3D και γενικότερα της *Ανάλυσης Κυρίαρχων Συνιστωσών* στον χώρο και στον χρόνο σε γλώσσα προγραμματισμού *C* ή *C++*.
Συμπερασματικά, οι κυριότερες εφικτές μελλοντικές κατευθύνσεις που πυροδοτούνται από την παρούσα διπλωματική εργασία συνοψίζονται στα παρακάτω σημεία:

- Συγχώνευση τοπικών χωροχρονικών χαρακτηριστικών από κλασικά σχήματα ανίχνευσης - περιγραφής και περιγραφέων υφής των εικόνων για την εκμάθηση και ταξινόμηση ανθρώπινων δράσεων μέσω μηχανών SVM με πολυκαναλικό πυρήνα.

- Αναζήτηση τρισδιάστατων τοπικών αναπαραστάσεων - περιγραφέων που θα προκύπτουν από χαρακτηριστικά αποδιαμόρφωσης ύστερα από Ανάλυση Κυρίαρχων Συνιστωσών επί του αρχικού όγκου βίντεο τόσο στον χώρο όσο και στον χρόνο.

- Εξέταση εναλλακτικών επιλογών σχεδιασμού των συστοιχιών φίλτρων που εφαρμόζονται στον χώρο και στόν χρόνο. Αξιολόγηση της δυνατότητας χρήσης μικρότερου αριθμού καναλιών για τη συστοιχία χωρικών φίλτρων προκειμένου να μειωθεί ο υπολογιστικός φόρτος.

- Ρύθμιση των παραμέτρων του ανιχνευτή dca3D με χρήση επαναληπτικών μεθόδων βελτιστοποίησης.

- Υλοποίηση της Ανάλυσης Κυρίαρχων Συνιστωσών στον χώρο και στον χρόνο καθώς και του ανιχνευτή dca3D σε γλώσσα προγραμματισμού *C* ή *C++* για την μείωση των χρόνων εκτέλεσης.

Βιβλιογραφία

[1] A. Bobick and J. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[2] M. M. Ullah, S.N. Parizi, and I. Laptev, "Improving bag-of-features action recognition with non-local cues," in *British Machine Vision Conference*, 2010.

[3] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *British Machine Vision Conference*, p. 127, sep 2009.

[4] P. Maragos, *Image Analysis and Computer Vision*. N.T.U.A., 2005.

[5] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976–990, 2010.

[6] G. Johansson, "Visual perception of biological motion and a model for its analysis," *Attention, Perception, & Psychophysics*, vol. 14, no. 2, pp. 201–211, 1973.

[7] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision and Image Understanding*, vol. 104, no. 2-3, pp. 249–257, 2006.

[8] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *IEEE International Conference on Computer Vision*, (Nice, France), pp. 726–733, 2003.

[9] I. Laptev and T. Lindeberg, "Space-time interest points," in *IEEE International Conference on Computer Vision (ICCV 2003)*, pp. 432–439, 2003.

[10] C. Harris and M. Stephens, "A combined corner and edge detector," in *Alvey vision conference*, vol. 15, p. 50, Manchester, UK, 1988.

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

[12] I. Laptev, B. Caputo, C. Schuldt, and T. Lindeberg, "Local velocity-adapted motion events for spatio-temporal recognition," *Computer Vision and Image Understanding*, vol. 108, no. 3, pp. 207–229, 2007.

[13] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005.*, pp. 65–72, 2005.

[14] A. Oikonomopoulos, I. Patras, and M. Pantic, "Spatiotemporal salient points for visual recognition of human actions," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics,*, vol. 36, no. 3, pp. 710–719, 2006.

[15] T. Kadir and M. Brady, "Saliency, scale and image description," *International Journal of Computer Vision*, vol. 45, no. 2, pp. 83–105, 2001.

[16] K. Rapantzikos, Y. Avrithis, and S. Kollias, "Spatiotemporal saliency for event detection and representation in the 3d wavelet domain: potential in human action recognition," in *Proceedings of the 6th ACM international conference on Image and video retrieval*, pp. 294–301, ACM, 2007.

[17] K. Rapantzikos, Y. Avrithis, and S. Kollias, "Dense saliency-based spatiotemporal feature points for action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, pp. 1454–1461, IEEE.

[18] S. Wong and R. Cipolla, "Extracting spatiotemporal interest points using global information," in *IEEE International Conference on Computer Vision (ICCV 2007)*, pp. 1–8, IEEE, 2007.

[19] M. Bregonzio, S. Gong, and T. Xiang, "Recognising action as clouds of space-time interest points," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, pp. 1948–1955, IEEE, 2009.

[20] A. Gilbert, J. Illingworth, and R. Bowden, "Fast realistic multi-action recognition using mined dense spatio-temporal features," in *IEEE International Conference on Computer Vision*, pp. 925–931, IEEE, 2009.

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

[22] T. Lindeberg, "Feature detection with automatic scale selection," *International Journal of Computer Vision*, vol. 30, no. 2, pp. 79–116, 1998.

[23] W. Förstner and E. Gülch, "A fast operator for detection and precise location of distinct points, corners and centres of circular features," in *Proc. ISPRS Intercommission Conference on Fast Processing of Photogrammetric Data*, pp. 281–305, 1987.

[24] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, pp. 1–8, IEEE, 2008.

[25] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, pp. 2929–2936, IEEE, 2009.

[26] T. Tuytelaars, "Dense interest points," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR2010)*, pp. 2281–2288, 2010.

[27] D. Lowe, "Object recognition from local scale-invariant features," in *The Proceedings of the IEEE International Conference on Computer Vision, 1999.*, vol. 2, pp. 1150–1157, Ieee, 1999.

[28] J. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *International Journal of Computer Vision*, vol. 79, no. 3, pp. 299–318, 2008.

[29] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proceedings of the 15th international conference on Multimedia*, pp. 357–360, ACM, 2007.

[30] A. Kläser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *British Machine Vision Conference*, pp. 995–1004, Citeseer, 2008.

[31] Y. Ke, R. Sukthankar, and M. Hebert, "Efficient visual event detection using volumetric features," in *International Conference on Computer Vision*, vol. 1, pp. 166–173, October 2005.

[32] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, vol. 1, pp. 886–893, IEEE, 2005.

[33] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," *Computer Vision–ECCV 2006*, pp. 428–441, 2006.

[34] I. Laptev, "Improvements of object detection using boosted histograms," in *British Machine Vision Conference*, vol. 3, pp. 949–958, 2006.

[35] H. Teager and S. Teager, "Evidence for nonlinear sound production mechanisms in the vocal tract," *Speech production and speech modelling*, vol. 55, 1990.

[36] J. Kaiser, "On a simple algorithm to calculate the energy of a signal," in *International Conference on Acoustics, Speech, and Signal Processing, 1990. ICASSP-90.*, pp. 381–384, IEEE, 1990.

[37] P. Maragos, J. Kaiser, and T. Quatieri, "Energy Separation in Signal Modulations with Application to Speech Analysis," *IEEE Transactions on Signal Processing*, vol. 41, no. 10, 1993.

[38] P. Maragos and A. Bovik, "Image demodulation using multidimensional energy separation," *Journal of the Optical Society of America A*, vol. 12, no. 9, pp. 1867–1876, 1995.

[39] A. Bovik, N. Gopal, T. Emmoth, and A. Restrepo, "Localized measurement of emergent image frequencies by gabor wavelets," *IEEE Transactions on Information Theory,*, vol. 38, no. 2, pp. 691–712, 1992.

[40] J. Havlicek, D. Harding, and A. Bovik, "The multicomponent am-fm image representation," *IEEE Transactions on Image Processing,*, vol. 5, no. 6, pp. 1094–1100, 1996.

[41] J. Havlicek, D. Harding, and A. Bovik, "Multidimensional quasi-eigenfunction approximations and multicomponent am-fm models," *IEEE Transactions on Image Processing,*, vol. 9, no. 2, pp. 227–242, 2000.

[42] I. Kokkinos, G. Evangelopoulos, and P. Maragos, "Texture analysis and segmentation using modulation features, generative models, and weighted curve evolution," *IEEE transactions on Pattern Analysis and Machine Intelligence*, pp. 142–157, 2009.

[43] D. Dimitriadis and P. Maragos, "Continuous energy demodulation methods and application to speech analysis," *Speech communication*, vol. 48, no. 7, pp. 819–837, 2006.

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

[45] N. Cristianini and J. Shawe-Taylor, "An introduction to support vector machines and other kernel-based learning methods," *An Introduction to Support Vector Machines and Other Kernelbased Learning Methods*, vol. 3, no. 1, 2000.

[46] S. Haykin, "Neural networks and learning machines," *McMaster University, Canada*, 2007.

[47] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *International Journal of Computer Vision*, vol. 73, no. 2, pp. 213–238, 2007.