ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ, ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ & ΔΙΚΤΥΩΝ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

"Video analysis for event and activity detection and recognition"

ΣΤΕΡΓΙΟΣ ΠΟΥΛΑΡΑΚΗΣ

Επιβλέποντες καθηγητές: κ. Χούστης Ηλίας
κ. Κατσαβουνίδης Ιωάννης
κα. Χούστη Αικατερίνη
κα. Μπριασούλη Αλεξία

Βόλος, Ιούλιος 2010

Στην αγάπη,

και σ' όλους όσους μας συνδέει

το όμορφο αυτό συναίσθημα.

Περίληψη

Σε αυτήν την μεταπτυχιακή διπλωματική εργασία ασχοληθήκαμε με το ζήτημα της ανίχνευσης, αναγνώρισης και ανάλυσης κίνησης σε ακολουθίες εικόνων *(video)*.

Στο πρώτο κεφάλαιο κάνουμε μια γενική εισαγωγή στο αντικείμενο της Τεχνητής Όρασης και παρουσιάζουμε κάποιες εφαρμογές και ανοιχτά ζητήματα, καθώς και την δομή ενός *απλού* συστήματος Τεχνητής Όρασης.

Στο δεύτερο κεφάλαιο παρουσιάζουμε κάποιες από τις βασικές έννοιες και αρχές της επεξεργασίας και ανάλυσης video, και παρουσιάζουμε τα απαραίτητα εργαλεία καθώς και την αντίστοιχη σημειογραφία *(notation)* που θα χρησιμοποιήσουμε σε όλα τα υπόλοιπα κεφάλαια.

Στο τρίτο κεφάλαιο ασχολούμαστε με το ζήτημα της ανίχνευσης κίνησης και εισάγουμε την έννοια των ενεργών frames, όπου δηλαδή παρατηρείται κάποια *σημαντική* κίνηση. Προτείνουμε τρεις μεθόδους και παρουσιάζουμε τόσο τα πειραματικά αποτελέσματα όσο και τα συμπεράσματά μας.

Στο τέταρτο κεφάλαιο ασχολούμαστε με την εξαγωγή δύο βασικών ιδιοτήτων κίνησης. Προτείνουμε δύο μεθόδους για κάθε ιδιότητα, και τις αξιολογούμε πειραματικά, οδηγούμενοι σε άριστα αποτελέσματα.

Στο πέμπτο κεφάλαιο ασχολούμαστε με την σειριακή ανίχνευση αλλαγών με την χρήση της δημοφιλούς μεθόδου CUSUM. Τροποποιούμε την μέθοδο και προτείνουμε έναν τρόπο μείωσης του χώρου αναζήτησης. Αξιολογούμε πειραματικά την μέθοδο για Translational κινήσεις και διαπιστώνουμε μια άριστη συμπεριφορά. Δυστυχώς, η μέθοδος αυτή δεν αποδίδει για no Translational κινήσεις, σκιαγραφούμε όμως μια πιθανή λύση, εντάσσοντάς την στα άμεσα μελλοντικά μας σχέδια.

Στο έκτο κεφάλαιο ασχολούμαστε με την λεπτομερή ανάλυση της κίνησης. Διευρύνουμε τον ορισμό του στιγμιοτύπου πλήρους κίνησης (FAI) και των σημάτων που ακολουθούν την κίνηση και αξιολογούμε 23 σήματα ως προς την δυσκολία ανίχνευσης FAI. Τα συμπεράσματά μας περιλαμβάνουν την εφαρμογή της ανάλυσης κίνησης τόσο πριν όσο και μετά την αναγνώριση κίνησης.

Στο έβδομο κεφάλαιο ασχολούμαστε με την αναγνώριση κίνησης. Παρουσιάζουμε τις πιο πρόσφατες προσπάθειες και διατυπώνουμε κάποια συμπεράσματα σχετικά με την απαραίτητη ποσότητα πληροφορίας που χρειάζεται για την αναγνώριση.

Τέλος, διατυπώνουμε τα συμπεράσματα μας και περιγράφουμε κάποιους από τους στόχους μελλοντικής εργασίας μας.

Ευχαριστίες

Σ’ αυτό το σημείο θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή κύριο Χούστη Ηλία, για την πραγματικά πολύτιμη βοήθειά του καθ’ όλη την διάρκεια των σπουδών μου. Η σημαντική καθοδήγηση του ήταν καθοριστικής σημασίας για την εκπόνηση της παρούσας μεταπτυχιακής διπλωματικής εργασίας.

Θα ήθελα ακόμη να ευχαριστήσω την κυρία Χούστη Αικατερίνη για την εξίσου σημαντική συμβολή της στην ολοκλήρωση της παρούσας εργασίας.

Επίσης, θα ήθελα να ευχαριστήσω τον κύριο Κατσαβουνίδη Ιωάννη για τις πολύτιμες συμβουλές του αλλά και για τις νέες βάσεις μάθησης που έθεσε καθ’ όλη την διάρκεια των μεταπτυχιακών μου σπουδών.

Θα ήθελα ακόμη να πω ένα μεγάλο ευχαριστώ στην κυρία Μπριασούλη Αλεξία, χάρη στην οποία ολοκληρώθηκε η διπλωματική μου διατριβή. Οι χρήσιμες συμβουλές, η συνεχής υποστήριξη, και κυρίως η δημιουργία κλίματος εμπιστοσύνης και κατανόησης, με βοήθησαν να αντιμετωπίσω θαρρετά όλα τα δύσκολα σημεία.

Σίγουρα χρωστάω ένα μεγάλο ευχαριστώ στους γονείς μου και τον αδερφό μου για την διαρκή τους στήριξη και ενθάρρυνση σε όλες τις στιγμές, στις καλές και στις δύσκολες. Θα ήθελα να ευχαριστήσω ιδιαίτερα τον αδερφό μου για την καθημερινή υπομονή και κατανόηση που επέδειξε σε μια φορτωμένη για μένα περίοδο.

Τέλος, θα ήθελα να ευχαριστήσω ιδιαίτερα την Χρύσα για την διαρκή υποστήριξη και συμπαράσταση καθ’ όλη την διάρκεια των σπουδών μου καθώς και τους συμφοιτητές μου και φίλους μου για την συνεχή τους ενθάρρυνση: Μαρία, Μάρθα, Κωνσταντίνα, Κώστα, Σεραφείμ, Θάνο, Βαλάντη, Χάρη, Ηλία, Νίκο, Γιώργο, Αποστόλη.

Περιεχόμενα

Κεφάλαιο 1

“Εισαγωγή στην ανίχνευση, αναγνώριση και ανάλυση κίνησης”

Σελίδα σκόπιμα κενή.

1.1 Εισαγωγή

Η Τεχνητή Όραση *(ή Όραση Υπολογιστών - Computer Vision)* είναι η περιοχή της Πληροφορικής που ασχολείται με την εξαγωγή χρήσιμων πληροφοριών από διδιάστατες προβολές (π.χ. φωτογραφίες) μιας σκηνής. Η Τεχνητή Όραση προσπαθεί να απαντήσει ερωτήσεις του τύπου «Που βρίσκεται ένας δρόμος σε μια αεροφωτογραφία;» ή «Υπάρχει κάποιος άνθρωπος μέσα σε αυτήν την εικόνα;». Για να το πετύχει αυτό προχωρεί σε ανάλυση των διαθέσιμων ψηφιακών πολυμέσων *(εικόνων και video).*

Τα τελευταία χρόνια η ανάλυση ψηφιακών πολυμέσων γίνεται συνεχώς όλο και πιο σημαντική καθώς χρησιμοποιείται σε όλο και περισσότερες εφαρμογές της καθημερινής ζωής. Ενδεικτικά αναφέρουμε τα συστήματα παρακολούθησης χώρων, την αυτόματη παρακολούθηση ηλικιωμένων ανθρώπων, την υποστήριξη ατόμων με ειδικές ανάγκες (π.χ. αναγνώριση της νοηματικής γλώσσας), την επικοινωνία ανθρώπου-μηχανής, τον σημασιολογικό ιστό, τον έλεγχο του κυκλοφοριακού συστήματος και την *διαιτησία* αθλημάτων.

Ιδιαίτερα ενθαρρυντικό είναι το γεγονός πως στις περισσότερες από τις εφαρμογές μας ενδιαφέρει περισσότερο η αναγνώριση και κατανόηση της ανθρώπινης συμπεριφοράς, κάτι που μαρτυρά μια στροφή προς έναν πιο ανθρωποκεντρικό ρόλο της επιστήμης της Πληροφορικής.

1.2 Γενική ανασκόπηση ενός συστήματος αναγνώρισης κίνησης

Παρόλο που η αναγνώριση κίνησης είναι ίσως το σημαντικότερο πρόβλημα της Τεχνητής Όρασης, ένα *σύστημα αναγνώρισης κίνησης* αποτελείται από πολλά υποσυστήματα που προσπαθούν να απαντήσουν σε ποικίλα άλλα ερωτήματα, όπως «Υπάρχει κάποιος άνθρωπος στην σκηνή;» ή «Προς ποια κατεύθυνση κινείται αυτός ο άνθρωπος;».

Πιστεύουμε μάλιστα πως η αποτελεσματικότερη αντιμετώπιση αυτών των προβλημάτων θα οδηγήσει σε καλύτερα αποτελέσματα τις υπάρχουσες και μελλοντικές μεθόδους αναγνώρισης κίνησης.

Στην συνέχεια παρουσιάζουμε συνοπτικά καθένα από τα υποσυστήματα ενός πλήρους συστήματος αναγνώρισης κίνησης.

Ανίχνευση ενεργών frames

Σε αυτό το στάδιο ανιχνεύονται τα frames στα οποία πραγματοποιείται κάποια σημαντική κίνηση. Λόγω θορύβου (π.χ. κίνηση κάμερας, κίνηση φύλλων, καιρικά φαινόμενα), υπάρχει η πιθανότητα ένα frame να «φαίνεται» πως είναι σημαντικό χωρίς να είναι στην πραγματικότητα. Το αρχικό video διαιρείται έτσι σε πολλά μικρά video, καθένα από τα οποία χαρακτηρίζεται ως *"ενεργό"* ή ως *"στατικό"*, ανάλογα με το αν πραγματοποιείται κάποια σημαντική κίνηση ή όχι. Τα επόμενα βήματα εφαρμόζονται στα *ενεργά* video καθώς τα στατικά περιέχουν μόνο θόρυβο. Με την ανίχνευση ενεργών frames ασχολούμαστε αναλυτικά στο κεφάλαιο 3.

Ανίχνευση ανθρώπου

Σε αυτό το βήμα προσπαθούμε να βρούμε αν υπάρχει κάποιος άνθρωπος στην σκηνή και σε ποια pixels ακριβώς εμφανίζεται. Οι μέθοδοι χωρίζονται σε δύο κατηγορίες: σε εκείνες που απαιτούν αφαίρεση background [43, 44, 45] και σε εκείνες που δεν απαιτούν κάποια τέτοια προεπεξεργασία αλλά προσπαθούν να ανιχνεύσουν τον άνθρωπο απευθείας από τα δεδομένα [46, 47, 48]. Μια εκτενής επισκόπηση αυτών των μεθόδων μπορεί να βρεθεί στο [49].

Ανίχνευση ενεργών pixels

Σε αυτό το στάδιο ανιχνεύονται τα ενεργά pixels στα οποία πραγματοποιείται η κίνηση, τα οποία και διαχωρίζονται από τα στατικά. Έχουν προταθεί διάφορες μέθοδοι [5, 4]. Μερικές από τις γνωστότερες μεθόδους ανίχνευσης ενεργών pixels παρουσιάζονται συνοπτικά στο κεφάλαιο 2.

Εξαγωγή ιδιοτήτων κίνησης

Σε αυτό το στάδιο εξάγουμε ορισμένες σημαντικές ιδιότητες της κίνησης, όπως αν πρόκειται για κίνηση κατά μήκος του άξονα *x* (π.χ. περπάτημα, τρέξιμο) ή όχι (π.χ. επιτόπιο άλμα) καθώς και η κατεύθυνση της κίνησης (αριστερά / δεξιά). Με την εξαγωγή των παραπάνω ιδιοτήτων κίνησης ασχολούμαστε αναλυτικά στο κεφάλαιο 4.

Ανίχνευση αλλαγών κίνησης

Σε αυτό το στάδιο προσπαθούμε να εντοπίσουμε τα χρονικά σημεία στα οποία μεταβαίνουμε από μία κίνηση σε μία άλλη διαφορετική. Κατ' αυτόν τον τρόπο, το

αρχικό video χωρίζεται σε μικρότερα video, στα οποία πραγματοποιείται μόνο μία κατηγορία κίνησης. Με την ανίχνευση αλλαγών κίνησης ασχολούμαστε αναλυτικά στο κεφάλαιο 5.

Αναγνώριση κίνησης

Σε αυτό το στάδιο πραγματοποιείται η αναγνώριση των κινήσεων που εκτελεί ο κινούμενος άνθρωπος. Υπάρχουν πάρα πολλές μέθοδοι αναγνώρισης [1, 2] με αρκετά καλά αποτελέσματα. Με την αναγνώριση κίνησης ασχολούμαστε πιο αναλυτικά στο κεφάλαιο 7.

Ανάλυση κίνησης

Σε αυτό το στάδιο πραγματοποιείται λεπτομερής ανάλυση της κίνησης, μέσω της οποίας προσπαθούμε να απαντήσουμε σε ερωτήματα όπως: «Πόσα βήματα κάνει αυτός ο άνθρωπος;» ή «Πόσο διαρκεί το πρώτο βήμα και πόσο το δεύτερο;». Σημαντικές εδώ είναι οι έννοιες του «στιγμιοτύπου πλήρους κίνησης - FAI» και των «σημάτων που ακολουθούν την κίνηση». Με την ανάλυση κίνησης ασχολούμαστε αναλυτικά στο κεφάλαιο 6.

Αναγνώριση ανθρώπων

Το κύριο ερώτημα εδώ είναι: «ποιος είναι αυτός ο άνθρωπος που κινείται;». Για να δοθεί απάντηση, χρησιμοποιούνται διάφορα βιομετρικά χαρακτηριστικά του κινούμενου ανθρώπου (π.χ. χαρακτηριστικά προσώπου, τρόπος βαδίσματος). Τα βιομετρικά χαρακτηριστικά μπορεί να έχουν υπολογιστεί από πριν, με δεδομένα εκπαίδευσης, και να αποθηκευτούν σε βάση δεδομένων. Το είδος των χαρακτηριστικών που χρησιμοποιούνται εξαρτάται από την εφαρμογή, για παράδειγμα αν παρέχονται καθαρές εικόνες προσώπων, είναι δυνατή η εφαρμογή μεθόδων αναγνώρισης προσώπων *(face recognition)* [50]. Αν, αντίθετα, είναι διαθέσιμες ακολουθίες εικόνων των ανθρώπων να περπατάνε από μακριά, έχει νόημα να χρησιμοποιηθούν μέθοδοι αναγνώρισης πόζας ή βαδίσματος *(pose, gait recognition)* [51]. Όταν υπολογιστούν τα σχετικά χαρακτηριστικά του βίντεο, ακολουθεί αναζήτηση σε μια βάση δεδομένων ώστε να βρεθεί αν ο άνθρωπος είναι γνωστός *(εάν υπάρχει ήδη στην βάση)* ή άγνωστος.

Εξαγωγή συμπερασμάτων

Αν και αυτό το βήμα ανήκει στον τομέα της Τεχνητής Νοημοσύνης και όχι στην Τεχνητή Όραση αναφέρεται για λόγους πληρότητας. Με βάση τα *χαρακτηριστικά χαμηλού επιπέδου* (είδος βαδίσματος, χαρακτηριστικά προσώπου κλπ), μπορούν να εξαχθούν συμπεράσματα για τα συμβάντα που λαμβάνουν χώρα και να ληφθούν αντίστοιχα οι αποφάσεις μελλοντικής δράσης του συστήματος, υλοποιώντας έτσι κατά κάποιον τρόπο την *«Λειτουργική Όραση» (Functional Vision)* του υπολογιστή.

Η ερμηνεία των αποτελεσμάτων της αναγνώρισης κίνησης είναι ορισμένες φορές αρκετά δύσκολη και λεπτή υπόθεση καθώς επηρεάζεται και από προσωπικούς ή πολιτισμικούς παράγοντες (για παράδειγμα, το νεύμα «όχι» των Ινδών ερμηνεύεται ως «ναι» από τους Έλληνες) καθώς και από χωρο-χρονικούς *(“Είναι το ίδιο μια γροθιά στο ρινγκ του μποξ και μια γροθιά μέσα σε μια τράπεζα;”)* και επομένως εξαρτάται από τους σκοπούς της κάθε εφαρμογής.

1.3 Γενικά προβλήματα στην αναγνώριση κίνησης

Κάποια από τα προβλήματα που καλείται να αντιμετωπίσει κάθε μέθοδος αναγνώρισης κίνησης είναι τα εξής:

Αξιοπιστία

Η αξιοπιστία μιας μεθόδου καθορίζει και το πόσο χρήσιμη θα είναι η συγκεκριμένη μέθοδος. Για παράδειγμα, μια μέθοδος που έχει σφάλμα μη επιτυχημένης αναγνώρισης 0.0001% θεωρείται εξαιρετικά χρήσιμη και έχει μεγάλες πιθανότητες να χρησιμοποιηθεί σε ιδιαίτερα κρίσιμες εφαρμογές (π.χ. συστήματα ασφαλείας). Αντίθετα, μια μέθοδος με σφάλμα 15% δεν μπορεί να χρησιμοποιηθεί σε κρίσιμες εφαρμογές και χρειάζεται οπωσδήποτε επιβεβαίωση των αποτελεσμάτων της από κάποιον άνθρωπο.

Επίλυση σε πραγματικό χρόνο

Αυτό είναι μάλλον το πιο κρίσιμο πρόβλημα. Γενικά μας ενδιαφέρει το σύστημά μας να είναι όσο πιο γρήγορο γίνεται, παρόλο που οι απαιτήσεις στην ταχύτητα εξαρτώνται και από τις απαιτήσεις κάθε εφαρμογής. Για παράδειγμα, αν θέλουμε να φτιάξουμε μια εφαρμογή που να ταξινομεί διάφορα video σε μια βάση δεδομένων για

μελλοντική αναζήτηση, η ταχύτητα εκτέλεσης δεν είναι τόσο κρίσιμος παράγοντας. Αν όμως θέλουμε να αναπτύξουμε μια ρομποτική εφαρμογή για ένα σύστημα ασφαλείας, τότε απαιτούμε πάρα πολύ μεγάλη ταχύτητα εκτέλεσης.

Ανεξαρτησία από συγκεκριμένο περιβάλλον κίνησης

Κάποιες φορές μπορεί να αναπτύξουμε μια μέθοδο ειδικά για κάποια συγκεκριμένη εφαρμογή. Αυτή η λύση ίσως να είναι εύκολο να υλοποιηθεί, παρουσιάζει όμως δυσκολίες προσαρμογής σε νέες μελλοντικές αλλαγές. Για παράδειγμα, αν έχουμε φτιάξει ένα σύστημα που αναγνωρίζει τις κινήσεις ενός παίκτη του τένις σε ένα γήπεδο με χορτάρι *(πράσινο τερραίν)*, το σύστημα αυτό δεν θα μπορέσει να λειτουργήσει σε ένα άλλο γήπεδο με χώμα *(καφέ τερραίν)*.

Ανεξαρτησία από την θέση της κάμερας

Είναι σημαντικό για μια μέθοδο να είναι ανεξάρτητη από την θέση της κάμερας λήψης. Για παράδειγμα, αν μια μέθοδος αναγνωρίζει επιτυχημένα μόνο κάποιον άνθρωπο που περπατάει σε γωνία 0^o σε σχέση με την κάμερα λήψης, τότε η μέθοδος αυτή δεν θα έχει μεγάλη πρακτική σημασία για τις περιπτώσεις που ακόμη και ο ίδιος άνθρωπος θα περπατάει σε γωνία 35^o .

Κινούμενη κάμερα / zoom

Η κίνηση της κάμερας δημιουργεί πρόσθετα προβλήματα και γενικά απαιτεί διαφορετικές προσεγγίσεις από τις περιπτώσεις μη κινούμενης κάμερας. Για παράδειγμα, καθώς η κάμερα κινείται παρουσιάζονται μεταβολές στην φωτεινότητα της εικόνας που μπορεί να εκληφθούν λανθασμένα σαν κίνηση χωρίς να υπάρχει όμως καμία κίνηση από τον άνθρωπο. Αντίστοιχα προβλήματα προκαλεί και το zoom.

Χειρισμός Occlusion (Occlusion handling)

Occlusion έχουμε όταν δύο ή περισσότεροι άνθρωποι καλύπτουν κοινά σημεία σε μια ακολουθία εικόνων. Σε αυτήν την περίπτωση είναι δύσκολο να ανιχνευθεί αποτελεσματικά ο κάθε άνθρωπος και φυσικά είναι ακόμη πιο δύσκολο να αναγνωριστούν επαρκώς οι κινήσεις τους. Ένα καλό σύστημα θα πρέπει να λαμβάνει κάποια μέριμνα σχετικά με το occlusion.

Θόρυβος

Όσο προσεκτικοί και να είμαστε κατά την διάρκεια της λήψης ενός video, πάντα θα υπάρχει παρουσία θορύβου. Ο θόρυβος μπορεί να αρκετά ορατός (κακή ποιότητα λήψης λόγω μη ποιοτικής μηχανής λήψης, παρεμβολή *άσχετων* αντικειμένων ή κινήσεων στην σκηνή λήψης, κακές καιρικές συνθήκες, έντονες φωτοσκιάσεις κτλ) ή και σχεδόν αόρατος (ανεπαίσθητη κίνηση κάμερας, μικρές αλλαγές στον φωτισμό κτλ). Η μέθοδος αναγνώρισης κίνησης πρέπει να είναι σε θέση να απομακρύνει ή να αντιμετωπίζει όσο πιο αποτελεσματικά γίνεται τον θόρυβο. Μια κύρια προσέγγιση στην βιβλιογραφία είναι να θεωρείται ότι ο θόρυβος είναι ανεξάρτητος από pixel σε pixel και ότι ακολουθεί την Κανονική Κατανομή (Gaussian).

1.4 Εφαρμογές της ανίχνευσης και αναγνώρισης ανθρώπινης κίνησης

Τα τελευταία χρόνια η ανάλυση ψηφιακών πολυμέσων γίνεται συνεχώς όλο και πιο σημαντική καθώς χρησιμοποιείται σε όλο και περισσότερες εφαρμογές της καθημερινής ζωής. Σε αυτήν την ενότητα παρουσιάζουμε συνοπτικά κάποιες από τις σημαντικότερες εφαρμογές των συστημάτων αναγνώρισης κίνησης.

Παρακολούθηση χώρων (Security and Surveillance)

Η αναγνώριση κίνησης είναι ιδιαίτερα κρίσιμη σε συστήματα ασφάλειας και παρακολούθησης χώρων (*Security and Surveillance Systems*). Συνήθως αυτά τα συστήματα βασίζουν την λειτουργία τους σε ένα δίκτυο από βιντεοκάμερες οι οποίες χειρίζονται από κάποιον υπεύθυνο ασφαλείας. Ωστόσο, πολλές προσπάθειες επιχειρούν την αντικατάσταση ή την υποβοήθηση του ανθρώπινου παράγοντα. Η αυτόματη αναγνώριση παράξενων ενεργειών *(anomalies)* είναι ένα από τα ζητήματα που έχουν απασχολήσει αρκετούς ερευνητές [53, 54].

Επικοινωνία ανθρώπου-μηχανής

Πριν την ανάπτυξη της Τεχνητής Όρασης, η επικοινωνία ανθρώπου-μηχανής γινόταν με τις κλασικές συσκευές εισόδου-εξόδου (π.χ. πληκτρολόγιο / ποντίκι – οθόνη / εκτυπωτής). Η Τεχνητή Όραση οδήγησε σε τρόπους επικοινωνίας περισσότερο ανθρωποκεντρικούς, καθώς και σε συνδυασμό τόσο λεκτικής όσο και μη λεκτικής επικοινωνίας (νεύματα, στάση του σώματος, εκφράσεις του προσώπου).

Behavioural Biometrics

Η Βιομετρία (Biometrics) ασχολείται με την μελέτη μεθόδων και αλγορίθμων με σκοπό την *μοναδική ταυτοποίηση (unique identification)* των ανθρώπων με βάση τα φυσικά χαρακτηριστικά τους (δακτυλικά αποτυπώματα, πρόσωπο, ίριδα των ματιών). Όλες αυτές οι μέθοδοι απαιτούν την συνεργασία των ανθρώπων για την συλλογή αυτών των φυσικών χαρακτηριστικών.

Πρόσφατα, έγινε αρκετά δημοφιλής ένας νέος τύπος βιομετρίας που ονομάζεται *Συμπεριφορική Βιομετρία* (*Behavioural Biometrics*) και μελετάει την ανθρώπινη συμπεριφορά (π.χ. βάδισμα). Το κύριο χαρακτηριστικό αυτής της προσέγγισης είναι πως η ανθρώπινη συνεργασία δεν είναι πλέον απαραίτητη και επιπλέον η αναγνώριση μπορεί να εκτελείται χωρίς να διακόπτεται ή να εμποδίζεται η δραστηριότητα του ανθρώπου για την συλλογή αυτών των στοιχείων.

Αυτόματη παρακολούθηση παιδιών και ηλικιωμένων ανθρώπων

Στην αυτόματη παρακολούθηση παιδιών και ηλικιωμένων ανθρώπων μας ενδιαφέρει πολύ η ανίχνευση συγκεκριμένων κινήσεων, όπως π.χ. η λήψη ενός φαρμάκου ή του φαγητού. Επιπλέον σημαντικό ρόλο έχει και η μη ανίχνευση μιας συγκεκριμένης κίνησης. Για παράδειγμα, η μη ανίχνευση της λήψης φαρμάκου πέραν μιας συγκεκριμένης ώρας θα έπρεπε να σημάνει κάποιον συναγερμό.

Υποστήριξη ατόμων με ειδικές ανάγκες

Η τεχνητή όραση μπορεί να φανεί ιδιαίτερα χρήσιμη στην υποστήριξη ατόμων με ειδικές ανάγκες. Στο πεδίο της επικοινωνίας ανθρώπου-μηχανής έχουν ήδη αναπτυχθεί αρκετά συστήματα για την υποστήριξη κατάλληλων τρόπων επικοινωνίας, όπως η κατανόηση της νοηματικής γλώσσας (π.χ. στο [52] χρησιμοποιούνται HMMs). Επιπλέον, οι γενικές αρχές των συστημάτων παρακολούθησης παιδιών και ηλικιωμένων ανθρώπων μπορούν να εφαρμοστούν και στην παρακολούθηση ατόμων με ειδικές ανάγκες.

Σημασιολογικός ιστός (semantic web)

Στόχος του σημασιολογικού ιστού είναι να γίνει επεξεργάσιμος ο τεράστιος όγκος των ψηφιακών πολυμεσικών δεδομένων που υπάρχουν στο διαδίκτυο. Τελικός σκοπός είναι να υπάρχουν δεδομένα καθορισμένα και συνδεδεμένα σημασιολογικά έτσι ώστε διάφορες εφαρμογές να μπορούν να τα επεξεργαστούν, να τα συνδυάζουν

και να τα επαναχρησιμοποιούν. Η αποτελεσματική ανάλυση και επεξεργασία ψηφιακών πολυμέσων θεωρείται απαραίτητη για την επίτευξη αυτού του στόχου.

Content-Based Video Analysis

Η συνεχής εξάπλωση της χρήσης του διαδικτύου (*internet*) έχει κάνει τον διαμοιρασμό video (*video sharing*) μια αρκετά δημοφιλής διαδικασία. Οι ιστοσελίδες που ασχολούνται με το διαμοιρασμό video (*video sharing websites*) γνωρίζουν όλο και μεγαλύτερη άνθηση, καθιστώντας έτσι επιτακτική την ανάγκη ανάπτυξης αποδοτικών μεθόδων ταξινόμησης και αποθήκευσης *(indexing and storage schemes)* ώστε να βελτιωθεί η αλληλεπίδραση με τον χρήστη. Ο στόχος αυτός οδηγεί στην περιγραφή ενός video με βάση το περιεχόμενό του (*content-based video summarization*), η οποία επηρεάζεται από την ανάκτηση εικόνας με βάση το περιεχόμενο της (*content-based image retrieval, CBIR*) [55]. Μια από τις πιο εμπορικές εφαρμογές της Content-Based Video Analysis είναι στα αθλητικά video [56].

Ρομποτικές εφαρμογές

Σήμερα όλο και πληθαίνουν οι προσπάθειες ανάπτυξης ρομποτικών κατασκευών με δυνατότητες επικοινωνίας με τον άνθρωπο. Η *όραση* ενός ρομπότ υλοποιείται συνήθως με μία ή περισσότερες κάμερες. Είναι επομένως αναγκαία η ευρεία χρήση ανάλυσης video και εικόνων, η αναζήτησή τους σε μια βάση δεδομένων κτλ

Έλεγχος του κυκλοφοριακού συστήματος

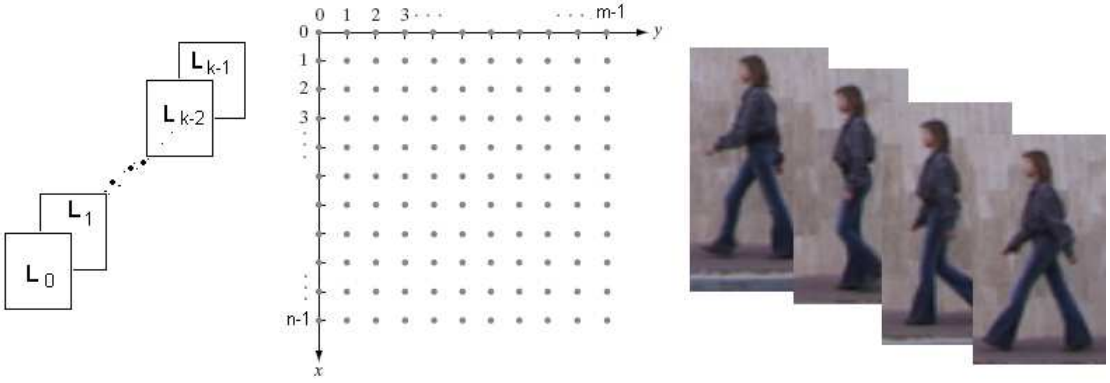
Πολλές φορές ο έλεγχος του κυκλοφοριακού συστήματος μπορεί να γίνεται αυτόματα μέσω συστημάτων παρακολούθησης χώρων. Ήδη χρησιμοποιούνται εφαρμογές για την καταγραφή παραβιάσεων του κώδικα οδικής κυκλοφορίας σε σημεία υψηλού κινδύνου εκδήλωσης ατυχημάτων ή κατά την διάρκεια της νύχτας.

Διαιτησία αθλημάτων

Η ανάλυση ψηφιακών πολυμέσων μπορεί να χρησιμοποιηθεί πολύ επιτυχημένα στην *διαιτησία* αθλημάτων. Η υψηλή ακρίβεια που μπορεί να επιτευχθεί μέσω των ψηφιακών συστημάτων μπορεί να αποδειχθεί ιδιαίτερα χρήσιμη κατά την λήψη κρίσιμων αποφάσεων σε περιπτώσεις που το ανθρώπινο μάτι δεν προλαβαίνει να *συλλάβει* ολόκληρη την κίνηση.

Κεφάλαιο 2

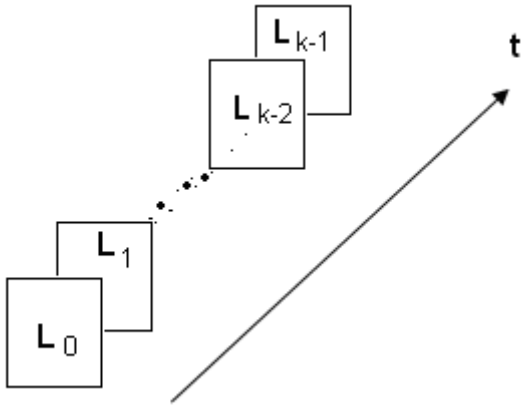
"Βασικές έννοιες επεξεργασίας και ανάλυσης video"



Σελίδα σκόπιμα κενή.

2.1 Το video σαν μονοδιάστατο σήμα

Ένα video* L μπορεί να θεωρηθεί σαν μια <u>*ακολουθία εικόνων*</u> $L = \{L_0, L_1, ..., L_{k-1}\}$. Το μήκος του video λέμε τότε ότι είναι k *frames*, όπου κάθε frame είναι μια εικόνα.

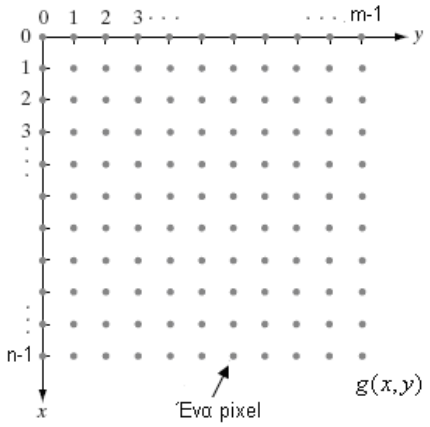


Επομένως, μια ακολουθία εικόνων (ή ένα video) θα μπορούσε να θεωρηθεί σαν ένα <u>μονοδιάστατο</u> σήμα που εξελίσσεται στον διακριτό χρόνο t, με τιμές $f(t) = L_t, t \in [0, k-1]$.

2.2 Το video σαν τρισδιάστατο σήμα

Ένα frame είναι μια εικόνα, ένα σύνολο επομένως από τιμές έντασης της φωτεινότητας κάθε εικονοστοιχείου *(pixel)*.

Όλα τα frames ενός video έχουν τις ίδιες ακριβώς διαστάσεις, έστω m ως προς τον οριζόντιο άξονα y και n ως προς τον κατακόρυφο άξονα *x*, με συνολικό αριθμό pixels $n \cdot m$ σε κάθε frame.



$$g(x,y) = \begin{vmatrix} g(0,0) & g(0,1) & \cdots & g(0,m-1) \\ g(1,0) & g(1,1) & \cdots & g(1,m-1) \\ \vdots & \vdots & & \vdots \\ g(n-1,0) & g(n-1,1) & \cdots & g(n-1,m-1) \end{vmatrix}$$

Ένα frame είναι ένας δισδιάστατος πίνακας n x m

*Αναφερόμαστε σε ασπρόμαυρο video, με ακέραιες τιμές στο διάστημα [0, 255]. Στις περιπτώσεις που έχουμε έγχρωμο video θα πρέπει πρώτα να το μετατρέψουμε σε ασπρόμαυρο.

Ένα frame λοιπόν είναι ένα διδιάστατο σήμα με τιμές

$$g(x, y), \forall x \in [0, n-1] \text{ και } y \in [0, m-1].$$

Εννοείται βέβαια πως κάθε frame L_t έχει την δική του συνάρτηση

$$L_t(x, y) = g_t(x, y).$$

Αν προσθέσουμε σαν τρίτη διάσταση τον διακριτό χρόνο t μπορούμε να πούμε πως ένα video είναι ένα τρισδιάστατο σήμα με τιμές

$h(x, y, t), \forall x \in [0, n-1],\ y \in [0, m-1]$ και $t \in [0, k-1]$

Η σύνδεση μεταξύ των *h* και *L* είναι η εξής:

$$h(x, y, t) = L_t(x, y)$$

$$\forall x \in [0, n-1],\ \ y \in [0, m-1] \ \kappa\alpha\iota\ t \in [0, k-1]$$

2.3 Χρονική διάρκεια του video

Η χρονική διάρκεια του video εξαρτάται από την χρονική διάρκεια ενός frame. Αυτή βρίσκεται από τον ρυθμό frame *(frame rate)*, που είναι ο αριθμός των frames σε ένα δευτερόλεπτο *(second).*

Τυπικές τιμές του frame rate είναι 24 frame/s, 25 (πρότυπο PAL), 29.97 (πρότυπο NTSC) μέχρι και 120 η και περισσότερο από σύγχρονες επαγγελματικές βιντεοκάμερες.

Αξίζει να σημειωθεί πως η χαμηλότερη τιμή για να εξαπατηθεί το ανθρώπινο μάτι είναι περίπου 15 frame/s.

Για όλη την υπόλοιπη εργασία θα μετράμε τον χρόνο σε frames, χωρίς να κάνουμε καμία υπόθεση για την τιμή του frame rate.

2.4 Παράγωγος ακολουθίας εικόνων

Είδαμε λοιπόν πως μια ακολουθία εικόνων μπορεί ισοδύναμα να θεωρηθεί σαν μια συνάρτηση είτε μιας μεταβλητής *(διάστασης)* είτε τριών.

Αν χρησιμοποιήσουμε την προσέγγιση των 3 διαστάσεων *x, y* και *t*, μπορούμε να ορίσουμε την μερική παράγωγο ως προς κάθε μια από αυτές τις διαστάσεις.

- Η μερική παράγωγος της $h(x, y, t)$ ως προς την διάσταση x σε κάποιο pixel (x, y, t) ορίζεται ως:

$$h_x(x\ , y\ , t\ ) = \frac{h(x+1, y, t) - h(x, y, t)}{(x+1) - x}$$

- Η μερική παράγωγος ως προς y ορίζεται ως:

$$h_y(x\ , y\ , t\ ) = \frac{h(x, y+1, t\ ) - h(x, y, t)}{(y+1) - y}$$

- Η μερική παράγωγος ως προς t ορίζεται ως:

$$h_t(x\ , y\ , t\ ) = \frac{h(x, y, t+1) - h(x, y, t)}{(t+1) - t}$$

Λόγω του ότι οι παράγωγοι ορίζονται πάντα σε γειτονικά pixels, οι παρονομαστές είναι πάντοτε ίσοι με την μονάδα και επομένως μπορούμε να γράψουμε πιο απλά :

- $h_x(x, y, t) = h(x+1, y, t) - h(x, y, t) = L_t(x+1, y) - L_t(x, y)$
- $h_y(x, y, t) = h(x, y+1, t) - h(x, y, t) = L_t(x, y+1) - L_t(x, y)$
- $h_t(x, y, t) = h(x, y, t+1) - h(x, y, t) = L_{t+1}(x, y) - L_t(x, y)$

Παρατηρούμε πως στην θέση της αρχικής ακολουθίας εικόνων $L = (L_1, L_2, ..., L_k)$ έχουμε τρεις νέες ακολουθίες εικόνων, τις L_x, L_y και L_t, όπου:

$$L_x = \{L_{x1}, L_{x2}, ..., L_{xp}\},\ p = k-1,$$

με

$$L_{xi} = \bigcup_{x=0,1,...,n-2}^{y=0,1,...,m-1} [L_i(x+1,y) - L_i(x,y)], \quad i = 0,1,\ldots,k-1$$

$$L_y = \{L_{y1}, L_{y2}, ..., L_{yp}\},\ p = k-1,$$

με

$$L_{yi} = \bigcup_{x=0,1,...,n-1}^{y=0,1,...,m-2} [L_i(x,y+1) - L_i(x,y)], \quad i = 0,1,\ldots,k-1$$

$$L_t = \{L_{t1}, L_{t2}, ..., L_{tp}\},\ p = k-2$$

με

$$L_{ti} = \bigcup_{x=0,1,...,n-1}^{y=0,1,...,m-1} [L_{i+1}(x,y) - L_i(x,y)], \quad i = 0,1,\ldots,k-2$$

2.5 Σήμα ενέργειας παραγώγου ακολουθίας εικόνων EL_t

Για κάθε μια από τις **k-1** εικόνες της ακολουθίας L_t μπορούμε να βρούμε την ενέργεια της ως εξής:

$$E\{L_{ti}\} = \sum_{x=0,1,...,n-1}^{y=0,1,...,m-1} \left| L_{ti}(x,y) \right|^2$$

Αν ορίσουμε μια ακολουθία EL_t ως εξής

$$EL_t(i) = E\{L_{ti}\},\ \ i = 0,\ 1,\ldots,k-2$$

παίρνουμε σαν αποτέλεσμα ένα μονοδιάστατο σήμα που δείχνει την χρονική εξέλιξη της ενέργειας της παραγώγου *(ως προς t)* της αρχικής ακολουθίας εικόνων. Το σήμα αυτό έχει ορισμένες αρκετά πολύ ενδιαφέρουσες ιδιότητες με τις οποίες θα ασχοληθούμε αναλυτικά στο κεφάλαιο 6.

2.6 Οπτική ροή (Optical flow)

Η παράγωγος ως προς τον χρόνο t, περιγράφει την μεταβολή της φωτεινότητας ενός pixel της θέσης (x, y) από την χρονική στιγμή t_i στην χρονική στιγμή t_{i+1}. Αντίθετα, η οπτική ροή *(optical flow)* περιγράφει την μετακίνηση του pixel της θέσης (x, y) στην νέα θέση (x', y').

Η οπτική ροή επιστρέφει λοιπόν δύο ακολουθίες εικόνων, sv_x και sv_y, μήκους $K-1$, όπου K το μήκος της αρχικής ακολουθίας. Η μετακίνηση του *pixel* της θέσης (x, y) από την χρονική στιγμή t_i στην χρονική στιγμή t_{i+1} περιγράφεται από το διάνυσμα $\langle sv_x(x, y, t_i),\ sv_y(x, y, t_i)\rangle$.

Η δημοφιλέστερη μέθοδος υπολογισμού της οπτικής ροής προτάθηκε το 1981 από τους Lucas και Kanade *[3]* και βασίζεται στις εξής τρεις υποθέσεις:

1. Διατήρηση της φωτεινότητας
   Η φωτεινότητα ενός pixel δεν αλλάζει καθώς αυτό κινείται από frame σε frame.
2. Μικρές μεταβολές κίνησης
   Τα pixel μετακινούνται αργά από frame σε frame.
3. Χωρική συνάφεια
   Τα γειτονικά pixel ανήκουν στην ίδια επιφάνεια και έχουν παρόμοια κίνηση.

2.7 Video με θόρυβο

Πολύ συχνά μπορεί να περιέχεται θόρυβος σε ένα video. Στην βιβλιογραφία, ο θόρυβος θεωρείται προσθετικός, ανεξάρτητος από pixel σε pixel και ότι ακολουθεί την Κανονική (Gaussian) με μηδενική μέση τιμή και διακύμανση σ^2. Στο [5] δείχνεται πως ακόμα κι αν ο θόρυβος δεν ακολουθεί στην πραγματικότητα την Κανονική Κατανομή, αυτή η υπόθεση είναι αρκετά ικανοποιητική.

Αν λοιπόν το frame περιέχει θόρυβο, τότε μπορούμε να πούμε πως το video αναπαρίσταται σαν $f(t) = L_t + z_t, t \in [0, k-1]$ όπου n_t είναι ο θόρυβος και $z_t \sim N(0, \sigma^2)$.

2.8 Motion Energy Images (MEI)

Είδαμε πως αν έχουμε μια ακολουθία εικόνων $L(x,y,t)$ μπορούμε να ορίσουμε την παράγωγό της $L_k(x,y,t)$ ως προς κάθε μια από τις διαστάσεις $k=\{x,y,t\}$. Επιπρόσθετα, μπορούμε να ορίσουμε μια δυαδική ακολουθία εικόνων $D(x,y,t)$ τέτοια ώστε:

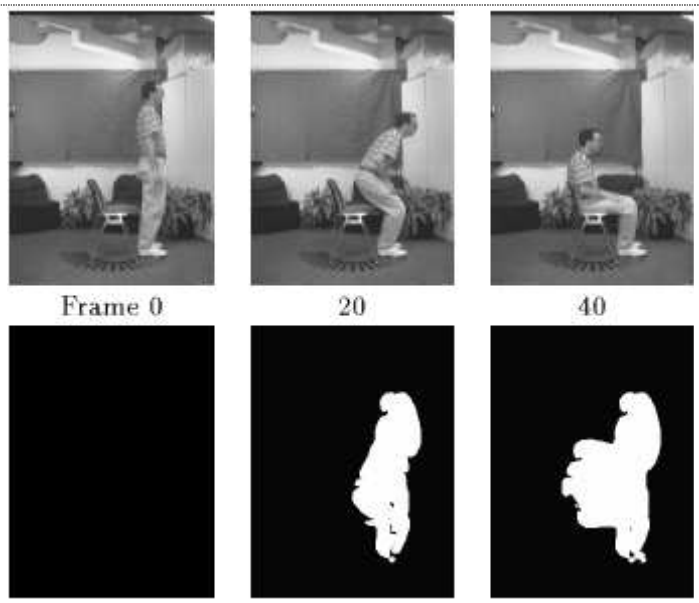
$$D(x,y,t)=\begin{cases}0, & \alpha\nu\ L_t(x,y,t)=0\\ 1, & \alpha\lambda\lambda ου\end{cases}$$

Τότε μπορεί να οριστεί το MEI $E_{\mathrm{T}}(x,y,t)$ της ακολουθίας ως εξής:

$$E_{\mathrm{T}}(x,y,t)=\bigcup_{i=1}^{\mathrm{T}} D(x,y,t-i)$$

Το MEI είναι μια εικόνα, όπως βλέπουμε και στο σχήμα 2.1, έχουμε δηλαδή αντιστοίχηση μιας ακολουθίας εικόνων σε μία μόνο εικόνα.

Ουσιαστικά λοιπόν το MEI προσπαθεί να βρει το σύνολο των ενεργών pixels στο σύνολο των frames του video. Περισσότερες λεπτομέρειες για τα MEI μπορούν να βρεθούν στο [4].



Σχήμα 2.1 Το MEI μιας ακολουθίας εικόνων όπου κάποιος άνθρωπος κάθεται σε μια καρέκλα

2.9 Motion History Images (MHI)

Το MHI είναι επέκταση του MEI και δείχνει την μεταβολή της κίνησης στην ακολουθία video. Σε ένα MHI η τιμή κάθε pixel είναι μια συνάρτηση της ιστορίας κίνησης του συγκεκριμένου pixel.

Μια συνάρτηση που χρησιμοποιείται συχνά είναι η εξής:

$$H_T(x,y,t) = \begin{cases} T, & \alpha\nu\ D(x,y,t)=1 \\ \max\ (\ 0,\ H(x,y,t-1)-1), & \alpha\lambda\lambda ου \end{cases}$$

Το αποτέλεσμα είναι μια εικόνα όπου τα πιο pixel με την πιο πρόσφατη κίνηση είναι φωτεινότερα. Κάποια παραδείγματα MHI φαίνονται στο σχήμα 2.2.

Περισσότερες λεπτομέρειες για τα MHI μπορούν να βρεθούν στο [4].

Αρχικό video	**MHI**
	
	
	

Σχήμα 2.2 ΤΟ MHI για κάποια παραδείγματα κίνησης.

2.10 Activity Areas

Τα Activity Areas (AA) προτάθηκαν στο [5] και μοιάζουν με τα MEI μιας και έχουμε αντιστοίχηση μιας ακολουθίας εικόνων σε μία μόνο εικόνα. Διαφέρουν ωστόσο στον τρόπο σχηματισμού της εικόνας.

Τα MEI σχηματίζονται από την απλή ένωση των σημείων της παραγώγου της ακολουθίας εικόνων $L(x, y, t)$. Αυτή η διαδικασία είναι πολύ απλή και δεν αναμένεται να είναι αποτελεσματική σε περιπτώσεις θορύβου, μεταβαλλόμενου φωτισμού και τρεμουλιάσματος της κάμερας. Τα Activity Areas αντίθετα χρησιμοποιούν μια πιο σύνθετη επεξεργασία για την επιλογή των pixel.

Η βασική ιδέα στα Activity Areas είναι η εξής:

Η αλλαγή στην τιμή ενός pixel μπορεί να οφείλεται είτε σε πραγματική κίνηση είτε σε θόρυβο. Επομένως έχουμε είτε *pixel κίνησης* είτε *στατικά pixel* που λόγω του θορύβου φαίνεται πως κινούνται. Είναι προφανές πως μας ενδιαφέρουν μόνο τα pixel κίνησης.

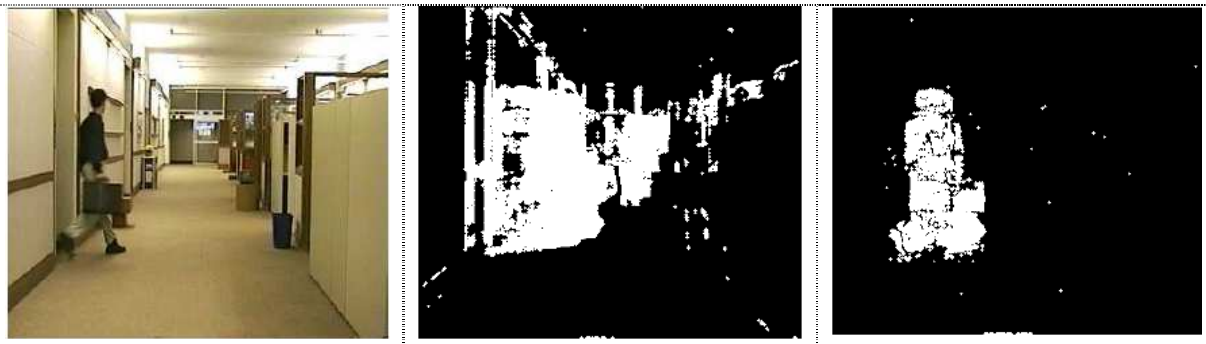
Για τον διαχωρισμό των pixel σε στατικά (που ακολουθούν την Κανονική Κατανομή) και κίνησης (που δεν ακολουθούν την Κανονική Κατανομή) ελέγχεται η μη-κανονικότητα (non-gaussianity) των δεδομένων [6]. Το κλασικό μέτρο μη-κανονικότητας μιας τυχαίας μεταβλητής είναι η Κύρτωση (kurtosis), η οποία ορίζεται ως εξής:

$kurt(X) = E[X^4] - 3(E[X^2])^2$ για την τ.μ. Χ.

Η κύρτωση είναι μηδέν αν η κατανομή της τυχαίας μεταβλητής είναι η Κανονική και διάφορη του μηδενός αν δεν είναι η Κανονική.

Σύγκριση με τα MEI

Τα Activity Areas γενικά υπερτερούν των MEI καθώς χρησιμοποιούν μια αρκετά καλή προσέγγιση του θορύβου, απομακρύνοντάς τον πολύ πιο αποτελεσματικά. Όπως φαίνεται και στο σχήμα 2.3, το Activity Area συγκεντρώνεται σχεδόν αποκλειστικά στην περιοχή κίνησης.



(α) *(β)* *(γ)*

Σχήμα 2.3. Σύγκριση MEI - Activity Area. (α) αρχική σκηνή, (β) MEI, (γ) Activity Area

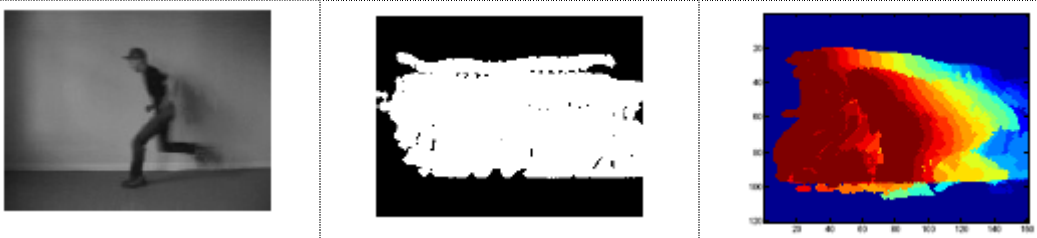
2.11 Activity History Areas

Τα Activity History Areas (AHA) προτάθηκαν στο [5] και σχετίζονται με τα MHI όπως σχετίζονται τα Activity Areas με τα MEI. Για τον σχηματισμό τους χρησιμοποιούμε τα Activity Areas, ως εξής:

Αν η τιμή του Activity Area AA στο pixel $\bar{r}$ είναι $AA(\bar{r})$ στο frame t, τότε το Activity History Area AHA είναι:

$$AHA(\bar{r},t)=\begin{cases}\tau, & \alpha\nu \quad AA(\bar{r})=1\\ \max(0,\ AHA(r,\ t-1)-1), & \alpha\lambda\lambda o\acute{u}\end{cases}$$

Τα AHA λοιπόν δίνουν περισσότερη πληροφορία σχετικά με την κίνηση σε σχέση με τα AA. Επιπλέον, είναι πολύ πιο ανθεκτικά στην παρουσία θορύβου σε σχέση με τα MHI.



(α) *(β)* *(γ)*

Σχήμα 2.4 (α) Χαρακτηριστικό frame από video με έναν άνθρωπο που τρέχει
(β) Το Activity Area του ίδιου video, (γ) Το Activity History Area του ίδιου video

2.12 Fourier Descriptor

Ο Fourier Descriptor [40] είναι μια μέθοδος περιγραφής του σχήματος μιας κλειστής καμπύλης που χρησιμοποιείται συχνά για την αναγνώριση εικόνας.

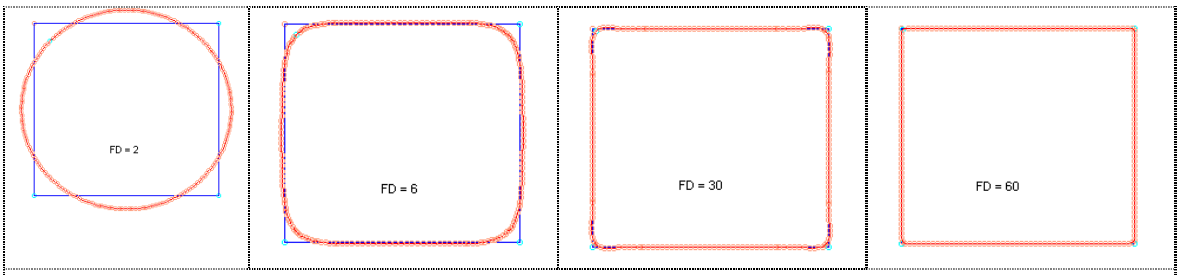
Αν το περίγραμμα της καμπύλης αποτελείται από N pixels με συντεταγμένες $(x_0, y_0), (x_1, y_1), \ldots, (x_{N-1}, y_{N-1})$, μπορούμε να ορίσουμε την ακολουθία μιγαδικών αριθμών $s_i = x_i + j \cdot y_i$, $i = 0,\ 1, ..., N-1$. Ο Διακριτός Μετασχηματισμός Fourier της ακολουθίας s_i σχηματίζει την ακολουθία F_k:

$$F_k = \frac{1}{N} \sum_{k=0}^{N-1} s_i \cdot \exp\left[\frac{-j2\pi ki}{N}\right],\ k = 0,1,..., N-1$$

Ο Fourier Descriptor σχηματίζεται κρατώντας το μέτρο των συντελεστών και διαιρώντας με το μέτρο του DC όρου F_0:

$$\overline{f} = \left[\frac{|F_1|}{|F_0|},\ \frac{|F_2|}{|F_0|},\ ...,\frac{|F_{N-1}|}{|F_0|}\right]$$

Ο όρος F_0 δηλώνει την θέση του σχήματος στον χώρο και επομένως δεν χρησιμοποιείται. Η διαίρεση ωστόσο με το $|F_0|$ εξασφαλίζει αμεταβλητότητα *(invariance)* ως προς την αλλαγή κλίμακας *(scaling)*. Επίσης, η χρήση του μέτρου εξασφαλίζει αμεταβλητότητα ως προς την περιστροφή *(rotation)*. Διαπιστώθηκε ακόμα πως οι χαμηλότερες συχνότητες αντιστοιχούν στην περιγραφή του μέσου σχήματος *(average shape)* ενώ οι υψηλότερες περιγράφουν λεπτομέρειες του σχήματος [41]. Επομένως, η χρήση λιγότερων από N συντελεστές επαρκεί για την αναγνώριση παρόμοιων σχημάτων. Στο σχήμα 2.5 φαίνεται η προσέγγιση ενός τετραγώνου χρησιμοποιώντας 2, 6, 30 και 60 συντελεστές. Η χρήση περισσότερων συντελεστών συνεπάγεται μεγαλύτερες απαιτήσεις ομοιότητας.



Σχήμα 2.5 Η προσέγγιση ενός τετραγώνου χρησιμοποιώντας 2, 6, 30 και 60 συντελεστές.

2.13 Datasets

Για την αξιολόγηση των μεθόδων ανίχνευσης, αναγνώρισης και ανάλυσης κίνησης πραγματοποιούνται μετρήσεις πάνω σε διεθνώς αποδεκτά video benchmarks. Τα video αυτά χωρίζονται σε βάσεις *(Datasets)* και δημιουργήθηκαν κυρίως για ερευνητικούς σκοπούς από ερευνητικές ομάδες που ήθελαν ακριβώς κάποια δείγματα για να εφαρμόσουν τις μεθόδους τους.

Δύο από τα πλέον γνωστά Datasets στην περιοχή της αναγνώρισης κίνησης είναι το **KTH** και το **Weizmann**.

2.13.1 KTH dataset

Το KTH Dataset πρωτοχρησιμοποιήθηκε στο [11] και περιέχει τους εξής 6 τύπους ανθρώπινης κίνησης: boxing, hand-waving, handclapping, jogging, running και walking. Οι κινήσεις εκτελούνται αρκετές φορές από 25 ανθρώπους σε τέσσερα διαφορετικά σενάρια συνθηκών: outdoors (d1), outdoors with scale variation/zoom (d2), outdoors with different clothes (d3) and indoors (d4). Κάθε video περιέχει μόνο έναν τύπο κίνησης. Συνολικά η βάση περιέχει 599 ασπρόμαυρα *(grayscale)* video, 300-500 frames το καθένα.

Οι τεχνικές αναγνώρισης κίνησης σε αυτό το dataset επιτυγχάνουν ποσοστά επιτυχίας μέχρι και 94% [12, 13], χρησιμοποιώντας όμως τεράστιες ποσότητες πληροφορίας εκπαίδευσης.

2.13.2 Weizmann dataset

Το Weizmann Dataset πρωτοχρησιμοποιήθηκε στο [15] και περιέχει τους εξής 10 τύπους ανθρώπινης κίνησης: Walk, Jump, Run, Pjump *(Jump in place)*, Bend, Jack, Side *(Gallop sideways)*, Skip, Wave1 *(One-hand wave)*, Wave2 *(Two-hands wave)*.

Συνολικά η βάση περιέχει 93 έγχρωμα video, 40-120 frames το καθένα.

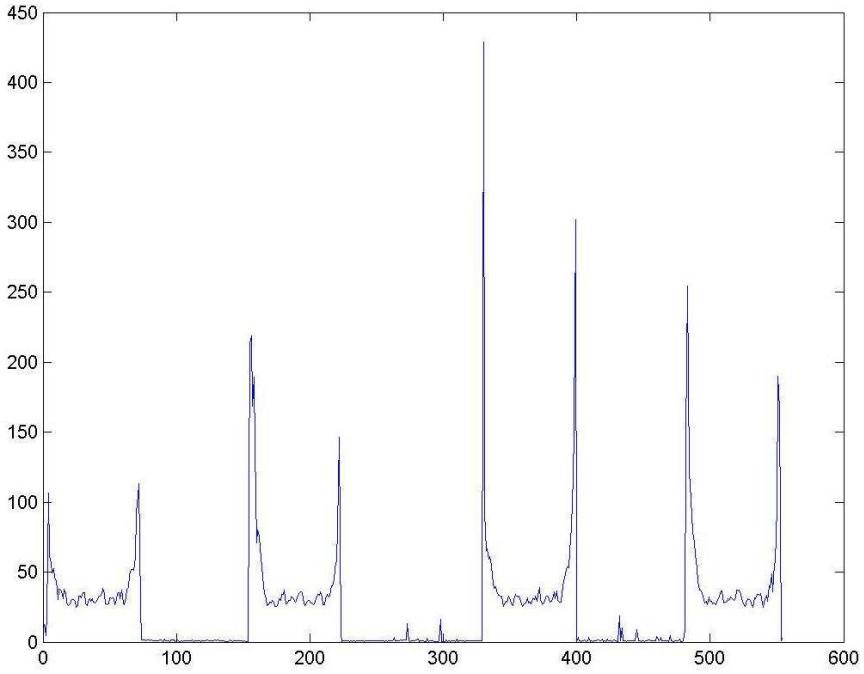
Οι τεχνικές αναγνώρισης κίνησης σε αυτό το dataset επιτυγχάνουν ποσοστά επιτυχίας μέχρι και 100% [14], χρησιμοποιώντας όμως τεράστιες ποσότητες πληροφορίας εκπαίδευσης.

Το Weizmann σίγουρα δεν έχει τον πλούτο διαφορετικών περιπτώσεων του KTH dataset. Επιπλέον, η κάμερα είναι στατική και το περιβάλλον *(background)* απλό. Ωστόσο, περιλαμβάνει περισσότερους τύπους κίνησης, γεγονός που το κάνει το ίδιο δημοφιλές. Επιπλέον, τα video του είναι έγχρωμα, και επομένως προσφέρεται και για εφαρμογή μεθόδων που λαμβάνουν υπόψη τους τον RGB χώρο αναπαράστασης.

Σελίδα σκόπιμα κενή.

Κεφάλαιο 3

“Ενεργά και μη ενεργά frames”



Σελίδα σκόπιμα κενή.

3.1 Εισαγωγικά

Όπως είπαμε και σε προηγούμενο κεφάλαιο, ένα video μπορεί να θεωρηθεί σαν μια ακολουθία εικόνων (frames). Είναι προφανές πως σχεδόν ποτέ δεν πρόκειται να υπάρξει μια άπειρη ακολουθία στην οποία θα πραγματοποιούνται συνεχώς *ενδιαφέρουσες* κινήσεις. Ο όρος "ενδιαφέρουσες" αλλάζει φυσικά από εφαρμογή σε εφαρμογή, αλλά γενικά μπορούμε να πούμε πως αναφέρεται σε κινήσεις για τις οποίες έχει νόημα και χρησιμότητα η ανίχνευση και αναγνώρισή τους.

Για παράδειγμα, σε ένα θαλασσινό τοπίο, η κίνηση των κυμάτων μπορεί να θεωρηθεί απλά σαν κινούμενο background ενώ αντίθετα το κολύμπι ενός ανθρώπου μπορεί να έχει πολύ περισσότερο νόημα για περαιτέρω ανάλυση.

Ορίζουμε λοιπόν την έννοια των ενεργών (active) frames ως εξής:

> Ένα frame ονομάζεται *ενεργό (active)* όταν πραγματοποιείται σε αυτό κάποια κίνηση ή δραστηριότητα, *πιθανώς ενδιαφέρουσα* για την συγκεκριμένη εφαρμογή που εξετάζουμε.

Αξίζει να προσέξουμε την έννοια της «*πιθανώς ενδιαφέρουσας δραστηριότητας*»: Δεν γνωρίζουμε σίγουρα αν η δραστηριότητα είναι σίγουρα ενδιαφέρουσα αλλά σίγουρα αξίζει τον κόπο να την εξετάσουμε καλύτερα.

Παρόλο που αυτός ο ορισμός ίσως μοιάζει πολύ γενικός, στην πραγματικότητα μας επιτρέπει να καλύψουμε ένα μεγάλο, προσαρμοζόμενο κάθε φορά, εύρος δραστηριοτήτων. Επιπλέον, η ίδια η έννοια της απόλυτης βεβαιότητας μπορεί να αποκτηθεί μόνο όταν πραγματοποιηθεί και η φάση της αναγνώρισης κίνησης, η οποία τελικά θα πραγματοποιηθεί μόνο στα ενεργά frames.

Προκύπτει λοιπόν το πρόβλημα της ανίχνευσης των *ενεργών* και *μη ενεργών* (ή *στατικών*) frames με όσο το δυνατόν πιο υψηλή ακρίβεια και όσο το δυνατόν μικρότερη χρονική καθυστέρηση.

Σε αυτό το κεφάλαιο θα επιχειρήσουμε μια διερεύνηση αυτού του προβλήματος και θα καταλήξουμε σε κάποια ενδιαφέροντα αποτελέσματα και συμπεράσματα.

3.2 Χρησιμότητα

Η χρησιμότητα της ανίχνευσης των ενεργών frames σε video μπορεί να γίνει περισσότερο εμφανής μέσα από τα επόμενα δύο παραδείγματα:

Παράδειγμα 1

Ας φανταστούμε πως σχεδιάζουμε ένα σύστημα ασφαλείας ενός κτιρίου. Μια κάμερα εγκαθίσταται στην κεντρική είσοδο και στέλνει εικόνες σε κάποιον υπολογιστή, στο εσωτερικό του κτιρίου. Κάποιες από τις εικόνες θα περιέχουν ανθρώπους οι οποίοι θα έρχονται κοντά στην κεντρική είσοδο ενώ οι περισσότερες θα δείχνουν κάποια «σχεδόν ακίνητη» εικόνα. Είναι προφανές πως όταν θα βρέχει, η κίνηση της βροχής θα μπερδεύει το σύστημα, το οποίο θα σημαίνει κάποιον λανθασμένο συναγερμό (false alarm). Το ζητούμενο είναι να ρυθμίσουμε το σύστημα ώστε να σημαίνει όσο το δυνατόν λιγότερους λανθασμένους συναγερμούς.

Παράδειγμα 2

Ας φανταστούμε πως θέλουμε να αναγνωρίσουμε την κίνηση που πραγματοποιείται σε ένα video. Είναι προφανές πως αν δεν πραγματοποιείται καμία κίνηση σε κάποια frames, τότε δεν έχει νόημα να ξεκινήσει η διαδικασία αναγνώρισης, καθώς θα οδηγήσει πιθανότατα σε λανθασμένα ή παραπλανητικά αποτελέσματα.

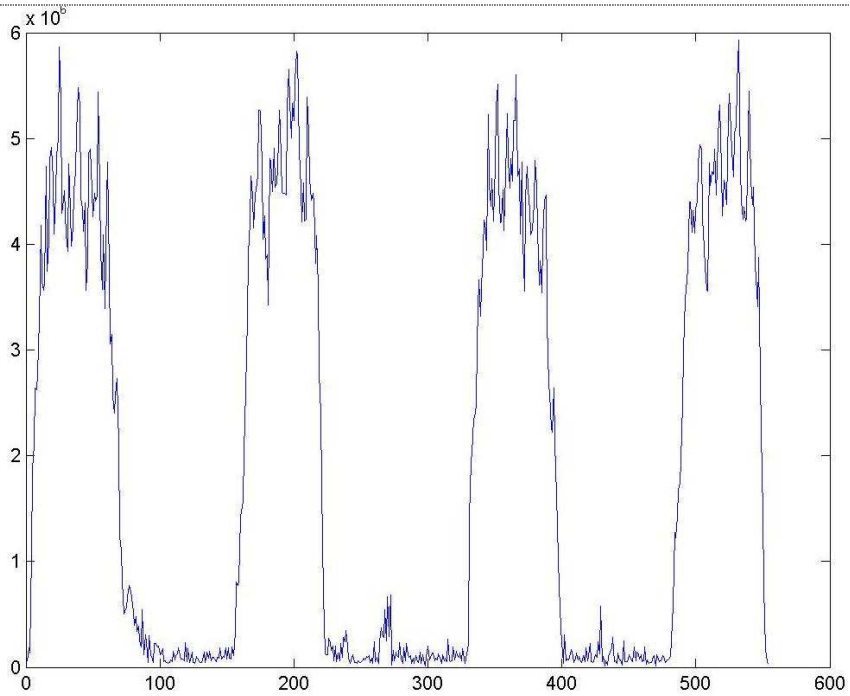
Στο πρώτο παράδειγμα η ανίχνευση των ενεργών frames είναι ίσως το πιο σημαντικό τμήμα της εφαρμογής. Αντίθετα, στο δεύτερο παράδειγμα, η ανίχνευση είναι ένα ενδιάμεσο στάδιο, το οποίο χρησιμεύει στην παραγωγή πιο ποιοτικών αποτελεσμάτων σε κάποιο από τα επόμενα βήματα (στην αναγνώριση κίνησης).

3.3 Ανίχνευση με την μέθοδο της ενέργειας

3.3.1 Περιγραφή της μεθόδου

Εφόσον στα ενεργά frames πραγματοποιούνται ενδιαφέρουσες δραστηριότητες, μπορούμε γενικά να υποστηρίξουμε πως τα ενεργά frames έχουν γενικά πολύ μεγαλύτερη κινητική ενέργεια από τα μη ενεργά frames. Για παράδειγμα, η βροχή προκαλεί πολύ μικρές μεταβολές στην εικόνα σε σχέση για παράδειγμα με έναν άνθρωπο που περπατάει.

Η κινητική ενέργεια σχετίζεται κάπως με το σήμα *ELt*. Το σήμα *ELt* ενός video από το KTH dataset φαίνεται στο σχήμα 3.1.

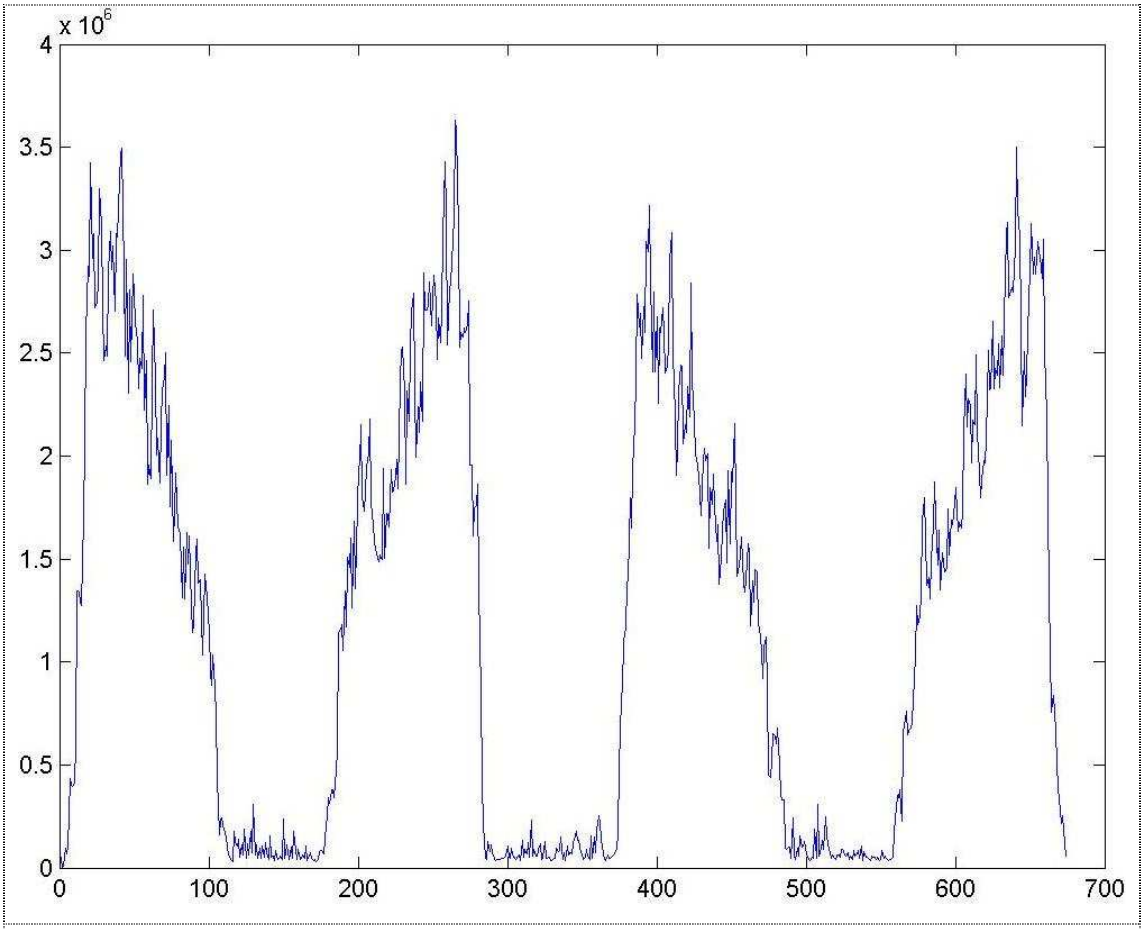


Σχήμα 3.1. Το σήμα ELt του video person01_walking_d1 του KTH dataset.

Παρατηρούμε πως στα frames όπου δεν έχουμε κίνηση αλλά μόνο κάποιο θορυβώδες background, οι τιμές του *ELt* είναι πολύ χαμηλές. Αντίθετα, στα frames όπου έχουμε κίνηση οι τιμές του *ELt* είναι πολύ υψηλές και κινούνται γύρω από μία κεντρική τιμή. Σε αυτήν την περίπτωση λοιπόν, η διάκριση μεταξύ ενεργών και μη ενεργών frames είναι μια αρκετά εύκολη υπόθεση.

Ωστόσο, όταν εξετάζουμε πιο πολύπλοκα video όπως είναι αυτά της κατηγορίας d2, η διάκριση είναι δυνατή αλλά όχι το ίδιο εύκολη. Στο σχήμα 3.2 βλέπουμε πως το

σήμα *ELt* παίρνει τιμές σε ένα πολύ μεγαλύτερο διάστημα τιμών ενώ πλέον δεν κινείται γύρω από κάποια κεντρική τιμή.



Σχήμα 3.2. Το σήμα ELt του video person01_walking_d2 του KTH dataset.

Βλέπουμε λοιπόν πως το σήμα *ELt* μας επιτρέπει να σχεδιάσουμε ένα σύστημα ανίχνευσης ως εξής:

- Υπολογίζουμε τα σήματα *ELt* διάφορων ανθρώπων που εκτελούν διάφορες δραστηριότητες σε εκπαιδευτικά βίντεο (training videos). Από αυτά τα σήματα προσδιορίζουμε ένα κατώφλι T (μπορεί να σχετίζεται με το καθολικό ελάχιστο ή με τα στατιστικά των σημάτων) για να βρίσκουμε αν υπάρχει έντονη δραστηριότητα σε ένα βίντεο.
- Χρησιμοποιούμε αυτό το κατώφλι σε άγνωστο video και χαρακτηρίζουμε όσα frames το υπερβαίνουν ως “ενεργό” και τα υπόλοιπα ως “μη ενεργά”.

Η μέθοδος αυτή είναι κατάλληλη για χρήση σε πραγματικό χρόνο, καθώς απαιτεί μόνο την καθυστέρηση ενός frame, λόγω της χρήσης της παραγώγου.

3.3.2 Πειραματικά αποτελέσματα στο KTH dataset

Χρησιμοποιήσαμε τα video ενός ανθρώπου Α που εκτελεί τρεις κινήσεις (walk, jog, run) για εκπαίδευση *(training)* και τα αντίστοιχα video ενός άλλου ανθρώπου Β για έλεγχο *(testing)*. Στην φάση της εκπαίδευσης *(training)*, βρήκαμε την μέση τιμή των ενεργών frames για κάθε video του ανθρώπου Α και επιλέξαμε την μικρότερη για τιμή του κατωφλίου Τ. Στην φάση του ελέγχου *(testing)*, ανιχνεύσαμε τα ενεργά frames των video του ανθρώπου Β χρησιμοποιώντας το κατώφλι Τ που επιλέξαμε στην φάση της εκπαίδευσης.

Η μέθοδος πέτυχε ένα πολύ υψηλό ποσοστό σωστής ανίχνευσης των ενεργών frames κατά 95.86%. Αναλυτικά τα αποτελέσματα φαίνονται στον πίνακα 3.1.

Πίνακας 3.1. Απόδοση της μεθόδου της ενέργειας σε 4 διαφορετικά περιβάλλοντα, χρησιμοποιώντας και τα 4 περιβάλλοντα στην εκπαίδευση.

	Περιβάλλον d1	***Περιβάλλον d2***	***Περιβάλλον d3***	***Περιβάλλον d4***	***Μέση συμπεριφορά***
Ποσοστό (%)	97.88	94.84	97.56	93.16	95.86

Παρατηρούμε πως τα χαμηλότερα ποσοστά σημειώθηκαν στα περιβάλλοντα d2-d4 που έχουν πολύ θορυβώδες background *(zoom, φωτοσκιάσεις)*. Κρίνονται όμως αρκετά ικανοποιητικά.

Ωστόσο, παρά το υψηλό ποσοστό, το training set που χρησιμοποιήσαμε ήταν ίσως πολύ μεγάλο, καθώς χρησιμοποιήσαμε 12 video (3 κινήσεις x 4 περιβάλλοντα). Επιπλέον, το περιβάλλον δεν ήταν *άγνωστο*, καθώς στα εκπαιδευτικά video είχαμε τουλάχιστον ένα video για κάθε τύπο περιβάλλον. Θελήσαμε λοιπόν να μειώσουμε το μέγεθος του training set σε 3 video, καλύπτοντας πάλι το σύνολο των διαφορετικών κινήσεων αλλά χρησιμοποιώντας μόνο έναν τύπο περιβάλλοντος σαν γνωστό. Τα αποτελέσματα φαίνονται στον πίνακα 3.2.

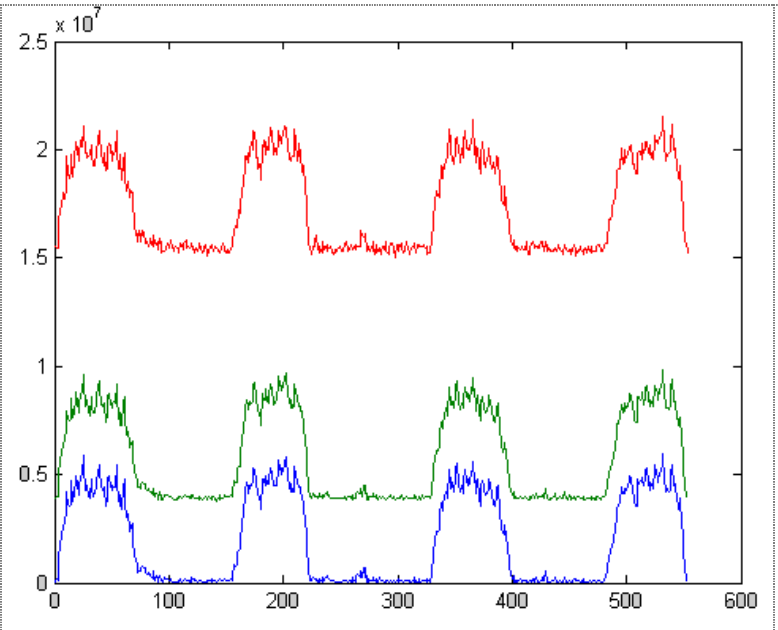
Παρατηρούμε πως τα ποσοστά γενικά είναι πολύ μικρότερα. Εξαίρεση αποτελεί η περίπτωση του d2, που έχει τα ίδια ποσοστά και φαίνεται πως ήταν το περιβάλλον με την μικρότερη μέση τιμή ενέργειας ενεργών frames. Το χαμηλότερο μέσο ποσοστό είναι το 85.61%, το οποίο ωστόσο είναι σχετικά υψηλό και επομένως η μέθοδος της ενέργειας μπορεί να λειτουργεί αρκετά καλά ακόμα και σε παντελώς άγνωστο περιβάλλον..

*Πίνακας 3.2. Απόδοση της μεθόδου της ενέργειας σε 4 διαφορετικά περιβάλλοντα, χρησιμοποιώντας κάθε φορά μόνο ένα περιβάλλον στην εκπαίδευση. Με * σημειώνεται η απόδοση ανίχνευσης χωρίς την χρήση του training background στην φάση του ελέγχου.*

Training/Testing background	***d1***	***d2***	***d3***	***d4***	***Μέση συμπεριφορά****	***Μέση συμπεριφορά***
d1	94.60	71.66	93.13	83.05	82.61	***85.61***
d2	97.88	94.84	97.56	93.16	96.20	***95.86***
d3	95.71	79.51	94.35	86.78	87.33	***89.09***
d4	95.71	79.51	94.35	86.78	89.86	***89.09***

3.3.3 Μειονεκτήματα της μεθόδου της ενέργειας

Η μέθοδος της ενέργειας αποδίδει πολύ καλά σε περιβάλλοντα με όσο γίνεται πιο στατικό background, χωρίς πολύ θόρυβο. Όταν όμως έχουμε θόρυβο, πρέπει να προσαρμόζουμε το κατώφλι με βάση τα χαρακτηριστικά του θορύβου. Δεν μπορούμε να χρησιμοποιήσουμε το ίδιο κατώφλι γιατί ο θόρυβος μεταβάλλει πολύ την ενέργεια τόσο των ενεργών όσο και των μη ενεργών frames, σε σημείο που τα στατικά frames με θόρυβο να έχουν πολύ μεγαλύτερη ενέργεια σε σχέση με τα ενεργά frames χωρίς θόρυβο, όπως φαίνεται και στο σχήμα 3.3.



Σχήμα 3.3. Το σήμα ELt του ίδιου video με την εφαρμογή προσθετικού Gaussian θορύβου $N(0,\sigma^2)$

Απαιτείται λοιπόν να λαμβάνουμε υπόψη μας τον θόρυβο και να προσαρμόζουμε ανάλογα το κατώφλι. Στην βιβλιογραφία ο θόρυβος θεωρείται προσθετικός και ότι ακολουθεί την Κανονική (Gaussian) με μηδενική μέση τιμή και διακύμανση σ^2. Αν λοιπόν το frame περιέχει θόρυβο, τότε μπορούμε να πούμε πως το video αναπαρίσταται σαν $f(t) = L_t + z_t, t \in [0, k-1]$ όπου n_t είναι ο θόρυβος και $z_t \sim N(0, \sigma^2)$.

Το σήμα ELt ενός θορυβώδους video λοιπόν είναι:

$$E\{L_{ti}'\} = \sum_{x=0,1,...,n-1}^{y=0,1,...,m-1} \left| L_{ti}(x,y) + z_{ti}(x,y) \right|^2$$

$$= \sum \left| L_{ti}(x,y) \right|^2 + \left| z_{ti}(x,y) \right|^2 + 2 * L_{ti}(x,y) * z_{ti}(x,y)$$

$$= \sum \left| L_{ti}(x,y) \right|^2 + \sum \left| z_{ti}(x,y) \right|^2 + \sum 2 * L_{ti}(x,y) * z_{ti}(x,y)$$

$$= E\{L_{ti}\} + E\{z_{ti}\} + 2 * \sum L_{ti}(x,y) * z_{ti}(x,y)$$

Παρατηρούμε πως η ενέργεια του θορυβώδους frame είναι ίση με την ενέργεια του frame χωρίς τον θόρυβο συν την ενέργεια του θορύβου συν τον όρο διασυσχέτισης $2 * \sum L_{ti}(x,y) * z_{ti}(x,y)$.

Ο όρος διασυσχέτισης οφείλεται στην μη γραμμική συμπεριφορά της ενέργειας και η ύπαρξη του καθιστά την ανάλυση σε θορυβώδες περιβάλλον από πολύ δύσκολη έως και αδύνατη. Αν δεν υπήρχε αυτός ο όρος, θα μπορούσαμε να υπολογίσουμε στατιστικά κάποιες τιμές της ενέργειας του θορύβου για ένα πεδίο τιμών της διακύμανσης και στην συνέχεια να ανιχνεύουμε τα στατιστικά του θορύβου και να αφαιρούμε την αντίστοιχη τιμή προτού προχωρήσουμε στην σύγκριση με το κατώφλι $\overline{T} = T + \sigma^2$. Ο όρος αυτός όμως καθιστά αυτήν την λύση επιρρεπή σε λάθη, αφού δεν μπορούμε να ελέγξουμε τις τιμές του.

Μια διαφορετική προσέγγιση θα ήταν να εκπαιδεύαμε το σύστημα προσθέτοντας θόρυβο στα video του training set. Η λύση αυτή θα δούλευε καλά με την προϋπόθεση πως θα είμαστε σε θέση να ανιχνεύουμε με ακρίβεια τα χαρακτηριστικά του θορύβου στο testing video.

3.4 Ανίχνευση με την μέθοδο του γραμμικού αθροίσματος

3.4.1 Παρουσίαση της μεθόδου

Μπορούμε αντί για την ενέργεια να χρησιμοποιήσουμε μια γραμμική συνάρτηση, όπως το απλό άθροισμα ή ο απλός μέσος όρος της αλλαγής φωτεινότητας (παραγώγου, οπτικής ροής). Τα frames με έντονη κίνηση πιθανότατα θα έχουν μεγαλύτερο μέσο όρο σε σχέση με τα frames χωρίς έντονη κίνηση. Θεωρητικά θα έπρεπε να έχουμε ανοχή σε προσθετικό θόρυβο μηδενικής μέσης τιμής, εφόσον είχαμε όμως και άπειρες τιμές των δεδομένων. Στην πράξη όμως δημιουργούνται αρκετά προβλήματα, που οφείλονται σε ακραίες τιμές *(outliers)*, οπότε και απαιτείται κάποια αντιμετώπιση του θορύβου. Θα δοκιμάσουμε λοιπόν να αντιμετωπίσουμε τον θόρυβο αναπροσαρμόζοντας το κατώφλι ως εξής:

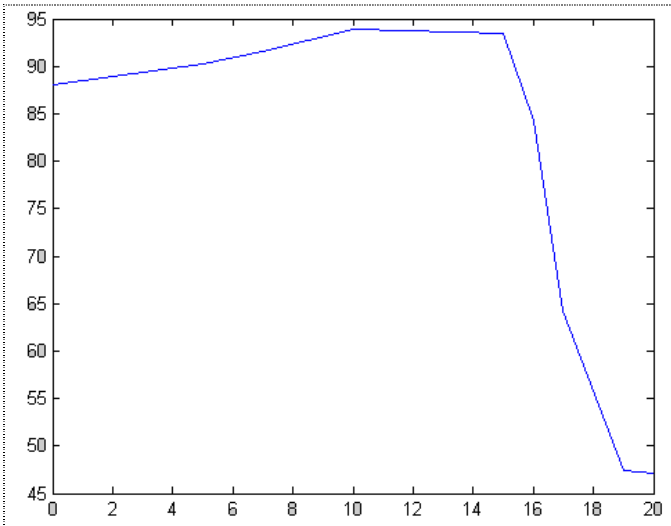
$$\overline{\mathrm{T}} = T + \sigma$$

Στην πράξη βέβαια, το απλό άθροισμα της μεταβολής φωτεινότητας των pixels δεν εκφράζει επαρκώς την ποσότητα της κίνησης, καθώς υπάρχει πάντα η πιθανότητα να έχουμε μικρό άθροισμα ακόμα και σε frames με έντονη κίνηση λόγω πολλών αντίθετων τιμών *(με διαφορετικό πρόσημο)*. Η ποσότητα της κίνησης εκφράζεται καλύτερα με το άθροισμα των απολύτων τιμών, που όμως δεν είναι γραμμική συνάρτηση και επομένως θα οδηγούμασταν σε παρόμοια προβλήματα με την ενέργεια. Για τον λόγο αυτό, στη συνέχεια εξετάζουμε την απόδοση της χρήσης του απλού αθροίσματος της αλλαγής φωτεινότητας.

3.4.2 Πειραματικά αποτελέσματα στο KTH dataset

Η μέθοδος αυτή αποδίδει καλύτερα αλλά μέχρι συγκεκριμένες τιμές θορύβου. Στο σχήμα 3.4 βλέπουμε την μέση απόδοση της μεθόδου σε σχέση με την αύξηση του θορύβου, όπως την αξιολογήσαμε στο KTH dataset. Παρατηρούμε πως η μέθοδος αποδίδει αρκετά καλά, αλλά δυστυχώς όταν ο θόρυβος γίνεται πολύ έντονος ($\sigma > 16$), η απόδοση πέφτει απότομα. Επομένως, η προσαρμογή του κατωφλίου $\overline{\mathrm{T}} = T + \sigma$ είναι στην πραγματικότητα μια προσέγγιση της βέλτιστης προσαρμογής.

Τα αναλυτικά αποτελέσματα αυτής της μεθόδου φαίνονται στον πίνακα 3.3.



Σχήμα 3.4.
Η μέση απόδοση της μεθόδου των μέσων όρων σε σχέση με την τυπική απόκλιση του θορύβου .

Πίνακας 3.3. Απόδοση της μεθόδου των μέσων όρων σε 4 διαφορετικά περιβάλλοντα, παρουσία προσθετικού θορύβου $N(0,\sigma^2)$

	d1	***d2***	***d3***	***d4***	***Μέση συμπεριφορά***
σ=0	87.11	89.97	83.06	91.94	***88.02***
σ=5	95.93	81.38	94.83	89.07	***90.30***
σ=10	97.05	90.69	96.47	91.35	***93.89***
σ=15	93.97	94.29	91.03	94.62	***93.47***
σ=17	59.03	71.40	53.14	73.47	***64.26***
σ=20	40.43	56.74	37.63	53.55	***47.09***

Βέβαια, η ανοχή στον θόρυβο μπορεί να θεωρηθεί αρκετά υψηλή, καθώς ο θόρυβος αλλοιώνει πολύ τα frames του video, όπως φαίνεται και στο σχήμα 3.5. Στην πράξη ίσως να συναντάμε σπάνια τόσο μεγάλο θόρυβο, γεγονός όμως που σε καμία περίπτωση δεν μπορεί να θεωρηθεί κανόνας και να επιτρέψει την ασφαλή χρήση της μεθόδου χωρίς κάποιον έλεγχο του θορύβου.

Παρόλο που η μέθοδος αυτή λειτουργεί καλύτερα και προσφέρει μια αρκετά υψηλή ανοχή στον θόρυβο (μέχρι *σ = 16*), πάλι θα πρέπει πάντα να αντιμετωπίζουμε επιτυχώς το πρόβλημα της ακριβούς γνώσης των χαρακτηριστικών του θορύβου.

Στην επόμενη ενότητα θα παρουσιάσουμε μια μέθοδο ανίχνευσης ενεργών frames η οποία είναι ανεξάρτητη από τα χαρακτηριστικά του θορύβου.

		
(α)	(β)	(γ)
		
(δ)	(ε)	(στ)

Σχήμα 3.5. Η επίδραση του προσθετικού Gaussian θορύβου στο ίδιο frame (α) ενός video. (β) σ = 5, (γ) σ = 10, (δ) σ = 15, (ε) σ = 17, (στ) σ = 20

3.5 Ανίχνευση με την μέθοδο της κύρτωσης

3.5.1 Παρουσίαση της μεθόδου

Η μέθοδος αυτή βασίζεται στην ιδέα των Activity Areas και στηρίζεται στην χρήση της κύρτωσης. Η κύρτωση *(kurtosis)* μιας τ.μ. Χ. ορίζεται ως εξής:

$$kurt(X) = E[X^4] - 3(E[X^2])^2$$

ή ισοδύναμα:

$$kurt(X) = \frac{E[X^4]}{(E[X^2])^2} - 3$$

Μπορεί να αποδειχθεί ότι κάτω από ορισμένες συνθήκες ανεξαρτησίας, η κύρτωση είναι γραμμική συνάρτηση, δηλαδή ότι για τις τ.μ. Χ, Υ ισχύει: $kurt(X+Y) = kurt(X) + kurt(Y)$

Η σημαντικότερη όμως ιδιότητα της κύρτωσης είναι ότι έχει μηδενική τιμή όταν η τ.μ. Χ ακολουθεί την Κανονική κατανομή, με οποιαδήποτε χαρακτηριστικά, δηλαδή: $kurt(X) = 0$ όταν $X \sim N(\mu, \sigma^2)$.

Αυτή η συμπεριφορά της κύρτωσης την καθιστά πραγματικά ανεξάρτητη από τον θόρυβο, αφού η κύρτωση του Gaussian θορύβου θα είναι πάντα 0. Έχει αποδειχθεί μάλιστα ότι ακόμα και για μη Gaussian θόρυβο, η κύρτωση πάλι παίρνει υψηλότερες

τιμές για μετρήσεις που προέρχονται από θόρυβο, και όχι από σήμα [7], καθώς είναι ευαίσθητη στην παρουσία ακραίων τιμών (outliers). Στην δική μας περίπτωση, όταν υπάρχει κίνηση και όχι μόνο θόρυβος μετρήσεων σε ένα pixel, το σήμα της αλλαγής φωτεινότητας διαφέρει αρκετά από αντίστοιχο σήμα που προέρχεται μόνο από θόρυβο, αποτελεί δηλαδή τιμή outlier.

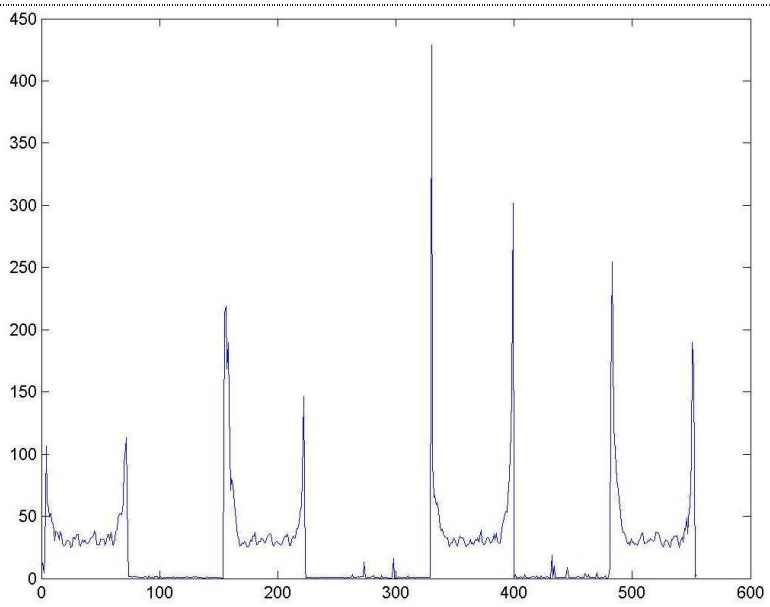
Η κύρτωση ενός θορυβώδους frame λοιπόν είναι:

$$kurt(L_{ti} + z_{ti}) = kurt(L_{ti}) + kurt(z_{ti}) = kurt(L_{ti}) + 0 = kurt(L_{ti})$$

Στην πράξη η κύρτωση παίρνει τιμές πολύ κοντά στο μηδέν, αλλά όχι πάντα ακριβώς μηδέν. Γι' αυτό ακριβώς και χρησιμοποιούμε ένα κατώφλι T κοντά στο 0, με βάση το οποίο αποφασίζουμε για το αν το frame είναι θορυβώδες ή όχι.

Τα στατικά frames έχουν γενικά πολύ μικρές τιμές κύρτωσης, που πλησιάζουν πολύ το μηδέν. Αντίθετα, τα ενεργά frames έχουν γενικά μεγάλες τιμές. Ο κανόνας αυτός ισχύει για ένα μεγάλο ποσοστό frames αλλά όχι παντού, γεγονός που μας οδηγεί κάποιες φορές σε σφάλματα. Ωστόσο, η μέθοδος της κύρτωσης δεν απαιτεί κάποια φάση εκπαίδευσης.

Στο σχήμα 3.6 βλέπουμε την κύρτωση των frames ενός video κίνησης walk. Παρατηρούμε ότι τα στατικά frames έχουν μικρές τιμές ενώ τα ενεργά υψηλές. Εντύπωση μας προκαλούν οι έντονες αλλαγές στα σημεία όπου εμφανίζεται και εξαφανίζεται από την σκηνή ο κινούμενος άνθρωπος. Η διερεύνηση αυτού του χαρακτηριστικού ίσως να οδηγήσει σε πολύ υψηλά ποσοστά επιτυχίας, ωστόσο αφήνεται για μελλοντική εργασία.



Σχήμα 3.6. Η κύρτωση των frames του video person01_walking_d1 του KTH dataset.

3.5.2 Πειραματικά αποτελέσματα στο KTH dataset

Ελέγξαμε την ανίχνευση active frames με την μέθοδο της κύρτωσης στο KTH dataset. Χρησιμοποιήσαμε video και από τους 4 τύπους περιβάλλοντος. Εισάγαμε ακόμη προσθετικό Gaussian θόρυβο $N(0,\sigma^2)$, για $\sigma = 0,\ 1,\ 2,\ ...,\ 20$.

Κάναμε πειράματα με διάφορες τιμές κατωφλίου μέσα στο διάστημα [0, 7]. Ελέγξαμε την απόδοση της μεθόδου τόσο συγκεκριμένα σε καθένα από τα 4 περιβάλλοντα όσο και συνολικά, ανεξαρτήτως περιβάλλοντος.

Αρχικά, επιβεβαιώσαμε την μη αποδοτικότητα του απόλυτου μηδέν σαν κατώφλι για την κύρτωση. Τα ποσοστά επιτυχίας φαίνονται στον πίνακα 3.4 και είναι μάλλον αναμενόμενα, καθώς στην πράξη δεν έχουμε ποτέ ούτε άπειρες τιμές μιας τυχαίας μεταβλητής ούτε τέλεια Gaussian θόρυβο.

Πίνακας 3.4. Απόδοση της μεθόδου της κύρτωσης σε διαφορετικά περιβάλλοντα χρησιμοποιώντας σαν κατώφλι Τ το απόλυτο μηδέν.

	Περιβάλλον d1	***Περιβάλλον d2***	***Περιβάλλον d3***	***Περιβάλλον d4***	***Μέση συμπεριφορά***
Ποσοστό (%)	46.77	59.65	48.81	65.96	55.30

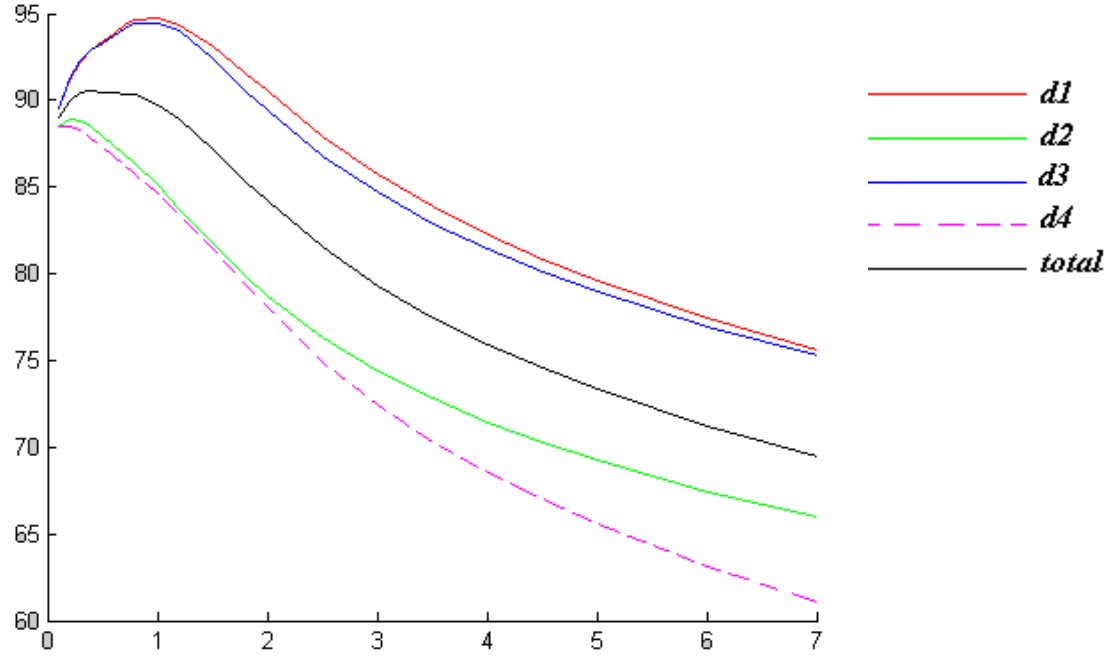
Παρατηρούμε πως τα ποσοστά είναι αρκετά χαμηλά ενώ το καλύτερο ποσοστό πετυχαίνεται στο περιβάλλον d4 (φωτισμός εσωτερικού δωματίου).

Αυξάνοντας ωστόσο την τιμή του κατωφλίου στο 0.1 είδαμε τα πρώτα θετικά αποτελέσματα. Η συμπεριφορά της μεθόδου τόσο σε καθένα από τα 4 περιβάλλοντα όσο και συνολικά, φαίνεται στο σχήμα 3.7.

Παρατηρούμε ότι σε κάθε περιβάλλον έχουμε διαφορετική συμπεριφορά, κάτι που ασφαλώς περιμέναμε. Τα περιβάλλοντα d1,d3 έχουν μεγάλες ομοιότητες και έτσι εξηγούμε την ομοιότητα των καμπυλών τους. Στα περιβάλλοντα d2,d4, ο θόρυβος που εισάγεται από το zoom και από τις φωτοσκιάσεις οδηγεί σε πολύ μικρότερα ποσοστά επιτυχούς ανίχνευσης. Η μέση συμπεριφορά κινείται μεταξύ των δύο αυτών ομάδων καμπυλών.

Παρατηρούμε ωστόσο ότι η βέλτιστη τιμή κατωφλίου είναι μικρότερη του 2, κάτι που μας δείχνει πως παρόλο που η κύρτωση δεν είναι ακριβώς μηδέν, έχει σίγουρα τιμές πολύ κοντά στο μηδέν. Ενδεικτικά για ένα video, στο σχήμα 3.6, βλέπουμε πως η κύρτωση των ενεργών frames παίρνει τιμές κοντά στο 50, επομένως η διαφορά της

κύρτωσης των ενεργών frames σε σχέση με την τιμή του κατωφλίου είναι ικανοποιητικά μεγάλη.



Σχήμα 3.7 Η συμπεριφορά της μεθόδου της κύρτωσης για τιμές κατωφλίου στο διάστημα [0.1, 7].

Το υψηλότερο μέσο ποσοστό επιτυχούς ανίχνευσης αντιστοιχεί στην τιμή *T=0.4* ήταν **90.50%**, το οποίο κρίνουμε αρκετά ενθαρρυντικό, δεδομένου της ποικιλίας των συνθηκών background που ελέγξαμε (εξωτερικό περιβάλλον, φωτισμός εσωτερικού χώρου, κάμερα με μεταβαλλόμενο zoom) καθώς και της ανοχής στον θόρυβο ($\sigma \in [0,20]$). Η συμπεριφορά της μεθόδου και στα άλλα περιβάλλοντα για *T=0.4* φαίνεται στον πίνακα 3.5 και όπως παρατηρούμε, τα ποσοστά είναι αρκετά υψηλά.

Πίνακας 3.5 Η απόδοση της μεθόδου της κύρτωσης σε 4 διαφορετικά περιβάλλοντα με χρήση του κατωφλίου T = 0.4.

	Περιβάλλον d1	***Περιβάλλον d2***	***Περιβάλλον d3***	***Περιβάλλον d4***	***Μέση συμπεριφορά***
Ποσοστό (%)	92.92	88.48	92.87	87.76	90.51

Τα καλύτερα ποσοστά σε κάθε περιβάλλον και οι αντίστοιχες τιμές κατωφλίου φαίνονται στον πίνακα 3.6. Τα ποσοστά είναι υψηλότερα από την περίπτωση T=0.4, αλλά η διαφορά δεν είναι πολύ μεγάλη (περίπου +2%). Παρατηρούμε ακόμη την ομαδοποίηση των περιβαλλόντων d1, d3 και d2, d4, η οποία ήταν εμφανής και από την απλή παρατήρηση των καμπυλών.

Πίνακας 3.6 Η καλύτερη απόδοση της μεθόδου της κύρτωσης σε 4 διαφορετικά περιβάλλοντα και οι αντίστοιχες τιμές του κατωφλίου T.

	Περιβάλλον d1	***Περιβάλλον d2***	***Περιβάλλον d3***	***Περιβάλλον d4***	***Μέση συμπεριφορά***
Ποσοστό (%)	94.7181	88.9095	94.4725	88.5288	90.5058
Κατώφλι T	1	0.2	1	0.2	0.4

3.6 Συμπεράσματα

Σε αυτό το κεφάλαιο ήρθαμε αντιμέτωποι με το πρόβλημα της ανίχνευσης ενεργών frames σε video. Η σημασία της ανίχνευσης κρίνεται ιδιαίτερα σημαντική, τόσο γιατί είναι το πρώτο βήμα για την ανάλυση και αναγνώριση κίνησης όσο και επειδή επηρεάζει άμεσα την απόδοση των επόμενων βημάτων.

Αρχικά παρουσιάσαμε μια μέθοδο βασισμένη στην ενέργεια, η οποία αποδίδει εξαιρετικά ακόμα και σε άγνωστα περιβάλλοντα, απαιτώντας πολύ μικρό σύνολο εκπαίδευσης *(training set)*. Ωστόσο, η μέθοδος αυτή δεν αποδίδει καλά σε περιβάλλοντα με θόρυβο, λόγω της μη γραμμικής της συμπεριφοράς.

Στην συνέχεια, παρουσιάσαμε την μέθοδο των μέσων όρων, η οποία παρέχει κάποια ανοχή στον θόρυβο, χωρίς ωστόσο να λύνει το πρόβλημα.

Τέλος, παρουσιάσαμε την μέθοδο της κύρτωσης, η οποία εκτός του ότι δεν απαιτεί φάση εκπαίδευσης, λειτούργησε αρκετά αποτελεσματικά (90.5%) σε πολλά διαφορετικά περιβάλλοντα, παρουσιάζοντας παράλληλα εκπληκτική ανοχή στον θόρυβο.

Και οι τρεις μέθοδοι είναι κατάλληλες για χρήση σε εφαρμογές πραγματικού χρόνου, καθώς απαιτούν μόνο την καθυστέρηση ενός frame, λόγω της χρήσης της παραγώγου.

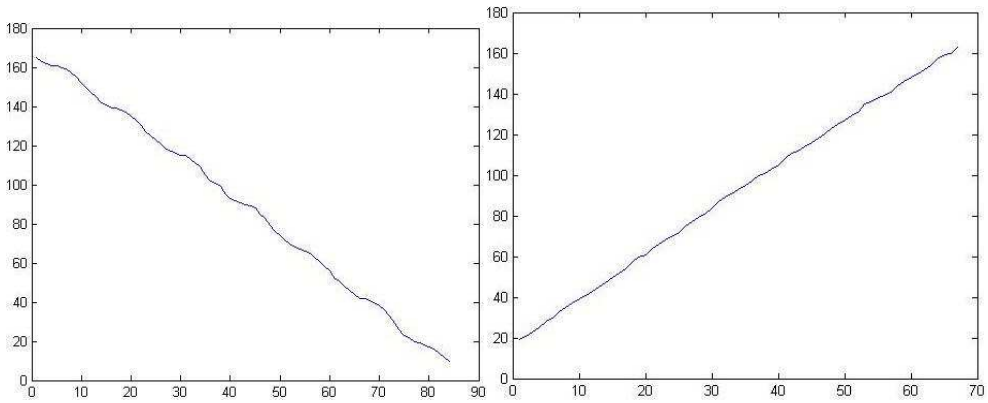
Στα σχέδια μελλοντικής εργασίας μας εντάσσεται τόσο η διερεύνηση της χρήσης των έντονων αλλαγών στα σημεία αλλαγής σκηνικού *(όπου εμφανίζεται και εξαφανίζεται ο κινούμενος άνθρωπος από την σκηνή)* που παρουσιάζει η μέθοδος της

κύρτωσης, όσο και η βελτίωση των αποτελεσμάτων της μεθόδου με χρήση διάφορων τεχνικών *(ομαδοποίηση γειτονικών frames, φιλτράρισμα)*. Τέλος, στα άμεσα σχέδιά μας εντάσσεται η διερεύνηση της χρήσης της κύρτωσης για την ανίχνευση *"ενεργών ομάδων pixels" (blocks)* σε ένα frame.

Σελίδα σκόπιμα κενή.

Κεφάλαιο 4

"Εξαγωγή ιδιοτήτων κίνησης"



Σελίδα σκόπιμα κενή.

4.1 Εισαγωγή

Σε αυτό το κεφάλαιο θα ασχοληθούμε με την εξαγωγή κάποιων σημαντικών ιδιοτήτων/χαρακτηριστικών της κίνησης που πραγματοποιείται σε κάποιο video. Οι ιδιότητες με τις οποίες θα ασχοληθούμε είναι ο χαρακτηρισμός της κίνησης ως Translational ή μη Translational και η κατεύθυνση του κινούμενου ανθρώπου (στην περίπτωση Translational κίνησης). Παρόλο που για την παρουσίαση και πειραματική αξιολόγηση των μεθόδων χρησιμοποιήσαμε video με ανθρώπους, πιστεύουμε πως οι μέθοδοι είναι πιο γενικές και θα μπορούσαν να εφαρμοστούν για την εξαγωγή ιδιοτήτων και άλλων κινούμενων υποκειμένων (π.χ. αυτοκίνητα), υπόθεση που θα ελέγξουμε σε μελλοντική μας εργασία.

4.1.1 Σημασία των ιδιοτήτων κίνησης

Η εξαγωγή ιδιοτήτων της κίνησης είναι ένα σημαντικό βήμα τόσο για τον πληρέστερο χαρακτηρισμό της κίνησης όσο και για την δημιουργία ομάδων κίνησης, γεγονός που είναι ιδιαίτερα χρήσιμο στα επόμενα στάδια του συστήματος ανάλυσης video.

Συγκεκριμένα, όσον αφορά την αναγνώριση κίνησης, είναι σημαντικό να διαβλέπει κανείς «ομάδες κίνησης» γιατί έτσι:

- Βελτιώνεται το ποσοστό επιτυχίας των μεθόδων αναγνώρισης, καθώς μειώνονται οι διαστάσεις και η εντροπία του αρχικού προβλήματος αναγνώρισης.
- Ακόμη και τα λάθη της αναγνώρισης έχουν πλέον περισσότερο νόημα, καθώς πρόκειται για λανθασμένη αναγνώρισης μιας παρόμοιας κίνησης.
- Η διαδικασία αναγνώρισης επιταχύνεται καθώς πλέον το πεδίο αναζήτησης είναι μικρότερο.

4.2 Διαχωρισμός σε Translational και non-Translational κινήσεις

4.2.1 Ορισμοί

Translational κινήσεις εννοούμε τις κινήσεις στις οποίες το υποκείμενο κινείται κατά μήκος του οριζόντιου άξονα *x*. Για παράδειγμα, οι κινήσεις walk, jog, run, jump, skip, side είναι Translational κινήσεις.

Non-Translational κινήσεις είναι οι κινήσεις που δεν είναι Translational. Για παράδειγμα, οι κινήσεις bend, wave1, wave2, pjump, box είναι non-Translational κινήσεις.

4.2.2 Διαχωρισμός με χρήση των Activity Areas

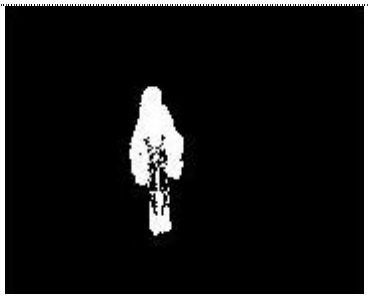
4.2.2.1 Περιγραφή της μεθόδου

Για να πετύχουμε τον διαχωρισμό σε Translational και non-Translational κινήσεις κάνουμε τα εξής βήματα [39]:

1. Βρίσκουμε το activity area του video χρησιμοποιώντας την απλή παράγωγο ως προς τον χρόνο t.
2. Αθροίζουμε την εικόνα του activity area κατά στήλες.
3. Μετράμε το πλήθος των μηδενικών τιμών Z.
4. Μετράμε το πλήθος των μη μηδενικών τιμών N.
5. Αν Z>N τότε η κίνηση είναι non-Translational, αλλιώς είναι Translational.

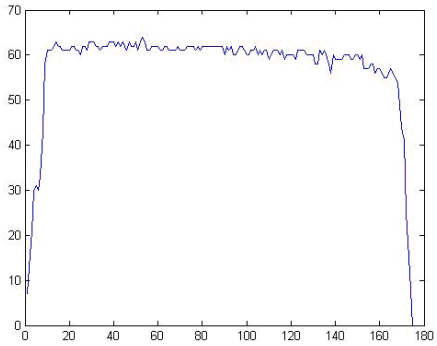
Παρακάτω βλέπουμε τα activity areas δύο διαφορετικών κινήσεων, του walk και του pjump, όταν αυτές οι κινήσεις εκτελούνται από τον ίδιο άνθρωπο. Παρατηρούμε πως στην translational κίνηση τα ενεργά pixel καταλαμβάνουν ένα μεγάλο μέρος της εικόνας, κατά μήκος του άξονα x. Αντίθετα, στην non-Translational κίνηση, τα ενεργά pixel καταλαμβάνουν πολύ μικρό ποσοστό του άξονα *x (σχήμα 4.1)*.

Σχήμα 4.1 (α) Το activity area του daria_walk από το Weizmann dataset

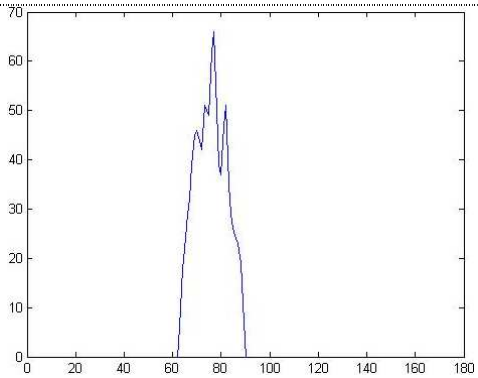


Σχήμα 4.1 (β) Το activity area του daria_pjump από το Weizmann dataset

Κοιτάζοντας το άθροισμα κατά στήλες των δύο αυτών κινήσεων έχουμε τα αποτελέσματα του σχήματος 4.2.



Σχήμα 4.2 (α) Το άθροισμα κατά στήλες του daria_walk από το Weizmann dataset



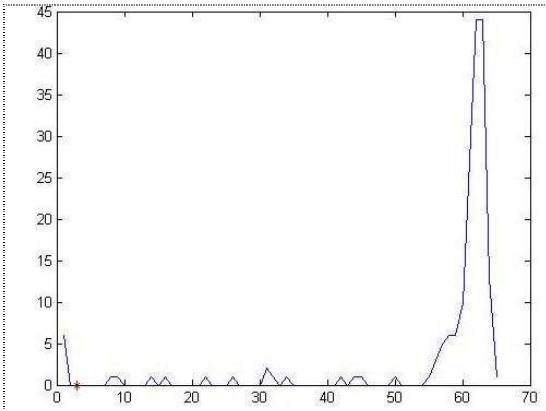
Σχήμα 4.2 (β) Το άθροισμα κατά στήλες του daria_pjump από το Weizmann dataset

4.2.2.2 Τιμές κοντά στο μηδέν

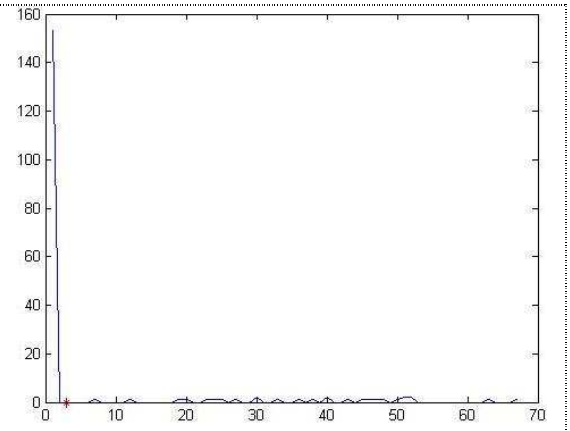
Ενδιαφέρον έχει ίσως η αντικατάσταση των μηδενικών τιμών με την έννοια των *«τιμών κοντά στο μηδέν»*. Λόγω θορύβου θα μπορούσαν να υπάρξουν περιπτώσεις όπου θα είχαμε πολλές τιμές κοντά στο μηδέν αλλά όχι απόλυτα μηδέν. Σε αυτήν την περίπτωση θα μπορούσαμε να χρησιμοποιήσουμε ένα κατώφλι T και να θεωρούμε ως *«τιμές κοντά στο μηδέν»* όσες είναι μικρότερες ή ίσες από αυτό το κατώφλι.

Η επιλογή του T είναι σημαντική και πρέπει να είναι γενικά κοντά στο μηδέν. Ωστόσο, αξίζει να σημειώσουμε πως η παρουσία του θορύβου είναι γενικά μικρή στα Activity Areas αφού εξ ορισμού το Activity Area χρησιμοποιεί την κύρτωση για να φιλτράρει αποτελεσματικά τον θόρυβο. Επομένως, το T είναι σχεδόν ανεξάρτητο από τον θόρυβο και θα μπορούσαμε να χρησιμοποιήσουμε μια πολύ μικρή τιμή (π.χ. 3) με αρκετά καλά αποτελέσματα.

Για τα πειράματά μας χρησιμοποιήσαμε την *μέση τιμή* του ιστογράμματος των τιμών αθροισμάτων κατά στήλες, η οποία έδωσε άριστα αποτελέσματα (100%).



Σχήμα 4.3 (α) Το ιστόγραμμα του daria_walk από το Weizmann dataset



Σχήμα 4.3 (β) Το ιστόγραμμα του daria_pjump από το Weizmann dataset

4.2.2.3 Πειραματικά αποτελέσματα

Ελέγξαμε πειραματικά την παραπάνω μέθοδο στο Weizmann Dataset. Χρησιμοποιώντας σαν κατώφλι T το 0 το ποσοστό επιτυχίας ήταν 96.67%, γεγονός που μαρτυρά μια μικρή παρουσία θορύβου. Χρησιμοποιώντας έπειτα σαν κατώφλι την μέση τιμή του ιστογράμματος το ποσοστό επιτυχίας αυξήθηκε στο 100%.

Παρατηρούμε λοιπόν πως πάντα υπάρχει κάποιος μικρός θόρυβος, τον οποίο μπορούμε ωστόσο εύκολα να αντιμετωπίσουμε με την χρήση του κατωφλίου T.

4.2.3 Διαχωρισμός με χρήση του κέντρου βάρους

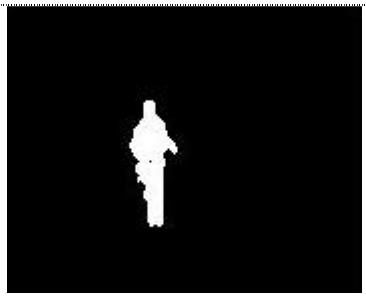
4.2.3.1 Περιγραφή της μεθόδου

Η μέθοδος που προτείνουμε ακολουθεί τα εξής βήματα:

1. Με *αφαίρεση background*, βρίσκουμε την σιλουέτα του κινούμενου ανθρώπου σε κάθε frame του video.
2. Βρίσκουμε το *κέντρο βάρους* κάθε σιλουέτας.
3. Δημιουργούμε το σήμα *sx(t)*, το οποίο δείχνει την συντεταγμένη *(ως προς τον άξονα x)* του κέντρου βάρους για την σιλουέτα του frame t.
4. Μετράμε το *εύρος* E του πεδίου τιμών του σήματος *sx(t)*.
5. Αν E>T, τότε η κίνηση είναι Translational, αλλιώς είναι non-Translational. T είναι ένα κατώφλι.

Η *αφαίρεση background* [43, 44, 45] οδηγεί στην εξαγωγή της σιλουέτας του κινούμενου ανθρώπου (σχήμα 4.4). Φυσικά, η επιτυχία της *αφαίρεσης background* είναι αρκετά σημαντική για την απόδοση της μεθόδου.

Σχήμα 4.4 (α) Σιλουέτα του 5^{ου} frame του daria_walk από το Weizmann dataset

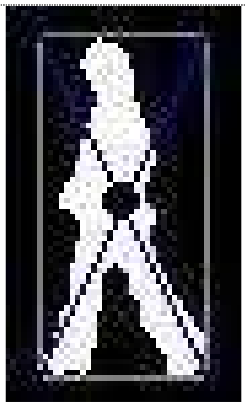


Σχήμα 4.4 (β) Σιλουέτα του 5^{ου} frame του daria_pjump από το Weizmann dataset

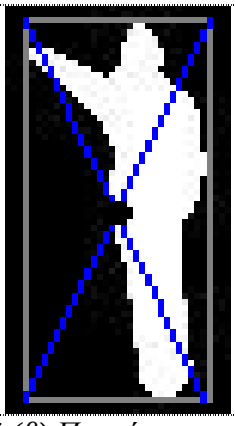
Η εύρεση του κέντρου βάρους απαιτεί γενικά μια σύνθετη διαδικασία. Προς το παρόν θα προσεγγίσουμε το κέντρο βάρος θεωρώντας πως συμπίπτει με το κέντρο βάρους του ορθογωνίου παραλληλογράμμου που περικλείει την σιλουέτα του ανθρώπου.

Η προσέγγιση αυτή προσεγγίζει με μεγάλη ακρίβεια το αληθινό κέντρο βάρους όταν η σιλουέτα μοιάζει με ορθογώνιο παραλληλόγραμμο. Διαφορετικά, το σφάλμα μπορεί να γίνει αρκετά μεγάλο και κάποιες φορές μπορεί να βρούμε κέντρο βάρους που ξεφεύγει από τα όρια της ίδιας της σιλουέτας.

Για παράδειγμα, στο σχήμα 4.5 (α) βλέπουμε πως το κέντρο βάρους δεν απέχει πολύ από το διαισθητικά σωστό. Αντίθετα, στο σχήμα 4.5 (β), παρατηρούμε πως απλωμένο χέρι εξωθεί το κέντρο βάρους σχεδόν στα όρια της σιλουέτας.



Σχήμα 4.5 (α) Προσέγγιση του κέντρου βάρους του daria_walk του Weizmann dataset



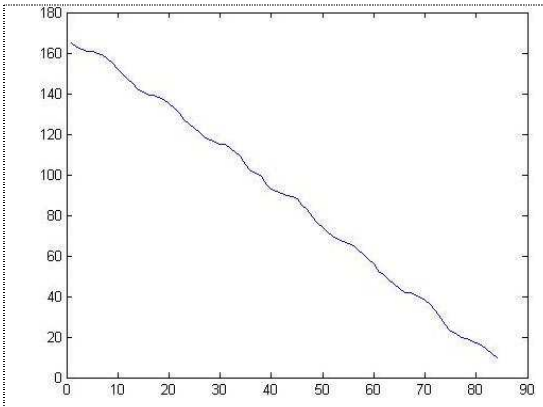
Σχήμα 4.5 (β) Προσέγγιση του κέντρου βάρους του daria_wave1 του Weizmann dataset

Ωστόσο, αυτή η προσέγγιση είναι αρκετή για την εξαγωγή των ιδιοτήτων που εξετάζουμε. Ο ακριβής υπολογισμός του κέντρου βάρους θα έδινε σίγουρα ακριβέστερα αποτελέσματα αυξάνοντας όμως πολύ το υπολογιστικό κόστος της μεθόδου. Μια απλή αντιπρόταση είναι να μην λαμβάνουμε υπόψη μας τα *θορυβώδη* μέρη του σώματος (χέρια, πόδια) και να εργαζόμαστε μόνο με τον κορμό του σώματος.

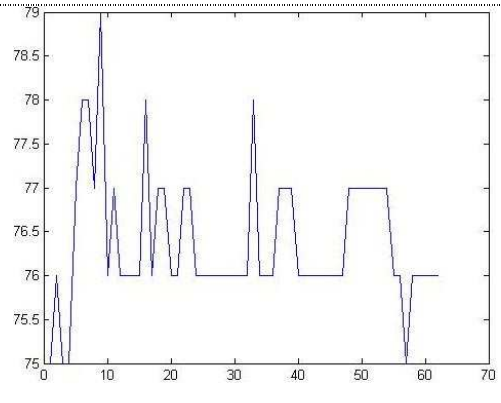
Το *εύρος* Ε του πεδίου τιμών του σήματος *sx(t)* είναι επίσης ένα σημαντικό θέμα. Μπορούμε να το μετρήσουμε ως την διαφορά ανάμεσα στην μέγιστη και ελάχιστη τιμή. Αυτή η λύση αγνοεί την κατανομή των τιμών, κάτι που την κάνει ευαίσθητη σε κάποιες ακραίες θορυβώδεις τιμές *(outliers)*. Προτιμούμε σαν καλύτερη λύση την τυπική απόκλιση *σ* των τιμών.

Τέλος, η επιλογή του κατωφλίου Τ πρέπει να γίνεται με βάση τον στόχο της κάθε εφαρμογής και τα χαρακτηριστικά του κινούμενου υποκειμένου (άνθρωπος, αυτοκίνητο, κτλ). Ωστόσο, για ανθρώπινες κινήσεις ένα κατώφλι μεταξύ του 5 και του 10 θεωρούμε πως είναι μάλλον αρκετό.

Στο σχήμα 4.6 βλέπουμε το σήμα *sx(t)* δύο διαφορετικών κινήσεων. Παρατηρούμε πως στην Translational κίνηση *(walk)* οι τιμές του σήματος εκτείνονται από το 0 μέχρι το 160. Η τυπική τους απόκλιση είναι περίπου 47. Αντίστοιχα, στην πο Translational κίνηση *(pjump)* οι τιμές του σήματος εκτείνονται από το 75 μέχρι το 79. Η τυπική τους απόκλιση είναι περίπου 0.77.



Σχήμα 4.6 (α) Το σήμα sx(t) του daria_walk από το Weizmann dataset



Σχήμα 4.6 (β) Το σήμα sx(t) του daria_wave1 από το Weizmann dataset

4.2.3.2 Πειραματικά αποτελέσματα

Παρόλο που χρησιμοποιήσαμε την απλή προσέγγιση του παραλληλογράμμου για την εύρεση του κέντρου βάρους, με την χρήση του κατωφλίου $T = 5$, καταφέραμε να πετύχουμε 100% διαχωρισμό σε Translational και non-Translational κινήσεις στο Weizmann Dataset.

4.2.4 Διαχωρισμός με επεξεργασία τμήματος της ακολουθίας

Παρόλο που και οι δύο μέθοδοι επέδειξαν άριστη συμπεριφορά, έχει ενδιαφέρον να εξετάσουμε το θέμα του ελάχιστου πλήθος frames που απαιτούνται για την αποτελεσματική εφαρμογή κάθε μεθόδου.

Επαναλάβαμε λοιπόν τα πειράματά μας, χρησιμοποιώντας από 5 έως και 40 frames (όπου αυτό ήταν δυνατό) για τον διαχωρισμό των κινήσεων. Τα αποτελέσματα φαίνονται στους πίνακες 4.1 και 4.2.

Πίνακας 4.1 Διαχωρισμός σε Translational και non-Translational κινήσεις στο Weizmann Dataset, με την μέθοδο των Activity Areas, χρησιμοποιώντας μικρό πλήθος frames.

	Κίνηση									
Πλήθος frames	***Walk***	***Jump***	***Run***	***Side***	***Skip***	***Pjump***	***Bend***	***Jack***	***Wave1***	***Wave 2***
5	*0*	*0*	*0*	*0*	*0*	*100*	*28.57*	*100*	*71.43*	*100*
10	*0*	*0*	*11.11*	*0*	*0*	*100*	*66.67*	*100*	*100*	*100*
15	*0*	*0*	*44.44*	*0*	*11.11*	*100*	*77.78*	*100*	*100*	*100*
20	*0*	*22.22*	*88.89*	*11.11*	*44.44*	*100*	*100*	*100*	*100*	*100*
25	*11.11*	*33.33*	*100*	*77.78*	*100*	*100*	*100*	*100*	*100*	*100*
30	*66.67*	*66.67*	*100*	*88.89*	*100*	*100*	*100*	*100*	*100*	*88.89*
35	*77.78*	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*88.89*
40	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*77.78*
full sequence	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*

Πίνακας 4.2 Διαχωρισμός σε Translational και non-Translational κινήσεις στο Weizmann Dataset, με την μέθοδο του κέντρου βάρους, χρησιμοποιώντας μικρό πλήθος frames.

	Κίνηση									
Πλήθος frames	***Walk***	***Jump***	***Run***	***Side***	***Skip***	***Pjump***	***Bend***	***Jack***	***Wave1***	***Wave 2***
5	*0*	*0*	*66.67*	*22.22*	*11.11*	*100*	*100*	*100*	*100*	*100*
10	*77.78*	*100*	*100*	*88.89*	*100*	*100*	*100*	*100*	*100*	*100*
15 -40	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*
full sequence	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*

Η μέθοδος των Activity Areas (όπως την περιγράψαμε στην ενότητα 4.2.2.1) είναι καθολική, εφαρμόζεται δηλαδή σε όλα τα pixels του Activity Area. Άλλωστε και το Activity Area είναι ουσιαστικά μια καθολική αναπαράσταση, καθώς συμπυκνώνει την κίνηση πολλών frames σε ένα μόνο frame. Αυτό σημαίνει πως είναι μάλλον άστοχο να προσπαθούμε να χρησιμοποιήσουμε λίγα frames με αυτήν την μέθοδο, γεγονός που αποτυπώνεται και στα αποτελέσματα του πίνακα 4.1.

Ίσως μελλοντικά να έχουμε καλύτερα αποτελέσματα αν προσπαθήσουμε να μετατρέψουμε την μέθοδο σε τοπική, αλλάζοντας δηλαδή το βήμα 5: “Αν Z>N τότε η κίνηση είναι non-Translational, αλλιώς είναι Translational.” ώστε να χρησιμοποιεί μόνο τα μη μηδενικά στοιχεία N που είναι άλλωστε και τα στοιχεία της κίνησης.

Αντίθετα, η μέθοδος του κέντρου βάρους είναι τοπική και επομένως μπορεί να έχει πολύ καλή συμπεριφορά χρησιμοποιώντας ακόμα και 15 frames, όπως φαίνεται στον πίνακα 4.2. Το γεγονός αυτό μας επιτρέπει να βγάζουμε συμπεράσματα ακόμα και από λίγα διαθέσιμα frames, δηλαδή ακόμα και σε πραγματικό χρόνο.

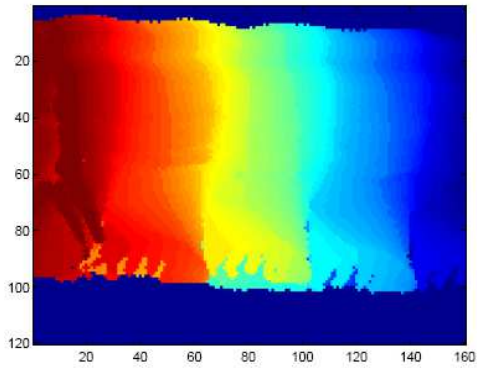
4.3 Εύρεση κατεύθυνσης κινούμενου ανθρώπου για Translational κινήσεις

4.3.1 Εύρεση κατεύθυνσης με χρήση Activity History Area

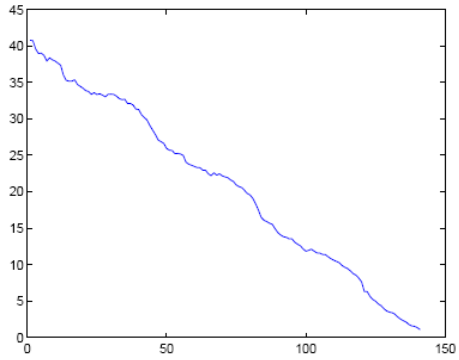
4.3.1.1 Περιγραφή της μεθόδου

Μπορούμε να βρούμε την κατεύθυνση του κινούμενου ανθρώπου χρησιμοποιώντας το *Activity History Area* ως εξής:

1. Βρίσκουμε το *Activity History Area* του video.
2. Βρίσκουμε τον μέσο όρο κάθε στήλης του *Activity History Area*
3. Σχηματίζεται ένα σήμα, από το οποίο υπολογίζουμε την κλίση του λ
4. Αν $\lambda > 0$ τότε η κατεύθυνση είναι προς τα δεξιά, αλλιώς είναι προς τα αριστερά.



Σχήμα 4.7 (α) Το Activity History Area για την κίνηση walk_left



Σχήμα 4.7 (β) Το σήμα των μέσων όρων των στηλών για το ίδιο video. Παρατηρούμε ότι η κλίση είναι αρνητική.

4.3.1.2 Πειραματικά αποτελέσματα

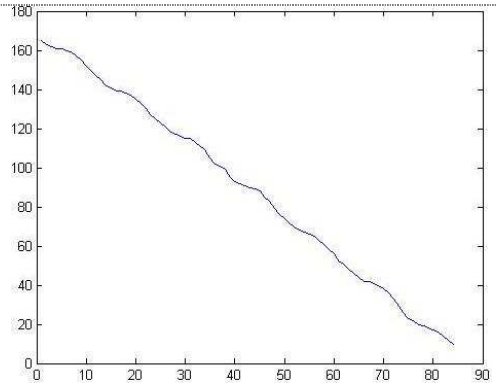
Αξιολογώντας πειραματικά την μέθοδο αυτή, καταφέραμε να βρίσκουμε σωστά στην κατεύθυνση κίνησης στο 100% του Weizmann Dataset.

4.3.2 Εύρεση κατεύθυνσης με χρήση του κέντρου βάρους

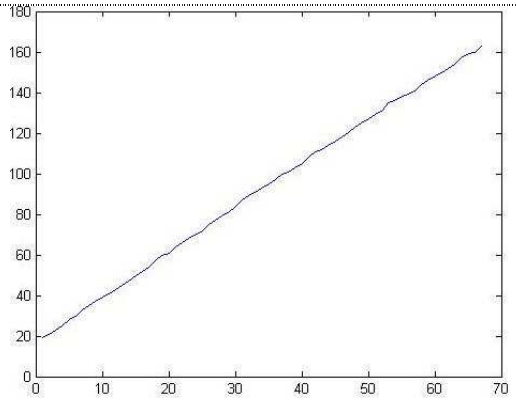
4.3.2.1 Περιγραφή της μεθόδου

Μπορούμε να βρούμε την κατεύθυνση του κινούμενου ανθρώπου χρησιμοποιώντας το *κέντρο βάρους* του κινούμενου ανθρώπου ως εξής:

1. Με *αφαίρεση background*, βρίσκουμε την σιλουέτα του κινούμενου ανθρώπου σε κάθε frame του video.
2. Βρίσκουμε το *κέντρο βάρους* κάθε σιλουέτας.
3. Δημιουργούμε το σήμα *sx(t)*, το οποίο δείχνει την συντεταγμένη ως προς τον άξονα x του κέντρου βάρους για την σιλουέτα του frame *t*.
4. Υπολογίζουμε την κλίση λ του σήματος *sx(t)*
5. Αν $\lambda > 0$ τότε η κατεύθυνση είναι προς τα δεξιά, αλλιώς είναι προς τα αριστερά.



Σχήμα 4.8 (α) Το σήμα sx(t) για την κίνηση walk_left



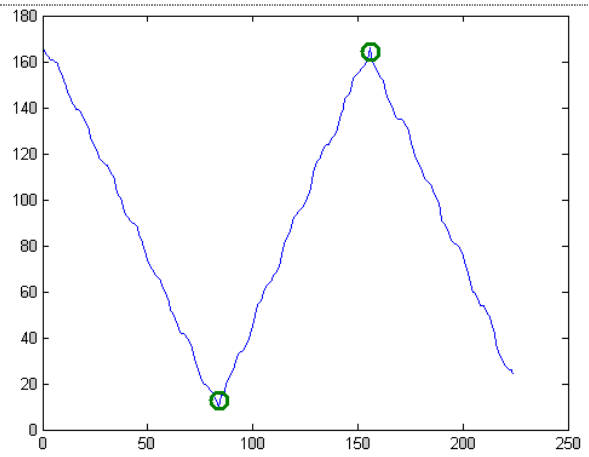
Σχήμα 4.8 (β) Το σήμα sx(t) για την κίνηση jump_right

4.3.2.2 Πειραματικά αποτελέσματα

Παρόλο που χρησιμοποιήσαμε την απλή προσέγγιση του παραλληλογράμμου για την εύρεση του κέντρου βάρους, καταφέραμε να βρίσκουμε σωστά στην κατεύθυνση κίνησης στο 100% του Weizmann Dataset.

4.3.3 Εύρεση αλλαγής κατεύθυνσης της κίνησης

Ενδιαφέρον παρουσιάζει η ανίχνευση των αλλαγών στην κατεύθυνση της κίνησης. Οι επικαλυπτόμενες κινήσεις *γράφουν* το ένα Activity History Area πάνω στο άλλο, διαγράφοντας έτσι την *μνήμη* του. Στο σχήμα 4.9 βλέπουμε το σήμα *sx(t) (κέντρου βάρους)* για ένα video με 3 διαφορετικές κατευθύνσεις. Το σήμα αυτό είναι θορυβώδες αλλά θα μπορούσε να προσεγγιστεί με 3 ευθύγραμμα τμήματα με την μέθοδο προσέγγισης σήματος που θα παρουσιάσουμε αναλυτικά στο κεφάλαιο 5.



Σχήμα 4.9 Το σήμα sx(t) για ένα video με την αλλαγή κατευθύνσεων left→ right→ left. Τα σημεία αλλαγής σημειώνονται με ***O****.*

4.4 Συμπεράσματα

Σε αυτό το κεφάλαιο είδαμε την σημασία εύρεσης δύο βασικών ιδιοτήτων κίνησης, αυτές του τύπου της κίνησης *(Translational-non-Translational)* και της κατεύθυνσης.

Για κάθε ιδιότητα χρησιμοποιήσαμε και αξιολογήσαμε πειραματικά τόσο μια ήδη υπάρχουσα μέθοδο *(Activity Area - Activity History Area)* όσο και μια νέα προτεινόμενη μέθοδο που βασίζεται στην χρήση του κέντρου βάρους του κινούμενου ανθρώπου και απαιτεί την εξαγωγή της σιλουέτας του κινούμενου ανθρώπου.

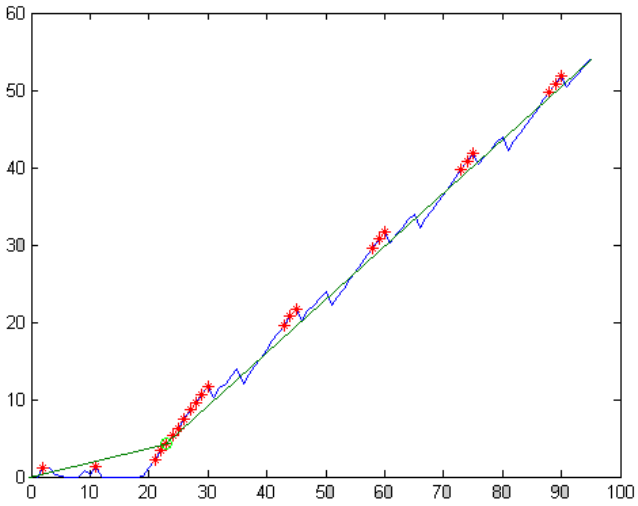
Και οι δύο μέθοδοι έχουν εξαιρετική απόδοση (100%), γεγονός που επιτρέπει την *ασφαλή* ενσωμάτωσή τους ως ενδιάμεσα στάδια ενός γενικότερου συστήματος κατανόησης κίνησης. Είδαμε ακόμη πως η μέθοδος του κέντρου βάρους είναι καταλληλότερη για επεξεργασία ακόμα και όταν έχουμε λίγα διαθέσιμα frames της κίνησης. Τέλος, προτείναμε έναν τρόπο ανίχνευσης αλλαγών στην κατεύθυνση της κίνησης με βάση το σήμα *sx(t)* του κέντρου βάρους.

Οι στόχοι μελλοντικής μας εργασίας περιλαμβάνουν την πειραματική αξιολόγηση των μεθόδων σε πιο απαιτητικά περιβάλλοντα *(όπως στο KTH Dataset)*, την βελτίωση της μεθόδου των Activity Areas ώστε να μπορεί να χρησιμοποιείται και σε λίγα διαθέσιμα frames, και την πειραματική αξιολόγηση της εύρεσης αλλαγών στην κατεύθυνση με σκοπό την παραγωγή πιο ποιοτικών Activity History Areas.

Σελίδα σκόπιμα κενή.

Κεφάλαιο 5

“Sequential Change Detection”



Σελίδα σκόπιμα κενή.

5.1 Εισαγωγικά

Η ανίχνευση αλλαγών προσπαθεί να εντοπίσει τα χρονικά σημεία στα οποία μεταβαίνουμε από μία κίνηση σε μία διαφορετική. Κατ' αυτόν τον τρόπο, το αρχικό video χωρίζεται σε μικρότερα video, στα οποία πραγματοποιείται μόνο μία κατηγορία κίνησης.

Στην συνέχεια, η περαιτέρω ανάλυση της κίνησης (π.χ. αναγνώριση, εύρεση περιόδου/άλλων χαρακτηριστικών, κτλ) πραγματοποιείται σε αυτά τα μικρότερα video. Η απόδοση των διάφορων μεθόδων αναγνώρισης και ανάλυσης θα είναι σίγουρα καλύτερη απ' όσο αν εργαζόμασταν σε frames με δύο ή περισσότερες κινήσεις, καθώς τότε θα υπήρχε σύγχυση ως προς το ποια από τις δύο κινήσεις πραγματοποιείται τελικά.

Παρόλο που έχει γίνει αρκετή έρευνα σχετικά με την ανίχνευση αλλαγών σκηνής σε βίντεο (*shot change detection*) [33, 34, 35, 36], δεν έχει δοθεί αρκετή έμφαση στην ανίχνευση αλλαγών κίνησης [37, 38]. Οι μέθοδοι εύρεσης αλλαγών σκηνής βασίζονται σε χαρακτηριστικά της εμφάνισης του βίντεο, και κατά συνέπεια δεν μπορούν να χωρίσουν το βίντεο σε τμήματα που αποτελούνται από διαφορετικές κινήσεις και, πιθανώς, διαφορετικά γεγονότα *(events)*. Για παράδειγμα, σε ένα βίντεο παρακολούθησης, μπορεί να μπει κάποιος στο χώρο χωρίς να αλλάξει σημαντικά η εμφάνιση της σκηνής. Η προτεινόμενη μέθοδος αποσκοπεί στην ανίχνευση τέτοιων αλλαγών, αλλά οι υπάρχουσες μέθοδοι για shot change detection δεν μπορούν να τις βρουν.

Σε αυτό το κεφάλαιο θα παρουσιάσουμε μια νέα προσέγγιση, που επιτρέπει την ανίχνευση αλλαγών κίνησης με βάση τα στατιστικά τους χαρακτηριστικά, και τον διαχωρισμό του βίντεο σε υπο-ακολουθίες που περιέχουν διαφορετικά γεγονότα.

5.2 Sequential Change Detection - CUSUM

Η σειριακή ανίχνευση αλλαγών *(sequential change detection)* είναι γνωστή και ως CUSUM, καθώς βασίζεται στο συσσωρευτικό άθροισμα του log-likelihood ratio *(cumulative sum of the log-likelihood ratio)* [8]. Θεωρούμε ότι τα δεδομένα στα οποία θέλουμε να βρούμε μια αλλαγή είναι μια ακολουθία δεδομένων κίνησης $\overline{\upsilon}_\kappa$ για το frame k, όπου $1 \le k \le N$, που αποτελείται από τα δεδομένα κίνησης μέχρι και αυτό το frame:

$$\overline{\upsilon}_k = [\upsilon_1, \upsilon_2, ..., \upsilon_\kappa].$$

Παρατηρούμε ότι χρησιμοποιούνται μόνο τα δεδομένα που είναι διαθέσιμα μέχρι την παρούσα χρονική στιγμή, για να λειτουργεί το προτεινόμενο σύστημα σε πραγματικό χρόνο. Αυτά τα δεδομένα ακολουθούν μια κατανομή $f_0(\overline{\upsilon}_\kappa)$ πριν την αλλαγή και $f_1(\overline{\upsilon}_\kappa)$ μετά την αλλαγή. Η στιγμή της αλλαγής δεν είναι γνωστή, αλλά μπορεί να υπολογιστεί με την μέθοδο CUSUM με απλή υπολογιστική πολυπλοκότητα, και μάλιστα σε πραγματικό χρόνο. Αυτό γιατί η προτεινόμενη μέθοδος λαμβάνει απόφαση για το αν έχει γίνει αλλαγή σε κάθε χρονική στιγμή χρησιμοποιώντας μόνο τα δεδομένα που είναι διαθέσιμα μέχρι εκείνη την χρονική στιγμή. Στη συνέχεια παραθέτουμε τα βασικά βήματα του σειριακού ελέγχου CUSUM.

Για να υλοποιηθεί η σειριακή ανίχνευση αλλαγής, χρησιμοποιείται το log-likelihood ratio, που ορίζεται ως:

$$T_k = \ln \frac{f_1(\overline{\upsilon}_k)}{f_0(\overline{\upsilon}_k)} = \ln \prod_{i=1}^{k} \frac{f_1(\upsilon_i)}{f_0(\upsilon_i)} = \sum_{i=1}^{k} \ln \frac{f_1(\upsilon_i)}{f_0(\upsilon_i)}$$

όπου έχει γίνει η παραδοχή ότι τα δεδομένα κίνησης μέχρι την χρονική στιγμή k είναι iid (independent identically distributed). Στην πράξη αυτό μπορεί να μην ισχύει πάντα με ακρίβεια, αλλά είναι μια απαραίτητη υπόθεση, καθώς η εξαγωγή της από κοινού συνάρτησης πυκνότητας πιθανότητας (σ.π.π.) έχει μεγάλο υπολογιστικό κόστος, και δεν είναι πάντα εφικτή. Επιπλέον, έχει αποδειχθεί ότι ακόμα και όταν δεν ισχύει πλήρως η υπόθεση για i.i.d. δεδομένα, ο CUSUM παραμένει ασυμπτωτικά βέλτιστος, δίνοντας μικρό σφάλμα στην ανίχνευση αλλαγών [9]. Αυτό επιβεβαιώνεται και από τα πειράματά μας, όπου οι αλλαγές σε μια ποικιλία από βίντεο ανιχνεύονται με καλή ακρίβεια.

Οι κατανομές των δεδομένων πριν και μετά την αλλαγή δεν είναι γνωστές εξ' αρχής και πρέπει να προσεγγιστούν από τα διαθέσιμα δεδομένα. Η αρχική κατανομή f_0 μπορεί να προσεγγιστεί με ικανοποιητική ακρίβεια από τα αρχικά δείγματα [59], κάνοντας την υπόθεση πως δεν συμβαίνει καμία αλλαγή στα πρώτα w_0 frames της ακολουθίας. Εφόσον η στιγμή της αλλαγής είναι άγνωστη, η f_1 προσεγγίζεται χρησιμοποιώντας τα w_1 πιο πρόσφατα δείγματα. Παρόλο που δεν υπάρχει κάποιος θεωρητικός τρόπος για να υπολογίσουμε τα βέλτιστα w_0 και w_1, τιμές μέσα στο διάστημα [5,15] δίνουν αρκετά καλά αποτελέσματα. Στην πράξη χρησιμοποιήσαμε παράθυρα μήκους $w_0 = 15$ και $w_1 = 10$ στα βίντεο που παρουσιάζονται σε αυτή την εργασία.

Καθώς το σύστημα στοχεύει να λειτουργεί και σε πραγματικό χρόνο, προτείνεται η χρήση μόνο των δεδομένων κίνησης που βρίσκονται μέσα στις περιοχές κίνησης *(Activity Areas)*, ώστε να μειωθεί το υπολογιστικό του κόστος. Άλλωστε, δεν έχει νόημα η εξέταση των ακίνητων pixels. Επίσης, μειώνεται και η πιθανότητα να υπάρξουν ψευδείς ανιχνεύσεις αλλαγών *(false alarms)*, καθώς δεν χρησιμοποιούνται ακίνητα pixels στα οποία θα μπορούσε να ανιχνευτεί εσφαλμένα αλλαγή.

Εξετάσαμε δύο περιπτώσεις για τα δεδομένα κίνησης που χρησιμοποιούνται στο τεστ CUSUM, και συγκεκριμένα την οπτική ροή και την ενέργεια της οπτικής ροής. Για το συγκεκριμένο πρόβλημα, διαπιστώσαμε ότι η ενέργεια της οπτικής ροής *(μια μορφή του σήματος ELt)* οδηγεί σε καλύτερα αποτελέσματα ανίχνευσης αλλαγών κίνησης απ'ό,τι η ίδια η οπτική ροή. Λεπτομερής ανάλυση αυτής της σύγκρισης βρίσκεται στην ενότητα 5.4.2. Σημειώνουμε ότι σύμφωνα με τα παραπάνω, η ενέργεια της οπτικής ροής υπολογίζεται στα pixels που κινούνται, δηλαδή στα pixels μέσα στο Activity Area. Στην περίπτωση που το τεστ υλοποιήθηκε με την οπτική ροή, πάλι χρησιμοποιήθηκαν μόνο οι τιμές της μέσα στα ενεργά pixels.

Η υλοποίηση του CUSUM απλοποιείται και επιταχύνεται σημαντικά με την παρακάτω αναδρομική σχέση, που πρότεινε ο Page [32], για i.i.d. δεδομένα:

$$T_k = \max\left(0, T_{k-1} + \ln\frac{f_1(\overline{\upsilon}_k)}{f_0(\overline{\upsilon}_k)}\right)$$

που στην περίπτωσή μας γίνεται:

$$T_k = \max\left(0, T_{k-1} + \sum_{i=1}^{k} \ln\frac{f_1(\upsilon_i)}{f_0(\upsilon_i)}\right)$$

Η ανίχνευση αλλαγής προκύπτει όταν η τιμή του Tk γίνει μεγαλύτερη από ένα προκαθορισμένο κατώφλι. Η βέλτιστη τιμή του κατωφλίου βρίσκεται πειραματικά εφαρμόζοντας τον CUSUM σε training video, δηλαδή video παρόμοια με αυτά που εξετάζουμε και επιλέγοντας κάθε φορά το κατώφλι που ανιχνεύει τα λιγότερα *false alarms* [10]. Βρήκαμε ότι η τιμή του κατωφλίου που οδηγεί σε καλά αποτελέσματα για μεγάλο εύρος βίντεο ανθρώπινης κίνησης δίνεται από την σχέση:

$$\eta = \mu_T + c \cdot \sigma_T\,,\ c \in [2,\ 3]\,.$$

Η τιμή του c καθορίστηκε μετά από πειράματα με δεδομένα εκπαίδευσης (training data), και βρέθηκε ότι τιμές από 2 έως 3 οδηγούν σε καλά αποτελέσματα.

5.3 Statistical Modeling

5.3.1 Εισαγωγικά

Η μέθοδος CUSUM βασίζεται σε log likelihood ratio που χρησιμοποιεί την κατανομή των δεδομένων. Χρειάζεται επομένως να ξέρουμε ποια κατανομή προσεγγίζει καλύτερα τα δεδομένα μας. Στην περίπτωση που εξετάζουμε, τα δεδομένα κίνησης εμφανίζουν απότομες αλλαγές, δηλαδή υπάρχουν τιμές που είναι outliers. Αυτό γιατί σε κάθε pixel στη διάρκεια του βίντεο, είναι πιθανό το δεδομένο κίνησης να μεταβεί απότομα από μια χαμηλή τιμή (αν αρχικά ήταν ακίνητο το αντικείμενο) σε μια υψηλή, ή το ανάποδο.

Για το λόγο αυτό συγκρίνουμε την μοντελοποίηση που πετυχαίνει η Gaussian κατανομή, που χρησιμοποιείται συχνά στην πράξη, με αυτήν της Laplace και της Exponential (εκθετικής) κατανομής, που είναι πιο κατάλληλες για δεδομένα που παρουσιάζουν μεγάλες διακυμάνσεις τιμών.

5.3.2 Οπτική ροή: Laplace fitting

Η οπτική ροή μοντελοποιείται συχνά με την κατανομή Laplace, η οποία έχει δειχθεί πως προσεγγίζει καλύτερα τα δεδομένα [38].

Η σ.π.π. της κατανομής Laplace δίνεται από:

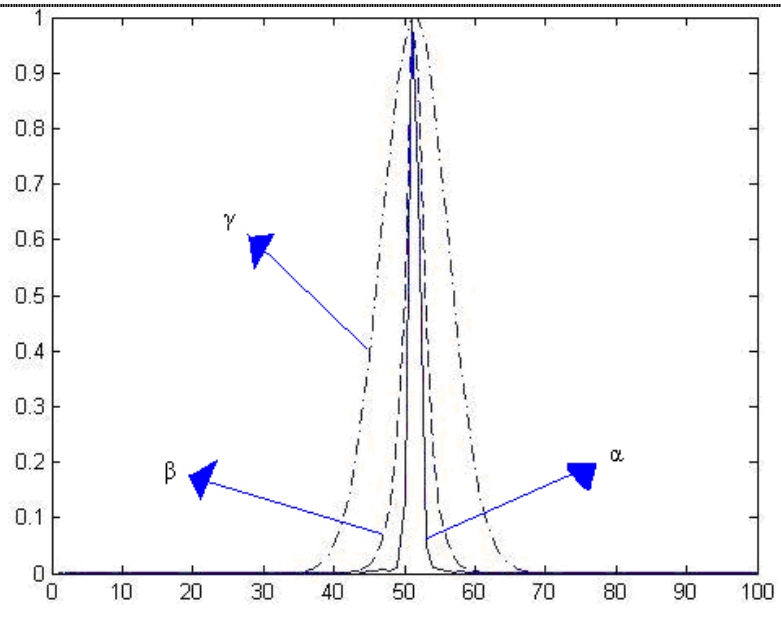
$$f(x) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right),$$

όπου μ είναι η μέση τιμή των δεδομένων και $b = \frac{\sigma}{\sqrt{2}}$, για τυπική απόκλιση σ.

Για την μοντελοποίηση εξετάζουμε βίντεο ανθρώπινης κίνησης που συμβαίνει κατά τον οριζόντιο άξονα *(Translational)*. Τότε, χρησιμοποιώντας σαν δεδομένα περιγραφής ενός frame την *x* συνιστώσα της οπτικής ροής, παίρνουμε τα αποτελέσματα του σχήματος 5.1. Παρατηρούμε πως η κατανομή Laplace προσεγγίζει καλύτερα τα δεδομένα μας, επομένως στην CUSUM θα χρησιμοποιούμε αυτήν την κατανομή. Πειράματα με περισσότερα βίντεο επιβεβαίωσαν ότι η Laplace κατανομή περιγράφει καλύτερα τα δεδομένα μας, και επομένως πρέπει να χρησιμοποιηθεί στο CUSUM. Επομένως, το στατιστικό τεστ παίρνει την μορφή:

$$T_k = \max\left(0, T_{k-1} + k\ln\frac{b_0}{b_1} + \sum_{i=1}^{k}\exp\left(-\frac{|\upsilon_i-\mu_0|}{b_0} + \frac{|\upsilon_i-\mu_1|}{b_1}\right)\right),$$

όπου τα μ_0, μ_1 είναι η μέση τιμή των δεδομένων πριν και μετά την αλλαγή και τα b_0, b_1 υπολογίζονται από τις αντίστοιχες τυπικές αποκλίσεις.



Σχήμα 5.1
Μοντελοποίηση της x συνιστώσας της οπτικής ροής ενός video.
(α) πραγματική κατανομή,
(β) Laplace,
(γ) Gaussian

5.3.3 Exponential fitting

Καθώς θέλουμε να ελέγξουμε την χρήση της ενέργειας της οπτικής ροής αντί για την ίδια την οπτική ροή, πρέπει να γνωρίζουμε την κατανομή που την προσεγγίζει καλύτερα. Επειδή η ενέργεια παίρνει μόνο θετικές τιμές, ελέγχουμε την προσέγγιση της Εκθετικής κατανομής *(Exponential)*.

Η σ.π.π. της Εκθετικής κατανομής δίνεται από:

$$f(x) = \begin{cases} \lambda \cdot \exp(-\lambda x), & x \geq 0 \\ 0, & x < 0 \end{cases}$$

όπου $\lambda = \frac{1}{\mu}$ και μ είναι η μέση τιμή των δεδομένων.

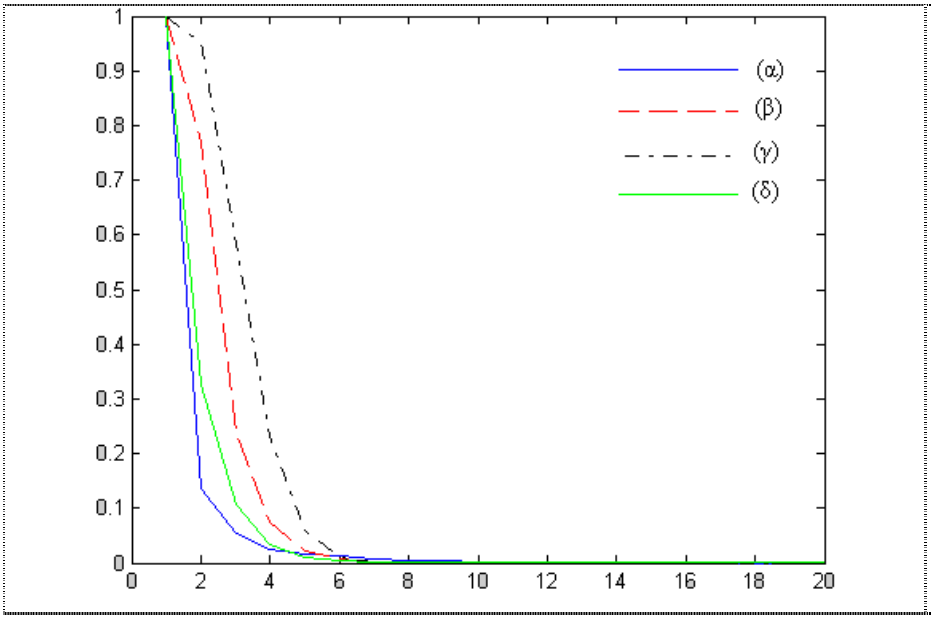
Για την μοντελοποίηση εξετάζουμε βίντεο ανθρώπινης κίνησης που συμβαίνει κατά τον οριζόντιο άξονα *(Translational)*. Τότε, χρησιμοποιώντας σαν δεδομένα περιγραφής ενός frame την *x* συνιστώσα της οπτικής ροής, παίρνουμε τα αποτελέσματα του σχήματος 5.2. Παρατηρούμε πως η Εκθετική κατανομή προσεγγίζει καλύτερα τα δεδομένα μας, επομένως στην CUSUM θα χρησιμοποιούμε αυτήν την κατανομή. Πειράματα με περισσότερα βίντεο επιβεβαίωσαν ότι η Laplace κατανομή περιγράφει καλύτερα τα δεδομένα μας, και επομένως πρέπει να

χρησιμοποιηθεί στο CUSUM. Στον πίνακα 5.1 φαίνονται οι αποστάσεις (μέσο τετραγωνικό σφάλμα) των κατανομών της ενέργειας τριών video σε σχέση με τις τρεις κατανομές που εξετάζουμε. Όπως βλέπουμε, η Εκθετική κατανομή προσεγγίζει πολύ καλύτερα τα δεδομένα μας.

Επομένως, το στατιστικό τεστ παίρνει την μορφή:

$$T_k = \max\left(0, T_{k-1} + \ln\frac{\mu_0}{\mu_1} + \frac{x}{\mu_0} - \frac{x}{\mu_1}\right),$$

όπου τα μ_0, μ_1 είναι η μέση τιμή των δεδομένων πριν και μετά την αλλαγή.



Σχήμα 5.2 Μοντελοποίηση της ενέργειας της x συνιστώσας της οπτικής ροής ενός video (α) πραγματική κατανομή, (β) Laplace, (γ) Gaussian (δ) Exponential

Πίνακας 5.1 Η προσέγγιση της κατανομής της ενέργειας της οπτικής ροής από τρεις κατανομές.

	Gaussian	***Laplace***	***Exponential***
daria_walk	0.0016	9.5275e-004	3.2834e-004
eli_jump	5.6531e-004	3.4468e-004	1.3718e-004
denis_run	0.0011	7.3060e-004	2.5519e-004

5.4 Η μέθοδος που προτείνουμε

5.4.1 Εισαγωγικά

Προτείνουμε μια μέθοδο που χρησιμοποιεί την CUSUM για την ανίχνευση αλλαγών κίνησης σε βίντεο ανθρώπινης κίνησης. Γίνεται σύγκριση της ακρίβειας των αποτελεσμάτων που παίρνουμε χρησιμοποιώντας την οπτική ροή και μια μορφή της ενέργειας του σήματος. Όπως θα φανεί στη συνέχεια, καταλήγουμε ότι μια μορφή του σήματος *ELt* (ενέργεια της *x* συνιστώσας της οπτικής ροής κάθε frame, για οριζόντιες κινήσεις) οδηγεί σε καλύτερο διαχωρισμό των κινήσεων.

Στην πράξη η CUSUM είναι πιθανό να βρει επιπλέον λανθασμένα πιθανά σημεία αλλαγής κίνησης *(false alarms)*. Στην περίπτωση των ανθρώπινων κινήσεων αυτό οφείλεται στο ότι κάθε κίνηση εμπεριέχει μικρότερες, συνήθως περιοδικές, κινήσεις, όπως τα χέρια που κινούνται στο περπάτημα. Για τον λόγο αυτό, έχουμε αναπτύξει δύο τεχνικές για την όσο γίνεται καλύτερη απομάκρυνση των σημείων λανθασμένης ανίχνευσης, ο συνδυασμός των οποίων οδηγεί σε αρκετά καλά αποτελέσματα.

5.4.2 Σύγκριση χρήσης οπτικής ροής με ενέργεια κίνησης

Ελέγξαμε την χρήση της *x* συνιστώσας της οπτικής ροής στο KTH Dataset. Χρησιμοποιήσαμε αποσπάσματα 25-150 frames από 12 διαφορετικά video καλύπτοντας 6 κινήσεις *(walk, jog, run)*, 2 κατευθύνσεις κίνησης *(αριστερά, δεξιά)* και 2 ανθρώπους. Σχηματίστηκαν έτσι 144 νέα video (καθένα με ένα ζεύγος κινήσεων), στα οποία εφαρμόσαμε την μέθοδό ανίχνευσης αλλαγών CUSUM με την κατανομή Laplace.

Η μέθοδος ανίχνευσε την ύπαρξη αλλαγής μόνο στο 66% των περιπτώσεων. Στις περιπτώσεις βέβαια που ανιχνεύθηκε αλλαγή, τα σημεία που βρέθηκαν απείχαν λίγο από τα πραγματικά σημεία αλλαγής. Χρησιμοποιώντας 15 frames σαν επιτρεπόμενο όριο απόκλισης, το ποσοστό σωστής ανίχνευσης ήταν αρκετά υψηλό (88.42%). Ωστόσο, το χαμηλό ποσοστό ανίχνευσης αλλαγής μειώνει πολύ την αξιοπιστία της μεθόδου (66% x 88.42% = 58.36%). Τα αποτελέσματα για κάποια επιτρεπόμενα σφάλματα φαίνονται στον πίνακα 5.2.

Πίνακας 5.2 Η απόδοση της CUSUM με την χρήση οπτικής ροής (x συνιστώσας) σε σχέση με το επιτρεπόμενο σφάλμα του σημείου ανίχνευσης αλλαγής.

Επιτρεπόμενο σφάλμα (σε frames)	5	10	***15***	20	25
Απόδοση (%)	16.67	57.64	***58.36***	58.36	63.89

Επιπλέον, η μέθοδος δεν ανιχνεύει αλλαγή όταν αλλάζει η κατεύθυνση της κίνησης, κάτι που διαπιστώσαμε επαναλαμβάνοντας το ίδιο πείραμα, χωρίς να λαμβάνουμε την κατεύθυνση της κίνησης *(αριστερά, δεξιά)* ως διαφορά κίνησης. Η μέθοδος ανίχνευσε την ύπαρξη αλλαγής στο 77% των περιπτώσεων. Στις περιπτώσεις βέβαια που ανιχνεύθηκε αλλαγή, τα σημεία που βρέθηκαν απείχαν λίγο από τα πραγματικά σημεία αλλαγής. Χρησιμοποιώντας 15 frames σαν επιτρεπόμενο όριο απόκλισης, το ποσοστό σωστής ανίχνευσης ήταν αρκετά υψηλό (97.30%). Και πάλι όμως, το χαμηλό ποσοστό ανίχνευσης αλλαγής μειώνει πολύ την αξιοπιστία της μεθόδου (77% x 97.30% = 74.92%). Τα αποτελέσματα για κάποια επιτρεπόμενα σφάλματα φαίνονται στον πίνακα 5.3.

Πίνακας 5.3 Η απόδοση της CUSUM με την χρήση οπτικής ροής (x συνιστώσας) σε σχέση με το επιτρεπόμενο σφάλμα του σημείου ανίχνευσης αλλαγής, χωρίς να λαμβάνουμε την κατεύθυνση της κίνησης (δεξιά / αριστερά) ως διαφορά κίνησης.

Επιτρεπόμενο σφάλμα (σε frames)	5	10	***15***	20	25
Απόδοση (%)	33.33	74.31	***75***	75	76.39

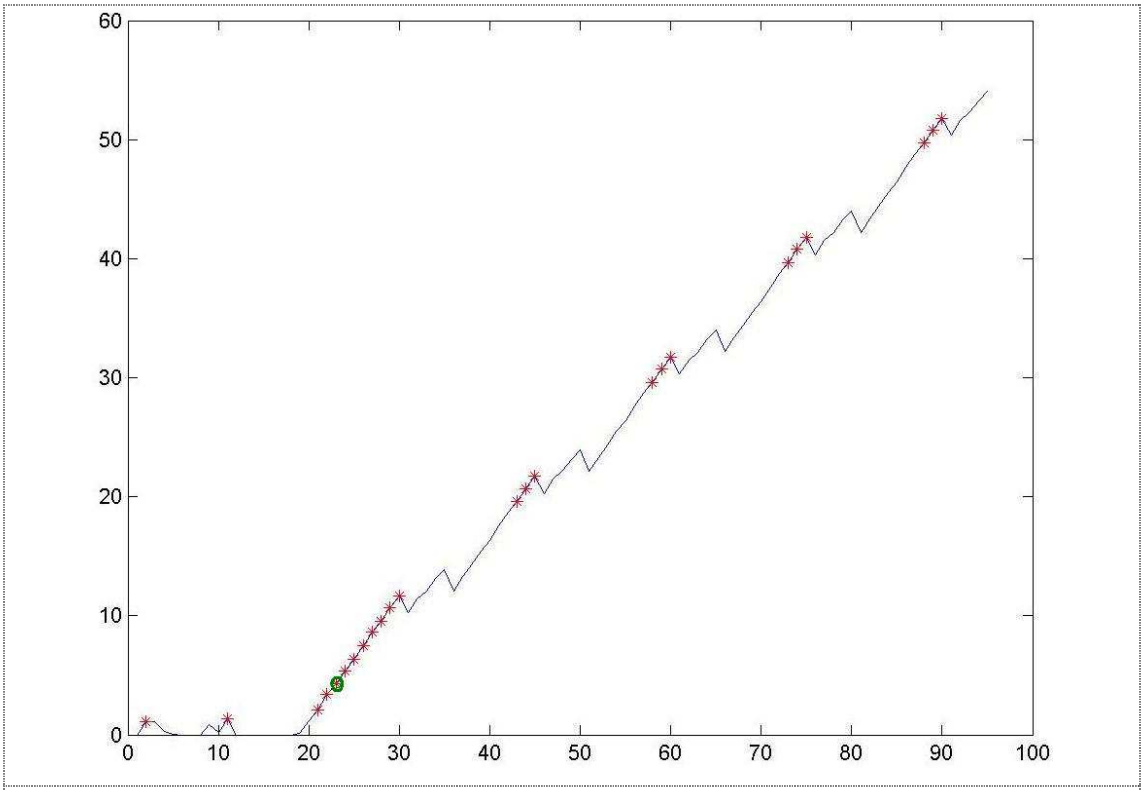
Συμπερασματικά λοιπόν, παρατηρούμε ότι η χρήση της οπτικής ροής δεν οδηγεί σε πολύ αξιόπιστα αποτελέσματα. Απαιτείται λοιπόν είτε κάποια μετα-επεξεργασία των αποτελεσμάτων είτε η χρήση άλλων δεδομένων στην CUSUM. Όπως θα δείξουμε παρακάτω, η χρήση μια μορφής του σήματος *ELt* (ενέργεια της x συνιστώσας της οπτικής ροής κάθε frame, για οριζόντιες κινήσεις), σε συνδυασμό με μια νέα τεχνική μετα-επεξεργασίας, οδηγεί σε πολύ υψηλές αποδόσεις, λαμβάνοντας υπόψη και την αλλαγή της κατεύθυνσης, πραγματοποιώντας έτσι λεπτομερέστερη ανάλυση.

5.4.3 Απομάκρυνση Λανθασμένων Ανιχνεύσεων Αλλαγών (false alarms)

Στο σχήμα 5.3 βλέπουμε το log likelihood ratio ενός video, όπου συμβαίνει αλλαγή κίνησης από *jog_left* σε *run_right* στο frame 24.

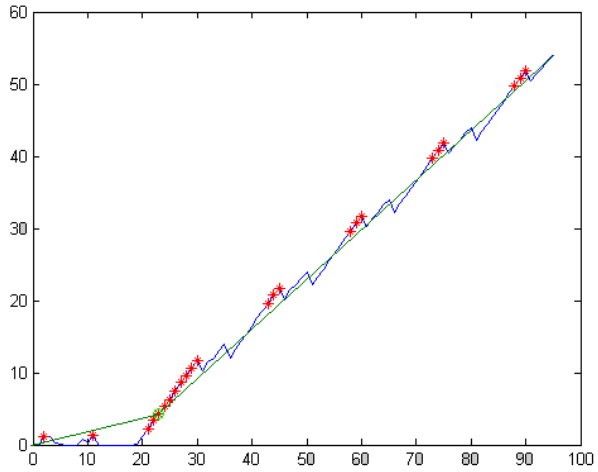
Παρατηρούμε πως υπάρχουν πολλά σημεία πιθανής αλλαγής συγκεντρωμένα σε ομάδες. Προφανώς δύο σημεία αλλαγής πρέπει να απέχουν μεταξύ τους κατά τουλάχιστον L frames, ώστε να προλάβει όντως να πραγματοποιηθεί η αλλαγή στην κίνηση. Έτσι, όταν βρίσκονται αλλαγές σε απόσταση μικρότερη από L frames, θεωρούμε ότι συμβαίνει αλλαγή στο κεντρικό σημείο που έχει εντοπιστεί.

Η τιμή του L εξαρτάται από την κάθε εφαρμογή αλλά κάποιες αποδοτικές τιμές του L μπορεί να ανήκουν στο πεδίο [5, 25]. Για τα πειράματά μας χρησιμοποιήσαμε L=25, το οποίο έδωσε πολύ καλά αποτελέσματα.



Σχήμα 5.3 Οι αλλαγές που ανιχνεύει η CUSUM σε ένα video jog_left → run_right. Η αληθινή αλλαγή σημειώνεται με o ενώ οι πιθανές με *.

Παρατηρούμε ακόμη πως το σήμα θα μπορούσε να προσεγγιστεί με δύο ευθύγραμμα τμήματα, ένα από την αρχή της κίνησης μέχρι το σημείο αλλαγής και ένα από το σημείο αλλαγής μέχρι το τέλος της κίνησης, όπως φαίνεται στο σχήμα 5.4. Τέλος, παρατηρούμε πως το σήμα έχει θόρυβο στις ψηλές συχνότητες, καθώς σχηματίζονται μικρές *ασυνέχειες* στα ευθύγραμμα τμήματα.



Σχήμα 5.4 Προσέγγιση του σήματος με δύο ευθύγραμμα τμήματα.

5.4.4 Αναλυτική παρουσίαση της μεθόδου

Με βάση τις παραπάνω παρατηρήσεις, προτείνουμε την εξής μέθοδο για την απαλοιφή των λανθασμένων σημείων αλλαγής και την διατήρηση του πραγματικού:

1. Υπολογίζουμε το στατιστικό τεστ T_k με την μέθοδο CUSUM.
2. Περνάμε το σήμα από ένα βαθυπερατό φίλτρο με συχνότητα αποκοπής ω_c, ώστε να απομακρύνουμε τον θόρυβο.
3. Μειώνουμε την πυκνότητα των σημείων πιθανής αλλαγής, ορίζοντας σαν L frames την μέγιστη επιτρεπόμενη απόσταση μεταξύ δύο σημείων αλλαγής.
4. Κανονικοποιούμε το σήμα T_k στο 1.
5. Προσεγγίζουμε το σήμα με ευθύγραμμα τμήματα, επιτρέποντας μια διαφορά εμβαδού Τ μεταξύ κάθε ευθύγραμμου τμήματος και του σήματος.
6. Εξετάζουμε τα σημεία ως προς τα χαρακτηριστικά της κίνησης για να επιλέξουμε τα πραγματικά σημεία αλλαγής *(αν υπάρχουν)*.

Στην συνέχεια, εξηγούμε αναλυτικά καθένα από τα παραπάνω βήματα.

5.4.4.1 Στατιστικό τεστ T_k με βάση την CUSUM.

Υπολογίζουμε το T_k όπως αναλύσαμε στην ενότητα 5.2. Τα δεδομένα κίνησης που χρησιμοποιούμε είναι μια μορφή του σήματος *Elt*, και συγκεκριμένα, η ενέργεια της *x* συνιστώσας της οπτικής ροής κάθε frame (επειδή εξετάζουμε κυρίως οριζόντιες κινήσεις). Αυτό αντιστοιχίζει σε κάθε frame μία μόνο τιμή που συνοψίζει τα χαρακτηριστικά της κίνησής του.

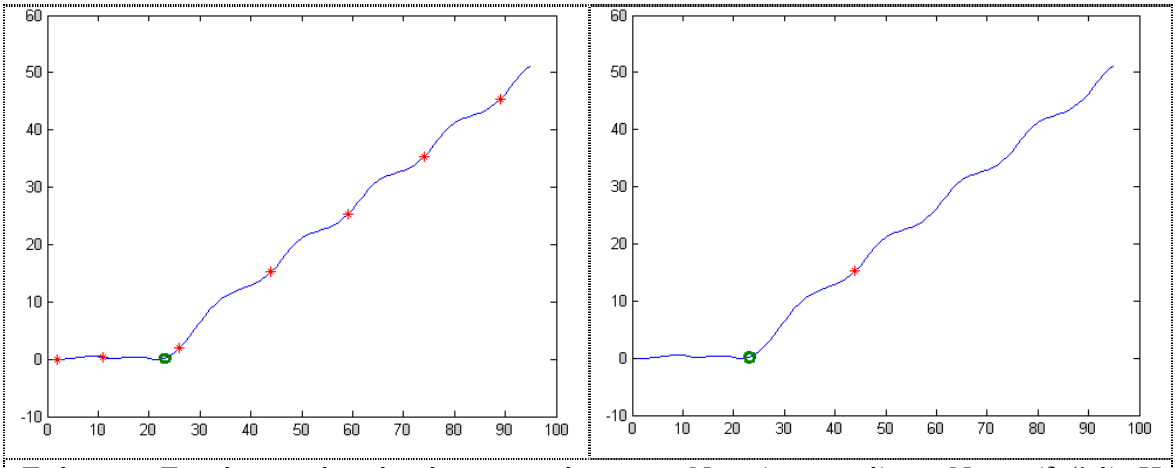
5.4.4.2 Βαθυπερατό φιλτράρισμα

Το βαθυπερατό φιλτράρισμα απομακρύνει μεγάλο μέρος του θορύβου που εμφανίζεται στο T_k και κάνει το σήμα αρκετά ομαλό, ώστε να λειτουργήσει καλύτερα η προσέγγισή του με ευθύγραμμα τμήματα. Η συχνότητα αποκοπής ω_c μπορεί να είναι αρκετά χαμηλή (δοκιμάσαμε $\omega_c = 0.2$ και είχαμε πολύ καλά αποτελέσματα).

5.4.4.3 Μείωση της πυκνότητας

Τα σημεία που απέχουν λιγότερο από L μεταξύ τους σχηματίζουν μια ομάδα σημείων, από την οποία κρατάμε μόνο το κεντρικό σημείο της ομάδας. Πρακτικά αυτό σημαίνει πως αν το πραγματικό σημείο ανήκει στην ομάδα αλλά δεν είναι το κεντρικό σημείο της ομάδας, τότε υφίσταται χρονική ολίσθηση, εισάγοντας κάποιο σφάλμα.

Η επιλογή του L είναι ιδιαίτερα σημαντική. Μικρό L $(L < 5)$ μπορεί να αφήσει πάρα πολλά σημεία και να δυσκολέψει τα επόμενα βήματα. Από την άλλη, μια μεγάλη τιμή για το L $(L > 25)$ μπορεί να προκαλέσει πολύ μεγάλη ολίσθηση στα αληθινά σημεία αλλαγής, εισάγοντας έτσι μεγάλο σφάλμα. Η συμπεριφορά αυτή αποτυπώνεται στο σχήμα 5.5.



Σχήμα 5.5 Το σήμα μετά από μείωση πυκνότητας με N=5 (αριστερά) και N=25 (δεξιά). Η μικρή τιμή του N αφήνει περισσότερα σημεία αλλά πλησιάζει καλύτερα την πραγματική αλλαγή. Αντίθετα, η υψηλή τιμή του N αποκόβει πολλά σημεία εισάγοντας μεγαλύτερο σφάλμα.

5.4.4.4 Κανονικοποίηση στο 1

Στο βήμα αυτό κάνουμε μια απλή κανονικοποίηση διαιρώντας κάθε τιμή του σήματος με την μέγιστη τιμή του, ώστε το σήμα να παίρνει τιμές στο διάστημα [0, 1]. Το βήμα αυτό είναι βοηθητικό για το επόμενο και μας επιτρέπει να επιλέξουμε πιο εύκολα την τιμή του κατωφλίου, όπως θα εξηγήσουμε στην συνέχεια.

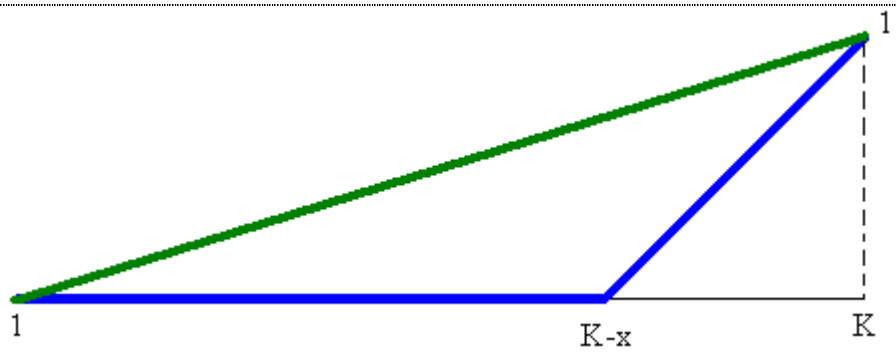
5.4.4.5 Προσέγγιση σήματος με ευθύγραμμα τμήματα

Στο βήμα αυτό προσεγγίζουμε το σήμα με ευθύγραμμα τμήματα. Από το αρχικό σήμα κρατάμε μόνο τα σημεία πιθανής αλλαγής, καθώς και το πρώτο και τελευταίο frame. Αυτά τα σημεία δημιουργούν το *σύνολο ενδιαφέροντων σημείων S*. Η διαδικασία που ακολουθείται περιγράφεται στον παρακάτω αλγόριθμο.

Αλγόριθμος 5.1: "Προσέγγιση σήματος με ευθύγραμμα τμήματα για ανίχνευση σημείων αλλαγής"
Είσοδος : το σήμα $x(t)$, ενδιαφέροντα σημεία *S*, μέγιστο σφάλμα *T* **Έξοδος** : S', όπου $S' \subseteq S$
$S' = \{\}$ Για $i = 1 : \lvert S \rvert - 2$ 1. $a = S_i$ 2. $b = S_{i+2}$ 3. $s = S_{i+1}$ 3. Υπολογίζουμε το ευθύγραμμο τμήμα $y(t) = \lambda \cdot t + \beta,\ t \in [a,b]$, όπου: $$\lambda = \frac{x(b) - x(a)}{b - a} \text{ και } \beta = x(b) - \lambda \cdot b$$ 4. Υπολογίζουμε το εμβαδόν E_x του σήματος $x(t), t \in [a,b]$ σε σχέση με τον οριζόντιο άξονα. 5. Υπολογίζουμε το εμβαδόν E_y του ευθύγραμμου τμήματος $y(t)$ σε σχέση με τον οριζόντιο άξονα. 6. $E = \left\lvert \frac{E_x - E_y}{b - a} \right\rvert$ 7. Αν $E > T$ τότε $S' = S' \cup \{s\}$ *Τέλος_επανάληψης*

Το σφάλμα Ε δείχνει την διαφορά στο εμβαδόν του σήματος T_k με τον οριζόντιο άξονα και του ευθυγράμμου τμήματος (που προσεγγίζει το T_k) με τον οριζόντιο άξονα. Γενικά, όσο πιο μικρή είναι η διαφορά, τόσο πιο καλά ταιριάζει το ευθύγραμμο τμήμα στην περιοχή του σήματος που εξετάζουμε.

Επειδή το Ε μπορεί να παίρνει τιμές αυθαίρετα μεγάλες, ειδικά αν εξετάζουμε βίντεο που είναι μεγάλο σε μήκος, διαιρούμε με το πλήθος των frames ώστε να έχουμε ανεξαρτησία από το μήκος της περιοχής που εξετάζουμε. Ακόμη, η κανονικοποίηση των τιμών του σήματος στο 1 *(ενότητα 5.4.4.4)* περιορίζει το σφάλμα στο διάστημα $\left[0,\ \frac{1}{2}\right]$. Πράγματι, η χειρότερη περίπτωση είναι αυτή του σχήματος 5.6, όπου το σφάλμα τείνει στο $\frac{1}{2}$.



Σχήμα 5.6 Η περίπτωση του μέγιστου σφάλματος

$$E_x = \frac{1}{2}\cdot(K-K+x)\cdot 1 = \frac{1}{2}\cdot x$$

$$E_y = \frac{1}{2}\cdot K\cdot 1 = \frac{1}{2}\cdot K$$

$$E = \frac{1}{K}\cdot(E_y - E_x) = \frac{1}{K}\cdot(\frac{1}{2}\cdot K - \frac{1}{2}\cdot x) = \frac{1}{2\cdot K}\cdot(K-x)$$

Το Ε γίνεται μέγιστο όταν $x=1$, *οπότε* $\max E = \frac{K-1}{2\cdot K} < \frac{1}{2}$

και $\lim_{K\to\infty} E = \lim_{K\to\infty} \frac{K-1}{2\cdot K} = \frac{1}{2}$

Το T είναι το μέγιστο επιτρεπόμενο σφάλμα ώστε η διαφορά εμβαδού να θεωρείται αμελητέα. Όταν $T=0$, απαιτούμε τέλεια προσέγγιση, οπότε επιστρέφονται όλα τα

σημεία του συνόλου S. Όταν $T \geq \frac{1}{2}$, επιτρέπουμε άπειρο σφάλμα, οπότε το S' περιέχει μόνο το αρχικό και τελικό σημείο του σήματος. Μια καλή επιλογή του *T (που έδωσε εξαιρετικά αποτελέσματα στα πειράματά μας)* είναι στο διάστημα [0.05, 0.15].

Μπορούμε ίσως να δούμε την προσέγγιση ενός σήματος με ευθύγραμμα τμήματα σαν μια μέθοδο συμπίεσης σήματος με απώλειες, καθώς από μια αναπαράσταση L δειγμάτων καταλήγουμε σε μια αναπαράσταση L' δειγμάτων, όπου $L' \leq L$.

5.4.4.6 Επιλογή σημείων αλλαγής με βάση τα στατιστικά της κίνησης

Η προσέγγιση με ευθύγραμμα τμήματα επιστρέφει πολύ λίγα σημεία, τα οποία όμως είναι ακόμη *πιθανά* σημεία αλλαγής, καθώς η αλλαγή στην πορεία του σήματος μπορεί να οφείλεται σε αλλαγές στην εμφάνιση του ανθρώπου, στο background, κτλ. Γι' αυτό επιβάλλεται η εξέταση των σημείων ως προς τα στατιστικά της κίνησης. Σημειώνουμε εδώ πως τα στατιστικά της κίνησης προκύπτουν από τις τιμές και όχι από την ενέργεια της οπτικής ροής, καθώς η ενέργεια δίνει σημαντική πληροφορία για την εξέλιξη της κίνησης αλλά όχι και για την επαρκή διάκριση των κινήσεων μεταξύ τους.

Η εξέταση γίνεται σε ένα παράθυρο *w* frames, πριν και μετά από κάθε σημείο αλλαγής. Σχηματίζονται έτσι δύο τρισδιάστατα σήματα - περιοχές (σαν μικρά video), μήκους *w* frames το καθένα, που περιέχουν την οπτική ροή των frames. Επειδή οι κινήσεις που εξετάζουμε εδώ είναι οριζόντιες και Translational, από την οπτική ροή κρατάμε μόνο την *x* συνιστώσα για την εξέταση.

Για κάθε περιοχή, εξάγουμε τα στατιστικά της, μέσο όρο μ και τυπική απόκλιση σ. Στην συνέχεια, υπολογίζουμε τις μεταβολές:

$$\Delta_1 = \left| \frac{\mu_1 - \mu_2}{\max(\mu_1, \mu_2)} \right| \text{ και } \Delta_2 = \left| \frac{\varepsilon_1 - \varepsilon_2}{\max(\varepsilon_1, \varepsilon_2)} \right|,$$

όπου τα ε_1, ε_2 είναι τα διανύσματα:

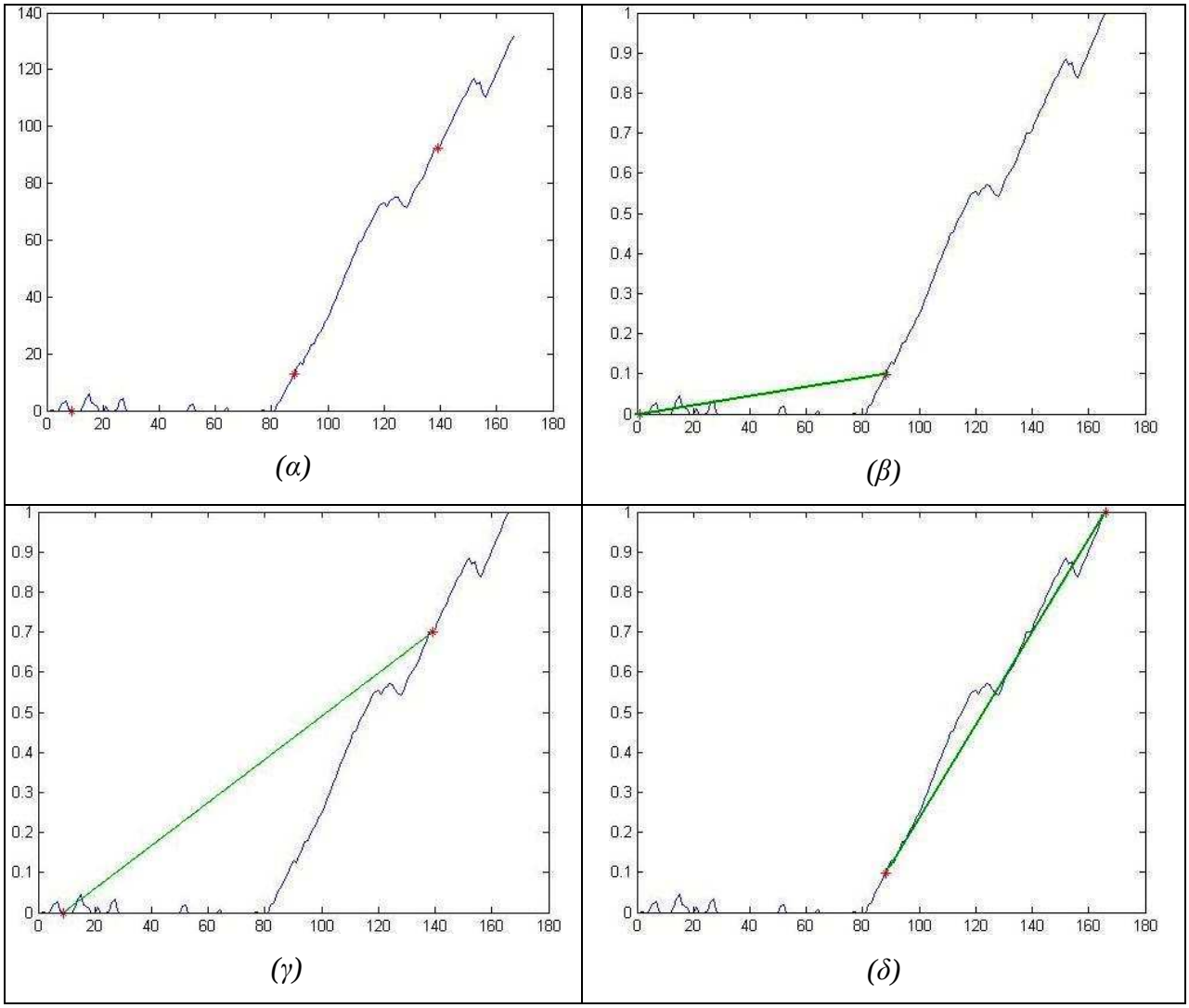
$$\varepsilon_1 = \left\langle \mu_1 - \sigma_1, \quad \mu_1 + \sigma_1 \right\rangle \text{ και } \varepsilon_2 = \left\langle \mu_2 - \sigma_2, \quad \mu_2 + \sigma_2 \right\rangle.$$

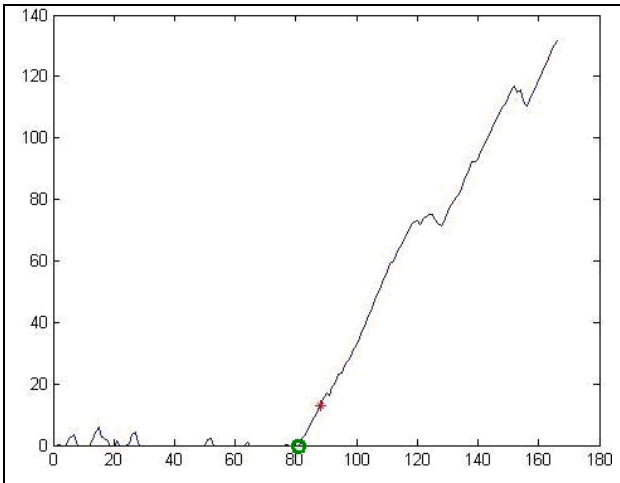
Η τελική μεταβολή είναι: $\Delta = \frac{\Delta_1 + \Delta_2}{2}$ και παράγει πιο καλά αποτελέσματα από όσο αν χρησιμοποιούσαμε μόνο κάποια από τις Δ_1, Δ_2. Αν $\Delta > \mathrm{T}_\Delta$ τότε θεωρείται ότι έχουμε πραγματική αλλαγή κίνησης.

Η επιλογή του παραθύρου *w* είναι σημαντική. Μικρό παράθυρο σημαίνει πως δεν θα εξεταστεί η κίνηση στο σύνολό της αλλά σε ένα μικρό μέρος της μόνο. Από την άλλη, ένα μεγάλο παράθυρο μπορεί να συλλάβει περισσότερες από δύο κινήσεις. Ιδανικά, θα θέλαμε το μήκος του παραθύρου να είναι ελάχιστα μεγαλύτερο από μια περίοδο της κίνησης (στην περίπτωση περιοδικών κινήσεων, όπως είναι οι περισσότερες ανθρώπινες κινήσεις), ώστε να συλλαμβάνουμε ολόκληρη την κίνηση, εξετάζοντας έτσι μόνο τα απολύτως απαραίτητα frames.

Η επιλογή του T_{Δ} είναι επίσης σημαντική γιατί καθορίζει το ποσοστό της μεταβολής που μπορούμε να ανεχτούμε σαν διαφορά μεταξύ διαφορετικών στιγμιοτύπων της ίδιας κίνησης. Μια τιμή κοντά στο 20-40% είναι ίσως μια αρκετά καλή επιλογή.

Ένα στιγμιότυπο εκτέλεσης του αλγορίθμου φαίνεται στο σχήμα 5.7.





(ε)

Σχήμα 5.7 Ένα στιγμιότυπο εκτέλεσης του αλγορίθμου επιλογής σημείων ανίχνευσης αλλαγής με προσέγγιση του σήματος με ευθύγραμμα τμήματα.

(α) το αρχικό σήμα με 3 σημεία αλλαγής

(β)-(δ) τα στάδια εξέτασης των ομάδων3 σημείων

(ε) το τελικό σήμα με 1 σημείο αλλαγής

** : σημείο αλλαγής*

o : πραγματικό σημείο αλλαγής

5.5 Πειραματικά αποτελέσματα

Ελέγξαμε την μέθοδό μας στο KTH Dataset. Χρησιμοποιήσαμε αποσπάσματα 25-150 frames από 12 διαφορετικά video καλύπτοντας 6 κινήσεις *(walk, jog, run)*, 2 κατευθύνσεις κίνησης *(αριστερά, δεξιά)* και 2 ανθρώπους. Σχηματίστηκαν έτσι 144 νέα video (καθένα με ένα ζεύγος κινήσεων), στα οποία εφαρμόσαμε την μέθοδό ανίχνευσης αλλαγών που περιγράψαμε στην ενότητα 5.4.4.

Η μέθοδος λειτούργησε αρκετά καλά, ανιχνεύοντας σωστά την ύπαρξη ή όχι αλλαγής κίνησης στο 98.61% των περιπτώσεων. Στις περιπτώσεις που ανιχνεύθηκε αλλαγή, τα σημεία που βρέθηκαν απείχαν λίγο από τα πραγματικά σημεία αλλαγής. Χρησιμοποιώντας 15 frames σαν επιτρεπόμενο όριο απόκλισης, το ποσοστό σωστής ανίχνευσης ήταν αρκετά υψηλό (92.25%).

Έτσι, συνολικά, το ποσοστό σωστής ανίχνευσης ήταν 98.61 x 92.95 = **90.97%**, χρησιμοποιώντας 15 frames σαν επιτρεπόμενο όριο απόκλισης. Η συμπεριφορά της μεθόδου για κάποια άλλα επιτρεπόμενα σφάλματα φαίνεται στον πίνακα 5.4.

Πίνακας 5.4 Η συμπεριφορά της μεθόδου σε σχέση με το επιτρεπόμενο σφάλμα του σημείου ανίχνευσης αλλαγής και σύγκριση με την ίδια μέθοδο χωρίς το βήμα της προσέγγισης με ευθύγραμμα τμήματα.

Επιτρεπόμενο σφάλμα (σε frames)	5	10	***15***	20	25
Απόδοση (%) της μεθόδου μας	40.28	63.89	***90.97***	96.53	97.92
Απόδοση (%) χωρίς την προσέγγιση με ευθύγραμμα τμήματα	36.11	52.08	***77.78***	83.33	84.03

Στον πίνακα 5.4 βλέπουμε ακόμη την απόδοση της μεθόδου χωρίς την προσέγγιση του σήματος με ευθύγραμμα τμήματα. Παρατηρούμε πως η προσέγγιση αυτή αυξάνει την απόδοση κατά περίπου 10-13%.

Επίσης, στον πίνακα 5.5 βλέπουμε τα αντίστοιχα αποτελέσματα αν είχαμε γνωστές τις κατανομές f_0 και f_1 *(αν για παράδειγμα ήταν γνωστές οι κινήσεις και ζητούσαμε να βρούμε σε ποιο σημείο ακριβώς έχουμε αλλαγή κίνησης)*. Παρατηρούμε πως τα ποσοστά είναι υψηλότερα κατά περίπου 2% στην περίπτωση της πλήρους μεθόδου και κατά περίπου 10% στην περίπτωση που δεν χρησιμοποιούμε την προσέγγιση με ευθύγραμμα τμήματα.

Η διαφορά 10% είναι ασφαλώς αναμενόμενη, καθώς η ακριβής γνώση των κατανομών οδηγεί σίγουρα σε πολύ πιο ακριβή αποτελέσματα. Ωστόσο, η πολύ μικρή διαφορά 2% αναδεικνύει σίγουρα την σπουδαιότητα της προσέγγισης του σήματος με ευθύγραμμα τμήματα, καθώς έτσι αναπληρώνεται κατά πολύ η απόκλιση που εισάγει η χρήση των εμπειρικών κατανομών από πεπερασμένο αριθμό δειγμάτων.

Πίνακας 5.5 Η συμπεριφορά της μεθόδου σε σχέση με το επιτρεπόμενο σφάλμα του σημείου ανίχνευσης αλλαγής και σύγκριση με την ίδια μέθοδο χωρίς το βήμα της προσέγγισης με ευθύγραμμα τμήματα θεωρώντας γνωστές τις κατανομές f_0 και f_1. .

Επιτρεπόμενο σφάλμα (σε frames)	5	10	***15***	20	25
Απόδοση (%) της μεθόδου μας	22.92	85.42	***93.06***	94.44	95.14
Απόδοση (%) χωρίς την προσέγγιση με ευθύγραμμα τμήματα	21.52	81.25	***88.19***	89.58	90.97

5.6 Ανίχνευση αλλαγών σε non-Translational κινήσεις

Η μέθοδος που προτείναμε στην ενότητα 5.4.4 δεν λειτουργεί αποδοτικά (25-35%) στην περίπτωση non-Translational κινήσεων, κάτι αναμενόμενο καθώς η ενέργεια της κίνησης δεν είναι τόσο περιγραφική σε αυτές τις κινήσεις, ώστε να διαχωρίζει την μία κίνηση από την άλλη. Προτείνεται λοιπόν η ανίχνευση αλλαγών με βάση το σχήμα του Activity Area, το οποίο αναμένουμε να είναι πολύ πιο περιγραφικό.

Στην πράξη, επιλέγεται ο τρόπος προσέγγισης του προβλήματος της αλλαγής κίνησης βρίσκοντας προηγουμένως αν η κίνηση είναι Translational ή non-

Translational. Η διαδικασία αυτή είναι αυτόματη και με πολύ μεγάλη ακρίβεια, οπότε η επίλυση του προβλήματος με την σωστή μέθοδο είναι αναμενόμενη.

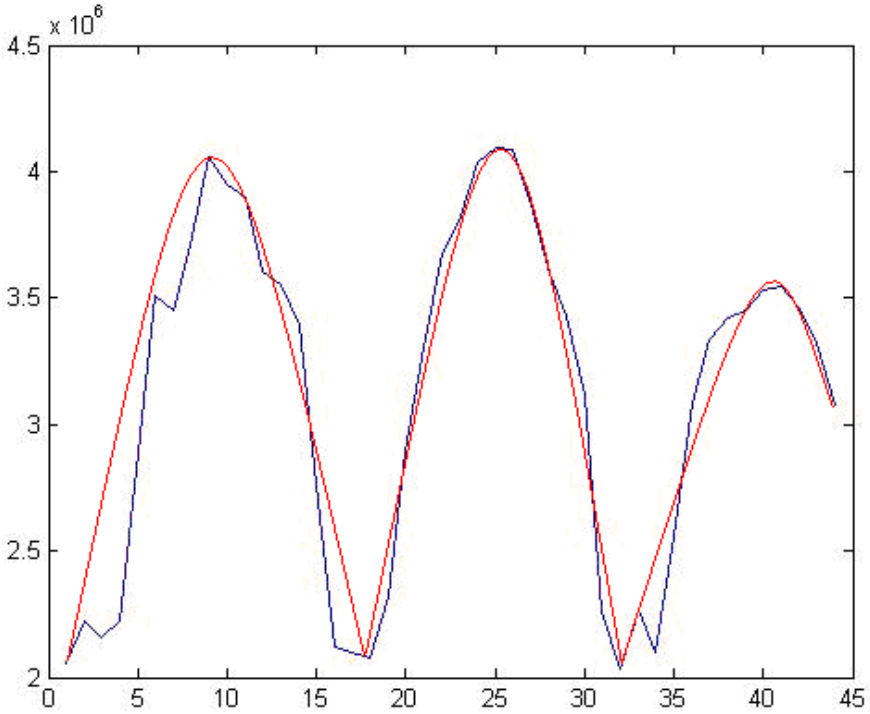
5.7 Συμπεράσματα

Σε αυτό το κεφάλαιο ασχοληθήκαμε με την σειριακή ανίχνευση αλλαγών και παρουσιάσαμε μία μέθοδο βασιζόμενη στην ενέργεια της κίνησης η οποία πειραματικά έδειξε αρκετά καλά αποτελέσματα (91%) με μικρό σχετικά σφάλμα (15 frames) στην περίπτωση των Translational κινήσεων.

Η ίδια μέθοδος, δεν έχει καλή απόδοση (25-35%) στις non-Translational κινήσεις, καθώς η ενέργειά τους δεν τις διαχωρίζει αποτελεσματικά. Παρόλα αυτά, η χρησιμότητά της δεν μειώνεται, καθώς η ανίχνευση αλλαγών έπεται σαν βήμα του διαχωρισμού σε Translational/non-Translational κίνηση, οπότε και χρησιμοποιείται διαφορετικό χαρακτηριστικό της κίνησης για τις non-Translational περιπτώσεις. Ωστόσο, η διερεύνηση της ανίχνευσης αλλαγών σε non-Translational κινήσεις εντάσσεται σίγουρα στους στόχους της μελλοντικής μας εργασίας.

Κεφάλαιο 6

"Σήματα που ακολουθούν την κίνηση"



Σελίδα σκόπιμα κενή.

6.1. Εισαγωγικά

Ορισμός 6.1

Σαν «στιγμιότυπο πλήρους κίνησης» (Full Action Instance - FAI) Π *ορίζουμε το βασικό μοτίβο κίνησης που επαναλαμβάνεται συνεχώς σε κάθε περιοδική κίνηση.*

Επομένως, μια ακολουθία εικόνων, στην οποία εμφανίζεται μόνο μια συγκεκριμένη περιοδική κίνηση, περιέχει επαναλήψεις του στιγμιοτύπου πλήρους κίνησης μιας περιόδου, Π, και μπορεί να περιγραφεί σαν $\mathrm{K} = \{\Pi, \Pi, ..., \Pi\}$.

Για παράδειγμα, στο περπάτημα το στιγμιότυπο πλήρους κίνησης είναι ένα βήμα, δηλαδή από την στιγμή που ο άνθρωπος στηρίζεται εξ ολοκλήρου στο ένα του πόδι και είναι έτοιμος να κινηθεί με το άλλο μέχρι την στιγμή που στηρίζεται στο άλλο του πόδι και είναι έτοιμος να κινηθεί με το πρώτο.

Βέβαια, σε μια κίνηση μπορούμε να έχουμε περισσότερους από έναν διαφορετικούς ορισμούς του στιγμιοτύπου πλήρους κίνησης. Για παράδειγμα, στο περπάτημα κάποιος θα μπορούσε να ορίσει το στιγμιότυπο πλήρους κίνησης από την στιγμή που ο άνθρωπος στηρίζεται εξ ολοκλήρου στο αριστερό του πόδι και είναι έτοιμος να κινηθεί με το δεξί μέχρι την στιγμή που στηρίζεται ξανά στο αριστερό του πόδι και είναι έτοιμος να κινηθεί με το δεξί. Σε αυτήν την περίπτωση λέμε ότι το στιγμιότυπο Π΄ είναι *διπλάσιο* από το Π.

Υπάρχουν δύο βασικοί στόχοι που πρέπει να ικανοποιεί ο ορισμός κάθε στιγμιοτύπου μιας κίνησης, χωρίς βέβαια να είναι δεσμευτικοί. Ο ορισμός πρέπει:

- να περιέχει όντως την έννοια του μοτίβου που επαναλαμβάνεται, να μην έχουμε δηλαδή περιγραφές του τύπου $\mathrm{K} = \{\Pi, \Pi, ..., \Theta, ..., \Pi\}$, όπου $\Theta \neq \Pi$, εκτός από την πρώτη και την τελευταία κίνηση, οι οποίες μπορεί να μην ήταν δυνατόν να ληφθούν ολοκληρωμένες, δηλαδή της μορφής $\mathrm{K} = \{\Theta, \Pi, \Pi, ..., ..., \Pi, \Theta\}$.
- να έχει πρακτική και λογική αξία. Για παράδειγμα, μπορούμε να ορίσουμε σαν αρχή ενός άλματος την στιγμή που άνθρωπος βρίσκεται στον αέρα, ο ορισμός μας όμως θα έχει μικρή πρακτική αξία καθώς σπάνια κάποιος θα όριζε έτσι ένα άλμα.

6.2 Το πρόβλημα του διαχωρισμού της κίνησης σε κύκλους

Το πρόβλημα του διαχωρισμού μιας περιοδικής κίνησης σε στιγμιότυπα *(κύκλους)* πλήρους κίνησης (FAI) ανήκει κυρίως στο πεδίο της ανάλυσης βηματισμού *(gait analysis)* και βρίσκει εφαρμογές στην Βιομετρική (αναγνώριση ανθρώπου από το περπάτημά του) [30, 31] αλλά και στην Ιατρική (κατανόηση της διαδικασίας του περπατήματος, ανίχνευση ανωμαλιών βαδίσματος). Ωστόσο, μέχρι τώρα δεν έχει εφαρμοστεί ευρέως σε άλλες κινήσεις εκτός από το περπάτημα. *Στα [28, 29]* δείξαμε ότι έχει νόημα να κάνουμε ανίχνευση κύκλων και σε άλλες κινήσεις, καθώς πολλές από τις κινήσεις του ανθρώπου είναι περιοδικές.

Σε αυτό το κεφάλαιο θα επιχειρήσουμε μια εξερεύνηση της κίνησης και των σημάτων που ακολουθούν την κίνηση. Πιστεύουμε πως η ανάλυσή σε κύκλους είναι αρκετά σημαντική καθώς μπορεί να δώσει απαντήσεις σε ερωτήματα όπως «Πόσα βήματα κάνει αυτός ο άνθρωπος;» ή «Πόσο διαρκεί το πρώτο βήμα και πόσο το δεύτερο;».

6.3 Σήματα που ακολουθούν την κίνηση

Ορισμός 6.2

Ένα σήμα ακολουθεί την κίνηση ενός κινούμενου ανθρώπου όταν μπορεί να δώσει πληροφορίες σχετικά με την έναρξη και την ολοκλήρωση των στιγμιοτύπων πλήρους κίνησης (FAI).

Το πιο ευρέως χρησιμοποιούμενο τέτοιο σήμα είναι το *foreground sum signal*. Στα [28,29] παρουσιάσαμε ένα άλλο σήμα, το σήμα ενέργειας παραγώγου ακολουθίας εικόνων EL_t , το οποίο δείξαμε πως έχει παρόμοιες ιδιότητες. Το ίσως πιο ενδιαφέρον είναι πως υπάρχουν πολλά ακόμη σήματα με παρόμοιες ιδιότητες, μερικά από τα οποία θα παρουσιάσουμε στην συνέχεια.

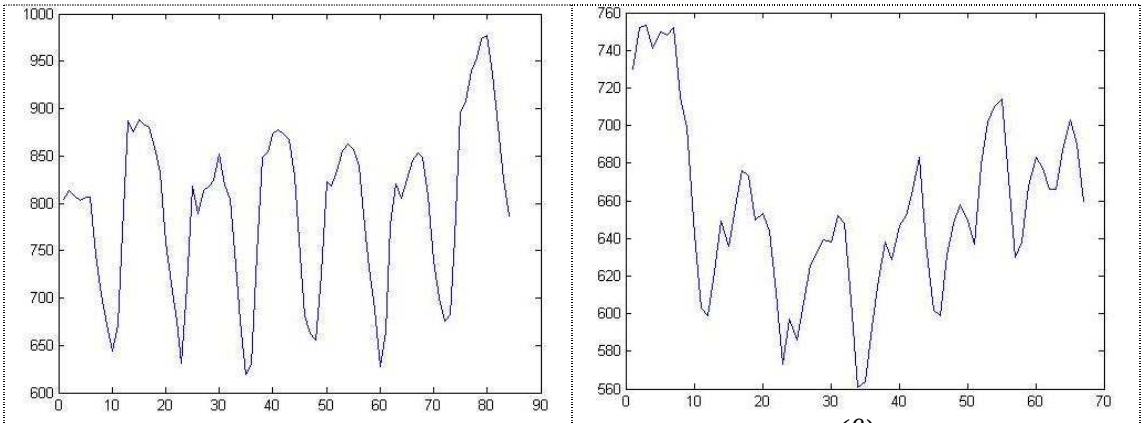
6.3.1 foreground sum signal

Το πιο ευρέως χρησιμοποιούμενο τέτοιο σήμα είναι το foreground sum signal, που ορίζεται ως εξής:

$$S(n) = \sum_{x,y} I(x, y, n), n = 0,...., N-1$$

όπου *I(x,y,n)* είναι η *n*-στη εικόνα περιγράμματος (silhouette image) του κινούμενου ανθρώπου. Αυτό το σήμα είναι συνήθως πολύ θορυβώδες εξαιτίας του θορυβώδους background και γι' αυτό οι περισσότερες μέθοδοι κάνουν πρώτα εξομάλυνσή του σήματος και στην συνέχεια ανιχνεύουν τα FAI *[30, 31]*.

Κοιτάζοντας τα σήματα S(n) δύο video παρατηρούμε πως όντως το σήμα S(n) ακολουθεί την κίνηση (σχήμα 6.1).



(α) *(β)*

Σχήμα 6.1 Το σήμα S(n) για το video daria_walk (α) και για το video daria_jump (β) του Weizmann Dataset

6.3.2 Σήμα ενέργειας παραγώγου ακολουθίας εικόνων EL_t

Για κάθε μια από τις **k-1** εικόνες της ακολουθίας L_t μπορούμε να βρούμε την ενέργεια της ως εξής:

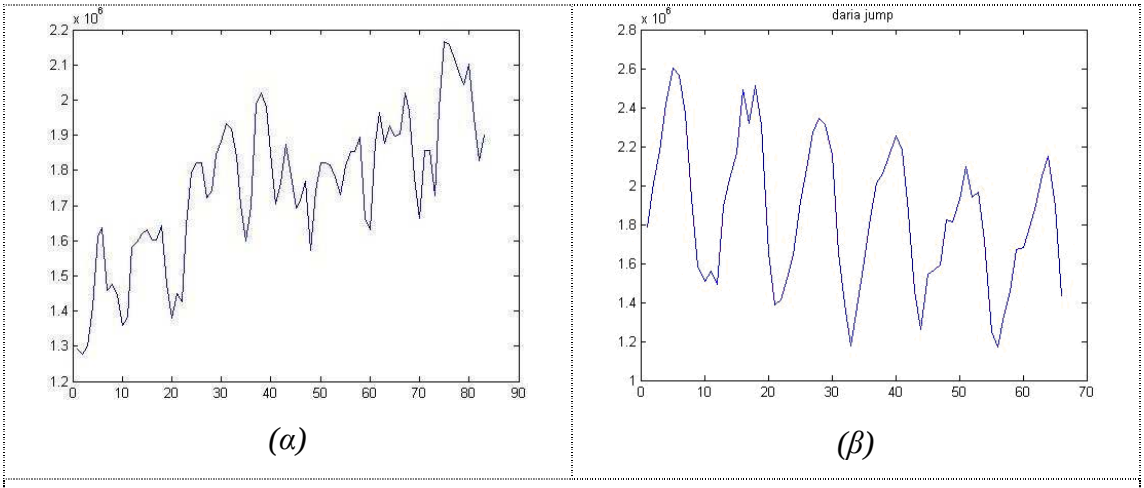
$$E\{L_{ti}\} = \sum_{x=0,1,\ldots,n-1}^{y=0,1,\ldots,m-1} \left| L_{ti}(x,y) \right|^2$$

Αν ορίσουμε μια ακολουθία EL_t ως εξής

$$EL_t(i) = E\{L_{ti}\}, i = 0,1,\ldots,k-2$$

παίρνουμε σαν αποτέλεσμα ένα μονοδιάστατο σήμα που δείχνει την χρονική εξέλιξη της ενέργειας της παραγώγου (ως προς t) της αρχικής ακολουθίας εικόνων.

Κοιτάζοντας τα σήματα EL_t δύο video παρατηρούμε πως όντως το σήμα EL_t ακολουθεί την κίνηση (σχήμα 6.2).



(α) *(β)*

Σχήμα 6.2 Το σήμα EL_t για το video daria_walk (α) και για το video daria_jump (β) του Weizmann Dataset

Στο [28] δείξαμε ότι το σήμα *ELt* είναι ένα αρκετά καλό σήμα κίνησης και περιγράψαμε μια μέθοδο υπολογισμού των άκρων των FAI με αρκετά μικρό σφάλμα *(3 frames)* και με αρκετά μεγάλη επιτυχία ανίχνευσης (89%).

Όπως θα δούμε σε αυτήν την εργασία, υπάρχουν πολλά ακόμη σήματα που ακολουθούν την κίνηση έχοντας παρόμοιες ιδιότητες.

6.4 Αξιολόγηση των σημάτων που ακολουθούν την κίνηση

Η αξιολόγηση των σημάτων που ακολουθούν την κίνηση κρίνεται απαραίτητη ώστε να έχουμε γνώση για την συμπεριφορά του σήματος αλλά και για το ποιο σήμα είναι καταλληλότερο για κάθε εφαρμογή.

Γενικά, προσδοκούμε πως η αξιολόγηση θα καταλήξει είτε σε ένα μοναδικό σήμα, το οποίο θα αναδειχθεί το καλύτερο σε όλα τα κριτήρια, είτε σε ένα σύνολο σημάτων, καθένα από τα οποία θα είναι καταλληλότερο για μια συγκεκριμένη εφαρμογή.

6.4.1 Ιδιότητες που θα πρέπει να έχει ένα καλό σήμα κίνησης

Ένα σήμα που ακολουθεί την κίνηση θα πρέπει να έχει κάποιες ιδιότητες ώστε να μπορεί να χρησιμοποιηθεί για την εξαγωγή χρήσιμων συμπερασμάτων. Κάποιες από αυτές είναι οι εξής:

α) Το σήμα πρέπει να μας δίνει την δυνατότητα να υπολογίζουμε τα άκρα των FAI με ικανοποιητική ακρίβεια.

β) Το σήμα πρέπει να μας δίνει την δυνατότητα να ανιχνεύουμε χρονικές στιγμές στις οποίες δεν γίνεται κάποια σημαντική κίνηση και να είναι αρκετά ανθεκτικό στην παρουσία θορύβου ως προς αυτό.

γ) Το σήμα πρέπει να μας δίνει την δυνατότητα να παρακολουθήσουμε ακριβώς την κίνηση. Πρέπει επομένως να διαχωρίζει με κάποιον τρόπο τα FAI της μορφής Θ που περιγράψαμε στην αρχή, και τα οποία αντιστοιχούν σε κινήσεις οι οποίες μπορεί να μην ήταν δυνατόν να ληφθούν ολοκληρωμένες, δίνοντας έτσι περιγραφές της μορφής $K = \{\Theta, \Pi, \Pi, ..., ..., \Pi, \Theta\}$.

Από τις τρεις παραπάνω ιδιότητες, θα ασχοληθούμε μόνο με την πρώτη. Παρόλο που έχουμε ενδείξεις πως πολλά από τα σήματα που θα παρουσιάσουμε στην συνέχεια ικανοποιούν ως έναν βαθμό και τις δύο άλλες ιδιότητες, η σε βάθος πειραματική διερεύνηση του ζητήματος αποτελεί έναν από τους στόχους μελλοντικής εργασίας μας.

6.5 Ποια σήματα θα αξιολογήσουμε

Όπως αναφέραμε και προηγουμένως, υπάρχουν πάρα πολλά σήματα τα οποία ακολουθούν την κίνηση. Στα πλαίσια αυτής της εργασίας θελήσαμε να εξετάσουμε 23 τέτοια σήματα.

Καθένα από αυτά τα σήματα θα μπορούσαμε να πούμε πως αποτελείται από τα δεδομένα που χρησιμοποιούμε, από έναν μετασχηματισμό που εφαρμόζεται στα δεδομένα και από μια συνάρτηση.

Τα δεδομένα που χρησιμοποιούμε είναι τα εξής:

Α) Οι δυαδικές μάσκες της σιλουέτας του κινούμενου υποκειμένου

Β) Η παράγωγος των frames (ως προς τον χρόνο *t*)

Γ) Η οπτική ροή *(optical flow)*

Οι μετασχηματισμοί που χρησιμοποιούμε είναι οι εξής:

Α) Ο ταυτοτικός μετασχηματισμός I

Β) Το φιλτράρισμα με την δυαδική μάσκα της σιλουέτας του κινούμενου υποκειμένου

Οι συναρτήσεις που χρησιμοποιούμε είναι οι εξής:

Α) Το άθροισμα των απολύτων τιμών των δεδομένων

Β) Το άθροισμα της ενέργειας των δεδομένων

Γ) Η από κοινού εντροπία (joint entropy) των δεδομένων

Για παράδειγμα, το σήμα *ELt* χρησιμοποιεί την συνάρτηση B, τον μετασχηματισμό A και τα δεδομένα B.

Τα σήματα που αξιολογούμε αναγράφονται στον πίνακα 6.1.

Πίνακας 6.1 23 σήματα που ακολουθούν την κίνηση.

Κωδικός σήματος	***Περιγραφή σήματος***
1.	Foreground Sum Signal
2.	Ενέργεια παραγώγου ως προς t, *ELt*
3.	Ενέργεια παραγώγου ως προς x, *ELx*
4.	Ενέργεια παραγώγου ως προς y, *ELy*
5.	Από κοινού εντροπία (joint entropy) της παραγώγου ως προς *t*
6.	Άθροισμα παραγώγου ως προς t
7.	Άθροισμα παραγώγου ως προς x
8.	Άθροισμα παραγώγου ως προς y
9.	Ενέργεια της συνισταμένης της οπτικής ροής
10.	Ενέργεια της συνιστώσας x της οπτικής ροής
11.	Ενέργεια της συνιστώσας y της οπτικής ροής
12.	Άθροισμα της συνισταμένης της οπτικής ροής
13.	Άθροισμα της συνιστώσας x της οπτικής ροής
14.	Άθροισμα της συνιστώσας y της οπτικής ροής
15.	Ενέργεια φιλτραρισμένης παραγώγου ως προς t
16.	Άθροισμα φιλτραρισμένης παραγώγου ως προς t
17.	Από κοινού εντροπία της φιλτραρισμένης παραγώγου ως προς *t*
18.	Ενέργεια της συνισταμένης της φιλτραρισμένης οπτικής ροής
19.	Ενέργεια της συνιστώσας x της φιλτραρισμένης οπτικής ροής
20.	Ενέργεια της συνιστώσας y της φιλτραρισμένης οπτικής ροής
21.	Άθροισμα της συνισταμένης της φιλτραρισμένης οπτικής ροής
22.	Άθροισμα της συνιστώσας x της φιλτραρισμένης οπτικής ροής
23.	Άθροισμα της συνιστώσας y της φιλτραρισμένης οπτικής ροής

6.6 Κριτήρια αξιολόγησης των σημάτων

Στο [28] εξερευνήσαμε το σήμα ELt και είδαμε ότι έχει κάποιες χρήσιμες ιδιότητες:

I) Τα άκρα των FAI είναι τοπικά ελάχιστα του σήματος με αρκετά μικρή απόκλιση (3 frames).

I I) Ένα στιγμιότυπο πλήρους κίνησης (FAI) έχει συνήθως την μορφή καμπάνας.

Η ιδιότητα I είναι πολύ σημαντική γιατί μειώνει τον χώρο στον οποίο μπορούμε να αναζητήσουμε τα άκρα των FAI. Όπως βλέπουμε στον πίνακα 6.2, η ιδιότητα ισχύει για το 90% των άκρων με σφάλμα 1 frame και για το 97.8% των άκρων με σφάλμα 3 frames, το οποίο είναι ένα αρκετά αποδεκτό σφάλμα.

Πίνακας 6.2 Ισχύς της ιδιότητας I με βάση το επιτρεπόμενο σφάλμα στα άκρα των FAI.

Σφάλμα σε *frames*	0	1	2	3	4	5	6
Ισχύς (%)	44.4	90.3	96.9	97.8	98.8	98.9	99.6

Επηρεασμένοι από αυτήν την ιδιότητα, αναπτύξαμε μια μέθοδο ανίχνευσης των άκρων των FAI [28]. Η μέθοδος αυτή εξερευνεί τα τοπικά ελάχιστα του σήματος *ELt* και ανιχνεύει τα άκρα των FAI. Στο [29] δείξαμε ότι η απόδοση της μεθόδου φθίνει με την αύξηση του θορύβου, η οποία και οδηγεί σε ανάλογη αύξηση της πυκνότητας των τοπικών ελαχίστων του σήματος (πίνακας 6.3).

Η πυκνότητα εκφράζεται ως:

$$\delta_{\min} = \frac{\Pi\lambda\acute{\eta}\theta o\varsigma\ \ \tau o\pi\iota\kappa\acute{\omega}\nu\ \ \varepsilon\lambda\alpha\chi\acute{\iota}\sigma\tau\omega\nu}{\Pi\lambda\acute{\eta}\theta o\varsigma\ \ \acute{\alpha}\kappa\rho\omega\nu\ \ \tau\omega\nu\ \ FAI}$$

Εφόσον η μέθοδός μας ανιχνεύει τα άκρα των FAI μέσα από τα τοπικά ελάχιστα του σήματος, είναι ίσως αρκετά προφανές πως μεγάλη πυκνότητα σημαίνει πολλά τοπικά ελάχιστα και επομένως μεγαλύτερη πιθανότητα σφάλματος αλλά και μεγαλύτερη δυσκολία επιλογής των σωστών άκρων.

σ_{Noise}	Avg. δ_{min}	Avg. Perf. (%)
0	1.69	89.33
1	1.72	89.08
3	1.8	88.83
5	2.02	88.59
7	2.22	86.6
9	2.44	88.83
11	2.64	83.37
13	2.75	76.43
15	2.97	75.19
17	3.11	73.45
19	3.11	65.76

Πίνακας 6.3 Η μέση πυκνότητα $\delta_{\min}$ ελαχίστων του σήματος αυξάνει καθώς αυξάνει η διακύμανση προσθετικού Gaussian θορύβου $N(0,\sigma^2)$, γεγονός που οδηγεί σε πτώση της απόδοσης της μεθόδου ανίχνευσης των FAI.

Ιδανικά θα θέλαμε μια πυκνότητα ίση με 1, οπότε θα ανιχνεύαμε άμεσα τα τοπικά ελάχιστα ως άκρα των FAI. Οι επιλογές είναι είτε να φιλτράρουμε τα δεδομένα ώστε να μειώσουμε τον θόρυβο είτε να κάνουμε κατευθείαν την ανίχνευση.

Πρακτικά όμως ποτέ δεν θα μπορέσουμε να έχουμε μηδενικό θόρυβο στα δεδομένα μας, επομένως θα πρέπει αναγκαστικά να αναπτύξουμε μεθόδους που να αντιμετωπίζουν δεδομένα με θόρυβο. Επιπλέον, το σήμα κίνησης περιέχει θόρυβο σε μια σχετικά μεγάλη περιοχή συχνοτήτων, επομένως οι κλασικές τεχνικές φίλτρων, αν δεν αποτυγχάνουν, απαιτούν τουλάχιστον την ακριβή εκτίμηση των συχνοτήτων αποκοπής.

Στην παρούσα εργασία εξετάζουμε τα σήματα τόσο ως προς την ισχύ της ιδιότητας I όσο και ως προς την πυκνότητά τους, τόσο συνολικά *(μέση πυκνότητα σήματος)* όσο και ειδικά *(πυκνότητα σήματος σε συγκεκριμένη κατηγορία κίνησης)*.

6.7 Αποτελέσματα της αξιολόγησης των σημάτων ως προς την ισχύ της ιδιότητας I

Αρχικά αξιολογήσαμε τα 23 σήματα ως προς την ισχύ της ιδιότητας I, ώστε να δούμε αν όντως το σήμα ακολουθεί την κίνηση. Η αξιολόγηση έγινε σε όλα τα video του Weizmann Dataset, χρησιμοποιώντας για ορισμούς των FAI τους ορισμούς που δώσαμε στο [28]. Για κάθε σήμα προσδιορίσαμε το ποσοστό των άκρων των FAI περιλαμβανόταν στα τοπικά ελάχιστα του σήματος.

Με απόκλιση κανενός (0) frame, τα ποσοστά ήταν από 15% έως 45%. Ακόμα όμως και με απόκλιση 1 frame, αρκετά σήματα παρουσιάσανε ένα αρκετά υψηλό ποσοστό (80-90%).

Αυτό είναι αναμενόμενο γιατί μικρά σφάλματα μπορούν να προκύψουν από θόρυβο μετρήσεων στο βίντεο *(measurement noise)* αλλά επίσης και από αριθμητικά σφάλματα στις πράξεις. Όμως η απόκλιση κατά μηδέν ή ένα frames δεν είναι απαραίτητη, αφού και το ανθρώπινο μάτι δεν είναι ευαίσθητο σε αλλαγές λίγων frames, της τάξης των 10 – 20.

Με απόκλιση 3 frames, σε όλα τα σήματα το ποσοστό ήταν αρκετά ικανοποιητικό (περίπου 90%). Όταν το σφάλμα ανέβαινε στα 5 frames, το ποσοστό ανέβαινε κοντά στο 97% ενώ γινόταν *(ή έτεινε για ορισμένα σήματα στο)* 100% για σφάλμα 10 frames.

6.8 Αποτελέσματα της αξιολόγησης των σημάτων ως προς την πυκνότητα

Η αξιολόγηση έγινε σε όλα τα video του Weizmann Dataset, χρησιμοποιώντας για ορισμούς των FAI τους ορισμούς που δώσαμε στο [28].

6.8.1 Μέση πυκνότητα

Αξιολογήσαμε ως προς την μέση πυκνότητα τα 23 σήματα και καταλήξαμε στα αποτελέσματα του πίνακα 6.4.

Πίνακας 6.4 Αποτελέσματα της αξιολόγησης των 23 σημάτων ως προς την μέση πυκνότητα.

Κωδικός	***Πυκνότητα***	***Περιγραφή σήματος***
1	2.56	Foreground Sum Signal
2	1.63	Ενέργεια παραγώγου ως προς t, ELt
3	2.38	Ενέργεια παραγώγου ως προς x, ELx
4	2.50	Ενέργεια παραγώγου ως προς y, ELy
5	2.25	Από κοινού εντροπία (joint entropy) της παραγώγου ως προς t
6	1.77	Άθροισμα παραγώγου ως προς t
7	2.93	Άθροισμα παραγώγου ως προς x
8	3.17	Άθροισμα παραγώγου ως προς y
9	2.70	Ενέργεια της συνισταμένης της οπτικής ροής (optical flow)
10	2.88	Ενέργεια της συνιστώσας x της οπτικής ροής
11	2.91	Ενέργεια της συνιστώσας y της οπτικής ροής
12	2.62	Άθροισμα της συνισταμένης της οπτικής ροής (optical flow)
13	2.74	Άθροισμα της συνιστώσας x της οπτικής ροής
14	2.93	Άθροισμα της συνιστώσας y της οπτικής ροής
15	2.04	Ενέργεια φιλτραρισμένης παραγώγου ως προς t
16	2.06	Άθροισμα φιλτραρισμένης παραγώγου ως προς t
17	2.25	Από κοινού εντροπία (joint entropy) της φιλτραρισμένης παραγώγου ως προς t
18	2.73	Ενέργεια της συνισταμένης της φιλτραρισμένης οπτικής ροής
19	2.82	Ενέργεια της συνιστώσας x της φιλτραρισμένης οπτικής ροής
20	2.93	Ενέργεια της συνιστώσας y της φιλτραρισμένης οπτικής ροής
21	2.47	Άθροισμα της συνισταμένης της φιλτραρισμένης οπτικής ροής
22	2.51	Άθροισμα της συνιστώσας x της φιλτραρισμένης οπτικής ροής
23	2.75	Άθροισμα της συνιστώσας y της φιλτραρισμένης οπτικής ροής

Από τα αποτελέσματα του πίνακα 6.4 βλέπουμε ότι το σήμα *ELt* έχει την μικρότερη πυκνότητα και επομένως αναμένεται να δίνει αρκετά αξιόπιστα αποτελέσματα για πολλά είδη κίνησης.

6.8.2 Πυκνότητα σε κάθε κατηγορία κίνησης

Στην συνέχεια, μετρήσαμε την πυκνότητα των 23 σημάτων σε κάθε μια από τις 10 κατηγορίες κίνησης του Weizmann Dataset. Τα αποτελέσματα φαίνονται στον πίνακα 6.5. Με έντονο χρώμα φόντου φαίνεται η πυκνότητα του *καλύτερου* σήματος για κάθε κατηγορία κίνησης.

Πίνακας 6.5 Αποτελέσματα της αξιολόγησης των 23 σημάτων ως προς την μέση πυκνότητα ανά κατηγορία κίνησης. Με έντονο χρώμα φόντου σημειώνονται τα σήματα με τα καλύτερα αποτελέσματα ανά κατηγορία κίνησης.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
walk	1.92	2.76	2.73	3.47	2.73	2.76	3.24	3.73	3.37	3.51	3.45
jump	2.24	1.69	2.55	2.62	1.62	1.43	2.93	3.14	2.71	2.62	2.83
run	1.62	1.33	2.43	2.55	1.71	1.29	2.71	2.69	2.33	2.60	2.60
pjump	1.47	0.93	1.15	1.17	1.27	0.97	1.64	1.76	1.05	1.16	1.16
bend	6.11	1.74	6.30	5.81	4.78	3.15	7.37	7.74	6.07	6.93	6.48
jack	2.15	1.51	1.77	2.18	1.79	1.64	2.10	2.56	2.66	2.64	2.87
side	2.78	2.69	2.75	2.86	2.78	2.31	2.81	3.36	3.33	3.58	3.56
skip	2.10	1.90	3.41	3.27	2.39	1.93	3.63	3.56	2.98	3.00	2.98
wave1	3.69	1.30	2.15	2.30	2.91	1.83	3.30	3.70	2.80	3.06	3.07
wave2	3.46	1.23	1.48	1.35	2.23	1.44	2.52	2.46	2.23	2.56	2.75

	12	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**
walk	3.06	3.10	3.33	2.76	2.33	2.73	2.73	2.86	3.14	2.14	2.20	2.61
jump	2.31	2.21	2.71	1.88	1.83	1.62	3.00	2.98	3.12	2.71	2.69	2.88
run	2.10	2.19	2.33	1.55	1.38	1.71	2.07	2.29	2.36	1.81	1.81	2.10
pjump	0.99	1.11	1.24	1.12	1.09	1.27	1.20	1.28	1.33	1.15	1.15	1.32
bend	6.22	6.78	6.89	3.19	3.48	4.78	5.44	6.22	5.81	5.07	5.74	5.81
jack	2.46	2.51	2.79	2.00	2.26	1.79	2.59	2.92	2.72	2.30	2.57	2.52
side	3.39	3.36	3.53	2.86	2.86	2.78	3.58	3.42	3.64	3.39	3.03	3.36
skip	2.71	2.80	2.93	2.02	1.90	2.39	2.78	2.63	2.85	2.34	2.34	2.41
wave1	3.20	3.41	3.43	2.35	2.54	2.91	3.19	3.11	3.37	3.04	3.02	3.35
wave2	2.37	2.65	2.85	1.75	2.04	2.23	2.92	2.83	3.17	2.77	2.71	3.23

6.8.3 Ερμηνεία των αποτελεσμάτων

Τα αποτελέσματα αυτά θεωρούμε ότι είναι πολύ σημαντικά γιατί μας επιβεβαιώνουν την ορθότητα της μέχρι τώρα χρήσης του Foreground Sum Signal στην ανάλυση περπατήματος *(walk)*.

Μας δείχνουν ωστόσο ακόμη πως ενώ το Foreground Sum Signal μπορεί να χρησιμοποιείται άφοβα στο περπάτημα, δεν ισχύει το ίδιο και για τις υπόλοιπες κατηγορίες κίνησης. Ειδικά σε κινήσεις όπως η bend, το Foreground Sum Signal εμφανίζει μια υπερβολικά υψηλή πυκνότητα, παρόλο που τελικά υπάρχουν και άλλα σήματα με μεγαλύτερη πυκνότητα για την ίδια κίνηση. Εντύπωση προκαλεί επίσης το πολύ χαμηλό ποσοστό του σήματος *ELt* στην ίδια κίνηση, το οποίο μάλιστα ξεφεύγει πολύ από τον μέσο όρο των υπόλοιπων σημάτων. Θεωρούμε πως αυτή η διαφορά οφείλεται στο γεγονός ότι το σήμα *ELt* σχετίζεται πολύ πιο άμεσα με την κινητική ενέργεια του κινούμενου ανθρώπου και επομένως είναι σε θέση να εκφράζει καλύτερα την κίνηση που παρακολουθεί σε περισσότερες περιπτώσεις.

6.8.4 Πειραματική επιβεβαίωση των συμπερασμάτων μας

Για να ενισχύσουμε την άποψη μας ότι όντως το Foreground Sum Signal είναι καλύτερο στο walk ενώ το *ELt* καλύτερο στις άλλες κινήσεις, εφαρμόσαμε μια μέθοδο ανίχνευσης FAI για καθένα από τα δύο σήματα. Τα αποτελέσματα που προέκυψαν φαίνονται στον πίνακα 6.6.

Πίνακας 6.6 Απόδοση των σημάτων Foreground Sum Signal και ELt ως προς την ανίχνευση FAI χρησιμοποιώντας την μέθοδο ανίχνευσης του [28].

Κίνηση	***Foreground Sum Signal***	***Σήμα ELt***
walk	100.0 %	72.3 %
jump	45.5 %	100.0 %
run	100.0 %	94.4 %
pjump	4.5 %	89.4 %
bend	16.7 %	88.9 %
jack	50.0 %	100.0 %
side	37.0 %	44.4 %
skip	21.9 %	88.9 %
wave1	24.4 %	100.0 %
wave2	16.3 %	100.0 %
Μέση απόδοση (%)	***40%***	***89%***

Παρατηρούμε ότι η μέση απόδοση με το Foreground Sum Signal ήταν 40.15% ενώ με το *ELt* 89%. Αυτό είναι κάτι που το περιμέναμε καθώς το Foreground Sum Signal έχει πολύ μεγαλύτερη πυκνότητα από το *ELt*.

Παρατηρούμε όμως ακόμη πως στο walk το Foreground Sum Signal έχει άριστη απόδοση (100%) ενώ, στην ίδια κίνηση, το *ELt* έχει την δεύτερη χειρότερη απόδοσή του. Αυτό το γεγονός ενισχύει την άποψή μας πως η απόδοση ενός σήματος εξαρτάται από την πυκνότητα των τοπικών ελαχίστων.

Στις υπόλοιπες κινήσεις παρατηρούμε πως το *ELt* αποδίδει πάντοτε καλύτερα, με εξαίρεση την κίνηση run, όπου όμως η διαφορά είναι πολύ μικρή.

6.9 Συμπεράσματα

Σε αυτό το κεφάλαιο δώσαμε τον ορισμό των FAI και δείξαμε πως υπάρχουν πάρα πολλά σήματα που μπορούν να παραχθούν από ένα video και τα οποία να ακολουθούν την κίνηση του κινούμενου ανθρώπου. Δείξαμε πως τα άκρα των FAI είναι τοπικά ελάχιστα αυτών των σημάτων και επιχειρήσαμε μια αξιολόγηση των σημάτων ως προς την πυκνότητα των τοπικών ελαχίστων, την οποία και συνδέσαμε με την απόδοση ενός σήματος, όταν χρησιμοποιείται για την ανίχνευση των άκρων των FAI.

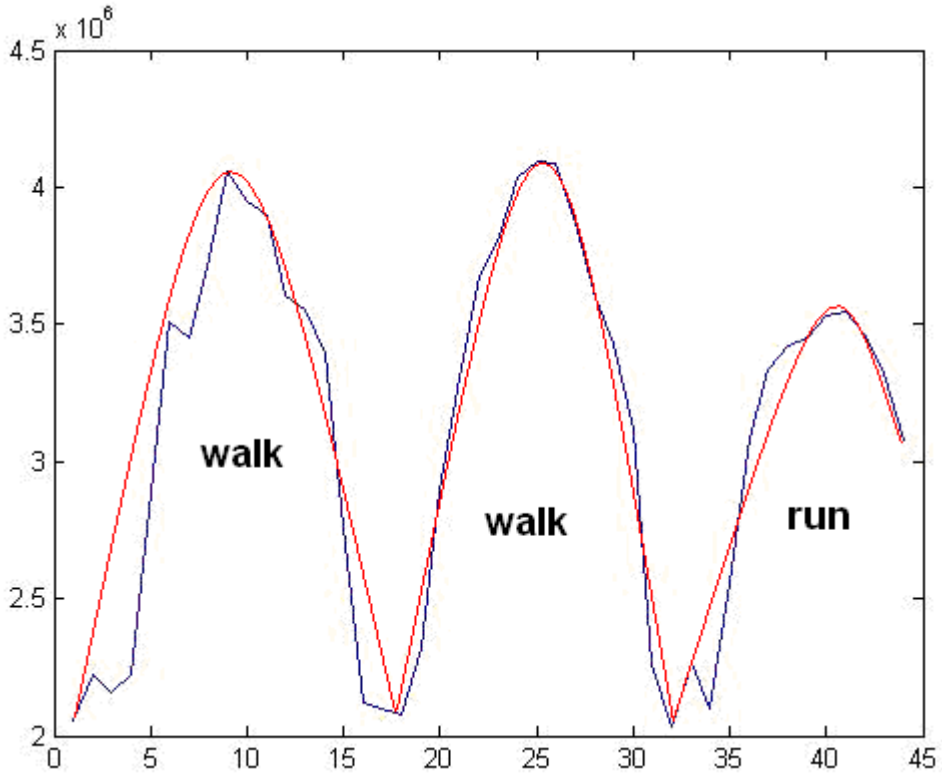
Τα αποτελέσματα που προέκυψαν είναι ιδιαίτερα χρήσιμα για την ανάλυση της κίνησης, καθώς επιτρέπουν την επιλογή του καταλληλότερου σήματος για ένα μεγάλο σύνολο κινήσεων (π.χ. για το walk επιλέγουμε το *Foreground Sum Signal*) ενώ αν δεν γνωρίζουμε το είδος της κίνησης χρησιμοποιούμε το σήμα *ELt*. Επομένως, ανάλογα με την εφαρμογή, η ανάλυση κίνησης μπορεί να πραγματοποιηθεί είτε πριν είτε μετά την αναγνώριση κίνησης.

Ωστόσο, υπάρχουν ακόμη πολλές ανεξερεύνητες ιδιότητες αυτών των σημάτων και πιστεύουμε πως η μελλοντική έρευνα θα πρέπει να δώσει περισσότερη προσοχή σε αυτό το πεδίο. Σίγουρα, η λεπτομερής ανάλυση της κίνησης θα μπορέσει να προσφέρει αναγνώριση κινήσεων σε εφαρμογές όπου απαιτείται η αναγνώριση *αρκετά όμοιων* κινήσεων *(π.χ. νοηματική γλώσσα, νεύματα)*.

Σελίδα σκόπιμα κενή.

Κεφάλαιο 7

"Αναγνώριση κίνησης"



Σελίδα σκόπιμα κενή.

7.1 Εισαγωγή

Η αναγνώριση κίνησης είναι ίσως το σημαντικότερο πρόβλημα που προσπαθεί να αντιμετωπίσει η Τεχνητή Όραση. Η επιτυχία αυτού του βήματος εξαρτάται από την επιτυχία των προηγούμενων βημάτων (ανίχνευση, εξαγωγή ιδιοτήτων) και επηρεάζει άμεσα τα επόμενα βήματα (ανάλυση κίνησης, εξαγωγή συμπερασμάτων, λήψη αποφάσεων).

Συνήθως, η αναγνώριση κίνησης αντιμετωπίζεται σαν ένα πρόβλημα κατηγοριοποίησης, με το ταίριασμα δηλαδή μιας άγνωστης ακολουθίας εικόνων με το μέλος μιας ομάδας προκατηγοριοποιημένων ακολουθιών που αναπαριστούν κάποιες τυπικές συμπεριφορές. Αυτή η προσέγγιση είναι αρκετά δημοφιλής, υπόκειται όμως στο βασικό μειονέκτημα κάθε κατηγοριοποίησης: δεν λαμβάνει υπόψη τις νέες συμπεριφορές / κατηγορίες κίνησης. Επομένως, είναι αρκετά αποδοτική για περιπτώσεις στις οποίες οι κινήσεις είναι μάλλον προβλεπόμενες και τα περιθώρια για εκδήλωση άγνωστης κίνησης είναι μικρά, αποδεικνύεται όμως ανεπαρκής για γενικευμένες περιπτώσεις.

Μια λύση σε αυτό το πρόβλημα είναι η Τεχνητή Μάθηση (machine learning), η *δια βίου* εκπαίδευση δηλαδή του συστήματος ώστε να είναι σε θέση να αναγνωρίζει και να *θυμάται* καινούργιες συμπεριφορές.

7.2 Οι κυριότερες προσπάθειες

Πολλοί ερευνητές ασχολήθηκαν με το πρόβλημα της αναγνώρισης κίνησης, συχνά με πολύ καλά αποτελέσματα. Η δημιουργία των benchmark Datasets KTH (2004) και Weizmann (2005) έδωσε νέα ώθηση καθώς πλέον οι διάφορες μέθοδοι μπορούν να συγκρίνονται μεταξύ τους στο ίδιο Dataset. Η πλειοψηφία των μεθόδων χρησιμοποιεί τεράστιες ποσότητες δεδομένων εκπαίδευσης και χρησιμοποιούν την τεχνική του διασταυρωτικού ελέγχου *(leave one out cross validation)*. Πλέον υπάρχουν μέθοδοι που καταφέρνουν να έχουν απόδοση ακόμα και στο 100% σε συγκεκριμένα όμως Datasets.

Στις πλέον κλασικές τεχνικές περιλαμβάνονται το Dynamic time warping [16, 17], οι μηχανές πεπερασμένων καταστάσεων *(FSMs)* [18], τα Hidden Markov Models *(HMMs)* [19, 20, 21], τα νευρωνικά δίκτυα χρονικής καθυστέρησης *(TDNNs)* [22,23], οι συντακτικές τεχνικές *(Syntactic techniques)* [24,25], τα μη ντετερμινιστικά

πεπερασμένα αυτόματα *(ΜΠΑ)* [26] και τα αυτοοργανωνώμενα νευρωνικά δίκτυα *(ΑΝΔ)* [27].

Από τις πιο σύγχρονες προσπάθειες, ξεχωρίζουν σίγουρα οι παρακάτω:

- Στο [15] εξάγεται η ανθρώπινη σιλουέτα σε κάθε frame και η κίνηση αναπαρίσταται σαν ένα σύνολο από χωρο-χρονικά σχήματα (space-time shapes). Η μέθοδος αυτή δοκιμάστηκε στο Weizmann Dataset με άριστα αποτελέσματα (97.83%).
- Στο [42] γίνεται μια προσπάθεια αναγνώρισης κίνησης από απόσταση, σε πολύ μικρές φιγούρες, χρησιμοποιώντας μια νέα περιγραφή κίνησης *(Motion Descriptor)*.
- Στο [7] χρησιμοποιείται η περιγραφή κίνησης του [4] για την εκπαίδευση ενός κατηγοριοποιητή (classifier) με χρήση AdaBoost.
- Στο [12] προτείνεται μια μέθοδος εμπνευσμένη από την βιολογική λειτουργία της όρασης χρησιμοποιώντας την απόκριση χωρο-χρονικών φίλτρων για ταίριασμα των κινήσεων *(spatio-temporal filter template matching)*.
- Στο [57] εξετάζεται η περίπτωση ύπαρξης μόνο ενός διαθέσιμου εκπαιδευτικού video για κάθε κίνηση. Για την κατηγοριοποίηση χρησιμοποιείται μια παραλλαγή των SVM, ενώ πραγματοποιούνται και πειράματα που συνδυάζουν δύο διαφορετικά Datasets (Weizmann, KTH) με αποτελέσματα εφάμιλλα με τις μεθόδους που χρησιμοποιούν τεράστια σύνολα εκπαίδευσης.
- Στο [39] παρουσιάζεται ένα σύστημα αποτελούμενο από τεχνικές ανίχνευσης ενεργών περιοχών (Activity Areas), σειριακή ανίχνευση αλλαγών και μια απλή μέθοδο κατηγοριοποίησης (σχήμα, ταχύτητα) που οδηγεί σε αρκετά καλά αποτελέσματα, χωρίς να απαιτεί ιδιαίτερη εκπαίδευση ή υψηλό υπολογιστικό κόστος.
- Στο [13] η απαραίτητη πληροφορία για αποτελεσματική αναγνώριση κίνησης συσχετίζεται το μήκος της ακολουθίας που εξετάζεται και αποδεικνύεται πειραματικά πως ακόμα και λίγα frames (1-10) μπορούν να οδηγήσουν σε άριστα αποτελέσματα (93.5 – 99.6% στο Weizmann Dataset). Αυτό το ζήτημα θα μας απασχολήσει σε όλο το υπόλοιπο κεφάλαιο.

7.3 Το ζήτημα της απαιτούμενης πληροφορίας

Βιολογικά, οι άνθρωποι αναγνωρίζουμε μια κίνηση σχεδόν άμεσα, χωρίς συνήθως να απαιτείται κάποια σημαντική χρονική καθυστέρηση. Προκύπτει επομένως το ερώτημα *"Πόση πληροφορία χρειάζεται για να μπορέσει ένας υπολογιστής να αναγνωρίσει μια κίνηση;"* .

Η απάντηση σε αυτό το ερώτημα θεωρούμε πως είναι ιδιαίτερα κρίσιμη για την σχεδίαση συστημάτων πραγματικού χρόνου καθώς το απαραίτητο πλήθος των frames καθορίζει την καθυστέρηση της έναρξης της διαδικασίας της αναγνώρισης.

Στο [13] η απαραίτητη πληροφορία για αποτελεσματική αναγνώριση κίνησης συσχετίζεται το μήκος της ακολουθίας που εξετάζεται. Αποδεικνύεται πειραματικά πως η συνδυασμένη πληροφορία τόσο του σχήματος *(Form)* όσο και την κίνησης *(Motion)* οδηγεί σε πολύ καλύτερα αποτελέσματα, ακόμα και όταν χρησιμοποιούνται λίγα *frames* (1-7).

Σε αυτήν την ενότητα θα εξετάσουμε το ερώτημα: "Επαρκεί η πληροφορία που περιέχεται σε 1 FAI για να αναγνωρίσουμε αποτελεσματικά την κίνηση;"

Βιολογικά, μια επανάληψη της κίνησης είναι συνήθως αρκετή για την αναγνώριση/κατανόηση, αν και αυτό δεν συμβαίνει πάντοτε: οι περισσότερες επαναλήψεις πιθανότατα ενισχύουν την αρχική μας γνώμη ως προς το ποια είναι τελικά η κίνηση που παρατηρήσαμε *(όπως συμβαίνει για παράδειγμα με τα replay videos των αμφισβητούμενων φάσεων στα αθλητικά παιχνίδια)*. Επομένως, σίγουρα δεν μπορούμε να περιμένουμε άριστα αποτελέσματα από ένα μόνο FAI ενώ αντίθετα αναμένουμε σημαντική βελτίωση στην αναγνώριση με την χρήση περισσότερων FAI.

Εμείς θα μελετήσουμε το ζήτημα μόνο σχετικά με τις non-Translational κινήσεις, χρησιμοποιώντας όμως μόνο το σχήμα, καθώς πιστεύουμε πως μπορεί να δώσει αρκετή πληροφορία στις non-Translational κινήσεις.

Χρησιμοποιούμε λοιπόν την εξής μέθοδο αναγνώρισης [39]:

1. Βρίσκουμε το Activity Area των frames και εξάγουμε το περίγραμμά του.
2. Υπολογίζουμε τους συντελεστές του Fourier Descriptor [40] του περιγράμματος.
3. Κρατάμε μόνο τους 20 πρώτους συντελεστές, κανονικοποιημένους ως προς το μέτρο του πρώτου συντελεστή.

Η χρήση του Fourier Descriptor για την περιγραφή του σχήματος μας εξασφαλίζει αμεταβλητότητα (invariance) ως προς την περιστροφή (rotation), την αλλαγή θέσης (translation) και την αλλαγή κλίμακας (scaling).

7.4 Πειραματικά αποτελέσματα

Χρησιμοποιήσαμε την μέθοδο των Activity Areas για να αναγνωρίσουμε non-Translational κινήσεις στο Weizmann και στο KTH Dataset. Χρησιμοποιήσαμε τόσο την χρήση όλης της ακολουθίας εικόνων όσο και ενός FAI για την δημιουργία των Activity Areas.

7.4.1 Αναγνώριση σε 1 FAI

Εδώ χρησιμοποιήσαμε σαν σύνολο εκπαίδευσης το Activity Area από το πρώτο FAI κάθε video. Στην συνέχεια, κάναμε αναγνώριση κίνησης σε κάθε FAI των video, χρησιμοποιώντας την μέθοδο του πλησιέστερου γείτονα (1-NN). Η διαδικασία επαναλήφθηκε για όλα τα video, φροντίζοντας να μην υπάρχει επικάλυψη μεταξύ των training και testing set.

Από το Weizmann Dataset χρησιμοποιήσαμε και τους 9 ανθρώπους ενώ από το KTH χρησιμοποιήσαμε μόνο τους 10 πρώτους ανθρώπους από το περιβάλλον *d1*.

Τα πειραματικά αποτελέσματα φαίνονται στους πίνακες 1 και 2.

Πίνακας 1. Πειραματικά αποτελέσματα αναγνώρισης κίνησης σε 1 FAI στο Weizmann Dataset. Η μέση απόδοση ήταν 66%.

	Pjump	***Bend***	***Jack***	***Wave1***	***Wave2***
Pjump	60.38	26.42	0	9.43	3.77
Bend	0	77.78	0	22.22	0
Jack	0	0	72.09	0	27.9
Wave1	0	16.67	0	63.89	19.44
Wave2	0	14.71	0	29.41	55.88

Πίνακας 2. Πειραματικά αποτελέσματα αναγνώρισης κίνησης σε 1 FAI στο KTH Dataset. Η μέση απόδοση ήταν 67.83%.

	Boxing	***HandClapping***	***HandWaving***
Boxing	48.48	51.52	0
HandClapping	20	80	0
HandWaving	15	10	75

7.4.2 Αναγνώριση σε όλη την ακολουθία εικόνων

Εδώ χρησιμοποιήσαμε σαν σύνολο εκπαίδευσης όλη την ακολουθία εικόνων από κάθε video. Στην συνέχεια, κάναμε αναγνώριση κίνησης σε κάθε video, χρησιμοποιώντας την μέθοδο του πλησιέστερου γείτονα (1-NN). Η διαδικασία επαναλήφθηκε για όλα τα video, φροντίζοντας να μην υπάρχει επικάλυψη μεταξύ των training και testing set.

Από το Weizmann Dataset χρησιμοποιήσαμε και τους 9 ανθρώπους ενώ από το KTH χρησιμοποιήσαμε τόσο τους 10 όσο και τους 23 πρώτους ανθρώπους από το περιβάλλον *d1*.

Τα πειραματικά αποτελέσματα φαίνονται στους πίνακες 3 και 4.

Πίνακας 3. Πειραματικά αποτελέσματα αναγνώρισης κίνησης σε ολόκληρο το video στο Weizmann Dataset. Η μέση απόδοση ήταν 77.78%.

	Pjump	***Bend***	***Jack***	***Wave1***	***Wave2***
Pjump	77.78	22.22	0	0	0
Bend	0	55.56	0	11.11	33.33
Jack	0	0	88.89	0	11.11
Wave1	0	0	0	88.89	11.11
Wave2	0	11.11	0	11.11	77.78

Πίνακας 4. Πειραματικά αποτελέσματα αναγνώρισης κίνησης σε ολόκληρο το video στο KTH Dataset. Η μέση απόδοση ήταν 60%.

	Boxing	***HandClapping***	***HandWaving***
Boxing	40	30	30
HandClapping	20	60	20
HandWaving	10	10	80

7.4.3 Ερμηνεία των αποτελεσμάτων

Παρατηρώντας τα αποτελέσματα των πινάκων 1 και 3, βλέπουμε πως η χρήση 1 FAI οδηγεί σε σχετικά μέτρια αποτελέσματα (66%), αλλά όχι σε χαμηλά, γεγονός που δείχνει πως 1 FAI περιέχει αρκετή πληροφορία για την κίνηση. Επιπλέον, η διαφορά στην απόδοση δεν είναι τόσο μεγάλη (περίπου 12%).

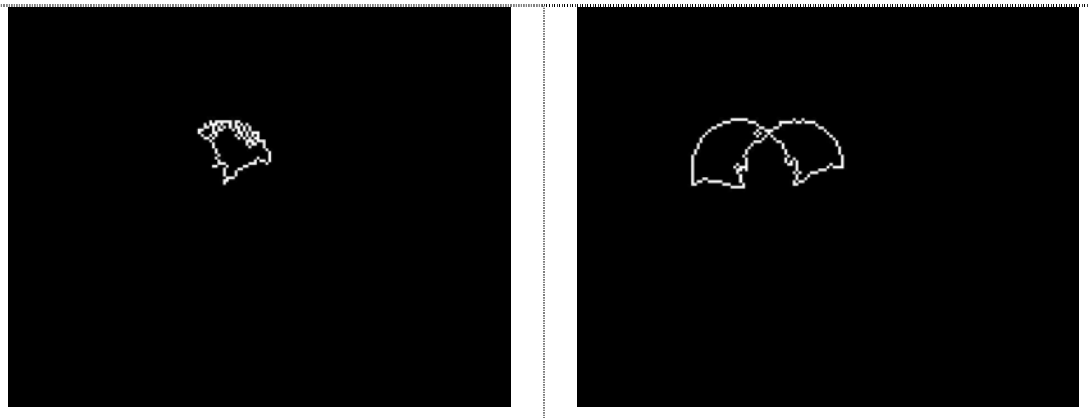
Βλέπουμε ακόμα πως η χρήση περισσότερων frames οδηγεί σε καλύτερα αποτελέσματα (77.78%). Το ίσως χαμηλό αυτό ποσοστό δεν μας απασχολεί εδώ καθώς οι κινήσεις που εξετάζουμε έχουν πολύ χαρακτηριστικό Activity Area (και MEI φυσικά) ώστε να γίνει αποτελεσματικά η κατηγοριοποίηση της κίνησης. Λόγω του ότι τα Activity Areas των κινήσεων μοιάζουν πολύ μεταξύ τους, ο Fourier Descriptor δεν μπορεί να τα ξεχωρίσει αποτελεσματικά και επομένως δεν μπορεί να γίνει καλή αναγνώριση της κίνησης.

Το γεγονός ωστόσο πως η χρήση περισσότερων frames οδηγεί σε καλύτερη απόδοση σε όλες τις κινήσεις εκτός από την κίνηση Bend, δεν μας επιτρέπει να συσχετίσουμε με ασφάλεια την απαιτούμενη πληροφορία με το πλήθος των frames.

Επιπλέον, καθώς χρησιμοποιούμε στατιστικές μεθόδους σε διακριτά δεδομένα, η χρήση περισσότερων δειγμάτων οδηγεί συνήθως σε καλύτερη συμπεριφορά.

Επίσης, στο Weizmann σπάνια έχουμε έντονες κινήσεις στο σώμα, εκτός από την κυρίως κίνηση. Επομένως, η χρήση περισσότερων frames ενισχύει την κυρίως κίνηση και εξασθενεί την *θορυβώδη* κίνηση (π.χ. κίνηση σώματος), παράγοντας έτσι καλύτερο Activity Area.

Στο σχήμα 1 βλέπουμε πως συχνά 1 FAI δεν επαρκεί για να *συλληφθεί* από το Activity Area η πλήρης κίνηση αλλά μόνο ένα τμήμα της κίνησης.



(α) *(β)*

Σχήμα 1. Activity Area από το ίδιο video του Weizman Dataset (denis wave2).
(α) 1 FAI (β) Πλήρης ακολουθία

Συμπερασματικά λοιπόν, μπορούμε να πούμε πως ένα FAI είναι πιθανό να περιέχει όλη την πληροφορία που χρειαζόμαστε για να αναγνωρίσουμε την κίνηση, αλλά αν η κίνηση εκτελεστεί με διαφορετικό τρόπο σε ένα FAI, η αναγνώριση κίνησης (είτε με 1 FAI είτε με πλήρη ακολουθία) μπορεί να αλλοιωθεί καθώς θα έχουμε μαγάλες αλλαγές στο σχήμα του Activity Area.

Παρατηρώντας τα αποτελέσματα στο KTH Dataset, διαπιστώνουμε ότι εδώ συμβαίνει το αντίθετο από την περίπτωση του Weizmann, καθώς οι άνθρωποι στα video του KTH πραγματοποιούν συνεχώς έντονες κινήσεις στο σώμα παράλληλα με την κυρίως κίνηση που εκτελούν (π.χ. ο κορμός κινείται μαζί με τα χέρια σε κινήσεις όπως το boxing). Επομένως, σε 1 FAI, η *θορυβώδης* κίνηση είναι μεν σημαντική (γεγονός που δικαιολογεί την χαμηλή απόδοση 67.83%), ενισχύεται όμως περισσότερο με την χρήση περισσότερων frames, οδηγώντας έτσι σε χειρότερη συμπεριφορά (60%).

Στο σχήμα 2 βλέπουμε ακριβώς την ενίσχυση της κίνησης του κορμού όταν αυξάνεται το πλήθος των frames.

(α)

(β)

Σχήμα 2. Activity Area από το ίδιο video του KTH Dataset (person07 boxing). (α) 1 FAI (β) Πλήρης ακολουθία

Βλέπουμε λοιπόν πως η χρήση λίγων ή περισσότερων frames άλλοτε ενισχύει την κυρίως κίνηση και άλλοτε τον θόρυβο, γεγονός που δεν μας επιτρέπει να διατυπώσουμε κάποιον γενικό κανόνα.

7.4.4 Σύνδεση με προηγούμενη εργασία

Στο [28] πραγματοποιήσαμε μια παρόμοια έρευνα και ελέγξαμε δύο μεθόδους (Εμπειρικές κατανομές [58] και Activity Areas) στο Weizmann Dataset. Ωστόσο,

ελέγξαμε και τις 10 κινήσεις χωρίς να κάνουμε διαχωρισμό σε Translational και non-Translational. Τα αποτελέσματα φαίνονται στους πίνακες 5 και 6.

Παρατηρούμε αρχικά ότι ο διαχωρισμός σε Translational και non- Translational κινήσεις οδηγεί όντως σημαντική βελτίωση της ίδιας μεθόδου (από 48.2% σε 66%), όπως επισημάναμε στο κεφάλαιο 4.

Πίνακας 5. Πειραματικά αποτελέσματα αναγνώρισης κίνησης σε ολόκληρο το video στο Weizmann Dataset με χρήση της μεθόδου των εμπειρικών κατανομών [58].

Κίνηση	Ποσοστό επιτυχίας (%)
Walk	37.5
Jump	74.3
Run	72.37
Pjump	91.5
Bend	78.3
Jack	92.6
Side	57.7
Skip	51.5
Wave1	68.1
Wave2	75
Μέσο ποσοστό: 69.9%	

Πίνακας 6. Πειραματικά αποτελέσματα αναγνώρισης κίνησης σε ολόκληρο το video στο Weizmann Dataset με χρήση της μεθόδου των Activity Areas.

Κίνηση	Ποσοστό επιτυχίας (%)
Walk	22.9
Jump	42.9
Run	30.3
Pjump	35.6
Bend	77.3
Jack	64.8
Side	28.0
Skip	30.3
Wave1	76.6
Wave2	72.7
Μέσο ποσοστό: 48.2%	

Παρατηρούμε ακόμη πως η μέθοδος των εμπειρικών κατανομών οδηγεί σε αρκετά υψηλά ποσοστά αναγνώρισης στις non-Translational κινήσεις (μέσο ποσοστό 81.10%). Το αποτέλεσμα αυτό μας δείχνει πως και η κίνηση *(Motion)* περιέχει αρκετή πληροφορία ακόμα και για τις non-Translational κινήσεις όπου θα περιμέναμε να αποδίδει καλύτερα το σχήμα *(Form)*. Το γεγονός αυτό, σε συνδυασμό και με τα χαμηλά ποσοστά των Activity Areas στις Translational κινήσεις, ενισχύει το συμπέρασμα του [13] ότι ο συνδυασμός *Form/Shape* και *Motion* οδηγεί σε καλύτερη απόδοση.

7.5 Συμπεράσματα

Σε αυτό το κεφάλαιο προσπαθήσαμε να ερευνήσουμε το ζήτημα της ποσότητας πληροφορίας που απαιτείται για την αναγνώριση κίνησης. Πραγματοποιήσαμε πειράματα σε δύο Datasets, μόνο σε non-Translational κινήσεις, χρησιμοποιώντας μόνο το σχήμα για την αναγνώριση και εξετάσαμε την χρήση της πληροφορίας που περιέχεται σε 1 FAI. Συμπεράναμε ότι η χρήση μόνο του σχήματος ή μόνο της πληροφορίας κίνησης δεν αρκεί για να επιτευχθούν υψηλά ποσοστά αναγνώρισης, είτε μέσα σε ένα FAI είτε σε όλο το βίντεο. Τα FAI παρόλα αυτά θα μπορούσαν να χρησιμοποιηθούν σε μέθοδο που συνδυάζει το σχήμα και την κίνηση, όπως το [13], για να καθορίσουν το πλήθος των frames που θα χρησιμοποιηθούν για την αναγνώριση. Η συνεισφορά τους σε μια τέτοια περίπτωση θα είναι σίγουρα σημαντική καθώς τα αποτελέσματά θα χρησιμεύσουν στον καλύτερο σχεδιασμό συστημάτων πραγματικού χρόνου.

Σελίδα σκόπιμα κενή.

Συνολικά Συμπεράσματα

Σε αυτήν την μεταπτυχιακή διπλωματική εργασία ασχοληθήκαμε με το ζήτημα της ανίχνευσης, αναγνώρισης και ανάλυσης κίνησης σε ακολουθίες εικόνων *(video)*.

Αρχικά κάναμε μια γενική εισαγωγή στο αντικείμενο της Τεχνητής Όρασης και παρουσιάζουμε κάποιες εφαρμογές και ανοιχτά ζητήματα, τις διάφορες μεθόδους επεξεργασίας εικόνας και video, καθώς και την δομή ενός *απλού* συστήματος Τεχνητής Όρασης. Επιμείναμε ιδιαίτερα στην σημασία της κατάλληλης προ-επεξεργασίας των δεδομένων για την βελτίωση των μεθόδων αναγνώρισης κίνησης.

Στο τρίτο κεφάλαιο ασχοληθήκαμε με το ζήτημα της ανίχνευσης κίνησης και εισάγαμε την έννοια των ενεργών frames, όπου δηλαδή παρατηρείται κάποια *σημαντική* κίνηση. Προτείναμε τρεις μεθόδους και δείξαμε πειραματικά πως η μέθοδος της ενέργειας μπορεί *με εκπαίδευση* να επιτύχει άριστη απόδοση σε video με ελάχιστο θόρυβο. Για την κατάλληλη αντιμετώπιση του θορύβου προτείναμε την μέθοδο της κύρτωσης, η οποία είχε αρκετά καλή συμπεριφορά *(90.5%)* χωρίς να απαιτεί κάποια φάση εκπαίδευσης.

Στο τέταρτο κεφάλαιο ασχοληθήκαμε με την εξαγωγή δύο βασικών ιδιοτήτων κίνησης. Προτείναμε την μέθοδο του κέντρου βάρους και δείξαμε πως έχει άριστη συμπεριφορά ακόμα και με ελάχιστα διαθέσιμα frames (15). Επίσης, περιγράψαμε πως η μέθοδος αυτή μπορεί να αντιμετωπίσει τα προβλήματα μνήμης των Activity History Areas.

Στο πέμπτο κεφάλαιο ασχοληθήκαμε με την σειριακή ανίχνευση αλλαγών με την χρήση της δημοφιλούς μεθόδου CUSUM. Τροποποιήσαμε την μέθοδο και προτείναμε μια μέθοδο προσέγγισης σήματος με ευθύγραμμα τμήματα για την μείωση του χώρου αναζήτησης. Αξιολογήσαμε πειραματικά την μέθοδο για Translational κινήσεις και διαπιστώνουμε μια άριστη συμπεριφορά. Δυστυχώς, η μέθοδος αυτή δεν αποδίδει για no Translational κινήσεις, σκιαγραφούμε όμως μια πιθανή λύση, εντάσσοντάς την στα άμεσα μελλοντικά μας σχέδια.

Στο έκτο κεφάλαιο ασχοληθήκαμε με την λεπτομερή ανάλυση της κίνησης. Διευρύναμε τον ορισμό του στιγμιοτύπου πλήρους κίνησης *(FAI)* και των σημάτων που ακολουθούν την κίνηση και αξιολογήσαμε 23 σήματα ως προς την δυσκολία ανίχνευσης FAI, καταλήγοντας έτσι στο καταλληλότερο σήμα για κάθε είδος κίνησης, τόσο πριν όσο και μετά την αναγνώριση κίνησης.

Στο έβδομο κεφάλαιο ασχοληθήκαμε με την αναγνώριση κίνησης. Παρουσιάσαμε τις πιο πρόσφατες προσπάθειες και διατυπώνουμε κάποια συμπεράσματα σχετικά με την ποσότητα πληροφορίας που περιέχεται σε ένα FAI.

Στόχοι μελλοντικής εργασίας

Σίγουρα υπάρχουν πολλές ιδέες για μελλοντική εργασία, τις οποίες θα θέλαμε να ελέγξουμε στα πλαίσια κάποιας μελλοντικής εργασίας μας.

Όσον αφορά την ανίχνευση κίνησης, η διερεύνηση της χρήσης των έντονων αλλαγών στα σημεία αλλαγής κίνησης που παρουσιάζει η μέθοδος της κύρτωσης και η βελτίωση των αποτελεσμάτων της μεθόδου με χρήση διάφορων τεχνικών (ομαδοποίηση γειτονικών frames, φιλτράρισμα), αλλά και η εκτενέστερη πειραματική αξιολόγηση των μεθόδων, σίγουρα θα οδηγήσουν σε αρκετά χρήσιμα συμπεράσματα.

Όσον αφορά την εξαγωγή ιδιοτήτων κίνησης, αναφέρουμε ενδεικτικά την πειραματική αξιολόγηση των μεθόδων σε πιο απαιτητικά περιβάλλοντα *(όπως στο KTH Dataset)*, την βελτίωση της μεθόδου των Activity Areas ώστε να μπορεί να χρησιμοποιείται και σε λίγα διαθέσιμα frames, και την πειραματική αξιολόγηση της εύρεσης αλλαγών στην κατεύθυνση με σκοπό την παραγωγή πιο ποιοτικών Activity History Areas.

Στον τομέα της ανίχνευσης αλλαγών κίνησης, θα είχε σίγουρα ενδιαφέρον η διερεύνηση της σειριακής ανίχνευσης αλλαγών σε non-Translational κινήσεις αλλά και η πειραματική αξιολόγηση άλλων μεθόδων εκτός της CUSUM.

Τα σήματα που ακολουθούν την κίνηση αποτελούν σίγουρα ένα συναρπαστικό αντικείμενο έρευνας, στο οποίο θα πρέπει σίγουρα να δοθεί περισσότερη προσοχή. Σίγουρα υπάρχουν ακόμη πολλές ανεξερεύνητες ιδιότητες αυτών των σημάτων οι οποίες θα μας οδηγήσουν σε μια αρκετά λεπτομερή ανάλυση της κίνησης. Ενδιαφέρον θα είχε επίσης και η πειραματική αξιολόγηση σε απαιτητικά περιβάλλοντα *(zoom, φωτισμός, κτλ).*

Τέλος, όσον αφορά την αναγνώριση της κίνησης, η διερεύνηση του ζητήματος της ελάχιστης απαιτούμενης πληροφορίας θα μας οδηγήσει στον καλύτερο σχεδιασμό συστημάτων πραγματικού χρόνου. Σίγουρα αξίζει τον κόπο η αξιολόγηση της χρήσης των FAI σε μεθόδους που συνδυάζουν το σχήμα και την κίνηση.

Βιβλιογραφία

[1] Wei Wei and An Yunxiao, **"Vision-based human motion recognition: A Survey,"** *2009 Second International Conference on Intelligent Networks and Intelligent Systems*

[2] Pavan Turaga, Rama Chellappa, V. S. Subrahmanian, and Octavian Udrea, **"Machine Recognition of Human Activities: A Survey,"** *IEEE Transactions On Circuits And Systems For Video Technology, Vol. 18, No. 11, November 2008*

[3] B. D. Lucas and T. Kanade, **"An iterative image registration technique with an application to stereo vision,"** *Proceedings of the 1981 DARPA Imaging Understanding Workshop (pp. 121–130), 1981.*

[4] Aaron F. Bobick and James W. Davis, **"Real-time Recognition of Activity Using Temporal Templates,"** *M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 386, 1996.*

[5] Alexia Briassouli, Vasileios Mezaris, and Ioannis Kompatsiaris, **"Combination of Accumulated Motion and Color Segmentation for Human Activity Analysis,"** *EURASIP Journal on Image and Video Processing, vol. 2008, Article ID 735141, 20 pages, 2008.*

[6] G.B. Giannakis and M. K. Tsatsanis, **"Time-domain tests for gaussianity and time-reversibility,"** *IEEE Transactions on Signal Processing, vol. 42, no. 12, pp. 3460 – 3472, Dec. 1994.*

[7] M. Hassouni, H. Cherifi, and D. Aboutajdine, **"Hos-based image sequence noise removal,"** *IEEE Transactions on Image Processing, vol. 15, no. 3, pp. 572–581, 2006.*

[8] D. Lelescu and D. Schonfeld, **"Statistical sequential analysis for real-time video scene change detection on compressed multimedia bitstream,"** *IEEE Transactions on Image Processing, vol. 5, no. 1, pp. 106–117, 2003.*

[9] R. K. Bansal and P. Papantoni-Kazakos, **"An algorithm for detecting a change in a stochastic process,"** *IEEE Transactions on Information Theory, vol. 32, no. 2, pp. 227–235, 1986.*

[10] G. V. Moustakides, **"Optimal stopping times for detecting changes in distributions,"** *Ann. Statist., vol. 14, p. 13791387, 1986.*

[11] C. Schuldt, L. Laptev, and B. Caputo, **"Recognizing human actions: a local SVM approach,"** *In ICPR, 2004*

[12] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, **"A biologically inspired system for action recognition,"** *In ICCV, 2007.*

[13] K. Schindler and L. Van Gool, **"Action snippets: How many frames does action recognition require?,"** *In CVPR, 2008.*

[14] A. Fathi and G. Mori, **"Action recognition by learning mid-level motion features,"** *In CVPR, 2008.*

[15] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, **"Actions as space-time shapes,"** *In ICCV, 2005.*

[16] K. Takahashi, S. Seki, H.Kojima, and R. Oka, **"Recognition of dexterous manipulations from time varying images,"** *in Proc. IEEEWorkshop Motion of Non-Rigid and Articulated Objects, Austin, TX, 1994, pp. 23–28.*

[17] A. F. Bobick and A. D.Wilson, **"A state-based technique to the representation and recognition of gesture,"** *IEEE Trans. Pattern Anal. Machine Intell., vol. 19, pp. 1325–1337, Dec. 1997.*

[18] A. D. Wilson, A. F. Bobick, and J. Cassell, **"Temporal classification of natural gesture and application to video coding,"** *in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 1997, pp. 948–954.*

[19] M. Brand, N. Oliver, and A. Pentland, **"Coupled hidden Markov models for complex action recognition,"** *in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 1997, pp. 994–999.*

[20] T. Starner, J. Weaver, and A. Pentland, **"Real-time American sign language recognition using desk and wearable computer-based video,"** *IEEE Trans. Pattern Anal. Machine Intell., vol. 20, pp. 1371–1375, Dec. 1998.*

[21] N. M. Oliver, B. Rosario, and A. P. Pentland, **"A Bayesian computer vision system for modeling human interactions,"** *IEEE Trans. Pattern Anal. Machine Intell., vol. 22, pp. 831–843, Aug. 2000.*

[22] M. Yang and N. Ahuja, **"Extraction and classification of visual motion pattern recognition,"** *in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 1998, pp. 892–897.*

[23] U. Meier, R. Stiefelhagen, J. Yang, and A.Waibel, **"Toward unrestricted lip reading,"** *Int. J. Pattern Recognit. Artificial Intell., vol. 14, no. 5, pp. 571–585, Aug 2000.*

[24] Y. A. Ivanov and A. F. Boblic, **“Recognition of visual activities and interactions by stochastic parsing,”** *IEEE Trans. Pattern Anal. Machine Intell., vol. 22, pp. 852–872, Aug. 2000.*

[25] M. Brand, **“Understanding manipulation in video,”** *in Proc. Int. Conf. Automatic Face and Gesture Recognition, 1996, pp. 94–99.*

[26] T.Wada and T. Matsuyama, **“Multi-object behavior recognition by event driven selective attention method,”** *IEEE Trans. Pattern Anal. Machine Intell., vol. 22, pp. 873–887, Aug. 2000.*

[27] N. Johnson and D. Hogg, **“Learning the distribution of object trajectories for event recognition,”** *Image Vis. Comput., vol. 14, no. 8, pp. 609–615, 1996.*

[28] Στέργιος Πουλαράκης, **"Στατιστική επεξεργασία video για χαρακτηρισμό και ανίχνευση ανθρώπινης κίνησης,"** *(Διπλωματική εργασία), Τμήμα Μηχανικών Ηλεκτρονικών Υπολογιστών, Τηλεπικοινωνιών & Δικτύων, Πολυτεχνική Σχολή, Πανεπιστήμιο Θεσσαλίας, Ιούλιος 2008*

[29] Stergios Poularakis, Alexia Briassouli, Ioannis Kompatsiaris, **"Full Action Instances for Motion Analysis,"** *10th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2009), May 6-8, London, UK, pp 37-40.*

[30] Jianyi Liu, Nanning Zheng, **“Partitioning Gait Cycles Adaptive to Fluctuating Periods and Bad Silhouettes,”** *ICB 2007: 347-355*

[31] Sundaresan, A., Roy Chowdhury, A.K., Chellappa, R., **“A hidden Markov model based framework for recognition of humans from gait sequences,”** *In: Proc. Int. Conf. Image Processing, vol. 2, pp. 14–17 (2003)*

[32] E. S. Page, **“Continuous inspection scheme,”** *Biometrika, vol. 41, pp. 100–115, 1954.*

[33] J. Zhou and X. Zhang, **“Video event detection using ICA Mixture Hidden Markov Models,”** *in Image Processing, 2006 IEEE International Conference on, Oct. 2006, pp. 3005–3008.*

[34] L. Xie, D. Xu, S. Ebadollahi, K.Scheinberg, S. Chang, and J. Smith, **“Detecting generic visual events with temporal cues,”** *in Signals, Systems and Computers, 2006. ACSSC ’06. Fortieth Asilomar Conference on, 2006, pp. 54–58.*

[35] N. Real, R. Dahyot, and A. Kokaram, **“Semantic event detection in sports through motion understanding,”** *in CIVR 2004: International conference on image and video retrieval, 2004, pp. 88–97.*

[36] R. Leonardi, P. Migliorati, and M. Prandini, **“Semantic indexing of soccer audio-visual sequences: A multimodal approach based on controlled markov chains,”** *IEEE Transactions on Circuits and Systems for Video Technology, vol. 14, no. 5, pp. 634–643, 2004.*

[37] D. Ajay, R. Radhakrishan, and K. Peker, **“Video summarization using descriptors of motion activity: a motion activity based approach to key-frame extraction from video shots,”** *J. Electronic Imaging, vol. 10, no. 4, pp. 909–916, 2001.*

[38] D. C. B. Han and L. Davis, **“Sequential kernel density approximation through mode propagation: applications to background modeling,”** *in Proc. IEEE Conf. Computer Vision Patt. Recog., June 2003, pp. 65–72.*

[39] Alexia Briassouli , Vagia Tsiminaki , Ioannis Kompatsiaris, **“Human motion analysis via statistical motion processing and sequential change detection,”** *Journal on Image and Video Processing, 2009, p.1-1, January 2009*

[40] M. Bober, **“MPEG-7 visual shape descriptors,”** *IEEE Transactions on Circuits and Systems for Video Technology, vol. 11, no. 6, pp. 716–719, 2001.*

[41] D. S. Zhang and G. Lu, **“A comparative study of Fourier descriptors for shape representation and retrieval,”** *in Proc. Of the Fifth Asian Conference on Computer Vision (ACCV02), Jan. 2002, pp. 646–651.*

[42] A. A. Efros, A. C. Berg, G. Mori, and J. Malik., **“Recognizing action at a distance,”** *In ICCV, 2003.*

[43] Donovan H. Parks, Sidney S. Fels, **"Evaluation of Background Subtraction Algorithms with Post-Processing,"** *avss, pp.192-199, 2008 IEEE Fifth International Conference on Advanced Video and Signal Based Surveillance, 2008*

[44] D. Toth and T. Aach., “**Detection and recognition of moving objects using statistical motion detection and fourier descriptors,”** *International Conference on Image Analysis and Processing, pages 430–435, 2003.*

[45] Lijun Jiang, Feng Tian, Lim Ee Shen, Shiqian Wu, Susu Yao, Zhongkang Lu, and Lijun Xu. **“Perceptual-based fusion of ir and visual images for human**

detection," *International Symposium on Intelligent Multimedia, Video and Speech Processing, pages 514– 517, 2004.*

[46] R. Cutler and L. S. Davis., **"Robust real-time periodic motion detection, analysis, and applications,"** *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8):781–796, 2000.*

[47] P. Viola, M. J. Jones, and D. Snow, "**Detecting pedestrians using patterns of motion and appearance,"** *IEEE International Conference on Computer Vision, 2:734–741, 2003.*

[48] Navneet Dalal and Bill Triggs, **"Histograms of oriented gradients for human detection,"** *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 1063–6919, 2005.*

[49] N. Ogale, **"A Survey of Techniques for Human Detection from Video,"** *Master's thesis, University of Maryland, 2006.*

[50] W. Zhao, R. Chellappa, A. Rosenfeld, P.J. Phillips, **"Face Recognition: A Literature Survey,**" *ACM Computing Surveys, 2003, pp. 399-458*

[51] Ling-Feng Liu, Wei Jia, Yi-Hai Zhu: Survey of Gait Recognition. ICIC (2) 2009: 652-659

[52] T. Starner, J. Weaver, and A. Pentland, **"Real-time American sign language recognition using desk and wearable computer-based video,"** *IEEE Trans. Pattern Anal. Machine Intell., vol. 20, pp. 1371–1375, Dec. 1998.*

[53] Y. Rui, T. S. Huang, and S. F. Chang, **"Image retrieval: Current techniques, promising directions and open issues,"** *J. Vis. Commun. Image Represent., vol. 10, no. 4, pp. 39–62, 1999.*

[54] S. F. Chang, **"The Holy Grail of content-based media analysis,"** *IEEE Multimedia Mag., vol. 9, no. 2, pp. 6–10, Apr. 2002.*

[55] H. Zhong, J. Shi, and M. Visontai, **"Detecting unusual activity in video,"** *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2004, pp. 819–826.*

[56] C. Stauffer and W. E. L. Grimson, **"Learning patterns of activity using real-time tracking,"** *IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 8, pp. 747–757, Aug. 2000.*

[57] Weilong Yang, Yang Wang, and Greg Mori, **"Human Action Recognition from a Single Clip per Action,"** *2nd International Workshop on Machine Learning for Vision-based Motion Analysis (at ICCV), 2009.*

[58] Lihi Zelnik-Manor and Michal Irani, **"Event-Based Analysis of Video,"** *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of*

the 2001 IEEE Computer Society Conference on Volume: 2, On page(s): II-123- II-130 vol.2

[59] S. Muthukrishnan, E. van den Berg, and Y. Wu., **"Sequential change detection on data streams,"** *In ICDM Workshopon Data Stream Mining and Management, Omaha NE, Oct. 2007.*