ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ
ΕΡΓΑΣΤΗΡΙΟ ΕΠΕΞΕΡΓΑΣΙΑΣ ΕΙΚΟΝΑΣ ΒΙΝΤΕΟ ΚΑΙ ΠΟΛΥΜΕΣΩΝ

Βάση Δεδομένων Αθλητικών Δράσεων Με Πληροφορία Εικόνας και Βάθους

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

Σοφίας Ν. Γούργαρη

Επιβλέπων : Στέφανος Κόλλιας
Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2013

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ
ΕΡΓΑΣΤΗΡΙΟ ΕΠΕΞΕΡΓΑΣΙΑΣ ΕΙΚΟΝΑΣ ΒΙΝΤΕΟ ΚΑΙ ΠΟΛΥΜΕΣΩΝ

Βάση Δεδομένων Αθλητικών Δράσεων Με Πληροφορία Εικόνας και Βάθους

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

Σοφίας Ν. Γούργαρη

Επιβλέπων : Στέφανος Κόλλιας
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 20η Μαρτίου 2013.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Στέφανος Κόλλιας Καθηγητής Ε.Μ.Π.	Ανδρέας-Γεώργιος Σταφυλοπάτης Καθηγητής Ε.Μ.Π.	Γεώργιος Στάμου Επίκουρος Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2013

.....
Σοφία Ν. Γούργαρη
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Ευχαριστίες

Κατ'αρχήν, θα ήθελα να ευχαριστήσω τον Δρ Κωσταντίνο Καρπούζη για την εμπιστοσύνη που μου έδειξε με την ανάθεση της παρούσας διπλωματικής εργασίας και για την καθοδήγηση του καθ'όλη τη διάρκεια της εκπόνησής της. Εν συνεχεία, θα ήθελα να εκφράσω την ευγνωμοσύνη μου απέναντι στον υποψήφιο Δρ Γεώργιο Γουδέλη για την αμέριστη συμπαράστασή του και τις πολύτιμες συμβουλές και ιδέες του ως προς την υλοποίηση της διπλωματικής. Επίσης, οφείλω να πω πως είμαι υπόχρεη στην Δρ Αναστασία Τσουρουφλή καθηγήτρια φυσικής αγωγής ΕΜΠ για την πολύτιμη συνδρομή της στη δημιουργία της βάσης, καθώς και στον Αναστάσιο Βενέτη και σε όλους όσους συμμετείχαν στις βιντεοσκοπήσεις για τη συλλογή του απαραίτητου υλικού για την δημιουργία της βάσης. Ακόμη, θα ήθελα να ευχαριστήσω τον Σταύρο Αποστόλου για την πολύτιμη βοήθεια που μου προσέφερε, καθώς επίσης και όλα τα μέλη του εργαστηρίου Ψηφιακής Επεξεργασίας Εικόνας, Βίντεο και Πολυμέσων που με βοήθησαν σε κάθε δυσκολία που αντιμετώπισα. Τέλος, θα ήθελα να ευχαριστήσω τους γονείς μου για την αμέριστη συμπαράστασή τους κάθε στιγμή όλα αυτά τα χρόνια.

Σύνοψη

Η Ανίχνευση και Κατηγοριοποίηση Ανθρωπίνων Κινήσεων, ως τομέας σύμπραξης της Όρασης Υπολογιστών και της Αναγνώρισης Προτύπων, χρησιμοποιείται σε ολοένα και περισσότερες εφαρμογές για την περιγραφή της ανθρώπινης δραστηριότητας, όπως η ανάλυση video με βάση το περιεχόμενο και η διαμόρφωση ευφυούς περιβάλλοντος με διαδραστικές εφαρμογές. Οι εφαρμογές αυτές απαιτούν αποδοτικές μεθόδους για την αυτόματη ανάλυση και ταξινόμηση των δεδομένων κίνησης, και αποτελούν ένα πολύ δραστήριο ερευνητικό πεδίο. Για την αντικειμενική αξιολόγηση και σύγκριση των διαφόρων μεθόδων ανάλυσης και κατηγοριοποίησης της ανθρώπινης δραστηριότητας, τα τελευταία χρόνια έχουν δημιουργηθεί συλλογές από video με καταγεγραμμένες κινήσεις.

Στόχος αυτής της Διπλωματικής εργασίας είναι να προσφέρει στην ερευνητική κοινότητα μια νέα βάση δεδομένων κίνησης που να μπορεί να χρησιμοποιηθεί στην αξιολόγηση και σύγκριση των διαφόρων μεθόδων ανάλυσης και κατηγοριοποίησης της ανθρώπινης δραστηριότητας. Η βάση μας περιλαμβάνει 8374 video, που περιέχουν 12 κινήσεις του αθλήματος της αντισφαίρισης εκτελεσμένων από 55 διαφορετικά άτομα καταγεγραμμένα με την κάμερα τρισδιάσταστης λήψης Kinect. Πιο συγκεκριμένα, περιλαμβάνει video που καταγράφουν την κάθε κίνηση στις τρεις διαστάσεις του χώρου, καθώς και την κίνηση του σκελετού του ανθρώπινου σώματος. Η συσκευή καταγραφής Kinect διαθέτει κάμερα υπερύθρων, επιτρέποντας έτσι, την εξαγωγή πληροφορίας σχετικά με το βάθος και τη θέση των αρθρώσεων του ανθρώπινου σώματος. Με αυτό τον τρόπο επιτυγχάνεται μια μοντελοποίηση του ανθρώπινου σκελετού σε τρεις διαστάσεις.

Στο πλαίσιο της εργασίας, εφαρμόζουμε δυο μεθόδους ανίχνευσης και ταξινόμησης των κινήσεων στα δεδομένα της βάσης μας. Ειδικότερα, εφαρμόζουμε τη μέθοδο «εντοπισμού σημείων ενδιαφέροντος στο χωροχρόνο» (Space-Time Interest Points) και τη μέθοδο «εντοπισμού πυκνών τροχιών κίνησης» (Dense Trajectories). Οι δύο μέθοδοι χρησιμοποιούν ως τοπικούς χωροχρονικούς περιγραφείς τα Ιστογράμματα Προσανατολισμένης Κλίσης (Histograms of Oriented Gradient-HOG), Ιστογράμματα Οπτικής Ροής (Histograms of Optical Flow - HOF) και Ιστογράμματα Ορίων Κίνησης (Motion Boundary Histograms). Η ταξινόμηση των video πραγματοποιείται με μια μηχανή διανυσμάτων υποστήριξης (Support Vector Machine- SVM) ως ταξινομητή πολλών κλάσεων. Τα αποτελέσματα της πειραματικής διαδικασίας δείχνουν ότι η βάση έχει δυναμική για τη χρησιμοποίηση της σε μελέτες για την ανάπτυξη εφαρμογών αναγνώρισης ανθρώπινων κινήσεων που παρουσιάζουν ιδιαίτερες προκλήσεις.

Λέξεις Κλειδιά

αναγνώριση ανθρωπίνων κινήσεων, kinect, ταξινόμηση video, αντισφαίριση, ιστόγραμμα προσανατολισμένης κλίσης, ιστόγραμμα οπτικής ροής, ιστόγραμμα ορίων κίνησης, μηχανή διανυσμάτων υποστήριξης, αλγόριθμος Κ- μέσων

Abstract

The detection and classification of human movements, as a joint field of Computer Vision and Pattern Recognition, is used with an increasing rate in applications designed to describe human activity, such content based video analysis interactive environments and applications such as smart rooms. Such applications require efficient methods and tools for the automatic analysis and classification of motion capture data, which constitute an active field of research. To facilitate the development and the benchmarking of methods for action recognition, several video collections have previously been proposed.

With this Diploma thesis we provide the research community with a new video database that can be used for an objective comparison and evaluation of different motion analysis and classification methods. Our database consists of 8374 video clips, which contain 12 different types of tennis actions performed by 55 individuals captured by the 3D motion capture device Kinect. To be more specific, the database contains video clips that capture the 3D motion of individuals. Kinect which is our motion capture device, is used as an infrared camera and provides us with the depth map of motion data and helps to extract the 3D skeletal joint connections from these depth maps. As a result, we achieve a 3D model of individuals' skeletal motion.

In the framework of this Diploma thesis, we apply two different methods of detection and action recognition, conducting experiments on our database. Particularly, we use the method of *Space-Time Interest Points* and the method *Dense Trajectories* for action recognition. These methods are based on the use of local spatio-temporal descriptors, such as Histograms of Oriented Gradient (HOG), Histograms of Optical Flow (HOF) and Motion Boundary Histograms (MBH). Moreover, we perform action classification on the video data, and the classification is carried out with a multiclass *support vector machine* (SVM) classifier. The accuracy rates attained with our experimental procedure show that this new action database could be used in research on human action recognition applications that introduce special challenges.

Keywords

Human action recognition, kinect, video classification, tennis, Histogram of Oriented Gradient, Histogram of Optical Flow, Motion Boundary Histograms, support vector machine, k-means algorithm.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

1.1 Συνεισφορά Διπλωματικής

Η αναγνώριση της ανθρώπινης δραστηριότητας από εικονοσειρές (video) αποτελεί ένα σημαντικό πρόβλημα του επιστημονικού πεδίου της όρασης υπολογιστών με εφαρμογές σε πολλούς τομείς, όπως η ανάκτηση video, η επιτήρηση δημοσίων χώρων και η αλληλεπίδραση ανθρώπου-υπολογιστή. Με την ευρεία εξάπλωση των μέσων επικοινωνίας υψηλής ταχύτητας, την εύκολη προσβασιμότητα σε μέσα αποθήκευσης μεγάλης χωρητικότητας και στον τεχνικό εξοπλισμό, πλέον το ψηφιακό video βρίσκεται στη διάθεση του καθενός. Κατά συνέπεια, η ζήτηση για αποτελεσματικές τεχνικές επεξεργασίας video και αναγνώρισης της δραστηριότητας αυξάνεται ραγδαία.

Τα χαρακτηριστικά που ορίζουν τι είναι η «δραστηριότητα», αρκετές φορές δεν είναι δυνατό να διατυπωθούν με σαφήνεια. Οι δραστηριότητες ή κινήσεις είναι σύνθετες οντότητες που διαφοροποιούνται ως προς τη διάρκεια· κάποιες είναι στιγμιαίες, ενώ άλλες έχουν παρατεταμένη διάρκεια. Επιπλέον, μπορεί να περιλαμβάνουν αλληλεπιδράσεις με άλλα άτομα ή αντικείμενα και να πραγματοποιούνται με τη συμμετοχή είτε όλου του σώματος, είτε κάποιων μελών του.

Αξίζει να σημειωθεί ότι υπάρχει σπουδαία βιβλιογραφία που αφορά στην αναγνώριση της ανθρώπινης δραστηριότητας από τον ηλεκτρονικό υπολογιστή, σε ποικίλα ερευνητικά πεδία, όπως η όραση υπολογιστών, η αναγνώριση προτύπων, η μηχανική μάθηση, ανάλυση σημάτων κ.α. Ανάμεσα στις διάφορες τεχνικές αναπαράστασης της πληροφορίας της κίνησης, οι πιο δημοφιλείς είναι τα σημεία ενδιαφέροντος χωροχρόνου (Space-Time Interest Points) και η χρήση πυκνών τροχιών κίνησης (Dense Trajectories).

Σε αυτό το σημείο είναι αναγκαίο να αναφέρουμε τρεις πολύ σημαντικές προκλήσεις στην αναγνώριση της ανθρώπινης δραστηριότητας. Κατ'αρχάς, σημαντικό πρόβλημα αποτελούν οι διαφοροποίησεις στην ίδια κλάση κινήσεων (intra – class) και η μεγάλη ομοιότητα διαφορετικών κλάσεων ανθρώπινης δραστηριότητας (inter –class). Απο τη μια πλευρά, τα άτομα μπορούν να εκτελέσουν την ίδια κίνηση με διαφορετικές κατευθύνσεις και διαφοροποιήσεις στην μετατόπιση των ανθρώπινων μελών, ενώ από την άλλη μπορεί να συμβεί, δυο κινήσεις που ανήκουν σε ξεχωριστές κλάσεις να διαχωρίζονται μόνο από δυσδιάκριτες λεπτομέρειες του χωροχρόνου. Δεύτερον, ο αριθμός των διαφορετικών κατηγοριών κίνησης είναι εξαιρετικά μεγάλος και τρίτον, οι διαφοροποιήσεις στις συνθήκες φωτισμού, στο παρασκήνιο (φόντο), στην οπτική γωνία λήψης, καθώς και η ύπαρξη σκιών μπορούν να αλλάξουν τον τρόπο με τον οποίο γίνεται αντιληπτή μια κίνηση.

Προκειμένου να διευκολυνθεί η ανάπτυξη μεθόδων για την αναγνώριση των κινήσεων, έχουν συγκεντρωθεί αρκετά σύνολα από video που περιέχουν διάφορα είδη κίνησης. Στην ενότητα 2.3 γίνεται εκτενής αναφορά στα πιο δημοφιλή σύνολα δεδομένων. Μέχρι τώρα η έρευνα έχει επικεντρωθεί στην αναγνώριση δραστηριότητας από video που έχουν καταγραφεί με κάμερες ορατού φωτός. Όμως, η χρήση συσκευών που καταγράφουν και το βάθος (range camera), αντιμετωπίζουν σε μεγάλο βαθμό τις προκλήσεις της τρίτης κατηγορίας που μειώνουν την αποδοτικότητα στην αναγνώριση της δραστηριότητας από απεικονίσεις 2D.

Φυσικά, μια κάμερα τέτοιου τύπου μπορεί να καταγράψει σημαντική πληροφορία για κινήσεις που περιέχουν αλλαγές στο βάθος. Παραδείγματος χάριν, ο διαχωρισμός της κίνησης «δείχνω» από την κίνηση «πιάνω» σε μια μπροστινή λήψη, θα ήταν πολύ πιο ακριβής από τις εικονοσειρές βάθους (depth map), παρά από εικονοσειρές RGB.

Παρ'όλα αυτά, οι παλαιότεροι αισθητήρες βάθους χαρακτηρίζονταν είτε από υψηλό κόστος είτε από χαμηλή ακρίβεια. Αντίθετα, η πρόσφατη κυκλοφορία της συσκευής Kinect της Microsoft αντιμετωπίζει αυτά τα ζητήματα και διαθέτει δυο κανάλια καταγραφής , κανάλι που καταγράφει βάθος και κανάλι που καταγράφει RGB εικόνα. Παρ'ότι πρωταρχικός στόχος της κυκλοφορίας του υπήρξε η αγορά της ψυχαγωγίας, το Kinect προκάλεσε το ενδιαφέρον της ερευνητικής κοινότητας της όρασης υπολογιστών, λόγω του εύρους των εφαρμογών που υποστηρίζει.

Για τους παραπάνω λόγους, αποφασίσαμε να προσφέρουμε στην ερευνητική κοινότητα ένα επιπλέον σύνολο δεδομένων από κινήσεις καταγεγραμμένες από τη συσκευή Kinect. Στόχος της νέας βάσης δεδομένων η οποία ονομάζεται THETIS (THree dimEnsional TennIs Shots) είναι να αποτελέσει ένα χρήσιμο εργαλείο αξιολόγησης για τους διάφορους αλγορίθμους αναγνώρισης κινήσεων, αλλά και γενικότερα να αποτελέσει εργαλείο για την ανάπτυξη εφαρμογών που σχετίζονται με την αναγνώριση της ανθρώπινης δραστηριότητας. Τα πλεονεκτήματα και οι δυνατότητες που προσφέρει η χρήση της συσκευής Kinect είναι πολλά και θα γίνει ιδιαίτερος λόγος για αυτά στην ενότητα 3.2. Ενδεικτικά, αναφέρουμε ότι η πληροφορία του βάθους που καταγράφεται από τη συσκευή Kinect, μας παρέχει όχι μόνο την αναπαράσταση της κίνησης στον τρισδιάστατο χώρο, αλλά και τη δυνατότητα εντοπισμού της θέσης των αρθρώσεων του ανθρώπινου σώματος και την αναπαράσταση της κίνησης του σκελετού στον χρόνο και στο χώρο 3D (XYZT). Με αυτή την αναπαράσταση προσεγγίζουμε το πρόβλημα της αναγνώρισης της ανθρώπινης δραστηριότητας με τρόπο λιγότερο σύνθετο σε σύγκριση με τη χρήση εικονοσειρών RGB. Επιπλέον, παρέχει ανεξαρτησία από τη γωνία λήψης και υψηλότερη ταχύτητα.

Στην παρούσα Διπλωματική εργασία πραγματοποιείται η παρουσίαση της βάσης δεδομένων THETIS, η οποία περιλαμβάνει 8374 video, που περιέχουν 12 κινήσεις του αθλήματος της αντισφαίρισης, εκτελεσμένες αρκετές φορές από 55 διαφορετικά άτομα. Επιπλέον, πραγματοποιείται η εφαρμογή δυο διαφορετικών μεθόδων, που αποτελούν τελευταία λέξη της τεχνολογίας για την εξαγωγή περιγραφέων των χαρακτηριστικών της κίνησης, στα video της βάσης THETIS και επιχειρείται η κατηγοριοποίηση των video με βάση το περιεχόμενο από μια μηχανή υποστήριξης ως ταξινομητή πολλών κλάσεων.

Συνοψίζονται τα κύρια θέματα που αποτελούν συνεισφορά αυτής της διπλωματικής εργασίας :

- Δημιουργία μεγάλης βάσης από video που καταγράφουν 12 κινήσεις της αντισφαίρισης, από 55 άτομα, εκτελεσμένες όλες αρκετές φορές από το κάθε άτομο. Για κάθε εκτέλεση παρέχονται 5 τύποι video:
    1. Video εικόνας RGB
    2. Video εικόνας βάθους (depth map)
    3. Video περιγράμματος/ σιλουέτας
    4. Video σκελετού 2D
    5. Video σκελετού 3D

- Εφαρμογή της μεθόδου σημείων ενδιαφέροντος στο χωροχρόνο [19] (Space-Time Interest Points) για την εξαγωγή των εξής τοπικών περιγραφέων : Ιστογράμματα Προσανατολισμένης Κλίσης (Histograms of Oriented Gradient-HOG) και Ιστογράμματα Οπτικής Ροής (Histograms of Optical Flow - HOF) στα video βάθους και σκελετού 3D, καθώς επίσης και στα δεδομένα του συνόλου KTH[53].

- Εφαρμογή της μεθόδου Dense Trajectories [73] για την εξαγωγή των εξής τοπικών περιγραφέων : Ιστογράμματα Προσανατολισμένης Κλίσης (Histograms of Oriented Gradient-HOG), Ιστογράμματα Οπτικής Ροής (Histograms of Optical Flow – HOF), Ιστογράμματα Ορίων Κίνησης (Motion Boundary Histograms-MBH) και εξαγωγή της τροχιάς της κίνησης στα video βάθους και σκελετού 3D, καθώς επίσης και στα δεδομένα του συνόλου KTH.

- Κβαντοποίηση των χαρακτηριστικών διανυσμάτων των video που προέκυψαν από τους περιγραφείς χρησιμοποιώντας τον αλγόριθμο συσταδοποίησης k-means και η δημιουργία ενός οπτικού λεξιλογίου (bag of features).

- Κατηγοριοποίηση των video βάθους και σκελετού 3D της βάσης THETIS και των video του συνόλου KTH. Για την ταξινόμηση χρησιμοποιήθηκε μια μηχανή μηχανή διανυσμάτων υποστήριξης (Support Vector Machine- SVM) ως ταξινομητή πολλών κλάσεων.

- Δημιουργία και χρήση συνοδευτικών προγραμμάτων για τη διεξαγωγή της πειραματικής διαδικασίας.

Όσον αφορά στην υλοποίηση, στο πλαίσιο της διπλωματικής εργασίας και για την κάλυψη των απαιτήσεων της πειραματικής διαδικασίας ασχολήθηκα με πληθώρα εφαρμογών και την ανάπτυξη κώδικα.

Αρχικά, πραγματοποιήθηκε η εξαγωγή του σκελετού 3D, του σκελετού 2D και της σιλουέτας από τα video βάθους και εικόνας RGB που καταγράφηκαν από το Kinect με χρήση έτοιμης υλοποίησης στο OpenNI Framework. Στη συνέχεια, τα video περικόπηκαν και προέκυψαν μικρότερα και περισσότερα στον αριθμό video, που το καθένα περιέχει ακριβώς μια πλήρη εκτέλεση της εκάστοτε κίνησης.

Έπειτα, διεξήχθησαν αρκετά πειράματα στο προγραμματιστικό περιβάλλον Matlab για την εξοικείωση με τα SVM και τις διάφορες υλοποιήσεις σε σύνολα δεδομένων, όπως KTH και Weizmann, με διάφορους πυρήνες, όπως ο γραμμικός και ο Gaussian. Σε αυτό το στάδιο, πραγματοποιήθηκε μιας μορφής αξιολόγηση των δεδομένων εργαλείων εκπαίδευσης, καθώς και μια προσπάθεια βελτιστοποίησης των διαφόρων παραμέτρων, με σκοπό την εξαγωγή των βέλτιστων τιμών ακρίβειας ταξινόμησης των δειγμάτων ελέγχου.

Τέλος, ασχολήθηκα με την υλοποίηση, σε προγραμματιστικό περιβάλλον Matlab, κώδικα για την κβαντοποίηση των περιγραφέων των video και για την υλοποίηση του πρωτοκόλλου ταξινόμησης των video, με χρήση SVM.

Πρέπει να αναφερθεί, ότι η διαδικασία εκπαίδευσης στο περιβάλλον Matlab, πραγματοποιήθηκε βάσει της βιβλιοθήκης Spider[1]. Επίσης, χρησιμοποιήθηκαν έτοιμα εκτελέσιμα αρχεία υλοποίησης των αλγορίθμων STIP και Dense Trajectories, όπως αυτά είχαν χρησιμοποιηθεί για τη διεξαγωγή πειραμάτων στα [19] και [73] αντίστοιχα.

[1] http://people.kyb.tuebingen.mpg.de/spider/main.html

1.2 Δομή Διπλωματικής

Η παρούσα διπλωματική δομείται στο πλαίσιο δυο βασικών θεμάτων, της παρουσίασης της νέας βάσης video THETIS και της διεξαγωγής πειραμάτων αναγνώρισης κινήσεων, με χρήση των μεθόδων STIP και Dense Trajectories και της μηχανής διανυσμάτων υποστήριξης SVM.

Στο κεφάλαιο 2, περιγράφεται το πρόβλημα της αναγνώρισης της ανθρώπινης δραστηριότητας από τον υπολογιστή. Γίνεται εκτενής αναφορά στις μεθόδους που έχουν προταθεί από την ερευνητική κοινότητα για την επίλυση του προβλήματος και στα σύνολα δεδομένων από video που έχουν χρησιμοποιηθεί για την αξιολόγηση και τη σύγκριση των μεθόδων.

Στο κεφάλαιο 3, πραγματοποιείται διεξοδική παρουσίαση της βάσης δεδομένων κίνησης THETIS. Πιο συγκεκριμένα, περιγράφονται με λεπτομέρεια τα μέσα και οι συνθήκες καταγραφής και δίνεται λεπτομερής αναφορά για το περιεχόμενο των video που αποτελούν τη βάση THETIS.

Στο κεφάλαιο 4, παρουσιάζεται λεπτομερώς η πειραματική διαδικασία. Παρουσιάζεται το θεωρητικό υπόβαθρο των μεθόδων STIP και Dense Trajectories για την εξαγωγή των περιγραφέων της κίνησης των video. Επίσης, περιγράφεται η διαδικασία κβαντοποίησης και ταξινόμησης των video με SVM. Ακόμη, παρουσιαζονται τα αποτελέσματα της πειραματικής διαδικασίας.

Τέλος, στο κεφάλαιο 5 συνοψίζουμε τα συμπεράσματα για το σύνολο δεδομένων THETIS και παρουσιάζονται νέες ιδέες στην προσπάθεια που αυτό τροφοδοτεί για νέα έρευνα.

ΚΕΦΑΛΑΙΟ 2

Αναγνώριση της ανθρώπινης δραστηριότητας(ΑΑΔ) από τον υπολογιστή

2.1 Έννοια της ΑΑΔ και εφαρμογές

2.1.1 Εισαγωγή

Η αναγνώριση της ανθρώπινης δραστηριότητας (human action recognition) καταλαμβάνει σημαντικό χώρο στην επιστημονική έρευνα που διεξάγεται στο πεδίο της όρασης υπολογιστών (computer vision). Σκοπό αυτής της ερευνητικής προσπάθειας αποτελεί η αυτόματη ανάλυση και κατ'επέκταση, η αναγνώριση των ανθρώπινων δραστηριοτήτων οι οποίες είναι καταγεγραμμένες σε εικονοσειρές (video).

Η επεξεργασία εικονοσειρών έχει προοδεύσει από το επίπεδο του εντοπισμού κάποιας κίνησης στο να αναγνωρίζει τις πράξεις και αλληλεπιδράσεις ως ξεχωριστά γεγονότα. Η αναγνώριση της ανθρώπινης δραστηριότητας από τον υπολογιστή περιλαμβάνει την κατανόηση της *ανθρώπινης κίνησης*, γεγονός που καθιστά την αναναγνώριση ένα ιδιαιτέρως πολύπλοκο αντικείμενο. Η δομή και το σχήμα του ανθρωπίνου σώματος δεν μπορεί να είναι σαφώς καθορισμένο, λόγω της ύπαρξης πολλών αρθρώσεων και λόγω της ύπαρξης των ενδυμάτων. Επίσης, οι αλλαγές στην φωτεινότητα της εικόνας καθώς και ο θόρυβος που προέρχεται από τις σκιές, δυσκολεύουν ακόμα περισσότερο τις προσπάθειες για αναγνώριση των ανθρωπίνων κινήσεων. Για παράδειγμα, η αναγνώριση δραστηριοτήτων σε εξωτερικούς χώρους επηρεάζεται σημαντικά από τις αλλαγές του καιρού και του φωτισμού.

2.1.2 Τύποι ανθρώπινης δραστηριότητας

Η κατανόηση της ανθρώπινης κίνησης, μπορεί να προσεγγιστεί με διάφορα επίπεδα λεπτομερειών, ανάλογα με την πολυπλοκότητα της εκάστοτε κίνησης. Η μοντελοποίηση και η αναγνώριση της ανθρώπινης συμπεριφοράς προϋποθέτει τον χαρακτηρισμό και την ταξινόμηση των διαφόρων ειδών δραστηριότητας. Μπορούμε να διακρίνουμε τέσσερις κατηγορίες ανθρώπινης δραστηριότητας με βάση το επίπεδο της πολυπλοκότητάς της. Στην πρώτη κατηγορία ανήκουν οι χειρονομίες (gestures), δηλαδή η μετακίνηση κάποιου μέρους του σώματος ενός ατόμου, παραδείγματος χάριν το σήκωμα του χεριού. Η δεύτερη κατηγορία απαρτίζεται από τις κινήσεις ενός μόνο ατόμου (actions) , που περιλαμβάνουν έναν αριθμό χειρονομιών. Κινήσεις θεωρούνται, για παράδειγμα, το τρέξιμο ,το περπάτημα κ.α. Με τον όρο δραστηριότητα, αναφερόμαστε στη σύνθετη ακολουθία κινήσεων που εκτελούν διάφορα άτομα όταν αλληλεπιδρούν (interaction) μεταξύ τους και είτε περιλαμβάνει κάποιο αντικείμενο είτε όχι. Από αυτά τα είδη ανθρώπινης δραστηριότητας αποτελείται η τρίτη κατηγορία, ενώ τέλος, υπάρχουν και οι ομαδικές δραστηριότητες (group activity) που πραγματοποιούνται από ομάδες ατόμων. Χαρακτηριστικό παράδειγμα ομαδικής δραστηριότητας αποτελεί μια ομάδα ατόμων που σχηματίζουν μια ουρά αναμονής.

Οι όροι κίνηση και δραστηριότητα συχνά συγχέονται. Συνήθως, οι δραστηριότητες χαρακτηρίζονται από μεγαλύτερη χρονική διάρκεια, όμως αυτό δεν είναι απόλυτο. Επίσης, δεν υπάρχει αυστηρή διαχωριστική γραμμή ανάμεσα στις δυο έννοιες. Για παράδειγμα, οι

χειρονομίες του μαέστρου μιας ορχήστρας θα μπορούσαν να χαρακτηριστούν ως κίνηση και δραστηριότητα ταυτόχρονα.

2.1.3 Εφαρμογές

Πολυάριθμες και πολύ σημαντικές είναι οι εφαρμογές που βασίζονται στην ικανότητα του υπολογιστή να αναγνωρίζει σύνθετες ανθρώπινες ενέργειες, οι οποίες συνήθως αποτελούνται από πιο απλές κινήσεις *(primative actions*) μέσω της επεξεργασίας και ανάλυσης των δεδομένων εισόδου μιας κάμερας. Σε αυτό το σημείο, θα παρουσιάσουμε κάποιες βάσικες εφαρμογές των συστημάτων αναγνώρισης της ανθρώπινης δραστηριότητας που τονίζουν τη σημασία αυτού του ερευνητικού πεδίου.

- ***Βιομετρικά δεδομένα που βασίζονται στη συμπεριφορά.*** Η συλλογή βιομετρικών δεδομένων συμπεριφοράς *(behavioural biometrics)* ασχολείται με την μελέτη μεθόδων για την αναγνώριση των ανθρώπων με βάση τα φυσικά τους χαρακτηριστικά ή/και την συμπεριφορά τους. Οι παραδοσιακές μέθοδοι συλλογής βιομετρικών δεδομένων, όπως το δακτυλικό αποτύπωμα και η ίριδα του ματιού στηρίζονται στα φυσικά χαρακτηριστικά του ατόμου *(physiological biometrics)* και απαιτούν την συνεργασία του ίδιου του ατόμου. Τελευταία όμως, το ενδιαφέρον για την συλλογή βιομετρικών δεδομένων από την συμπεριφορά του ατόμου έχει αυξηθεί καθώς δεν απαιτούν την συνεργασία του, ούτε παρεμβαίνουν στη δραστηριότητά του. Εφόσον, η παρατήρηση της ανθρώπινης συμπεριφοράς προϋποθέτει μεγαλύτερης διάρκειας παρακολούθηση του υποκειμένου, η αναγνώριση κινήσεων βοηθά στην επίλυση του προβλήματος.

- ***Ασφάλεια και επιτήρηση***. Συστήματα ασφάλειας *(security)* και επιτήρησης *(surveillance)*, τα οποία παραδοσιακά βασίζονται στην παρακολούθηση ενός δικτύου καμερών που καταγράφουν την δραστηριότητα των ανθρώπων, εξελίσσονται με την πρόοδο στην αναγνώριση ανθρώπινων κινήσεων. Σκοπός των εξελιγμένων συστημάτων επιτήρησης σε δημόσιους χώρους, όπως τα αεροδρόμια και οι σιδηροδρομικοί σταθμοί, οι τράπεζες (Σχήμα 2.1), είναι ο εντοπισμός σε πραγματικό χρόνο ασυνήθιστης ή ύποπτης ανθρώπινης δραστηριότητας, όπως κλοπή ή επίθεση, ώστε να παρέχεται δυνατότητα άμεσης αντίδρασης. Μια σχετική εφαρμογή περιλαμβάνει το ψάξιμο μιας συγκεκριμένης δραστηριότητας σε μεγάλες βάσεις δεδομένων μέσω της εκμάθησης προτύπων από μακράς διάρκειας video [28], [29].

- ***Διαδραστικά περιβάλλοντα και εφαρμογές***. Η κατανόηση της αλληλεπίδρασης μεταξύ ανθρώπου και υπολογιστή παραμένει μια διαρκής πρόκληση στο πρόβλημα του σχεδιασμού διαπροσωπειών ανθρώπου-υπολογιστή. Τα οπτικά ερεθίσματα συνιστούν την πιο σημαντική μορφή επικοινωνίας χωρίς ήχο. Επομένως, η αποτελεσματική χρήση αυτής της μορφής επικοινωνίας, όπως οι χειρονομίες και οι κινήσεις, και η επιτυχής αναγνώριση της ανθρώπινης δραστηριότητας υπόσχονται την δημιουργία συστημάτων και υπολογιστών που αλληλεπιδρούν καλύτερα με τους χρήστες. Επιπροσθέτως, παρόμοια διαδραστικά συστήματα που βασίζονται στην αναγνώριση δραστηριότητας συμβάλλουν στη διαμόρφωση ενός ευφυούς περιβάλλοντος (intelligent environment), κατάλληλου για ηλικιωμένους ή παιδιά, βελτιώνοντας την ποιότητα ζωής τους.

Σχήμα 2.1: Παράδειγμα στιγμιοτύπων από το video της προσομοίωσης ληστείας σε τράπεζα [68] (a) Ένα άτομο εισέρχεται στην τράπεζα, (b) Ο ληστής εισέρχεται στην τράπεζα. Άγνωστος εισέρχεται στο χρηματοκιβώτιο, (c) Ένας πελάτης φεύγει από την τράπεζα, (d) Ο ληστής εξέρχεται.

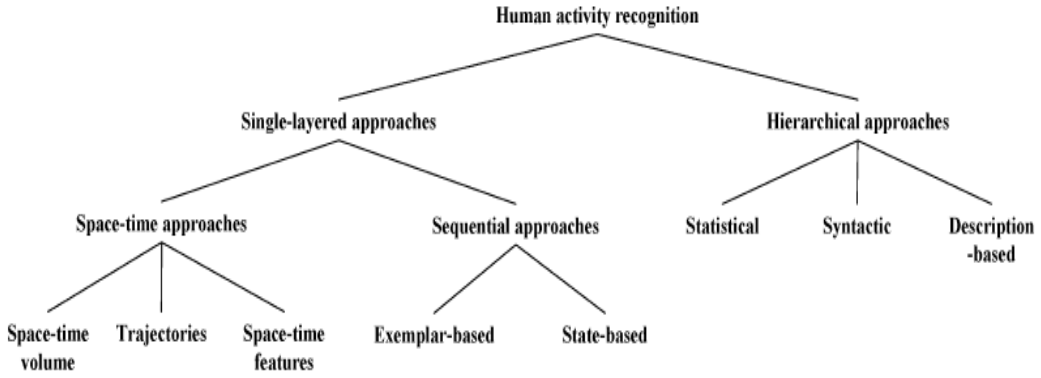
- ***Ανάλυση video με βάση το περιεχόμενο***. Τα video αποτελούν μέρος της καθημερινότητας των ανθρώπων και με την συνεχή εξάπλωση των ηλεκτρονικών κοινωνικών δικτύων που διαμοιράζουν πάσης φύσεως video κρίνεται αναγκαία η αποτελεσματική δημιουργία ευρετηρίου και αποθήκευση τους για την διευκόλυνση του χρήστη. Αυτή η διαδικασία απαιτεί την εκμάθηση προτύπων από video και την σύνοψη του περιεχομένου τους. Σε συνδυασμό με τις προόδους στην ανάκτηση εικόνας με βάση το περιεχόμενο *(content-based image retrieval)*, το ενδιαφέρον για έρευνα στο πρόβλημα της σύνοψης του περιεχομένου των video αυξήθηκε σημαντικά [27]. Η εμπορική εφαρμογή αυτής της τεχνολογίας είναι τα συστήματα που χρησιμοποιούνται στην ανάλυση αθλητικών αγώνων **(**sports play analysis). Η αναγνώριση των ενεργειών των μελών μιας αθλητικής ομάδας μπορεί να έχει πολλαπλές εφαρμογές, όπως η ανάλυση της τακτικής της, η εξαγωγή στατιστικών στοιχείων, ο αυτόματος σχολιασμός ενός αγώνα και ο αυτόματος έλεγχος μιας κάμερας αναμετάδοσης ενός αγώνα (Σχήμα 2.2).

Στην συνέχεια, θα δώσουμε έμφαση στις μεθόδους που έχουν χρησιμοποιηθεί στην αναγνώριση της ανθρώπινης δραστηριότητας σε υψηλό επίπεδο (high-level).

2.2 Μέθοδοι αναγνώρισης ανθρώπινων δραστηριοτήτων

Η πρόοδος στον τομέα της έρευνας που αφορά στην αναγνώριση της ανθρώπινης δραστηριότητας είναι αξιοσημείωτη και οι μεθοδολογίες που έχουν προταθεί από τους ερευνητές για την επίλυση του προβλήματος είναι πολλές και αξίζει να σημειωθεί ότι δε βασίζονται όλες στην ίδια προσέγγιση του προβλήματος. Σε αυτήν την ενότητα, θα περιγραφούν οι διάφορες μεθοδολογίες υψηλού επιπέδου αναγνώρισης κινήσεων, αλληλεπιδράσεων και ομαδικών δραστηριοτήτων. Επίσης, θα παρουσιαστεί μια ταξινόμησή τους που προτάθηκε από τους J.K Aggrarwal και M.S. Ryoo [1]. Η ταξινόμηση αυτή απεικονίζεται στο σχήμα 2.3 και όπως φαίνεται διακρίνονται δυο βασικές κατηγορίες μεθοδολογιών, οι single-layered ή μονής στιβάδας και οι ιεραρχικές και οι υποκατηγορίες τους που περιγράφονται λεπτομερώς στη συνέχεια.

Σχήμα 2.2: Διεπαφή χρήστη ενός συστήματος sport play analysis [1]. Τρεις παίκτες έχουν εντοπιστεί, και το σύστημα εστιάζει στον παίκτη της δεξιάς πλευράς.



Σχήμα 2.3: Ταξινόμηση των μεθοδολογιών με βάση τον τρόπο προσέγγισης του προβλήματος.

2.2.1 Μέθοδοι μονής στιβάδας

Ως single-layered (single-layered) χαρακτηρίζονται οι μέθοδοι που αναγνωρίζουν τις ανθρώπινες δραστηριότητες κατευθείαν από τα δεδομένα της ακολουθίας εικόνων. Κάθε δραστηριότητα αντιπροσωπεύει μια συγκεκριμένη κλάση από ακολουθίες εικόνων και στόχος των μεθόδων αυτού του είδους είναι να αναγνωρίσουν τη δραστηριότητα που περιλαμβάνεται σε μια άγνωστη ακολουθία εικόνων, κατατάσσοντάς την στη σωστή κλάση, με τη χρήση αλγορίθμων κατηγοριοποίησης. Αξίζει να σημειώσουμε, ότι όταν στη διαδικασία της εκπαίδευσης του αλγορίθμου εισαχθούν πρότυπα ακολουθιών από εικόνες που αντιπροσωπεύουν συγκεκριμένες κινήσεις ή δραστηριότητες, η επίδοση των μεθόδων single-

layered βελτιώνεται. Τέλος, κύριο αντικείμενο των μεθόδων αυτής της προσέγγισης αποτελεί η αναγνώριση σχετικά απλών διαδοχικών κινήσεων, όπως το χειροκρότημα και το τρέξιμο.

Υπάρχουν δυο κύριες υποκατηγορίες των single-layered προσεγγίσεων : οι προσεγγίσεις χώρου-χρόνου *(space-time)* και οι ακολουθιακές *(sequential).*

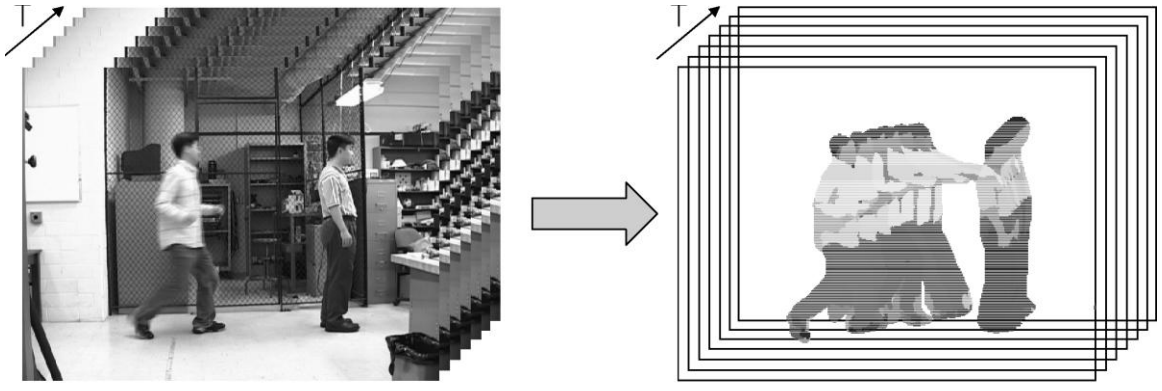
2.2.1.1 Μέθοδοι χωροχρόνου

Όπως είναι γνωστό, ένα video δεν είναι τίποτε άλλο παρά μια ακολουθία εικόνων τοποθετημένων σε χρονική σειρά. Οι εικόνες αποτελούν την προβολή της τρισδιάστασης πραγματικότητας σε δυο διαστάσεις και περιέχουν σχηματισμούς ανθρώπων και αντκειμένων. Επομένως, είναι δυνατή η αναπαράσταση ενός video με τον συνδυασμό της εικόνας στον χώρο και το χρόνο, ως όγκο στο χωροχρόνο (3D XYT space-time volume).

Μια τυπική μεθοδολογία αναγνώρισης ανθρώπινης δραστηριότητας που βασίζεται στον τρισδιάστατο όγκο χωροχρόνου ενός video και σε έναν αλγόριθμο ταιριάσματος προτύπων είναι η ακόλουθη. Αρχικά, κατασκευάζεται ένα μοντέλο 3D XYT space-time για κάθε δραστηριότητα που ανήκει στο σύνολο εκπαίδευσης. Στη συνέχεια, για κάθε άγνωστη ακολουθία εικόνων που δίνεται ως είσοδος στο σύστημα αναγνώρισης, κατασκευάζεται ο όγκος χωροχρόνου που την αντιπροσωπεύει. Τέλος, χρησιμοποιώντας έναν αλγόριθμο ταιριάσματων προτύπων, ο νέος όγκος χωροχρόνου συγκρίνεται με τα υπάρχοντα πρότυπα και επιλέγεται η δραστηριότητα εκείνη που το πρότυπό της ταιριάζει περισσότερο (Σχήμα 2.4).

Εκτός από την τυπική αναπαράσταση των video στο χωροχρόνο που μόλις παρουσιάστηκε, έχουν προταθεί και άλλες προσεγγίσεις του προβλήματος. Πρώτον, η δραστηριότητα ενός ατόμου ή μιας ομάδας ατόμων μπορεί να αναπαρασταθεί ως ένα σύνολο από τροχιές, δεδομένου ότι υπάρχει η δυνατότητα να εντοπιστούν σημεία ενδιαφέροντος, όπως παραδείγματος χάριν, η θέση των αρθρώσεων του ανθρώπινου σώματος. Δεύτερον, μια δραστηριότητα μπορεί να αποδοθεί ως ένα σύνολο από χαρακτηριστικά *(features)*, τα οποία έχουν εξαχθεί από τα δεδομένα που αναπαριστούν τον όγκο ή την τροχιά της κίνησης.

Ως προς τους αλγορίθμους αναγνώρισης που χρησιμοποιούνται για το ταίριασμα των όγκων, των τροχιών ή των χαρακτηριστικών τους, υπάρχουν επίσης αρκετές διαφορετικές προσεγγίσεις. Στη συνέχεια, γίνεται εκτενέστερη αναφορά στις βασικές μεθόδους αναπαράστασης της ακολουθίας εικόνων, καθώς και στους διάφορους αλγορίθμους που χρησιμοποιούνται στην αναγνώριση της ανθρώπινης δραστηριότητας.



Σχήμα 2.4: Παραδείγματα τρισδιάστατων όγκων XYT κατακευασμένα από : (a) ολόκληρες εικόνες και (b) blob εικόνων από εικονοσειρά που αναπαριστά την κίνηση «γρονθοκοπώ».

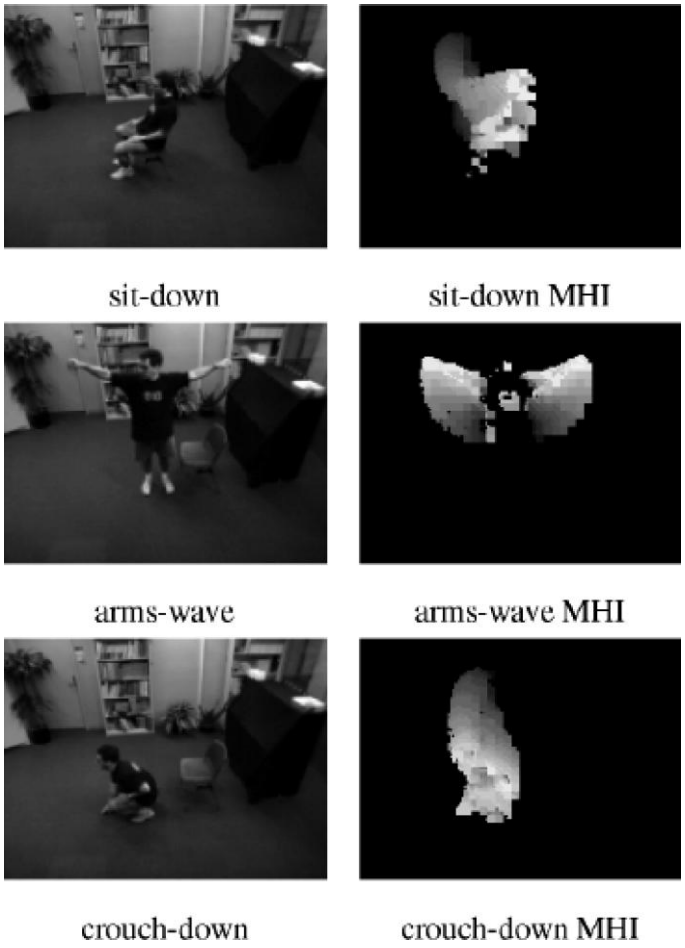
2.2.1.1.1 Όγκος χωροχρόνου

Η αναγνώριση κινήσεων μέσω της αναπαράστασης του όγκου στο χωροχρόνο (space-time volume) βασίζεται πρωτίστως στον υπολογισμό της ομοιότητας μεταξύ των όγκων που έχουν προκύψει από διαφορετικές ακολουθίες εικόνων. Επομένως, ένα τέτοιο σύστημα αναγνώρισης πρέπει να είναι σε θέση να υπολογίσει πόσο όμοιες είναι δυο ανθρώπινες κινήσεις που περιλαμβάνονται σε αυτές τις ακολουθίες εικόνων. Για την εξαγωγή συμπερασμάτων περί ομοιότητας έχουν προταθεί διαφορετικοί τύποι αναπαράστασης του όγκου στο χωροχρόνο, αλλά και διαφορετικοί τρόποι ταιριάσματος των όγκων για την αναγνώριση των κινήσεων.

Οι Bobick και Davis [2] πρότειναν ένα σύστημα αναγνώρισης κινήσεων πραγματικού χρόνου το οποίο χρησιμοποιεί ταίριασμα προτύπων. Σε αντίθεση με άλλα συστήματα που διατηρούν τον τρισδιάστατο όγκο του χωροχρόνου για κάθε κίνηση, το σύστημα αυτό αναπαριστά κάθε κίνηση με ένα πρότυπο που αποτελείται από δύο δισδιάστατες εικόνες : μια δυαδική εικόνα ενέργειας της κίνησης *(motion-energy image, MEI)* και μια εικόνα ιστορικού της κίνησης *(motion-history image, MHI)* (Σχήμα 2.6). Οι δύο εικόνες κατασκευάζονται από μια ακολουθία εικόνων στο μπροστινό πλάνο, οι οποίες αποτελούν ουσιαστικά δισδιάστατες προβολές (XY) του αρχικού τρισδιάστατου όγκου XYT στον χωροχρόνο. Στη συνέχεια, με τη χρήση μιας παραδοσιακής τεχνική ταιριάσματος προτύπων το σύστημα αυτό πραγματοποίησε επιτυχημένα αναγνώριση απλών κινήσεων, (π.χ. κάθομαι, σκύβω) με εφαρμογή σε διαδραστικό περιβάλλον για παιδιά με το όνομα Kids Room (Σχήμα 2.5).

Οι Shechtman και Irani [3] για την επίτευξη της αναγνώρισης της ανθρώπινης δραστηριότητας χρησιμοποιούν την οπτική ροή *(optical flow)* του τρισδιάστατου όγκου χωροχρόνου. Επιπλέον, μετρώντας την ομοιότητα που υπάρχει μεταξύ του εξαγώμενου όγκου ενός νέου video και των πρότυπων όγκου που έχουν στη διαθεσή τους, κατασκευάζουν μια συσχέτιση με τα πρότυπα video. Ο υπολογισμός της ομοιότητας γίνεται ως ακολούθως: σε κάθε σημείο του όγκου, ας πούμε *(x,y,t)*, εξάγεται ένα μικρό κομμάτι γύρω από το σημείο αυτό. Κάθε μικρό τεμάχιο όγκου περιέχει τη ροή της κίνησης στη συγκεκριμένη περιοχή και επομένως, η συσχέτιση ενός τμήματος από ένα πρότυπο με το τμήμα ενός video που βρίσκεται στην ίδια ακριβώς περιοχή, δίνει ένα τελικό τοπικό αποτέλεσμα ως προς την ομοιότητα. Αθροίζοντας όλα αυτά τα επιμέρους αποτελέσματα, τελικά υπολογίζεται η συνολική συσχέτιση ανάμεσα στα πρότυπα όγκου και τον όγκο του video που εξετάζει το σύστημα κάθε φορά. Έτσι, όταν δοθεί ένα άγνωστο video, το σύστημα υπολογίζει όλα τα πιθανά τρισδιάστατα τεμάχια όγκου με κέντρο κάθε *(x,y,t)* που ταιριάζουν περισσότερο με το πρότυπο. Η εφαρμογή του συστήματος πραγματοποιήθηκε επιτυχώς σε διάφορα είδη ανθρώπινης κίνησης όπως, καταδύσεις, κινήσεις μπαλέτου κ.α.

Σχήμα 2.5: Kids Room.



Σχήμα 2.6: Παραδείγματα αναπαράστασης της κίνησης στο χωροχρόνο, εικόνες MHI από τους Bobick, Davis [2].

Οι Ke et al. [4] αξιοποίησαν την κατάτμηση του όγκου στο χωροχρόνο για να μοντελοποιήσουν ανθρώπινες δραστηριότητες. Το σύστημά τους εφαρμόζει έναν ιεραρχικό αλγόριθμο *meanshift* για να κατηγοριοποιήσει τα ογκοστοιχεία ή voxels ανάλογα με το χρώμα τους, αποκτώντας έτσι κατατετμημένους όγκους. Η αναγνώριση της κίνησης επιτυγχάνεται ψάχνοντας για ένα υποσύνολο κατατετμημένων όγκων που ταιριάζουν περισσότερο με το πρότυπο της κίνησης. Το σύστημα εφαρμόστηκε στην αναγνώριση απλών κινήσεων της βάσης KTH [5], καθώς επίσης και σε αγώνες αντισφαίρισης σε video με πιο πολύπλοκο background.

Μια διαφορετική τεχνική χρησιμοποίησαν οι Rodriguez et al. [6] για την αναγνώριση κινήσεων, καθώς ανέλυσαν τους τρισδιάστατους όγκους στο χωροχρόνο με τη σύνθεση φίλτρων και συγκεκριμένα, των MACH (maximum average correlation height) φίλτρων. Για κάθε κίνηση, δημιουργείται ένας συνδυασμός φίλτρων που ταιριάζει με τον παρατηρούμενο όγκο και η ταξινόμηση των κινήσεων γίνεται εφαρμόζοντας το σύνθετο MACH φίλτρο κάθε κίνησης στο άγνωστο video και αναλύοντας την απόκρισή του. Πειράματα με χρήση της μεθόδου αυτής πραγματοποιήθηκαν πάνω στις βάσεις KTH και Weizmann [7].

Γενικά, το μεγαλύτερο μειονέκτημα των προσεγγίσεων που βασίζονται στον τρισδιάστατο όγκο του χωροχρόνου αποτελεί η δυσκολία αναγνώρισης κινήσεων όταν στη σκηνή είναι παρόντα πολλά άτομα. Το πρόβλημα αυτό συνήθως αντιμετωπίζεται με αλγορίθμους συρόμενου παραθύρου *(sliding-window)*, όμως το υπολογιστικό κόστος είναι

μεγάλο. Επιπλέον, η δυσκολία των προσεγγίσεων αυτών να αναγνωρίσουν κινήσεις που δεν μπορούν να τεμαχιστούν χωρικά αποτελεί ένα ακόμη μειονέκτημα.

2.2.1.1.2 Τροχιές χωροχρόνου

Για την αναγνώριση της ανθρώπινης δραστηριότητας υπάρχουν προσεγγίσεις που αντιλαμβάνονται την δραστηριότητα ως ένα σύνολο από τροχιές στο χωροχρόνο (space-time trajectories). Ένα άτομο αναπαρίσταται, συνήθως, ως σύνολο δισδιάστατων (XY) ή τρισδιάστατων (XYZ) σημείων που ανταποκρίνονται στις θέσεις των αρθρώσεων του. Επομένως, όταν το άτομο πραγματοποιεί μια κίνηση, οι αλλαγές στις θέσεις των αρθρώσεών του καταγράφονται ως τροχιές στο χωροχρόνο και τελικά, κατασκευάζεται μια αναπαράσταση σε τρεις (XYT) ή τέσσερις (XYZT) διαστάσεις.

Μερικές προσεγγίσεις για την αναπαράσταση και αναγνώριση των κινήσεων χρησιμοποιούν απ'ευθείας τις τροχιές. Παραδείγματος χάριν, οι Sheick et al. [8] αναπαριστούν μια κίνηση ως ένα σύνολο από 13 τροχιές σε έναν τετραδιάστατο XYZT χώρο, με σκοπό τον υπολογισμό της ομοιότητας μεταξύ δύο συνόλων από τροχιές ανεξάρτητα από την οπτική γωνία. Ομοίως, οι Yilmaz και Shah [9] κάνουν χρήση όμοιας αναπαράστασης για τη σύγκριση video από κάμερες που κινούνται.

Μια διαφορετική προσέγγιση εισήγαγαν οι Campbell και Bobick [10], οι οποίοι επιχειρούν την αναπαράσταση των ανθρώπινων κινήσεων ως καμπύλες σε χώρους φάσης χαμηλών διαστάσεων. Ο πυρήνας της μεθόδου τους είναι ότι όρισαν τη φάση χώρου ενός σώματος ως ένα χώρο όπου κάθε άξονας αποτελεί είτε μια ανεξάρτητη παράμετρο του σώματος(π.χ. γωνία αστραγάλου, γωνία γονάτου), είτε την πρώτη της παράγωγο. Στη φάση χώρου η στάσιμη κατάσταση του ατόμου σε κάθε κίνηση θεωρείται ένα σημείο και μια κίνηση αποτελείται από ένα σύνολο σημείων, όπως μια καμπύλη. Σύμφωνα με την προσέγγιση αυτή, η καμπύλη προβάλλεται σε πολλαπλούς δισδιάστατους υποχώρους και αποθηκεύεται για να αντιπροσωπεύσει την κίνηση. Τελικά, από όλες τις δυνατές καμπύλες των δισδιάστατων υποχώρων το σύστημα επιλέγει τις πιο αξιόπιστες που θα χρησιμοποιηθούν στη διαδικασία αναγνώρισης. Η αναγνώριση μιας κίνησης επιτυγχάνεται μετατρέποντας ένα άγνωστο video σε ένα σύνολο σημείων μέσα στο χώρο φάσης και έπειτα, το σύστημα είναι σε θέση να επιβεβαιώσει αν τα σημεία βρίσκονται πάνω στις προβολές των αποθηκευμένων καμπυλών. Η μέθοδος των Campbell και Bobick εφαρμόστηκε με επιτυχία σε βασικές κινήσεις μπαλέτου.

Σε αντίθεση με τις προηγούμενες μεθοδολογίες, όπου είναι απαραίτητη η διατήρηση της τροχιάς στο χωροχρόνο, οι Rao και Shah [11] εξήγαγαν χρήσιμα σχέδια καμπυλότητας από τις τροχιές. Το σύστημά τους εντοπίζει τις θέσεις των κορυφών των καμπυλωτών τροχιών, και αναπαριστά μια κίνηση με ένα σύνολο από κορυφές και ολοκληρώματα μεταξύ των, τα οποία είναι δε ανεξάρτητα από την οπτική γωνία. Έτσι, καθίσταται δυνατή η κατασκευή προτύπων κινήσεων και η αναγνώριση επιτυγχάνεται με αλγορίθμους ταιριάσματος προτύπων.

Το βασικό πλεονέκτημα των παραπάνω προσεγγίσεων είναι η ικανότητα να αναλύουν τις λεπτομέρειες των ανθρώπινων κινήσεων, με αποτέλεσμα να συμβάλλουν στην αναγνώριση κινήσεων διαφορετικών κλάσεων που παρουσιάζουν πολλές ομοιότητες μεταξύ τους. Επιπροσθέτως, οι περισσότερες μέθοδοι πους βασίζονται στην ανάλυση τροχιών είναι ανεξάρτητες από την οπτική γωνία. Παρ'όλα αυτά, για τον υπολογισμό των αρθρώσεων των ατόμων που εμφανίζονται στη σκηνή σε τρεις διαστάσεις XYZ, απαιτείται ένα ισχυρό low-level υπόβαθρο. Δηλαδή, οι παραπάνω προσεγγίσεις απαιτούν τη χρήση αποτελεσματικών αλγορίθμων τρισδιάσταστης ανίχνευσης και εντοπισμού των μελών του ανθρώπινου σώματος.

2.2.1.1.3 Χαρακτηριστικά χωροχρόνου

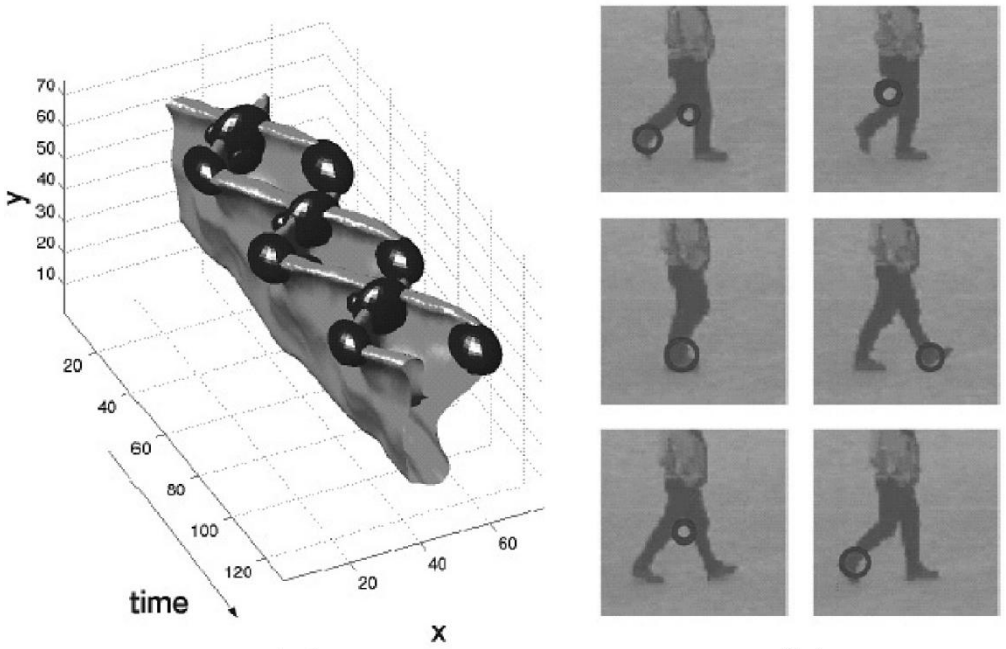
Οι μέθοδοι που ανήκουν σε αυτήν την κατηγορία χρησιμοποιούν τοπικά χαρακτηριστικά που εξάγονται από τους τρισδιάστατους όγκους στο χωροχρόνο για να αναπαραστήσουν και να αναγνωρίσουν την ανθρώπινη δρστηριότητα. Για να περιγραφεί επαρκώς μια μέθοδος τύπου χαρακτηριστικών χωροχρόνου (space-time features), είναι απαραίτητο να απαντηθούν τρια ερωτήματα που την αφορούν. Πρώτον, ποιά τοπικά χαρακτηριστικά εξάγει , δεύτερον, με ποόν τρόπο τα αξιοποιεί για να αναπαραστήσει μια κίνηση και τέλος, ποια μεθοδολογία χρησιμοποιεί για την ταξινόμηση των κινήσεων.

Οι Chomat και Crowley [12] χρησιμοποίησαν τοπικούς περιγραφείς εμφάνισης *(appearance descriptors)*. Στο σύστημα τους σε κάθε καρέ εντοπίζονται τοπικά χαρακτηριστικά εμφάνισης που περιγράφουν τον προσανατολισμό της κίνησης, με σκοπό την εξαγωγή πληροφορίας από μια ακολουθία εικόνων. Από αυτά τα τοπικά χαρακτηριστικά προκύπτουν έπειτα ιστογράμματα και εφαρμόζοντας τον κανόνα Bayes υπολογίζεται η πιθανότητα να εμφανιστεί μια συγκεκριμένη κίνηση. Παρόλο που το σύστημα αυτό αναγνωρίζει μόνον απλές χειρονομίες, εξαιτίας της απλότητας των πειγραφέων, αποδεικνύει πως οι αισθητήρες εμφάνισης μπορούν να συμβάλλουν στην αναγνώριση της ανθρώπινης δραστηριότητας.

Μια διαφορετική προσέγγιση που βασίζεται στη χρήση τοπικών χαρακτηριστικών σε πολλαπλές βαθμίδες πρότειναν οι Zelnic-Manor και Irani [13]. Η ανάλυση του όγκου των video σε χρονικές βαθμίδες αποσκοπεί στην αντιμετώπιση της διαφοροποίησης που μπορεί να υπάρξει στην ταχύτητα εκτέλεσης μιας δραστηριότητας. Για κάθε σημείο XYT του τρισδιάστατου όγκου, το σύστημά τους υπολογίζει την κανονικοποιημένη τοπική συνάρτηση κλήσης της έντασης *(local intensity gradient)*. Ακολούθως, εφαρμόζοντας στα ιστογράμματα κάποιον αλγόριθμο συσταδοποίησης *(clustering)* χωρίς επίβλεψη, το σύστημα αναγνώρισε επιτυχώς διάφορες δραστηριότητες, όπως καλαθοσφαίριση και αντισφαίριση σε εξωτερικούς χώρους.

Την εξαγωγή τοπικών χαρακτηριστικών υιοθέτησαν και οι Blank et al. [14], με τη διαφορά ότι η εξαγωγή πραγματοποιήθηκε με χρήση της συνάρτησης Poisson. Ακριβέστερα, για κάθε pixel κατασκευάζεται ένας τρισδιάστατος όγκος XYT του οποίου οι τιμές των pixel αποτελούν λύσεις της εξίσωσης Poisson. Με αυτόν τον τρόπο, εξάγονται χρήσιμες τοπικές ιδιότητες που αφορούν στον προσανατολισμό (orientation) και την υπεροχή (saliency). Έτσι, κάθε κίνηση αναπαρίσταται ως ένα σύνολο από γενικά χαρακτηριστικά, δηλαδή σταθμισμένα στιγμιότυπα των τοπικών γνωρισμάτων. Με την εφαρμογή ενός ταξινομητή κοντινότερου γείτονα (nearest neighbor), το σύστημα τους αναγνώρισε με επιτυχία κινήσεις από τη βάση Weizmann αλλά και χορευτικές φιγούρες μπαλέτου.

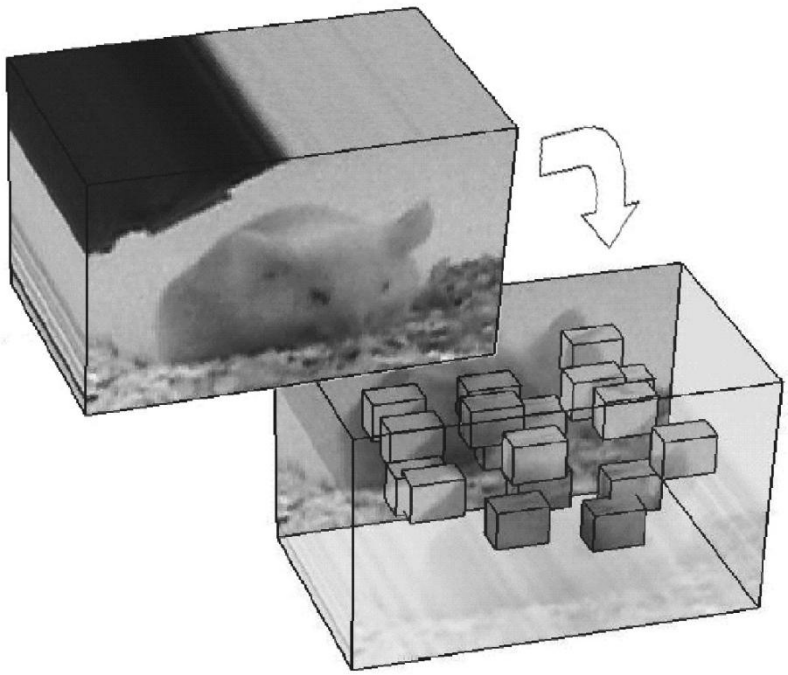
Σε αντίθεση με τις προσεγγίσεις που αναφέρθηκαν προηγουμένως, υπάρχουν άλλες όπου για την αναγνώριση της δραστηριότητας εξάγουν αραιά τοπικά χαρακτηριστικά (sparse local features). Χαρακτηριστικό παράδειγμα αποτελεί η μέθοδος των Laptev και Lindeberg [15], οι οποίοι εξάγουν αραιά σημεία ενδιαφέροντος (interest points) από τις ακολουθίες εικόνων. Συγκεκριμένα, ο ανιχνευτής τους εντοπίζει γωνίες στον τρισδιάστατο χώρο XYT, οι οποίες συλλαμβάνουν ποικίλες μεταβολές στην κίνηση, όπως αλλαγή κατέθυνσης ενός αντικειμένου, συγχώνευση ή διαχωρισμό της δομής μιας εικόνας (Σχήμα 2.7).



Σχήμα 2.7: Το Σχήμα 2.7 (a) παρουσιάζει τις συνδυασμένες επιφάνειες XYT του ποδιού ενός ατόμου και τα σημεία ενδιαφέροντος που εντοπίστηκαν με τη μέθοδο των Laptev and Lindeberg [15]. Το Σχήμα 2.7 (b) δείχνει τα ίδια σημεία ενδιαφέροντος σε μια ακολουθία πραγματικών εικόνων.

Οι μέθοδοι που στηρίζονται σε αραιά τοπικά σημεία ενδιαφέροντος έχουν γίνει ιδιαίτερα δημοφιλείς στην ερευνητική κοινότητα, διότι επικεντρώνονται στην εξαγωγή χαρακτηριστικών μόνο όταν υπάρχει κάποια μεταβολή του σχήματος του όγκου ή άλλου είδους μεταβολή που ξεχωρίζει, αντί να συλλέγουν χαρακτηριστικά από κάθε καρέ. Επιπροσθέτως,τα χαρακτηριστικά αυτά συνήθως, είναι ανεξάρτητα από την κλίμακα και την περιστροφή. Όπως και οι περιγραφείς που χρησιμοποιούνται στην αναγνώριση αντικειμένων, όπως ο SIFT descriptor του Lowe [16] .

Οι Dollar et al. [17] πρότειναν έναν νέο ανιχνευτή αραιών χαρακτηριστικών στο χωροχρόνο, ο οποίος εξάγει σημεία με τοπική περιοδική κίνηση και τα αντιστοιχίζει σε μικρούς τρισδιάστατους όγκους *(cuboids)* (Σχήμα 2.8). Στη συνέχεια, εφαρμόζει κατάλληλους μετασχηματισμούς για να προκύψουν τα τελικά τοπικά χαρακτηριστικά. Τέλος, το σύστημα μοντελοποιεί κάθε κίνηση με ένα ιστόγραμμα που χρησιμοποιείται στην τελική αναγνώριση εκφράσεων προσώπου, συμπεριφορών ποντικιών και ανθρώπινων δραστηριοτήτων.



Σχήμα 2.8: Κυβοειδή χαρακτηριστικά που εξήχθησαν από την κίνηση του ποντικιού, με τη μέθοδο [17].

Σε όλες τις προηγούμενες μεθόδους που στηρίζονται σε αραιά τοπικά χαρακτηριστικά, οι τοπικοί και χρονικοί συσχετισμοί μεταξύ των σημείων που εντοπίζονται αγνοούνται. Υπάρχουν, όμως, άλλες προσεγγίσεις που θεωρούν πολύ σημαντικούς αυτούς τους συσχετισμούς και επιχειρούν να μοντελοποιήσουν την κατανομή των χαρακτηριστικών στο χωροχρόνο για καλύτερες επιδόσεις στην αναγνώριση ανθρώπινης δραστηριότητας.

Παράδειγμα αυτής της προσπάθειας αποτελεί η μεθοδολογία που προτάθηκε από τους Savarese et al. [18] για την εξαγωγή της πληροφορίας περί συσχέτισης των χαρακτηριστικών. Ομοίως, οι Laptev et al. [19] πρότειναν την κατασκευή ιστογραμμάτων χώρου-χρόνου διαιρώντας τον όγκο σε πλέγματα. Με την τεχνική αυτή, υπολογίζεται η κατανομή των τοπικών περιγραφέων στον τρισδιάστατο XYT χωροχρόνο, αναλύοντας ποια χαρακτηριστικά πέφτουν σε ποιο πλέγμα. Η μέθοδος αυτή ελέγχθηκε στη βάση KTH, όπως και σε άλλα video περισσότερο ρεαλιστικά.

Στην ίδια κατεύθυνση βρίσκεται και η προσέγγιση των Ryoo και Aggarwal [20], η λεγόμενη «ταίριασμα σχέσεων χώρου-χρόνου» *(spatio-temporal relationship match- STR match)*. Εδώ μελετώνται οι δομικές ομοιότητες μεταξύ των ακολουθιών εικόνων και έχει εφαρμοστεί με επιτυχία τόσο σε απλες κινήσεις (KTH dataset), όσο και σε δραστηριότητες αλληλεπίδρασης.

Γενικά, οι μέθοδοι αναγνώρισης κινήσεων που χρησιμοποιούν τοπικά χαρακτηριστικά έχουν αρκετά πλεονεκτήματα. Συγκεκριμένα, δεν απαιτείται η αφαίρεση του πίσω σκηνικού ούτε η χρήση άλλων low-level εργαλείων. Επίσης, τα τοπικά χαρακτηριστικά είναι ανεξάρτητα από την κλίμακα και την περιστροφή στις πιο πολλές περιπτώσεις. Τέλος, είναι αρκετά κατάλληλες για την αναγνώριση απλών περιοδικών κινήσεων, όπως το βάδισμα καθότι οι περιοδικές κινήσεις παράγουν επαναλαμβανόμενα πρότυπα χαρακτηριστικών.

2.2.1.1.4 Σύγκριση

Επιχειρώντας μια σύγκριση στις διαφορετικές προσεγγίσεις space-time για την αναγνώριση δραστηριότητας διαπιστώνουμε ότι είναι όλες κατάλληλες για την αναγνώριση περιοδικών κινήσεων και χειρονομιών, όμως παρουσιάζουν δυσκολία στην αναγνώριση μεταβολών στην ταχύτητα εκτέλεσης μιας κίνησης.

Οι προσεγγίσεις που στηρίζονται στην εξαγωγή τροχιών, είναι συνήθως ανεξάρτητες από την οπτική γωνία και επιτυγχάνουν λεπτομερέστερη ανάλυση σε πολλαπλά επίπεδα. Εντούτοις, για την εφαρμογή τους απαιτείται η τρισδιάσταση προτυποποίηση των μερών του ανθρώπινου σώματος, πρόβλημα το οποίο παραμένει δύσκολο. Αντιθέτως, οι μέθοδοι που βασίζονται στην εξαγωγή τοπικών χαρακτηριστικών ολοένα κερδίζουν το ενδιαφέρον της επιστημονικής κοινότητας εξαιτίας της αξιοπιστίας τους ακόμη και σε συνθήκες όπου υπάρχει θόρυβος ή μεταβολές στο φωτισμό. Επιπλέον, πολλές από αυτές παρέχουν τη δυνατότητα αναγνώρισης διαφόρων δραστηριοτήτων χωρίς να απαιτείται απομόνωση του μπροστινού σκηνικού. Ωστόσο, αδυνατούν να αναγνωρίσουν πολύπλοκες δραστηριότητες και εξαρτώνται από τις μεταβολές της οπτικής γωνίας.

2.2.1.2 Ακολουθιακές μέθοδοι

Σ'αυτήν την κατηγορία ανήκουν οι μέθοδοι αναγνώρισης μέσω της ανάλυσης ακολουθιών από χαρακτηριστικά. Σε αυτά τα συστήματα ένα video θεωρείται ως μια ακολουθία από παρατηρήσεις (π.χ διανύσματα χαρακτηριστικών) και συμπεραίνουμε ότι μια δραστηριότητα λαμβάνει χώρα όταν παρατηρηθεί μια συγκεκριμένη ακολουθία που χαρακτηρίζει αυτή τη δραστηριότητα. Συνήθως, η ακολουθία εικόνων αρχικά μετατρέπεται σε ακολουθία από διανύσματα με χαρακτηριστικά, εξάγοντας χαρακτηριστικά (π.χ. γωνία της άρθρωσης του γονάτου) που περιγράφουν την κατάσταση ενός ατόμου σε κάθε καρέ. Στη συνέχεια, αναλύεται η ακολουθία για να υπολογιστεί πόση είναι η πιθανότητα τα διανύσματα χαρακτηριστικών να έχουν παραχθεί από το άτομο που εκτελεί μια συγκεκριμένη δραστηριότητα. Εάν η ομοιότητα ανάμεσα στην ακολουθία και την κλάση μιας δραστηριότητας είναι αρκετά μεγάλη, το σύστημα αποφασίζει ότι αυτή η ενέργεια έχει εκτελεστεί.

Οι ακολουθιακές (sequential) προσεγγίσεις χωρίζονται σε δυο υποκατηγορίες με βάση τη μεθοδολογία τους: αναγνώριση βάσει προτύπων και αναγνώριση βάσει μοντέλων κατάστασης. Στις επόμενες ενότητες περιγράφονται διάφορες μεθοδολογίες κάθε υποκατηγορίας.

2.2.1.2.1 Μέθοδοι βασισμένες σε πρότυπα

Τα συστήματα που χρησιμοποιούν πρότυπα (exemplar-based), αναπαριστούν μια ανθρώπινη δραστηριότητα διατηρώντας μια ακολουθία πρότυπο ή ένα σύνολο από δείγματα ακολουθιών που προέκυψαν από την εκτέλεση της δραστηριότητας. Όταν στο σύστημα δοθεί ένα άγνωστο video, συγκρίνεται η ακολουθία διανυσμάτων με τα χαρακτηριστικά από το video αυτό με την ακολουθία πρότυπο. Εάν υπάρχει ακρκετά μεγάλη ομοιότητα, το σύστημα συμπεραίνει ότι το video περιέχει την συγκεκριμένη δραστηριότητα.

Φυσικά, ο τρόπος και ο ρυθμός εκτέλεσης μιας κίνησης μπορεί να διαφέρει. Το σύστημα πρέπει να είναι ανεξάρτητο από τέτοιες μεταβολές. Για το λόγο αυτό, τεχνικές δυναμικής ευθυγράμμισης / σύγκρισης προτύπων *(dynamic time warping- DTW)* έχουν υιοθετηθεί για το ταίριασμα δυο ακολουθιών με χρονικές αποκλίσεις.

Οι Darrell και Pentland [21] πρότειναν μια μεθοδολογία που βασίζεται στην DTW για την αναγνώριση χειρονομιών. Το σύστημά τους διατηρεί πολλαπλά μοντέλα όψης *(view models)* ενός αντικειμένου υπό διαφορετικές συνθήκες. Μόλις δοθεί ως είσοδος ένα video, το αποτέλεσμα συσχέτισης μεταξύ της εικόνας σε κάθε καρέ και της κάθε όψης μοντελοποιείται σε μια συνάρτηση του χρόνου. Έπειτα, μια νέα παρατήρηση αντιστοιχίζεται στα πρότυπα μέσω του αλγορίθμου DTW και το σύστημα αναγνωρίζει διάφορες χειρονομίες που δίνονται ως είδοδοι.

Οι Gavrila και Davis [22] ανέπτυξαν μια μέθοδο εντοπισμού των μερών του ανθρώπινου σώματος η οποία χρησιμοποιεί τρισδιάστατα XYZ μοντέλα. Σκοπός της μεθόδου είναι ο υπολογισμός του τρισδιάστατου μοντέλου του σκελετού ενός ατόμου σε κάθε καρέ και η ανάλυση της κίνησης. Για να προκύψει το τρισδιάστατο μοντέλο χρησιμοποιούνται πολλές κάμερες λήψης. Τελικά, προκύπτει ένα μοντέλο ανθρώπινου σκελετού *(stick figure model)* με 17 βαθμούς ελευθερίας που καταγράφει τις τιμές των γωνιών που σχηματίζουν οι αρθρώσεις. Οι ακολουθίες των γωνιών αναλύονται με τον αλγόριθμο DTW και συγκρίνονται με την ακολουθία που χαρακτηρίζει την κάθε κίνηση.

Οι Yacoob και Black [23], αντιμετωπίζουν την είσοδο ως ένα σύνολο από σήματα αντί για διακριτές ακολουθίες που περιγράφουν διαδοχικές αλλαγές στις τιμές των χαρακτηριστικών. Για την ανάλυση των σημάτων χρησιμοποιούν ανάλυση ιδιαζουσών τιμών *(singular value decomposition - SVD)* και αναπαριστούν την δραστηριότητα σαν ένα γραμμικό συνδυασμό ενός συνόλου δραστηριοτήτων βάσης, οι οποίες είναι ουσιαστικά ένα σύνολο ιδιοδιανυσμάτων. Επομένως, για κάθε νέα είσοδο το σύστημα υπολογίζει τους συντελεστές των δραστηριοτήτων βάσης συμπεριλαμβάνοντας πιθανούς παράγοντες παραμόρφωσης όπως μεταβολές στην κλίμακα. Φυσικά, η ομοιότητα υπολογίζεται από την σύγκριση των συντελεστών που προκύπτουν με τους συντελεστές του προτύπου κάθε κίνησης.

Η αναγνώριση ανθρώπινων κινήσεων σε μεγάλη απόσταση όπου οι κινήσεις είναι δυσδιάκριτες, απασχόλησε τους Efros et al. [24] οι οποίοι χρησιμοποίησαν περιγραφείς κινήσεων που βασίζονται στην οπτική ροή σε κάθε καρέ. Το σύστημα υπολογίζει αρχικά τον όγκο στο χωροχρόνο για κάθε άτομο που ανιχνεύεται και εν συνεχεία, υπολογίζει τη δισδιάστατη οπτική ροή σε κάθε καρέ. Με τον τρόπο αυτό, το video μιας ανθρώπινης κίνησης μετατρέπεται σε μια ακολουθία περιγραφέων κίνησης που προκύπτουν από την οπτική ροή του ατόμου. Για την ταξινόμηση και την αναγνώριση των κινήσεων εφαρμόζεται η μέθοδος κοντινότερου γείτονα. Η προσέγγιση αυτή εφαρμόστηκε επιτυχώς σε κινήσεις μπαλέτου και σε αγώνες αντισφαίρισης και ποδοσφαίρου.

Μια διαφορετική προσέγγιση του προβλήματος επιχειρούν οι Lublinerman et al. [25], οι οποίοι μοντελοποιούν την ανθρώπινη δραστηριότητα με γραμμικά χρονικά αμετάβλητα συστήματα (linear-time-invariant – LTI). Το σύστημα αυτό μετατρέπει μια ακολουθία εικόνων σε μια ακολουθία από περιγράμματα *(silhouettes)* εξάγοντας δυο είδη σχηματικών αναπαραστάσεων: το πλάτος του περιγράμματος και τους περιγραφείς Fourier. Μια δραστηριότητα περιγράφεται σαν ένα LTI σύστημα που ανιχνεύει τη δυναμική των αλλαγών στα χαρακτηριστικά του περιγράμματος. Η ταξινόμηση σε αυτή την περίπτωση πραγματοποιείται με μηχανές διανυσμάτων υποστήριξης *(support vector machines- SVMs)*. Με την προσέγγιση αυτή, ταξινομήθηκαν κινήσεις όπως αργό βάδισμα, γρήγορο βάδισμα, και βάδισμα σε κεκλιμένο επίπεδο.

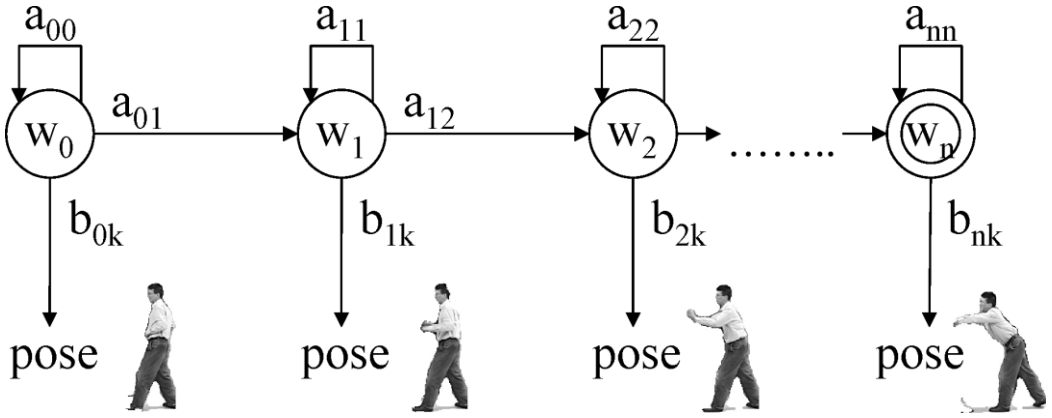
Σημαντική είναι η συμβολή των Veeraraghavan et al [26] στη σαφή μοντελοποίηση των ενδοατομικών και διατομικών διαφοροποήσεων που αφορούν στην ταχύτητα εκτέλεσης μιας δραστηριότητας. Τα άτομα είναι πιθανόν να αλλάζουν την ταχύτητα εκτέλεσης ενός μέρους μιας δραστηριότητας, χωρίς απαραίτητα να ισχύει το ίδιο για τα υπόλοιπα μέρη της δραστηριότητας. Οι Veeraraghavan et al. κατασκεύασαν ένα σύστημα που μαθαίνει τα μη γραμμικά χαρακτηριστικά των μεταβολών στην ταχύτητα εκτέλεσης. Πιο συγκεκριμένα,

μοντελοποιούν την εκτέλεση μιας κίνησης με δυο συναρτήσεις : (i) *συνάρτηση αλλαγών των χαρακτηριστικών με το χρόνο* και (ii) *χωρική συνάρτηση της πιθανής στρέβλωσης του χρόνου*. Επίσης, επέκτειναν τον αλγόριθμο DTW περιλαμβάνοντας τη συνάρτηση χρονικής στρέβλωσης για το ταίριασμα δυο ακολουθιών. Τέλος, αποτελέσματα υψηλής ακρίβειας προέκυψαν από την εφαρμογή της μεθόδου σε κινήσεις όπως άρση και ρίψη ενός αντικειμένου, σπρώξιμο κ.α.

2.2.1.2.2 Μέθοδοι βασισμένες σε μοντέλα κατάστασης

Οι μέθοδοι αναγνώρισης κινήσεων που στηρίζονται σε μοντέλα κατάστασης *(state model-based)* αναπαριστούν μια ανθρώπινη δραστηριότητα σαν ένα μοντέλο από ένα σύνολο καταστάσεων. Το κάθε μοντέλο εκπαιδεύεται στατιστικά ώστε να ανταποκρίνεται σε ακολουθίες από χαρακτηριστικά που ανήκουν στην κλάση της συγκεκριμένης δραστηριότητας. Ειδικότερα, το στατιστικό μοντέλο σχεδιάζεται ώστε να παράγει μια ακολουθία με συγκεκριμένη πιθανότητα. Για κάθε μοντέλο υπολογίζεται η πιθανότητα να παραχθεί η ακολουθία χαρακτηριστικών που παρατηρείται και με αυτόν τον τρόπο μετράται η ομοιότητα ανάμεσα στο μοντέλο της κίνησης και στην ακολουθία εικόνων που δίνεται ως είσοδος. Τέλος, για την αναγνώριση της δραστηριότητας κατασκευάζεται είτε ένας ταξινομητής - εκτιμητής μέγιστης πιθανοφάνειας *(maximum likelihood estimation – MLE)* είτε ένας ταξινομητής μέγιστης εκ των υστέρων πιθανότητας *(maximum a posteriori probability – MAP).*

Ευρέως διαδεδομένα εργαλεία των state model-based προσεγγίσεων αποτελούν τα κρυφά Μαρκοβιανά μοντέλα *(hidden Markov models – HMMs)* και τα δυναμικά δίκτυα Bayes *(dynamic Bayesian networks – DBNs).* Και στις δυο περιπτώσεις, μια δραστηριότητα αναπαρίσταται σαν ένα σύνολο κρυφών καταστάσεων. Το άτομο που εκτελεί μια δραστηριότητα σε κάθε καρέ, θεωρείται ότι βρίσκεται σε μια κατάσταση η οποία παράγει μια παρατήρηση (π.χ. διανύσματα χαρακτηριστικών). Στο επόμενο καρέ, το state model-based σύστημα μεταβαίνει σε άλλη κατάσταση λαμβάνοντας υπ'όψιν την πιθανότητα μετάβασης μεταξύ των καταστάσεων. Οι δραστηριότητες μπορούν να αναγνωριστούν επιλύοντας το πρόβλημα της αποτίμησης, δεδομένου ότι έχουν εκπαιδευτεί οι πιθανότητες μετάβασης και παρατήρησης για τα μοντέλα. Το πρόβλημα της αποτίμησης *(evaluation problem),* ορίζεται ως το πρόβλημα υπολογισμού της πιθανότητας που έχει μια δεδομένη ακολουθία να έχει παραχθεί από ένα συγκεκριμένο μοντέλο κατάστασης. Εάν αυτή η πιθανότητα είναι αρκετά μεγάλη, το σύστημα αποφασίζει ότι η δραστηριότητα που ανταποκρίνεται στο μοντέλο, λαμβάνει χώρα στο video που έχει δοθεί ως είσοδος. Το Σχήμα 2.9 παρουσιάζει ένα παράδειγμα ακολουθιακού HMM.



Σχήμα 2.9: Παράδειγμα κρυφού Μαρκοβιανού μοντέλου για την κίνηση «τεντώνω το χέρι».

Η εργασία των Yamato et al. [30] είναι η πρώτη όπου εφαρμόζονται μαρκοβιανά μοντέλα στην αναγνώριση κινήσεων. Πιο συγκεκριμένα, κατασκεύασαν ένα σύστημα που σε κάθε καρέ, μετατρέπει μια δυαδική εικόνα στην οποία έχει απομονωθεί το μπροστινό σκηνικό, σε έναν πίνακα από πλέγματα. Ο αριθμός των pixel σε κάθε πλέγμα θεωρείται ένα χαρακτηριστικό και κατά συνέπεια, ένα διάνυσμα χαρακτηριστικών εξάγεται για κάθε καρέ. Αυτό το διάνυσμα αντιμετωπίζεται ως μια ακολουθία παρατηρήσεων που δημιουργήθηκε από το μοντέλο δραστηριότητας, και κάθε δραστηριότητα αναπαρίσταται από ένα κρυφό μαρκοβιανό μοντέλο που ανταποκρίνεται - με βάση την πιθανότητα – σε συγκεκριμένες ακολουθίες διανυσμάτων με χαρακτηριστικά (π.χ. πλέγματα). Πιο συγκεκριμένα, οι παράμετροι των μαρκοβιανών μοντέλων εκπαιδεύονται με ένα σύνολο ταξινομημένων δεδομένων εφαρμόζοντας τον απλό αλγόριθμο εκμάθησης για τα HMM και τελικά, επιλύεται το πρόβλημα αποτίμησης. Το σύστημα που μόλις περιγράφηκε αναγνώρισε με επιτυχία κινήσεις αντισφαίρισης (tennis), και αποτέλεσε την πρώτη απόδειξη πως τα κρυφά μαρκοβιανά μοντέλα μπορούν να αναγνωρίσουν με αξιοπιστία τις αλλάγες στα χαρακτηριστικά των μοντέλων κατά την εκτέλεση ανθρώπινων δραστηριοτήτων.

Ομοίως, συμβατικά HMMs χρησιμοποιήθηκαν και από τους Starner και Pentland [31] με σκοπό την αναγνώριση της Αμερικάνικης νοηματικής γλώσσας (ASL). Η μέθοδός τους ανιχνεύει την θέση των χεριών και εξάγει χαρακτηριστικά που περιγράφουν τα σχήματα και την τοποθέτησή τους. Κάθε λέξη της ASL μοντελοποιείται σε ένα μαρκοβιανό μοντέλο, παράγοντας μια ακολουθία χαρακτηριστικών που περιγράφουν τα σχήματα των χεριών και τις θέσεις τους, όπως στην περίπτωση των Yamato et al. [30]. Στη συνέχεια, για τον υπολογισμό των πιθανοτήτων εφαρμόζεται ο αλγόριθμος του Viterbi που παρέχει μια αποτελεσματική προσέγγιση της απόστασης πιθανότητας, με αποτέλεσμα να μπορεί μια άγνωστη ακολουθία παρατηρήσεων να ταξινομηθεί στην κλάση της περισσότερο ταιριαστής λέξης.

Για την αναγνώριση δυο ειδών χειρονομίας – «χαιρετώ» και «δείχνω», οι Bobick και Wilson [32] έκαναν χρήση των μοντέλων κατάστασης. Στο σύστημά τους, μια χειρονομία παρουσιάζεται ως μια δισδιάσταση XY τροχιά που περιγράφει τις αλλαγές στη θέση των χεριών. Κάθε καμπύλη αναλύεται σε ακολουθίες διανύσματων και κατ'επέκταση σε ακολουθίες καταστάσεων που υπολογίζονται από ένα παράδειγμα εκπαίδευσης. Επιπροσθέτως, κάθε κατάσταση είναι ασαφής ώστε να λαμβάνονται υπ'όψιν τυχόν διαφοροποιήσεις στην ταχύτητα και τον τρόπο εκτέλεσης της ίδιας χειρονομίας. Η πιθανότητα μετάβασης από μια κατάσταση σε μια άλλη, ισοδυναμεί με το κόστος μετάβασης που υπολογίζεται από το σύστημα. Για την αναγνώριση των χειρονομιών εφαρμόζεται ένας αλγόριθμος δυναμικού προγραμματισμού. Κάποιες παραλλαγές των κρυφών μαρκοβιανών μοντέλων, καθώς και επεκτάσεις τους που χρησιμοποιήθηκαν για την αναγνώριση περισσότερο πολύπλοκων δραστηριοτήτων, αναφέρονται στις δυο επόμενες παραγράφους.

Οι Oliver et al.[33] κατασκεύασαν μια παραλλαγή του συμβατικού HMM (διπλό HMM) για την αναγνώριση αλληλεπιδράσεων ανθρώπου με άνθρωπο. Το μεγαλύτερο μειονέκτημα του συμβατικού HMM είναι η αδυναμία του να αναπαραστήσει δραστηριότητες που εμπλέκουν δύο ή περισσότερα άτομα, διότι το HMM είναι ένα ακολουθιακό μοντέλο όπου μια μόνο κατάσταση είναι ενεργή κάθε φορά. Έτσι, καθίσταται αδύνατο να αναπαρασταθούν οι δραστηριότητες πολλών ατόμων. Ουσιαστικά, το διπλό μαρκοβιανό μοντέλο κατασκευάστηκε συνδυάζοντας ανά δυο, απλά κρυφά μαρκοβιανά μοντέλα όπου κάθε απλό μοντέλο μοντελοποιεί την κίνηση ενός ατόμου. Πιο συγκεκριμένα, συνδυάστηκαν οι κρυφές καταστάσεις δυο διαφορετικών HMM προσδιορίζοντας τις εξαρτήσεις τους. Το τελικό σύστημα που κατασκευάστηκε αναγνώρισε σύνθετες αλληλεπιδράσεις μεταξύ δυο ατόμων, όπως συνδυασμοί των δραστηριοτήτων «προσέγγιση», «συνάντηση» και «συμπόρευση».

Οι Park και Aggarwal [34] χρησιμοποίησαν ένα δυναμικό δίκτυο Bayes (dynamic Bayesian network - DBN) για την αναγνώριση χειρονομιών δυο αλληλεπιδρώντων ατόμων. Με το σύστημά τους αναγνώρισαν χειρονομίες όπως «τεντώνω το χέρι» και «στρέφω το

κεφάλι αριστερά», κατασκευάζοντας ένα δενδροειδές δυναμικό δίκτυο Bayes. Ένα DBN αποτελεί μία προέκταση ενός HMM. Στην εργασία των Park και Aggarwal, μια χειρονομία μοντελοποιείται ως μεταβάσεις καταστάσεων των κρυμμένων κόμβων (π.χ. στάσεις μερών του σώματος) από ένα χρονικό σημείο σε άλλο χρονικό σημείο. Κάθε στάση παράγει ένα σύνολο χαρακτηριστικών που σχετίζονται με το αντίστοιχο μέρος του σώματος.

Οι Natarajan και Nevatia [35] ανέπτυξαν έναν αποδοτικό αλγόριθμο αναγνώρισης χρησιμοποιώντας διπλά κρυφά ημι-μαρκοβιανά μοντέλα, επεκτείνοντας το διπλό κρυφό μαρκοβιανό μοντέλο που αναφέρθηκε παραπάνω, μοντελοποιώντας με σαφήνεια τη χρονική διάρκεια παραμονής μιας δραστηριότητας σε κάθε κατάσταση. Η μοντελοποίηση της χρονικής διάρκειας της δραστηριότητας οδήγησε σε αποτελέσματα μεγαλύτερης ακρίβειας σε σύγκριση με τα αποτελέσματα άλλων πιο απλών στατιστικών μοντέλων.

2.2.1.2.3 Σύγκριση

Γενικά, οι ακολουθιακές προσεγγίσεις λαμβάνουν υπ'όψιν την ακολουθιακή σχέση μεταξύ των χαρακτηριστικών σε αντίθεση με τις προσεγγίσεις χώρου-χρόνου, και γι'αυτό καθιστούν δυνατή την αναγνώριση πιο σύνθετων δραστηριοτήτων.

Επιχειρώντας μια σύγκριση μεταξύ των ακολουθιακών προσεγγίσεων, παρατηρούμε ότι οι μέθοδοι που βασίζονται σε πρότυπα *(exemplar-based)* παρέχουν μεγαλύτερη ευελιξία σε σχέση με τις μεθόδους που στηρίζονται σε μοντέλα κατάστασης *(state-based)*, διότι μπορούν να διατηρούν πολλά δείγματα ακολουθιών που είναι πιθανώς πολύ διαφορετικά μεταξύ τους. Επιπροσθέτως, ο αλγόριθμος dynamic time warping DTW που χρησιμοποιείται συνήθως στα exemplar-based συστήματα παρέχει μια μη γραμμική μεθοδολογία ταιριάσματος, η οποία λαμβάνει υπ'όψιν τις διαφοροποιήσεις στο ρυθμό εκτέλεσης. Τέλος, οι μέθοδοι που βασίζονται σε πρότυπα αποδίδουν και με λιγότερα δεδομένα εκπαίδευσης.

Στον αντίποδα, οι μέθοδοι που βασίζονται σε μοντέλα κατάστασης μπορούν να υπολογίσουν την μεταγενέστερη πιθανότητα να συμβεί μια δραστηριότητα, παρέχοντας στο σύστημα τη δυνατότητα να την συγχωνεύσει με άλλες αποφάσεις. Η κυριότερη αδυναμία αυτών των μεθόδων, είναι ότι απαιτούν μεγάλο αριθμό δεδομένων για εκπαίδευση (video) αφού οι δραστηριότητες που σκοπεύουν να αναγνωρίσουν είναι περισσότερο σύνθετες.

2.2.2 Ιεραρχικές μέθοδοι

Ο πυρήνας των ιεραρχικών προσεγγίσεων στο πρόβλημα της αναγνώρισης ανθρώπινης δραστηριότητας είναι η αναγνώριση δραστηριοτήτων υψηλού επιπέδου με βάση τα αποτελέσματα αναγνώρισης άλλων απλούστερων δραστηριοτήτων. Παραδείγματος χάριν, μια αλληλεπίδραση υψηλού επιπέδου, όπως «παλεύω» μπορεί να αναγνωριστεί εντοπίζοντας μια ακολουθία από πράξεις όπως «γρονθοκοπώ» και «κλωτσώ». Επομένως, στις ιεραρχικές μεθόδους αναγνώρισης μια σύνθετη δραστηριότητα αναπαρίσταται ως ένα σύνολο από επιμέρους συμβάντα (subevents) έως ότου προκύψουν πολύ απλές κινήσεις.

Στις περισσότερες ιεραρχικές προσεγγίσεις, οι απλές ενέργειες (atomic or primitive actions) αναγνωρίζονται εφαρμόζοντας μεθοδολογίες αναγνώρισης μονής στιβάδας (single-layered). Για παράδειγμα, οι χειρονομίες «τεντώνω το χέρι» και «σηκώνω το χέρι» συμβαίνουν συχνά στην ανθρώπινη δραστηριότητα αποτελώντας έτσι, καλό παράδειγμα απλών ενεργειών για την αναπαράσταση ανθρώπινων δραστηριοτήτων όπως «χειραψία» ή «γρονθοκόπηση». Μέθοδοι μονής στιβάδας, όπως οι ακολουθιακές με τη χρήση κρυφών μαρκοβιανών μοντέλων, μπορούν να εφαρμοστούν για την αναγνώριση παρόμοιων χειρονομιών.

Το κυριότερο πλεονέκτημα των ιεραρχικών προσεγγίσεων είναι η ικανότητα αναγνώριση σύνθετων δομών. Ως εκ τούτου, είναι απολύτως κατάλληλες για την ανάλυση σε σημασιολογικό επίπεδο της αλληλεπίδρασης μεταξύ ατόμων ή/και αντικειμένων, καθώς και πολύπλοκων ομαδικών δραστηριοτήτων. Σε αυτή την ικανότητα συντελούν δυο στοιχεία των ιεραρχικών προσεγγίσεων: πρώτον, η ικανότητα να λειτουργούν με λίγα δεδομένα εκπαίδευσης και δεύτερον, η δυνατότητα να ενσωματώνουν προγενέστερη γνώση στην αναπαράσταση.

Κατ'αρχήν, ο όγκος των δεδομένων εκπαίδευσης που απαιτούν τα ιεραρχικά μοντέλα αναγνώρισης είναι σημαντικά μικρότερος από τον αντίστοιχο όγκο που απαιτείται στα μοντέλα μονής στιβάδας. Παραδείγματος χάριν, τα κρυφά μαρκοβιανά μοντέλα μονής στιβάδας, απαιτούν να μάθουν ένα μεγάλο αριθμό πιθανοτήτων μετάβασης και παρατήρησης, εφόσον ο αριθμός των κρυφών καταστάσεων αυξάνεται όσο οι δραστηριότητες γίνονται πιο σύνθετες. Περικλείοντας δομικά πολυάριθμες υποενέργειες που διαμοιράζονται ανάμεσα στις σύνθετες δραστηριότητες, οι ιεραρχικές προσεγγίσεις μοντελοποιούν τις δραστηριότητες με πολύ μικρότερο όγκο δεδομένων για εκπαίδευση και αναγνωρίζουν τις κινήσεις πιο αποτελεσματικά.

Επιπροσθέτως, η ενσωμάτωση προγενέστερης πληροφορίας στο σύστημα αναγνώρισης διευκολύνεται από την ιεραρχική μοντελοποίηση των δραστηριοτήτων υψηλού επιπέδου. Η ανθρώπινη γνώση μπορεί να συμπεριληφθεί στο σύστημα απαριθμώντας σημαντικές σημασιολογικά υποδραστηριότητες, που συνθέτουν μια δραστηριότητα υψηλού επιπέδου και/ή προσδιορίζοντας τις σχέσεις τους. Κατά την μοντελοποίηση των σύνθετων δραστηριοτήτων, οι μη ιεραρχικές τεχνικές τείνουν να χρησιμοποιούν πολύπλοκες δομές και χαρακτηριστικά τα οποία είναι δύσκολο να ερμηνευθούν αποτρέποντας την ενσωμάτωση προγενέστερης γνώσης. Από την άλλη μεριά, οι ιεραρχικές τεχνικές μοντελοποιούν την υψηλού επιπέδου δραστηριότητα σαν έναν οργανισμό από σημασιολογικά ερμηνευμένα υποσυμβάντα, καθιστώντας την ενσωμάτωση προγενέστερης γνώσης πιο εύκολη.

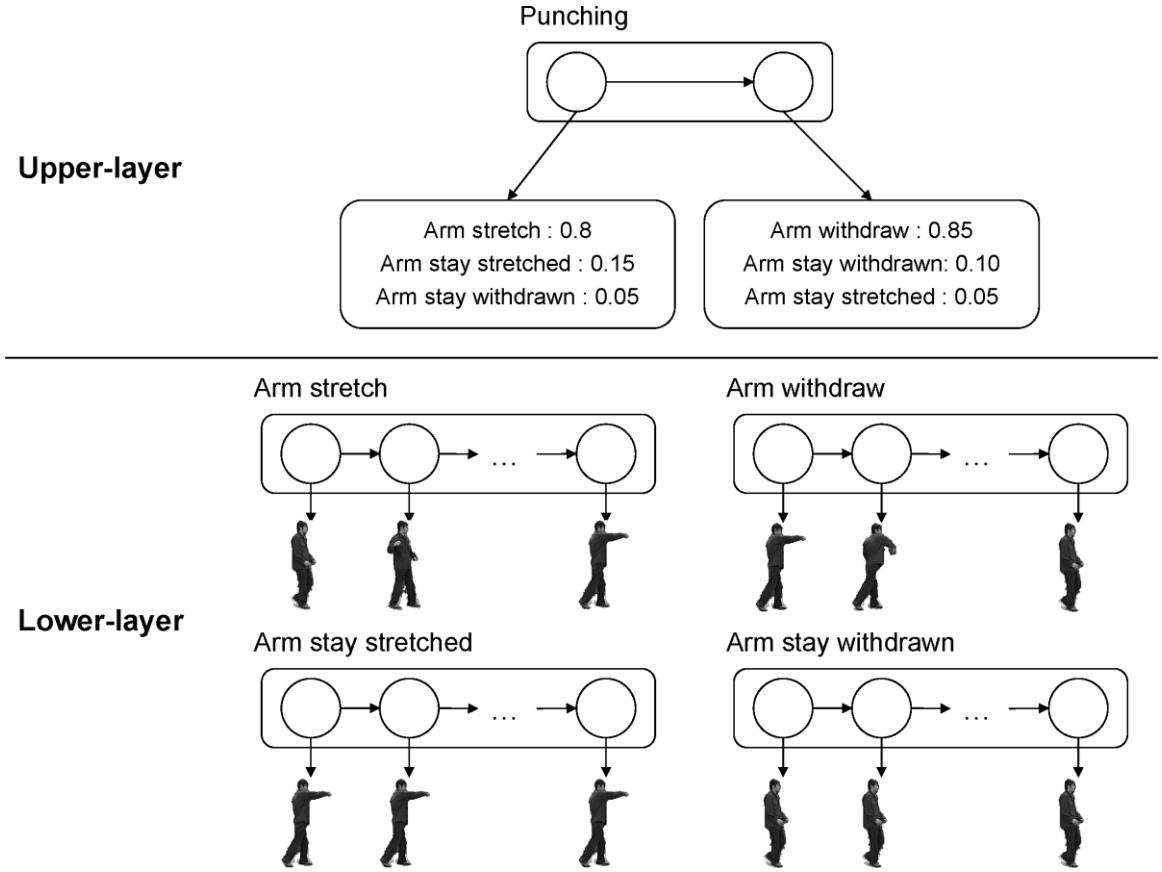
Οι ιεραρχικές μέθοδοι κατατάσσονται σε τρεις υποκατηγορίες με βάση τον τρόπο προσέγγισης: τις στατιστικές, τις συντακτικές και τις βασισμένες στην περιγραφή προσεγγίσεις.

2.2.2.1 Στατιστικές μέθοδοι

Η αναγνώριση ανθρώπινης δραστηριότητας μέσω των στατιστικών (statistical) μεθοδολογιών πραγματοποείται με τη χρήση στατιστικών μοντέλων καταστάσεων. Συγκεκριμένα, χρησιμοποιούνται πολλαπλά επίπεδα τέτοιων στατιστικών μοντέλων, όπως κρυφά Μαρκοβιανά μοντέλα *(hidden Markov models – HMMs)* και δυναμικά δίκτυα Bayes *(dynamic Bayesian networks – DBNs)* τα οποία στοχεύουν στην αναγνώριση δραστηριοτήτων με ακολουθιακή δομή. Στα χαμηλότερα επίπεδα, οι απλές ή ατομικές κινήσεις αναγνωρίζονται από ακολουθίες χαρακτηριστικών διανυσμάτων, όπως ακριβώς στις ακολουθιακές προσεγγίσεις μονής στιβάδας. Τα μοντέλα δευτέρου επιπέδου μεταχειρίζονται την ακολουθία των ατομικών κινήσεων ως παρατηρήσεις που δημιουργούνται από αυτά. Για κάθε μοντέλο, υπολογίζεται η πιθανότητα να δημιουργήσει μια ακολουθία από παρατηρήσεις και με αυτόν τον τρόπο, μετράται η ομοιότητα ανάμεσα σε μια δραστηριότητα και την ακολουθία εικόνων που έχει δοθεί ως είδοσος στο σύστημα. Το σχήμα 2.10 παρουσιάζει ένα παράδειγμα ενός τέτοιου μοντέλου για την αναγνώριση της δραστηριότητας «γρονθοκοπώ».

Οι Oliver et al. [36] εισήγαγαν μια από τις θεμελειώδεις μορφές των ιεραρχικών στατιστικών προσεγγίσεων, το πολυεπίπεδο κρυφό Μαρκοβιανό μοντέλο *(layered hidden Markov model – LHMM)*, η λειτουργία του οποίου παρουσιάστηκε στην προηγούμενη παράγραφο. Όπως γίνεται σαφές από την ίδια τη φύση του μοντέλου, όλα τα επιμέρους συμβάντα μιας δραστηριότητας θα πρέπει να είναι αυστηρώς διαδοχικά σε κάθε LHMM και

κάθε επίπεδο του HMM εκπαιδεύεται ξεχωριστά με πλήρως προσδιορισμένα δεδομένα καθιστώντας δυνατή την επανεκπαίδευση. Το μοντέλο αυτό αναγνώρισε με επιτυχία αλληλεπιδράσεις μεταξύ ανθρώπων σε μια αίθουσα συσκέψεων, συμπεριλαμβανομένων δραστηριοτήτων όπως, «ένα άτομο πραγματοποιεί μια παρουσίαση» και «συζήτηση πρόσωπο με πρόσωπο».



Σχήμα 2.10: Παράδειγμα ενός ιεραρχικού HMM για την αναγνώριση της κίνησης «γρονθοκοπώ». Το μοντέλο αποτελείται από δυο επίπεδα HMM..

Το παραπάνω παράδειγμα των μιμήθηκαν πολλοί ερευνητές, όπως ο Nguyen [37], οι οποίοι κατασκεύασαν ένα HMM δυο επιπέδων για την αναγνώριση πολύπλοκων ακολουθιακών δρατηριοτήτων, (π.χ. «ένα άτομο παίρνει ένα γεύμα» και «ένα άτομο τρώει ένα snack»). Ομοίως, οι Zhang et al.[38] χρησιμοποίησαν ένα πολυεπίπεδο HMM, που αποτελείται από HMM δυο επιπέδων για την αναγνώριση ομαδικών δραστηριοτήτων σε μια αίθουσα συνεδριάσεων.

Όπως αναφέρθηκε και προηγουμένως, ο ρόλος των *DBNs (dynamic Bayesian networks)* στις ιεραρχικές προσεγγίσεις είναι βασικός. Τα DBNs περιέχουν πολλαπλά επίπεδα κρυφών καταστάσεων τα οποία μπορούν να αναπαραστήσουν ιεραρχικές ανθρώπινες δραστηριότητες. Οι Gong και Xiang [39] επέκτειναν τα συμβατικά HMMs για την κατασκευή δυναμικών πιθανοτικών δικτύων (DPNs) με απώτερο σκοπό την αναπαράσταση δραστηριοτήτων με πολλούς συμμετέχοντες, όπως η τοποθέτηση και η αφαίρεση φορτίου σε φορτηγό όχημα. Επίσης, οι Dai et al.[40] χρησιμοποίησαν τα DBNs για την αναγνώριση ομαδικών δραστηριοτήτων. Παραδείγματος χάριν, «το διάλειμμα», «η παρουσίαση» και «η συζήτηση» αναγνωρίστηκαν με βάση απλές ενέργειες, όπως «μιλώ», «ρωτώ» κ.ο.κ. Τέλος οι Damen και Hogg [41], κατασκεύασαν Bayesian δίκτυα με τη χρήση αλυσίδας Monte Carlo του Markov (MCMC) για την αναγνώριση δραστηριοτήτων σχετικές με το ποδήλατο.

Η χρήση propagation networks (P-net) στις ιεραρχικές προσεγγίσεις προτάθηκε από τους Shi et al.[42]. Ένα P-net διαθέτει δομή παρόμοια με εκείνη ενός HMM, αφού μια δραστηριότητα αναπαρίσταται με πολλαπλούς κόμβους καταστάσεων, τις πιθανότητες μετάβασης από κόμβο σε κόμβο και της πιθανότητες παρατήρησης. Επιπροσθέτως, τα P-nets αποσυνθέτουν τις ενέργειες σε ατομικές κινήσεις και κατασκευάζουν το δίκτυο που περιγράφει την προσωρινή μεταξύ τους σειρά. Η κυριότερη διαφορά μεταξύ ενός P-net και ενός HMM είναι ότι τα P-nets επιτρέπουν την ενεργοποίηση πολλών κόμβων κατάστασης ταυτόχρονα. Η σημασία της ικανότητας αυτής των P-nets είναι μεγάλη, διότι έτσι επιτρέπεται η μοντελοποίηση δραστηριοτήτων υψηλού επιπέδου που αποτελούνται από υπο-συμβάντα που συμβαίνουν όχι μόνο διαδοχικά αλλά και ταυτόχρονα. Η εφαρμογή του P-net στην πράξη, είχε ως αποτέλεσμα την επιτυχή αναγνώριση την δραστηριότητα «διεξαγωγή ενός χημικού πειράματος».

Συμπερασματικά, οι στατιστικές προσεγγίσεις είναι κατάλληλες για την αναγνώριση ακολουθιακών δραστηριοτήτων. Υπό την προυπόθεση ότι υπάρχουν αρκετά δεδομένα για την εκπαίδευση, τα στατιστικά μοντέλα είναι σε θέση να προσδιορίζουν με αξιοπιστία δραστηριότητες ακόμη και αν υπάρχει θόρυβος στην είσοδο. Παρ'όλα αυτά, το σημαντικότερο μειονέκτημά τους είναι πως αδυνατούν να αναγνωρίσουν δραστηριότητες που αποτελούνται από ταυτόχρονα υπο-συμβάντα.

2.2.2.2 Συντακτικές μέθοδοι

Σε αυτή την κατηγορία ιεραρχικών μεθοδολογιών ανήκουν εκείνες που μοντελοποιούν τις ανθρώπινες δραστηριότητες ως συμβολοσειρές, όπου κάθε σύμβολο ανταποκρίνεται σε μια απλή ατομική κίνηση. Όπως και στις στατιστικές μεθοδολογίες που παρουσιάστηκαν προηγουμένως έτσι και στις συντακτικές (syntactic), απαιτείται αρχικά η αναγνώριση απλών κινήσεων μέσω των τεχνικών που αναφέρθηκαν. Σε γενικές γραμμές, η ανθρώπινη δραστηριότητα παρουσιάζεται σαν ένα σύνολο από κανόνες παραγωγής που δημιουργούν μια συμβολοσειρά από ατομικές ενέργειες και αναγνωρίζεται υιοθετώντας τεχνικές συντακτικής ανάλυσης από το πεδίο των γλωσσών προγραμματισμού. Γι'αυτόν τον σκοπό, οι ερευνητές χρησιμοποίησαν γραμματικές χωρίς συμφραζόμενα (context- free grammars -CFGs) και στοχαστικές γραμματικές χωρίς συμφραζόμενα (stochastic context-free grammars –SCFGs).

Οι Ivanon και Bobick [43] πρότειναν μια ιεραρχική μέθοδο με SCFG για την αναγνώριση πολύπλοκων ανθρώπινων κινήσεων. Ειδικότερα, κωδικοποίησαν ένα σημαντικό αριθμό από στοχαστικούς κανόνες παραγωγής για να εκφράσουν όλες τις πιθανές δραστηριότητες στο περιβάλλον ενδιαφέροντος. Τα HMM υψηλότερου επιπέδου αναλύουν συντακτικά μια συμβολοακολουθία από ατομικές ενέργειες που δημιουργήθηκε από τα χαμηλότερου επιπέδου HMM, αναγνωρίζοντας με αυτόν τον πιθανοτικό τρόπο διάφορες δραστηριότητες. Οι Moore και Essa [44] επέκτειναν την παραπάνω μέθοδο, αυξάνοντας την ακρίβεια και βελτιώνοντας την ανίχνευση λάθους. Αντιθέτως, το σύστημα ΑΑΔ των Minnen et. al.[45] επικεντρώθηκε στο πρόβλημα κατάτμησης πολλαπλών αντικειμένων, αποδεικνύοντας ότι η σημασιολογική επεξεργασία των κινήσεων με CFGs μπορεί να συμβάλλει στην αναγνώριση αντικειμένων. Εδώ εισάγεται για πρώτη φορά η ιδέα της παραίσθησης *(hallucination)* στην κατανόηση των αποτυχιών στον σαφή προσδιορισμό των απλών – ατομικών ενεργειών.

Επέκταση της γραμματικής SCFG επιχείρησαν οι Joo και Chellappa [46], δημιουργώντας μια Γραμματική Ιδιοτήτων (Attribute grammar) η οποία προσθέτει σημασιολογικές ετικέτες και συνθήκες στους κανόνες παραγωγής της γραμματικής SCFG. Η γραμματική αυτή έχει την ικανότητα να περιγράψει τους περιορισμούς στα χαρακτηριστικά αλλά και στο χρόνο που αντιστοιχούν στις ατομικές κινήσεις. Αυτό συνεπάγεται ότι μόνο

όταν ικανοποιούνται ταυτόχρονα η σύνταξη της SCFG και οι περιορισμοί αποφασίζεται από το σύστημα ότι μια δραστηριότητα έχει συμβεί. Η εφαρμογή ενός τέτοιου συστήματος σε ένα χώρο στάθμευσης παραδείγματος χάριν, μπορεί να οδηγήσει στο διαχωρισμό των φυσιολογικών δραστηριοτήτων, από τις μη φυσιολογικές.

Όσον αφορά στα μειονεκτήματα των συντακτικών προσεγγίσεων αξίζει να σημειωθεί ότι υπάρχει αδυναμία στην αναγνώριση δραστηριοτήτων που συμβαίνουν παράλληλα. Εφόσον, οι συντακτικές προσεγγίσεις μοντελοποιούν την υψηλού επιπέδου δραστηριότητα σαν συμβολοακολουθία από άλλες ατομικές κινήσεις, η χρονική διάταξη των τελευταίων πρέπει να είναι αυστηρώς ακολουθιακή.

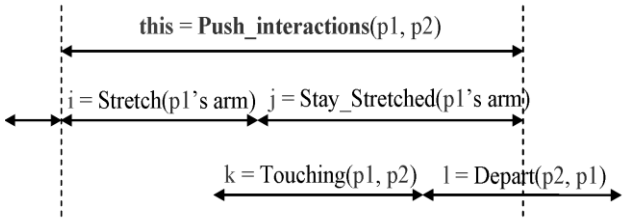
2.2.2.3 Περιγραφικές μέθοδοι

Οι περιγραφικές (description-based) ιεραρχικές μεθοδολογίες αναπαριστούν την ανθρώπινη δραστηριότητα υψηλού επιπέδου ως ένα σύνολο από πιο απλές κινήσεις (π.χ. υπο-γεγονότα), περιγράφοντας τις χωρικές, χρονικές και λογικές τους συσχετίσεις. Επομένως, η αναγνώριση μιας δραστηριότητας επιτυγχάνεται ψάχνοντας όλα τα υπο-γεγονότα που έχουν προσδιοριστεί κατά την αναπαράστασή της. Επίσης, είναι απαραίτητο να αναφερθεί πως οι περιγραφικές προσεγγίσεις μπορούν να διαχειριστούν δραστηριότητες με υποσυμβάντα που λαμβάνουν χώρα παράλληλα.

Θεμελιώδες στοιχείο των περιγραφικών τεχνικών αποτελούν τα χρονικά διαστήματα *(time intervals)* που χαρακτηρίζουν τα υπο-συμβάντα και καθορίζουν τις χρονικές σχέσεις μεταξύ τους. Προκειμένου να καθοριστούν αυτές οι σχέσεις, ορίστηκαν τα χρονικά κατηγορήματα του Allen[47], τα οποία είναι τα εξής : *before, meets, overlaps, during, starts, finishes* and *equals*. Αξίζει να σημειωθεί πως τα κατηγορήματα before και meets περιγράφουν διαδοχικές σχέσεις, ενώ τα υπόλοιπα ορίζουν σύγχρονες σχέσεις. Στο Σχήμα 2.11 (a) περιγράφεται μέσω χρονικών διαστημάτων η χρονική δομή της αλληλεπίδρασης «σπρώχνω».

Στις περιγραφικές μεθόδους χρησιμοποιείται επίσης, η γραμματική CFG ως τυπική σύνταξη για την αναπαράσταση της δραστηριότητας παρόλο που χρησιμοποιείται με διαφορετικό τρόπο από ότι στις συντακτικές προσεγγίσεις. Πιο συγκεκριμένα, στις συντακτικές προσεγγίσεις υπονοείται ότι οι ίδιες οι CFGs περιγράφουν τη σημασιολογία της δραστηριότας. Στον αντίποδα στις περιγραφικές μεθοδολογίες, η CFG θεωρείται ένα συντακτικό για την τυπική αναπαράσταση της δραστηριότητας ενώ η σημασιολογία της κωδικοποιείται με μια δομή παρόμοια με τις γλώσσες προγραμματισμού Σχήμα 2.11 (b). Επομένως, η CFG απλώς διαβεβαιώνει πως η αναπαράσταση της δραστηριότητας πληροί τους κανόνες της γραμματικής.

Οι Pinhanez και Bobick[48] πρότειναν ένα δίκτυο {παρελθόν, παρόν, μέλλον} (PNF-network), όπου τα υπο-γεγονότα ορίζονται ως κόμβοι και οι χρονικές σχέσεις μεταξύ τους περιγράφονται με ακμές. Επιπλέον, ανέπτυξαν έναν αλγόριθμο πολυωνυμικού χρόνου για την επεξεργασία του δικτύου και το εφήρμοσαν επιτυχώς στο περιβάλλον μιας κουζίνας σε δραστηριότητες μαγειρικής. Λίγο αργότερα, οι Intille και Bobick [49] χρησιμοποίησαν μια περιγραφική μέθοδο για την ανάλυση αγώνων Αμερικανικού ποδοσφαίρου, αναπαριστώντας την ανθρώπινη δραστηριότητα σε τρια επίπεδα: atomic, individual και team- level.



(a)

```
Push_interaction(p1, p2) = {
  list( def('i', Stretch(p1's arm)),
      list( def('j', Stay_stretched(p2's arm)),
          list( def('k', Touching(p1, p2)),
              def('l', Depart(p2, p1)) ))),
  and( meets('i', 'j'),
      and( during('k', 'this'),
          and( meets('k', 'l'),
              and( starts('i', 'this'),
                  finishes('j', 'this') ))))
};
```

(b)

Σχήμα 2.11: (a) Χρονικά διαστήματα της αλληλεπίδρασης «σπρώχνω» και των υπο-ενεργειών και (b) η αναπαράσταση τους σε γλώσσα προγραμματισμού.

Αξίζει να αναφερθεί η γλώσσα προγραμματισμού VERL που σχεδιάστηκε από τους Nevatia et al. [50], με σκοπό την περιγραφή της ανθρώπινης δραστηριότητας. Αρχικά όρισαν τρια επίπεδα ιεραρχίας για τις δραστηριότητες: «απλό γεγονός», «σύνθετο γεγονός» και «πολυνηματικό σύνθετο γεγονός». Στη συνέχεια, χρησιμοποιήθηκαν τα κατηγορήματα Allen και λογικά κατηγορήματα για την περιγραφή τους και τέλος, για την αναγνώριση Bayesian δίκτυα και HMM.

Το 2006 οι Ryoo και Aggarwall [51] πρότειναν μια προσέγγιση του προβλήματος όπου η GFG γραμματική τους επιτρέπει την αναπαράσταση της ανθρώπινων αλληλεπιδράσεων οποιουδήποτε ιεραρχικού επιπέδου και την περιγράφει ως λογικές πράξεις (and, or και not) ανάμεσα στις χρονικές και χωρικές σχέσεις των υπο-γεγονότων. Τέλος, υπήρξαν προσπάθειες για την χρήση τεχνικών τεχνητής νοημοσύνης στην αναγνώριση της ανθρώπινης δραστηριότητας όπως αυτή των Tran και Davis [52], οι οποίοι υιοθέτησαν Μαρκοβιανά λογικά δίκτυα για να αναγνωρίσουν ανθρώπινη δραστηριότητα σε χώρο στάθμευσης. Η προσπάθειά τους όμως, απείχε από την αναγνώριση δυναμικής αλληλεπίδρασης μεταξύ των δραστών.

2.2.2.4 Σύγκριση

Με βάση τα παραπάνω μπορούμε να συμπεράνουμε πως οι ιεραρχικές προσεγγίσεις είναι κατάλληλες για την αναγνώριση σύνθετων που αποτελούνται από επιμέρους πιο απλές κινήσεις. Επιπλέον, απαιτούν λιγότερα δεδομένα για την εκπαίδευσή τους. Επιχειρώντας μια σύγκριση όλων των ιεραρχικών προσεγγίσεων καταλήγουμε σε κάποιες διαπιστώσεις.

Πρώτον, οι στατιστικές και οι συντακτικές μεθοδολογίες παρέχουν ένα πιθανοτικό πλαίσιο για την αναγνώριση δραστηριοτήτων με αρκετό θόρυβο στην είσοδο.Παρ'όλα αυτά, αντιμετωπίζουν δυσκολία στην αναγνώριση κινήσεων που περιέχουν υπο-γεγονότα που συμβαίνουν παράλληλα.

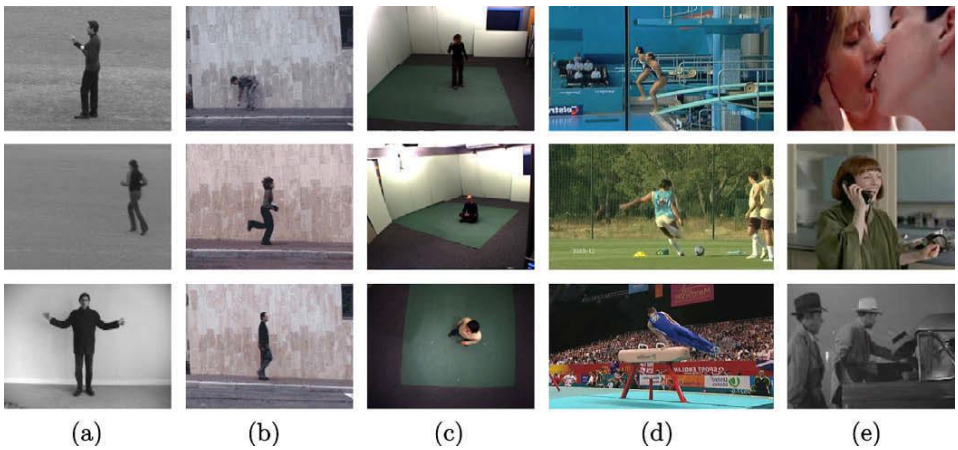
Στον αντίποδα, οι περιγραφικές μέθοδοι μπορούν να αναπαραστήσουν και να αναγνωρίσουν δραστηριότητες με πολύπλοκες χρονικές δομές, και να χειριστούν υπο-γεγονότα που συμβαίνουν διαδοχικά αλλά και ταυτόχρονα. Ωστόσο, παρουσιάζουν ένα σημαντικό μειονέκτημα που αφορά στην αποτυχία αναγνώρισης των πιο απλών κινήσεων που συνθέτουν μια δραστηριότητα (π.χ. αποτυχία αναγνώρισης χειρονομιών).

2.3 Βάσεις δεδομένων

2.3.1 Εισαγωγή

Η μεγάλη αξία των βάσεων δεδομένων ή αλλιώς datasets που περιέχουν καταγεγραμμένες ανθρώπινες δραστηριότητες σε videos, είναι αδιαμφισβήτητη διότι αποτελούν ένα κοινό κριτήριο για την μέτρηση και τη σύγκριση της ακρίβειας που προσφέρουν οι διάφορες μεθοδολογίες για την αναγνώριση της ανθρώπινης δραστηριότητας. Επομένως, η κατασκευή μιας τέτοιας βάσης δεδομένων συμβάλλει στην πρόοδο που συντελείται στην έρευνα για την αναγνώριση της ανθρώπινης δραστηριότητας.

Γενικά, οι βάσεις δεδομένων που αφορούν ανθρώπινες δραστηριότητες που είναι διαθέσιμες στο κοινό μπορούν να χωριστούν σε τρεις κατηγορίες. Στην πρώτη κατηγορία περιλαμβάνονται βάσεις δεδομένων όπως η βάση KTH [53] και η βάση Weizmann [54] που σχεδιάστηκαν για την ακαδημαική αξιολόγηση συστημάτων αναγνώρισης κινήσεων γενικού σκοπού. Περιλαμβάνουν video διαφόρων ατόμων που πραγματοποιούν απλές κινήσεις, όπως «περπατώ» και «γνέφω» σε ένα ελεγχόμενο περιβάλλον. Τα σύνολα δεδομενων της δεύτερης κατηγορίας είναι περισσότερο προσανατολισμένα στις εφαρμογές και προκύπτουν από ρεαλιστικά περιβάλλοντα, όπως αεροδρόμια. Παραδείγματος χάριν, το σύνολο δεδομένων PETS περιέχει δραστηριότητες όπως «κλοπή αποσκευών» και «πάλη» και στοχεύει σε εφαρμογές επιτήρησης. Τέλος, έχουν κατασκευαστεί και παρουσιαστεί βάσεις που προέκυψαν από τη συλλογή πραγματικών video από μέσα ενημέρωσης, όπως τηλεοπτικές εκπομπές και ταινίες. Στο σχήμα 2.12 απεικονίζονται στιγμιότυπα από δημοφιλή datasets.



Σχήμα 2.12: (a) KTH dataset, (b) Weizmann dataset, (c) Inria XMAS dataset, (d) UCF sports action dataset και (e) Hollywood human action dataset.

2.3.2 Σημαντικές Βάσεις δεδομένων

Ένα σημαντικό πλήθος ερευνητών έχουν δοκιμάσει το σύστημά τους στη βάση KTH. Η βάση KTH περιέχει 2391 videos, έξι δραστηριοτήτων εκτελεσμένων από 25 άτομα. Οι δραστηριότητες που περιλαμβάνονται είναι: Walking, jogging, running, boxing, hand-waving και hand-clapping. Τα videos έχουν καταγραφεί σε ελαφρώς διαφορετικές κλίμακες τόσο σε

εξωτερικό όσο και σε εσωτερικό περιβάλλον, με μόνο ένα άτομο στο σκηνικό. Κάθε video περιέχει επαναλαμβανόμενες εκτελέσεις μιας κίνησης σε ανάλυση 160x120, 25fps.

Η βάση Weizmann αποτελείται από 10 διαφορετικές κινήσεις, εκτελεσμένες από 9 διαφορετικά άτομα και περιέχει επομένως 90 videos. Εδώ το σκηνικό είναι απλό και παραμένει ίδιο για όλα τα video. Οι 10 κινήσεις που περιέχει είναι: running, walking, jumping-jack, jumping forward on two legs, skip, jumping in place on two legs, galloping sideways, waving one hand, waving two hands και bending**.** Η ανάλυση των videos είναι 180x144, 25fps και μόνο ένα άτομο εμφανίζεται κάθε φορά στο σκηνικό.

Ωφέλιμο είναι να επισημανθεί ότι οι παραπάνω βάσεις σχεδιάστηκαν για να αξιολογήσουν την ικανότητα ταξινόμησης των συστημάτων σε απλές κινήσεις. Γι' αυτόν το λόγο άλλωστε, σε κάθε video περιέχονται εκτελέσεις μιας απλής κίνησης και σκοπός είναι να αναγνωριστεί η κλάση της κίνησης του video, δεδομένου ότι αυτό ανήκει σε ένα περιορισμένο αριθμό γνωστών κλάσεων. Η δοκιμή μεθοδολογιών που χρησιμοποιούν χωροχρονικά τοπικά χαρακτηριστικά (ενότητα 2.2.1.1.3) είναι δημοφιλής, καθώς δεν απαιτούν απομόνωση του μπροστινού σκηνικού και αντιμετωπίζουν ικανοποιητικά τις αλλαγές στην κλίμακα. Επιπροσθέτως, είναι κατάλληλες για περιοδικές κινήσεις όπως όλες οι κινήσεις που αναφέρθηκαν προηγουμένως, αν εξαιρεθεί η κίνηση bend. Αυτό συμβαίνει διότι, τα χωροχρονικά χαρακτηριστικά θα εξαχθούν κατ'επανάληψη από τις περιοδικές κινήσεις.

Στον αντίποδα βρίσκονται βάσεις που στοχεύουν σε εφαρμογές που σχετίζονται με την επιτήρηση (surveillance datasets). Τα σύνολα δεδομένων PETS που έγιναν διαθέσιμα στα συνέδρια PETS 2004, 2006, 2007, αλλά και άλλα παρόμοια ,όπως το i-Lids αποτελούνται από ρεαλιστικά video σε μη ελεγχόμενα περιβάλλοντα, όπως σιδηροδρομικοί σταθμοί και αεροδρόμια. Η οπτική γωνία της κάμερας είναι παρόμοια με αυτήν των κλειστών κυκλωμάτων παρακολούθησης, ενώ σε μερικές βάσεις παρέχονται πολλές κάμερες, άρα και οπτικές γωνίες. Οι κάμερες είναι σταθερές, δίνοντας την εντύπωση πως το σκηνικό είναι στατικό και η κλίμακα σχεδόν σταθερή. Ένα επιπλέον χαρακτηριστικό σε αυτές τις βάσεις, είναι η ταυτόχρονη παρουσία πολλών ατόμων και αντικειμένων στο σκηνικό. Βασικός στόχος των βάσεων επιτήρησης είναι η αξιολόγηση της ικανότητας των συστημάτων αναγνώρισης να αναλύσουν συγκεκριμένες δραστηριότητες που παρουσιάζουν πρακτικό ενδιαφέρον, παραδείγματος χάριν εγκατάλειψη αποσκευής ή κλοπή αποσκευής.

Η βάση PETS 2004 γνωστή και ως CAVIAR περιλαμβάνει 6 κατηγορίες δραστηριοτήτων, όπου κάθε κατηγορία αποτελείται από μια ή περισσότερες κινήσεις: walking, browsing, resting-slumping-fainting, leaving bags behind, people meeting, walking together, splitting up, fighting. Κάθε κλάση έχει 3 έως 6 videos, καταλήγοντας σε ένα σύνολο από 28 videos με ανάλυση 384x288, 25fps. Τέλος, τα videos καταγράφηκαν σε περιβάλλον καταστήματος με μια κάμερα.

Στη βάση PETS 2006, για κάθε μια από τις τέσσερις οπτικές γωνίες έχουν καταγραφεί 7 σκηνές. Η βάση εστιάζει στο πρόβλημα εγκατάλειψης αποσκευών και κάθε σκηνή περιέχει την εγκατάλειψη μιας τσάντας σε ένα σιδηροδρομικό σταθμό. Σε κάθε δραστηριότητα συμμετέχουν ένα ή δυο άτομα, ενώ διάφοροι πεζοί είναι παρόντες στο σκηνικό. Και οι τέσσερις κάμερες διαθέτουν υψηλή ανάλυση 768x576, 25fps.

Παρομοίως, η βάση PETS 2007 ασχολήθηκε με την αλληλεπίδραση ανθρώπου – αποσκευών. Τα videos ελήφθησαν σε μια αίθουσα αεροδρομίου από τέσσερις κάμερες ίδιας ανάλυσης με τις παραπάνω, καταγράφοντας δυο σκηνές general loitering, τέσσερις σκηνές κλοπής αποσκευών, και δυο σκηνές εγκατάλειψης αποσκευών.

Μια ακόμη βάση δεδομένων κίνησης που ασχολείται με το πρόβλημα της εγκατάλειψης αποσκευών είναι η βάση i-Lids. Τα videos καταγράφηκαν σε ένα υπόγειο σταθμό τρένου στο

Λονδίνο από μια οπτική γωνία, σε συνωστισμένο περιβάλλον. Η ανάλυση των videos είναι 720x576, 25fps και δεν περιέχουν μόνο άτομα και αντικείμενα, αλλά και διερχόμενα τρένα όπου κόσμος επιβιβάζεται και αποβιβάζεται. Για σκοπούς εκπαίδευσης και επαλήθευσης, δημιουργήθηκαν τρία videos ενώ ένα μεγαλύτερης χρονικής διάρκειας video με έξι δραστηριότητες εγκατάλειψης αποσκευών χρησιμοποιήθηκε για δοκιμή.

Τέλος, όπως αναφέρθηκε και στην εισαγωγή της ενότητας, υπάρχουν βάσεις που προέκυψαν από τη συλλογή video τηλεοπτικών εκπομπών και ταινιών. Οι διαφορές τους με τις υπόλοιπες βάσεις είναι πως τα video καταγράφονται σε μη ελεγχόμενο περιβάλλον, ενώ υπάρχουν γρήγορες εναλλαγές στην οπτική γωνία και συνήθως δεν παρέχεται πληροφορία για το σκηνικό. Οι περισσότερες βάσεις από ταινίες [55][56][57], επικεντρώθηκαν σε απλές κινήσεις όπως για παράδειγμα «φιλώ» και «κτυπώ». Παρ'όλο που οι κινήσεις είναι απλές, κάθε video μιας κίνησης επιδεικνύει εξάρτηση από τις αλλαγές που αφορούν στο άτομο και στην οπτική γωνία. Επομένως η μεγαλύτερη πρόκληση είναι η αντιμετώπιση αυτών των εξαρτήσεων και όχι η αναγνώριση πολύπλοκων δραστηριοτήτων.

Το σύνολο δεδομένων κίνησης Hollywood2 [58][2] προέκυψε από την συλλογή video από 69 διαφορετικές ταινίες του Hollywood, και αποτελεί επέκταση της βάσης Hollywood. Περιλαμβάνει 12 κλάσεις κινήσεων: answering the phone, driving car, eating, fighting, getting out of car, hand shaking, hugging, kissing, running, sitting down, sitting up και standing up σε 10 διαφορετικές κλάσεις σκηνών, σε συνολικά 3669 video με διάρκεια 20.1 ώρες περίπου. Στόχος της συκγκεκριμένης βάσης είναι η αξιολόγηση των διαφόρων μεθοδολογιών για την Α.Α.Δ. σε ρεαλιστικές συνθήκες.

Μια ακόμη βάση με κινήσεις που δημιουργήθηκε από τη συλλογή ρεαλιστικών videos είναι η βάση UCF sport [59][3]. Περιέχει 10 κλάσεις κινήσεων : swinging (on the pommel horse and on the floor), diving, kicking (a ball), weight-lifting, horse-riding, running, skateboarding, swinging (at the high bar), golf swinging και walking και αποτελείται συνολικά από 200 video με ανάλυση 720x480. Οι κινήσεις προέρχονται από ποικίλλες αθλητικές δραστηριότητες που έχουν παρουσιαστεί σε τηλεοπτικές εκπομπές των δικτύων BBC και ECPN.

Τέλος, το σύνολο κινήσεων YouTube [60][4] περιέχει 11 κατηγορίες κινήσεων basketball shooting, biking/cycling, diving, golf swinging, horse back riding, soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking, και walking with a do και συνολικό αριθμό 1168 video. Ειδικά αυτή η βάση αποτελεί πρόκληση εξαιτίας των μεγάλων διαφοροποιήσεων στην κίνηση της κάμερας, στην εμφάνιση και την τοποθέτηση των αντικειμένων, την κλίμακα, την οπτική γωνία, τις συνθήκες φωτισμού και στις αλλαγές του σκηνικού. Μερικά επίσης δημοφιλή σύνολα με δεδομένα κίνησης, παρουσιάζονται στο Σχήμα 2.13 μαζί με μια συνοπτική περιγραφή τους.

Ονομασία Συνόλου Δεδομένων	**Κλάσεις**	**Videos**	**Συνθήκες Καταγραφής**	**Περιγραφή δεδομένων**
UMD[26]	10	100	Εργαστήριο: 1 άτομο, πολλές επαναλήψεις	Ανάλυση 300px
IXMAS[61]	11	110	Εργαστήριο: 10 άτομα, 5	Ανάλυση 100-200 px, πολύ μικρής διάρκειας λήψεις

[2] http://lear.inrialpes.fr/data
[3] http://vision.eecs.ucf.edu/
[4] http://crcv.ucf.edu/data/UCF_YouTube_Action.php

			γωνίες λήψεις με πολλές κάμερες	
Olympic_games[62]	17	166	Video από τους Ολυμπιακούς Αγώνες	5065 frames: υψηλή διαφοροποίηση στην ίδια κλάση, σημαντική κίνηση της κάμερας, θολή εικόνα λόγω κίνησης, διαφοροποιήσεις στην εμφάνιση
UFC[63]	2	20min	Video από τηλεοπτικές εκπομπές	Αλλαγές σε εμφάνιση, γωνία λήψης, κίνηση κάμερας, ταυτόχρονη εκτέλεση από πολλά άτομα
ADL[64]	10	150	Εργαστήριο : 5 άτομα, 3 επαναλήψεις	Σύνθετες δραστηριότητες , στατικό background, υψηλή ανάλυση, 240x450px,24fps
High-Five[65]	4	300	23 διαφορετικά TV shows	30-600 frames, ρεαλιστική αλληλεπίδραση ατόμων
Youtube Olympic Sports[66]	16	800	Videos από το Youtube	Σύνθετες δραστηριότητες, 50 σενάρια για κάθε κλάση.
MSR II[67]	3	54	Σε περιβάλλον συνωστισμού	Πολλά άτομα, 203 στιγμιότυπα, 320x240px, 15fps

Σχήμα 2.13: Δημοφιλή σύνολα δεδομένων κινήσεων.

2.4 Προκλήσεις

Την τελευταία δεκαετία έχει συντελεστεί εντυπωσιακή πρόοδος στο πεδίο της ΑΑΔ. Στην παρούσα φάση, πειραματικά συστήματα διατίθενται σε αεροδρόμια και άλλα δημόσια μέρη, φαινόμενο που προβλέπεται να αυξηθεί στο μέλλον. Οι προκλήσεις, όμως, δεν παύουν να υπάρχουν. Παραδείγματος χάριν, το πρόβλημα αναγνώρισης της ανθρώπινης δραστηριότητας από μια κινούμενη πλατφόρμα, όπως ένα όχημα ή ένα μη επανδρωμένο αεροσκάφος – UAV, εισάγει νέες προκλήσεις. Ζητήματα θορύβου, εντοπισμού και διαχωρισμού προστίθενται στο υπάρχον πρόβλημα της αναγνώρισης της δραστηριότητας.

Η κατεύθυνση των μελλοντικών ερευνών υπαγορεύεται και ενθαρρύνεται οπωσδήποτε από τις εφαρμογές. Οι περισσότερο πιεστικές εφαρμογές είναι η επιτήρηση και η παρακολούθηση δημόσιων χώρων, όπως αεροδρόμια, σιδηροδρομικοί σταθμοί, νοσοκομεία, η παρακολούθηση δραστηριοτήτων από UAVs και άλλες παρόμοιες εφαρμογές. Όλες αυτές οι εφαρμογές προσπαθούν να κατανοήσουν τις δραστηριότητες ενός ατόμου ή ενός συνόλου ατόμων και προβλέπεται να απασχολούν την ερευνητική κοινότητα για πολλά χρόνια ακόμη.

Όπως αναφέρθηκε προηγουμένως, ο ταυτόχρονος διαχωρισμός και εντοπισμός διαφόρων ατόμων στα videos αποτελούν πρόβλημα δυσκολότερο απ’ό,τι φαίνεται. Η δυσκολία του έγκειται στην έλλειψη φωτισμού, τις εικόνες που περιέχουν θόρυβο και τις μετατοπίσεις της κάμερας. Έτσι, αναπτύσσονται εναλλακτικές προσεγγίσεις για τον διαχωρισμό των μερών του σώματος με βάση τους τρισδιάστατους XYT όγκους.

Ένα ακόμη ζήτημα που απασχολεί τους ερευνητές είναι η εφαρμογή των διαφόρων μεθοδολογιών σε εφαρμογές πραγματικού χρόνου. Το ζήτημα αποδεικνύεται αρκετά απαιτητικό εξαιτίας της υπολογιστικής πολυπλοκότητας των μεθόδων που έχουν αναπτυχθεί. Μια πολλά υποσχόμενη κατεύθυνση για την υλοποίηση εφαρμογών πραγματικού χρόνου

είναι η μελέτη του υλικού (hardware). Άλλωστε, οι σύγχρονες CPUs και GPUs αποτελούνται από πολλούς πυρήνες και προκαλούν τους ερευνητές να εκμεταλλευτούν τις δυνατότητές τους.

Επιπροσθέτως, ερευνώνται πολλές άλλες καινοτόμες προσεγγίσεις. Παράδειγμα αποτελεί η προσέγγιση των Veeraraghavan et. al. [26] που εκμεταλλεύεται την πιθανή ύπαρξη εικόνων (σήματα πολλών διαστάσεων) σε πολλαπλότητες (manifolds) λιγότερων διάστασεων. Έτσι, ερευνητές ασχολούνται με θέματα που αφορούν στον χαρακτηρισμό των manifolds και στη διερεύνηση των σχέσεων μεταξύ των manifolds διαφορετικών δραστηριοτήτων εκτελεσμένων από το ίδιο άτομο, καθώς και των σχέσεων μεταξύ των manifolds της ίδιας δραστηριότητας εκτελεσμένης από διαφορετικά άτομα.

Τέλος, οι ιεραρχικές προσεγγίσειςαποτελούν αντικείμενο συστηματικής έρευνας, ιδίως για την αναγνώριση δραστηριότητας πολλών ατόμων. Στο εγγύς μέλλον, προβλέπεται πως θα μελετηθούν ακόμη περισσότερο ώστε να συναντήσουν τις απαιτήσεις των συστημάτων επιτήρησης καθώς και άλλων εφαρμογών.

ΚΕΦΑΛΑΙΟ 3

Παρουσίαση της Βάσης THETIS

3.1 Γενικά

Τα τελευταία χρόνια μεγάλες συλλογές από απλές αλλά και πιο σύνθετες καταγεγραμμένες κινήσεις, όπως παρουσιάστηκαν στην ενότητα 2.3 έχουν γίνει διαθέσιμες στο κοινό. Λόγω των διαρκώς αυξανόμενων αναγκών και των σύνθετων προκλήσεων που παρουσιάζονται σε εφαρμογές αναγνώρισης ανθρώπινης δραστηριότητας η ερευνητική κοινότητα έχει την ανάγκη νέων δεδομένων για συστηματική έρευνα και ανάλυση.

Σκοπός της βάσης THETIS είναι να προσφέρει στην ερευνητική κοινότητα ένα επιπλέον σύνολο δεδομένων από καταγεγραμμένες κινήσεις με συγκεκριμένο προσανατολισμό δίνοντας ερέθισμα για έρευνα πάνω σε πιο εξειδικευμένες εφαρμογές που αφορούν στην αναγνώριση ανθρώπινης δραστηριότητας. Η βάση THETIS περιλαμβάνει 12 βασικές κινήσεις του αθλήματος της αντισφαίρισης (tennis). Ακολουθώντας μια συγκεκριμένη διαδικασία που περιγράφεται στη συνέχεια, με τη χρήση της κάμερας τρισδιάστατης λήψης Kinect καταγράφηκαν συστηματικά video μερικών ωρών, που περιέχουν συγκεκριμένες κινήσεις αντισφαίρισης εκτελεσμένες από 55 διαφορετικά άτομα. Η βάση μας στην τελική της μορφή αποτελείται από 8374 video καταγεγραμμένης κίνησης. Ελπίζουμε ότι η βάση THETIS θα αποτελέσει ένα χρήσιμο εργαλείο αξιολόγησης και ανάλυσης των αλγορίθμων που προτείνονται για την επίλυση του προβλήματος της ΑΑΔ και πιο συγκεκριμένα, για εφαρμογές gaming, αυτοματοποιημένου σχολιασμού αθλητικών events κ.α.

Αξίζει να σημειωθεί πως προσφάτως έγιναν διαθέσιμα στην ερευνητική κοινότητα μερικά σύνολα δεδομένων κίνησης καταγεγραμμένα με τη συσκευή Kinect που περιέχουν πληροφορία βάθους. Χαρακτηριστικά αναφέρουμε τη βάση MSRDailyActivity3D[5] η οποία περιλαμβάνει κινήσεις της καθημερινότητας (π.χ. «τρώω», «μιλώ στο κινητό τηλέφωνο», «διαβάζω ένα βιβλίο» κ.α.) και τη βάση δεδομένων G3D [76] που περιλαμβάνει κινήσεις σχετικές με διάφορα αθλήματα. Κανένα από τα υπάρχοντα σύνολα δεδομένων κινήσεων δεν περιλαμβάνει όλο το φάσμα των βασικών κινήσεων του αθλήματος της αντισφαίρισης.

Σε αυτό το κεφάλαιο πραγματοποιείται η εκτενής παρουσίαση της βάσης δεδομένων THETIS. Στην ενότητα 3.2 παρουσιάζονται οι λεπτομέρειες που αφορούν στις συνθήκες και στα μέσα καταγραφής των δεδομένων κίνησης. Στη συνέχεια, στην ενότητα 3.3 δίνεται μια λεπτομερής περιγραφή όλων των δεδομένων κίνησης που περιλαμβάνονται στη βάση. Τέλος, η ενότητα 3.4 παρέχει πληροφορίες για τα εργαλεία που χρησιμοποιήθηκαν για την μετατροπή των αρχείων της βάσης στην τελική τους μορφή.

[5] http://research.microsoft.com/en-us/um/people/zliu/ActionRecoRsrc/default.htm

3.2 Καταγραφή των δεδομένων κίνησης

Υπάρχουν πολλοί διαφορετικοί τρόποι για την καταγραφή των δεδομένων κίνησης που απαιτείται για τη δημιουργία μιας βάσης δεδομένων κίνησης. Κάθε τεχνολογία έχει τα δικά της πλεονεκτήματα και μεινεκτήματα. Για τη δική μας βάση THETIS, χρησιμοποιήθηκε η συσκευή ανίχνευσης-καταγραφής κίνησης KINECT[6], της MICROSOFT που δημιουργήθηκε για την παιχνιδομηχανή XBOX 3600 και εμπεριέχει τεχνολογία λογισμικού της Microsoft και ενσωματωμένη κάμερα, τεχνολογίας PrimeSense.Οι λόγοι που οδήγησαν στην επιλογή του Kinect είναι σημαντικοί. Πρώτον,η πρόσβαση στη συσκευή είναι εύκολη διότι το κόστος της είναι χαμηλό και δεύτερον παρουσιάζει αυξημένο ερευνητικό ενδιαφέρον για τον τύπο των καταγραφών της και για την ανάπτυξη 3D εφαρμογών. Περισσότερες λεπτομέρειες για τις δυνατότητες της συσκευής Kinect παρουσιάζονται στην ενότητα 3.2.1.

Προκειμένου να επιτευχθεί η καταγραφή μέσω της συσκευής Kinect, χρησιμοποιήθηκε το πλαίσιο ανοικτού λογισμικού OpenNI 1.5.2 και το μεσολογισμικό NITE 1.5.2 της PrimeSense που περιέχει τον οδηγό του 3D αισθητήρα. Ο ρόλος τους παρουσιάζεται αναλυτικά στις ενότητες 3.2.2 και 3.2.3.

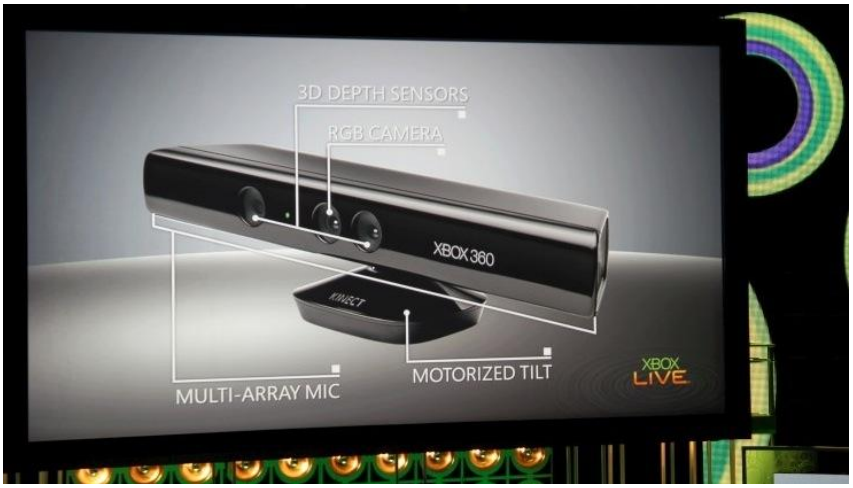
3.2.1 Συσκευή Καταγραφής

Το Kinect ως συσκευή λήψης εικόνων διαθέτει μια κάμερα RGB και μια IR κάμερα υπερύθρων με ειδικό microchip που ανιχνεύει και καταγράφει την κίνηση σε τρεις διαστάσεις. Αυτό το σύστημα τρισδιάστατης σάρωσης που ονομάζεται *Light Coding* επιτυγχάνει την ανακατασκευή της εικόνας σε τρεις διαστάσεις.

Πιο συγκεκριμένα, η συσκευή διαθέτει μια κάμερα RGB, έναν αισθητήρα βάθους (depth sensor) και μικρόφωνο για την καταγραφή ήχου. Ο αισθητήρας βάθους αποτελείται από ένα λέιζερ υπερύθρων σε συνδυασμό με ένα μονοχρωματικό αισθητήρα CMOS, που μπορεί να καταγράψει video τριων χωρικών διαστάσεων. Το εύρος του αισθητήρα προσαρμόζεται αυτόνομα από το λογισμικό που τον ρυθμίζει κατάλληλα με βάση το φυσικό περιβάλλον. Το σχήμα 3.1 επισημαίνει τα μέρη μιας συσκευής Kinect.

Ως προς τα τεχνικά χαρακτηριστικά και την ακρίβεια του Kinect πρέπει να αναφερθεί ότι καταγράφει video με frame rate 30 Hz, ενώ η ανάλυση του καναλιού RGB είναι 8-bit VGAC, 680x480 pixels και μπορεί να φτάσει και 1280x1024 σε χαμηλότερο frame rate. Το μονοχρωματικό κανάλι που καταγράφει το βάθος έχει ανάλυση 680x480 pixels και παρέχει 2048 επίπεδα ευαισθησίας. Ακόμη, υπάρχει η επιλογή της αποθήκευσης της εικόνας από την κάμερα υπερύθρων ως video σε ανάλυση 680x480 pixels ή 1280x1024 σε μικρό fps.

[6] http://www.microsoft.com/en-us/kinectforwindows/develop/resources.aspx



Σχήμα 3.1: Διαφάνεια από συνέδριο της Microsoft που παρουσιάζει τις τεχνολογίες που υποστηρίζει το Kinect.

Για τη σύμφωνη με τις προδιαγραφές λειτουργία της συσκευής Kinect υπάρχουν κάποιοι περιορισμοί ως προς την απόσταση μεταξύ της κάμερας και του αντικειμένου-υποκειμένου που καταγράφει. Ειδικότερα, η απόσταση ιδανικά πρέπει να είναι 0,8 έως 3,5 μέτρα. Ο αισθητήρας έχει εύρος λήψης 57º οριζόντια, 43º κάθετα και 70º διαγώνια. Το σύστημα στήριξης μπορεί να προσφέρει μετατόπιση στη γωνία λήψης 27º προς τα πάνω ή προς τα κάτω. Τέλος, για τη λειτουργία της συσκευής απαιτείται παροχή ρεύματος, που επιτυγχάνεται με το συνδυασμό δυο τύπων καλωδίου (ένα USB και ένα καλωδιο ρεύματος).

3.2.1.1 Τεχνολογία Light Coding

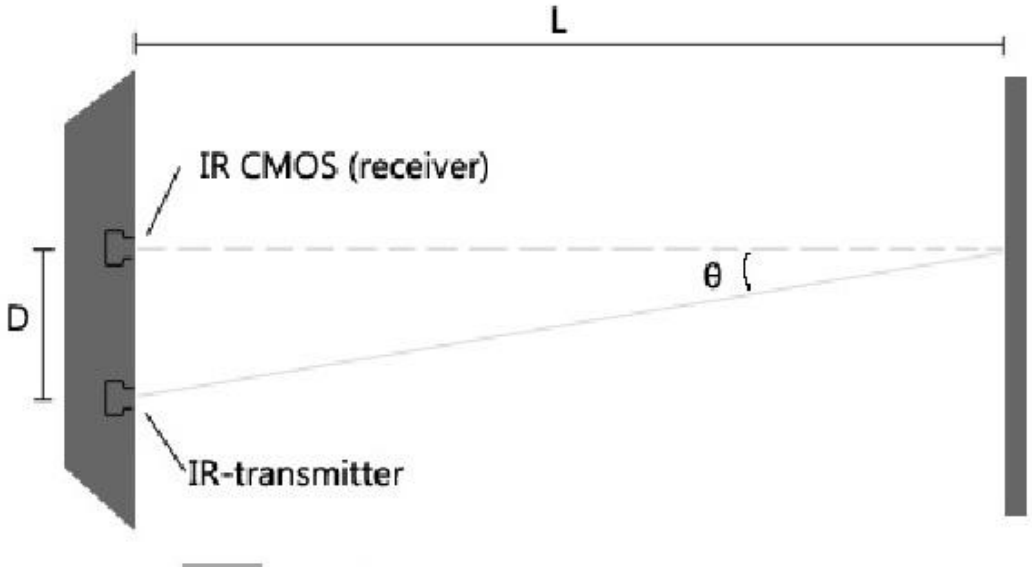
Οι περισσότερες τεχνολογίες που έχουν στόχο τον προσδιορισμό της απόστασης ενός αντικειμένου, μετρούν το χρόνο που χρειάζεται μια λάμψη φωτός για να ταξιδέψει μέχρι το αντικείμενο και να ανακλαστεί από την επιφάνειά του. Η τεχνολογία Light Coding χρησιμοποιεί μια εντελώς διαφορετική προσέγγιση, όπου η πηγή φωτός είναι μόνιμα αναμμένη, μειώνοντας την ανάγκη για ακριβείς μετρήσεις του χρόνου. Μια πηγή λέιζερ εκπέμπει μη-ορατό φως (προσεγγιστικά σε μήκος κύματος υπερύθρων), που περνά από ένα φίλτρο και σκεδάζεται σε ένα ημι-τυχαίο αλλά σταθερό σχέδιο από μικρές κουκκίδες που προβάλλεται στο περιβάλλον που βρίσκεται μπροστά στον αισθητήρα. Το ανακλώμενο σχέδιο στη συνέχεια, εντοπίζεται από μια κάμερα υπερύθρων (IR) και αναλύεται.

Για κάθε pixel στην εικόνα του βάθους, ανάλυσης 640x480, παρέχεται μια τιμή βάθους μέσα στο διάστημα [0-2048] (11bit). Προκειμένου να χρησιμοποιηθεί αυτή η πληροφορία, είναι απαραίτητο να οριστεί μια σχέση ανάμεσα σε αυτήν την τιμή του αισθητήρα και στην πραγματική απόσταση. Η προσέγγιση του προβλήματος φαίνεται στο σχήμα 3.2.

Η γωνία μπορεί να υπολογιστεί από την τριγωνομετρία ως εξής :

$$\theta = \arctan\left(\frac{D}{L}\right)$$

όπου η απόσταση μεταξύ του πομπού υπερύθρων και δέκτη είναι D=0.075 m και L, είναι η απόσταση του αισθητήρα από το αντικείμενο που μετρήθηκε. Συγκρίνοντας τις γωνίες που υπολογίζονται για όλες τις μετρήσεις με τις αντίστοιχες τιμές (N) του αισθητήρα, προκύπτει μια γραμμική σχέση :



Σχήμα 3.2: Διάγραμμα εφαρμογής της θεωρίας για την μέτρηση της απόστασης.

$$N = -4636{,}3\,\theta + 1092{,}5$$

Εισάγοντας την παραπάνω έξίσωση για το θ, σε αυτήν έχουμε :

$$N = -4636{,}3\ arctan\left(\frac{0{,}075}{L}\right) + 1092{,}5 \ \Leftrightarrow$$

$$L = -\frac{0.075}{\tan(0.0002157N - 0.2356)}[m]$$

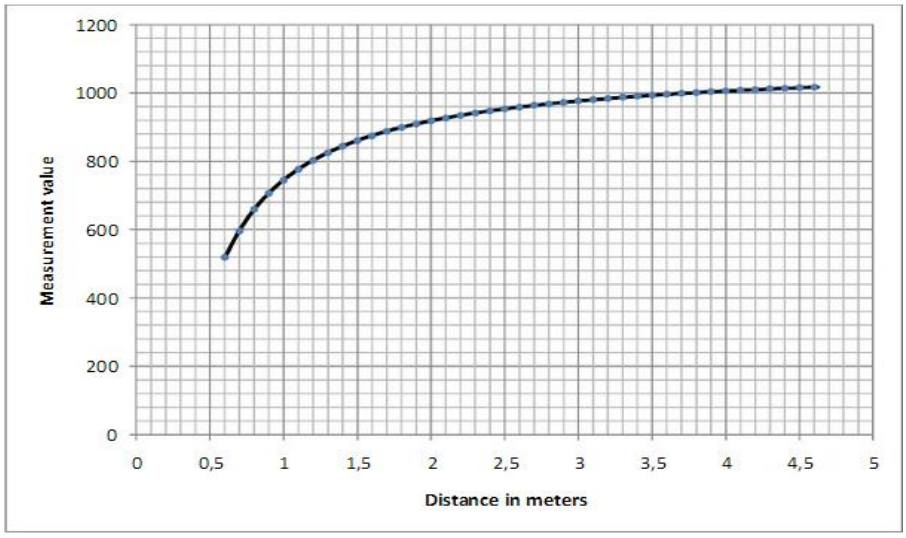
Η εξίσωση αυτή υπολογίζει την πραγματική απόσταση για μια δεδομένη τιμή N. Στο σχήμα 3.3 φαίνεται η γραφική αναπαράσταση της σχέσης ανάμεσα στην απόσταση και στις τιμές του αισθητήρα που μετρώνται και αυτές που υπολογίζονται. Ακόμα καλύτερα αποτελέσματα μπορούν να προκύψουν με τον επαναϋπολογισμό των N και θ, μετά από βαθμονόμηση.

3.2.2 Πλαίσιο λογισμικού OpenNI

Για την καταγραφή, σε συνδυασμό με τη συσκευή Kinect χρησιμοποιήθηκε το πλαίσιο (framework) ανοικτού λογισμικού OpenNI 1.5.2, που είναι κατάλληλο για την ανάπτυξη εφαρμογών μεσολογισμικού (middleware) και βιβλιοθηκών για αισθητήρες 3D.

Το OpenNI(Open Natural Interaction) αποτελεί ένα διαγλωσσικό και διαπλατφορμικό πλαίσιο που ορίζει μια διεπαφή για προγραμματισμό εφαρμογών (API) σχετικών με τη Φυσική Αλληλεπίδραση ανθρώπου-μηχανής (natural interaction). Ειδικότερα, επιτυγχάνει την επικοινωνία με :

- Οπτικοακουστικούς αισθητήρες
- Μεσολογισμικό που αναλύει τα οπτικά και ακουστικά δεδομένα που καταγράφει η συσκευή καταγραφής.



Σχήμα 3.3: Τιμές αισθητήρα που μετρήθηκαν (κουκκίδες) και υπολογίστηκαν (γραμμή) σε σχέση με την απόσταση.

Το σημαντικότερο πλεονέκτημα του OpenNI API είναι ότι επιτρέπει την ανάπτυξη και εφαρμογή αλγορίθμων στα ακατέργαστα δεδομένα που καταγράφονται, ανεξάρτητα από τη συσκευή ή αισθητήρα που τα έχει δημιουργήσει. Τέλος, επιτρέπει την ανίχνευση τρισδιάστατων σκηνών χρησιμοποιώντας μορφές δεδομένων που υπολογίζονται από τα δεδομένα εισόδου ενός αισθητήρα. Παραδείγματος χάριν, είναι δυνατή η αναπαράσταση ενός ανθρώπινου σώματος εντοπίζοντας τις αρθρώσεις του.

Αναλυτικότερα, το OpenNI API υποστηρίζει τον 3D αισθητήρα, την RGB κάμερα, την IR κάμερα υπερύθρων και την ακουστική συσκευή (μικρόφωνο). Επιπροσθέτως, υποστηρίζει τα εξής στοιχεία μεσολογισμικού:

1. Πλήρης Ανάλυση Σώματος: λογισμικό που επεξεργάζεται τα δεδομένα που προκύπτουν από τους αισθητήρες και δημιουργεί την αντίστοιχη πληροφορία που αφορά το ανθρώπινο σώμα (π.χ τα δεδομένα που περιγράφουν τις αρθρώσεις, το κέντρο μάζας κ.ο.κ)

2. Ανάλυση Χεριών

3. Αναγνώριση χειρονομιών (gesture detection): λογισμικό που αναγνωρίζει συγκεκριμένες χειρονομίες και εκκινεί ανάλογα διάφορες εφαρμογές.

4. Ανάλυση του σκηνικού: λογισμικό που αναλύει την εικόνα της κίνησης με σκοπό την παραγωγή πληροφορίας σχετικά με :

    - Το διαχωρισμό του προσκηνίου (foreground) από το παρασκήνιο (background).
    - Την κάτοψη του χώρου.
    - Την αναγνώριση μεμονομένων ατόμων στη σκηνή.

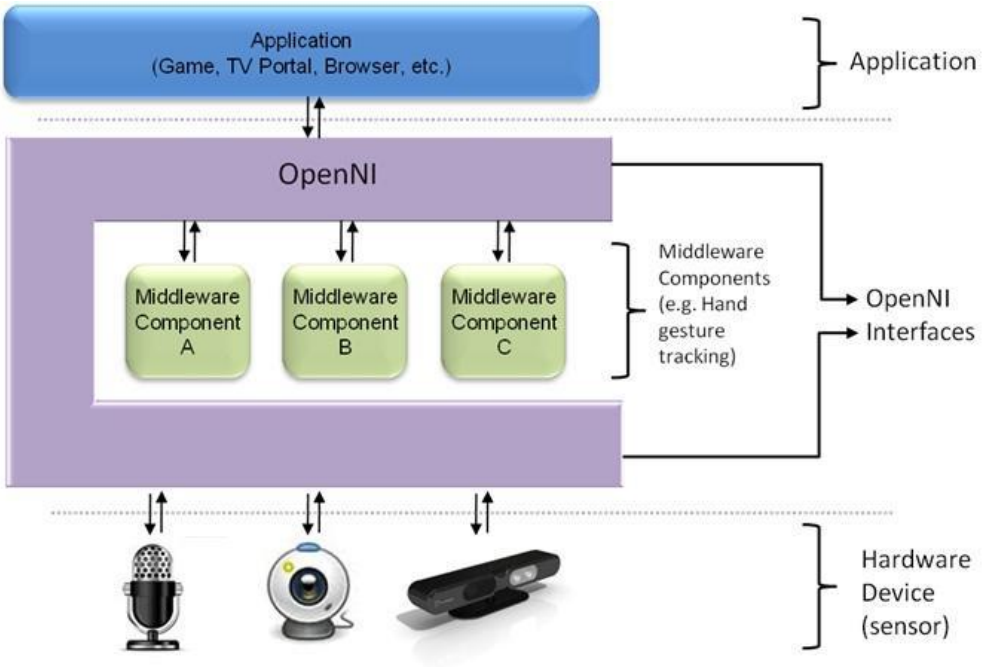
Από τα είδη μεσολογισμικού που αναφέρθηκαν παραπάνω, για τη δημιουργία της βάσης THETIS αξιοποιήθηκαν ιδιαίτερα η πρώτη και η τελευταία κατηγορία. Στο σχήμα 3.4 που

ακολουθεί, απεικονίζεται η σε τρία επίπεδα η λογική του OpenNI με κάθε επίπεδο να παρουσιάζει ένα ζωτικό στοιχείο.

- Κορυφή: Αναπαριστά το λογισμικό που υλοποιεί τις εφαρμογές αλληλεπίδρασης ανθρώπου-μηχανής.

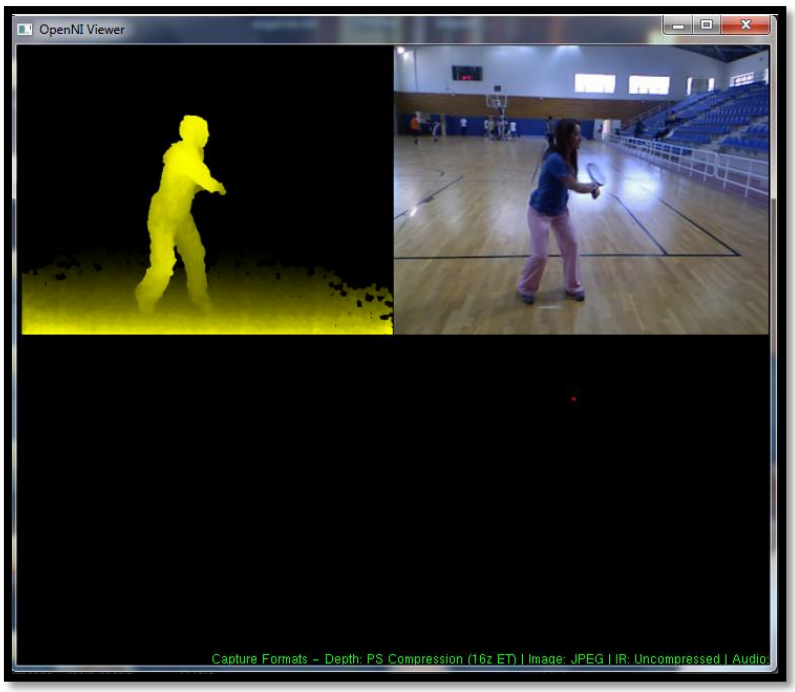
- Μέση: Αναπαριστά το OpenNI, που παρέχει διαπροσωπίες επικοινωνίας που αλληλεπιδρούν με τους αισθητήρες και με το μεσολογισμικό, το οποίο αναλύει τα δεδομένα των αισθητήρων.

- Βάση: Δείχνει τις συσκευές που καταγράφουν τα οπτικοακουστικά δεδομένα της σκηνής.



Σχήμα 3.4: Λογική τριών επιπέδων του OpenNI.

Η καταγραφή των δεδομένων κίνησης πραγματοποιήθηκε μέσω της εφαρμογής NiViewer του OpenNI ως περιγράφεται ακολούθως (Σχήμα 3.5). Τα δεδομένα που κατέγραψε ο αισθητήρας βάθους που ονομάζονται depth map, καθώς και τα δεδομένα που κατέγραψε η RGB κάμερα και ονομάζονται image map, συνδυάζονται και αποθηκεύονται σε ένα αρχείο τύπου ONI που είναι συμβατό με το OpenNI. Οι λόγοι μετατροπής των αρχικών αρχείων σε αρχεία τύπου AVI καθώς επίσης, και ο τρόπος μετατροπής τους περιγράφονται στην ενότητα 3.4.1.

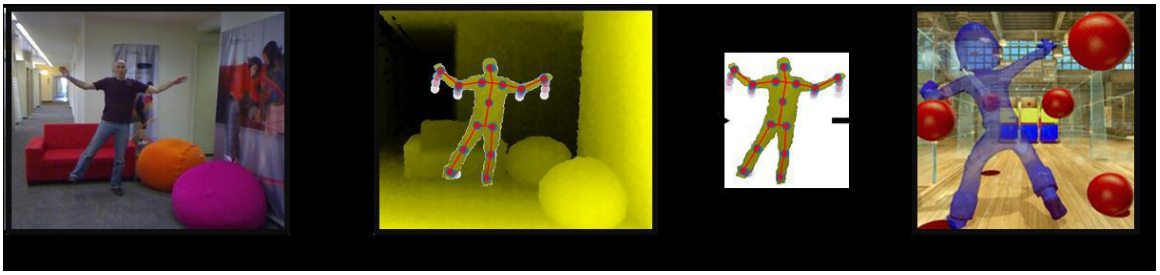


Σχήμα 3.5: Απεικόνιση του depth map και του image map από το NiViewer κατά τη διαδικασία καταγραφής.

3.2.3 Μεσολογισμικό NITE

NITE ή PrimeSense's Natural Interaction Technology for End-User ονομάζεται το μεσολογισμικό (middleware) που αντιλαμβάνεται τον κόμο σε 3 διαστάσεις με βάση τις εικόνες βάθους που καταγράφει ο 3D αισθητήρας και μεταφράζει τα δεδομένα αυτά σε δεδομένα με σημασία για τον άνθρωπο. Ο 3D αισθητήρας παρατηρεί το περιβάλλον του χρήστη, ενώ το μεσολογισμικό NITE λειτουργεί ως μηχανισμός αντίληψης που κατανοεί την αλληλεπίδραση του χρήστη με το περιβάλλον του. Συγκεκριμένα, το μεσολογισμικό NITE περιλαμβάνει αλγορίθμους όρασης υπολογιστή που καθιστούν δυνατό τον εντοπισμό των χρηστών και την αναγνώριση των κινήσεών τους.

Η ιδέα από την οποία προέκυψε ο σχεδιασμός του NITE βασίζεται σε δυο εφαρμογές. Η πρώτη ονομάζεται έλεγχος με τα χέρια (Hand Control) και επιτρέπει στο χρήστη να ελέγχει μια εφαρμογή με τις χειρονομίες των χεριών του. Η δεύτερη εφαρμογή σχετίζεται με εκείνα τα παιχνίδια όπου οι χρήστες ή ο χρήστης στέκονται μπροστά στην οθόνη της τηλεόρασης και επιδίδονται σε διάφορες διασκεδαστικές εφαρμογές. Στόχος της εφαρμογής, που ονομάζεται full body-based control είναι να εξάγει τα σημαντικά χαρακτηριστικά του σκελετού του χρήστη και να μπορεί να τα εντοπίζει σε όλη τη διάρκεια του παιχνιδιού. Αυτό έχει ως αποτέλεσμα η εφαρμογή να μην ασχολείται με τα δεδομένα που αφορούν το βάθος, παρά μόνο να λαμβάνει τα σημεία του σκελετού και τις χειρονομίες του σκελετού. Τελικά, επιτυγχάνει να εξάγει τις διαθέσεις του χρήστη με μικρότερο υπολογιστικό κόστος. Παράδειγμα μιας εφαρμογής full body-based control απεικονίζεται στο σχήμα 3.6.



Σχήμα 3.6: Η εφαρμογή Full body-based control.

3.2.4 Συνθήκες Καταγραφής

Η καταγραφή των κινήσεων αντισφαίρισης που αποτελούν την βάση THETIS πραγματοποιήθηκε σε δυο διαφορετικούς εσωτερικούς χώρους στο Αθλητικό Κέντρο Εθνικού Μετσόβιου Πολυτεχνείου και στον Όμιλο Αντισφαίρισης Γλυφάδας. Η επιλογή δυο εσωτερικών χώρων για την καταγραφή των video, οφείλεται στην ευαισθησία της συσκευής Kinect στο ηλιακό φως. Συγκεκριμένα, εξαιτίας της κάμερας IR υπέρυρθων ακτινών που χρησιμοποιεί το kinect, καθίσταται αδύνατη η καταγραφή των δεδομένων του βάθους όταν δέχεται άμεσα την ηλιακή ακτινοβολία. Επομένως, υπήρξε αναγκαία η διεξαγωγή των καταγραφών σε εσωτερικό χώρο για την αποφυγή του άμεσου ηλιακού φωτός και των παρεμβολών λόγω υπερύθρων που αυτό προκαλεί.

Στον κλειστό χώρο του Αθλητικού Κέντρου Ε.Μ.Π διεξήχθη η καταγραφή των κινήσεων αντισφαίρισης από 31 αρχάριους και 17 έμπειρους αντισφαιριστές. Για τους υπόλοιπους 7 έμπειρους αντισφαιριστές, τα δεδομένα κίνησης καταγράφηκαν στον κλειστό χώρο του Ομίλου Αντισφαίρισης της Γλυφάδας. Στη συνέχεια, περιγράφεται η διαδικασία που ακολουθήθηκε.

Η συσκευή Kinect αρχικά, τοποθετείται σε ύψος 1.6 μέτρων από το έδαφος. Η κάμερα παραμένει στατική. Σε απόσταση 1.5 μέτρου περίπου, ορίζεται το σημείο εκτέλεσης των 12 διαφορετικών κινήσεων της βάσης από τους συμμετέχοντες. Κάθε κίνηση επαναλαμβάνεται αρκετές φορές, ενώ η συσκευή καταγραφής παραμένει σταθερή.

Αναγκαίο κρίνεται να επισημανθεί ότι οι αρχάριοι αντισφαιριστές, παρακολουθούν αρχικά μια επίδειξη κίνησης από την εκπαιδεύτρια αντισφαίρισης του Αθλητικού Κέντρου του Ε.Μ.Π. Στη συνέχεια, επιχειρούν να μιμηθούν την ίδια κίνηση. Σχήμα 3.7.

Όσον αφορά στο παρασκήνιο (background) των video, πρέπει να σημειωθεί πως αυτό δεν παραμένει στατικό. Διαφοροποιείται τις περισσότερες φορές από άτομο σε άτομο, από κίνηση σε κίνηση για το ίδιο άτομο και τέλος, μπορεί να διαφοροποιείται κατά τη διάρκεια καταγραφής μιας κίνησης του ίδιου ατόμου

Σχήμα 3.7: Επίδειξη της κίνησης backhand από την εκπαιδεύτρια αντισφαίρισης.

Επιπροσθέτως, στα videos δεν εμφανίζεται μόνο το άτομο του οποίου η κίνηση μας ενδιαφέρει, αλλά και πλήθος άλλων ατόμων που διέρχονται στο παρασκήνιο, είτε συμμετεχόντων σε άλλου είδους δραστηριότητες, όπως καλαθοσφαίριση. Ακόμη, διαφοροποιήσεις ως προς τη γωνία λήψεις είναι πιθανό να υπάρχουν, όμως δεν κρίνονται υπολογίσιμες. Τέλος, υπάρχουν διαφοροποιήσεις στην απόσταση που χωρίζει τους αντισφαιριστές από τη συσκευή Kinect εξαιτίας του ότι για 7 από τα 55 άτομα που έχουν καταγραφεί, οι λήψεις πραγματοποιήθηκαν σε διαφορετικό χώρο και δεν ήταν δυνατή η διατήρηση του 1,5 μέτρου απόστασης από το Kinect. Η ανομοιογένεια στο background φαίνεται στο σχήμα 3.8.

Σχήμα 3.8: Διαφοροποιήσεις στο background κατά τη διάρκεια διαφορετικών λήψεων.

3.3 Δομή της Βάσης THETIS

3.3.1 Εισαγωγή

Η βάση δεδομένων κίνησης THETIS στην περιλαμβάνει 8374 video μορφής AVI συνολικής διάρκειας περίπου 7 ωρών και 15 λεπτών. Όπως αναφέρθηκε ήδη, 55 άτομα -31 αρχάριοι και 24 έμπειροι αντισφαιριστές συμμετείχαν στην κατασκευή της βάσης. Οι κινήσεις αντισφαίρισης που περιλαμβάνονται είναι οι παρακάτω. Η χρήση των αγγλικών ονομάτων προτιμάται καθώς συμπίπτει με τη διεθνώς αποδεκτή ορολογία για το άθλημα.

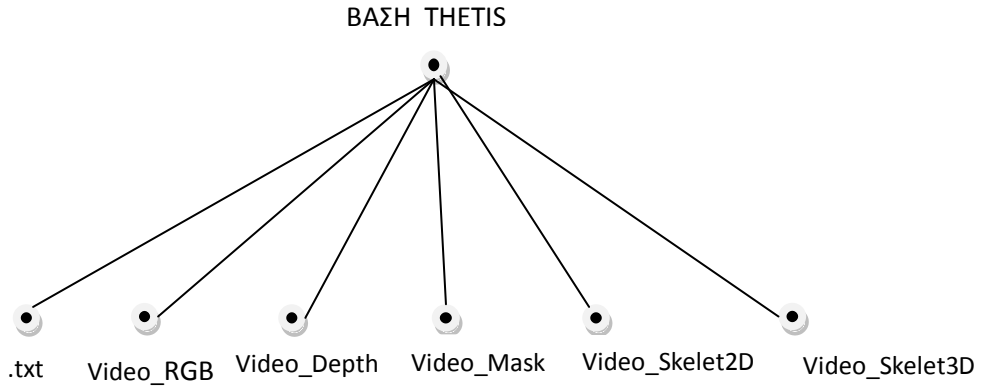
1. Backhand with two hands
2. Backhand
3. Backhand slice
4. Backhand volley
5. Forehand flat
6. Forehand open stands
7. Forehand slice
8. Forehand volley
9. Service flat
10. Service kick
11. Service slice
12. Smash

Αρχικά, κάθε άτομο εκτελεί κάθε μια από τις 12 κινήσεις αντισφαίρισης επαναλαμβάνοντας από δυο έως τέσσερις φορές Καταλήγουμε έτσι, σε 660 αρχεία τύπου ONI. Στη συνέχεια, μετατρέπονται τα αρχεία ONI σε αρχεία AVI με τη χρήση εφαρμογής που περιγράφεται στην ενότητα 3.4.1, βασισμένης στο πλαίσιο της OpenNI. Για κάθε αρχείο ONI δημιουργούνται ταυτόχρονα πέντε AVI αρχεία, ίσης διάρκειας. Συγκεκριμένα, δημουργούνται:

- Ένα αρχείο AVI που απεικονίζει την πληροφορία RGB του αρχικού αρχείου.
- Ένα αρχείο AVI που απεικονίζει την πληροφορία βάθους του αρχικού αρχείου.
- Ένα αρχείο AVI που απεικονίζει την σιλουέτα του ατόμου που απεικονίζεται στο αρχικό αρχείο.
- Ένα αρχείο AVI που απεικονίζει την κίνηση του σκελετού σε 2 διαστάσεις του ατόμου που απεικονίζεται στο αρχικό αρχείο.
- Ένα αρχείο AVI που απεικονίζει την κίνηση του σκελετού σε 3 διαστάσεις του ατόμου που απεικονίζεται στο αρχικό αρχείο.

Επομένως, προκύπτουν 3300 αρχεία AVI, που όμως δεν αποτελούν την τελική βάση δίοτι υφίστανται και άλλη επεξεργασία αφού κόπτονται σε επιμέρους video χειρωνακτικά. Στόχος της διαδικασίας κοψίματος είναι η δημιουργία από κάθε video τριών νέων, που το κάθε ένα θα περιέχει μόνο μια πλήρη επανάληψη της εκάστοτε κίνησης. Έτσι, προκύπτουν 1980 video RGB, 1980 video depth και 1980 video mask(silhouette). Όσον αφορά τα video που απεικονίζουν το σκελετό είτε σε δυο είτε σε τρεις διαστάσεις, δεν είναι πάντοτε διαθέσιμες τρεις επαναλήψεις. Το γεγονός αυτό οφείλεται στους περιορισμούς που υπάρχουν ώστε να αποκτήσει κανείς την πληροφορία σκελετού από ένα αρχικό αρχείο ONI. Συγκεκριμένα, ο χρήστης πρέπει να πάρει μια συγκεκριμένη θέση στην αρχή της

καταγραφής, που ονομάζεται calibration pose. Σε αντίθετη περίπτωση, δεν πραγματοποιείται η εξαγωγή του σκελετού. Δυστυχώς σε ορισμένες περιπτώσεις, η calibration pose των συμμετεχόντων δεν υπήρξε επιτυχής και αυτό είναι κάτι που δεν μπορεί προκαταβολικά να ελεγχθεί. Ακόμη, κάποιοι συμμετέχοντες πραγματοποίησαν με μεγάλη ταχύτητα την εκτέλεση κάποιων κινήσεων με αποτέλεσμα η εξαγωγή του σκελετού να επιτυγχάνεται στις τελευταίες επαναλήψεις μόνο. Για τους παραπάνω λόγους, προέκυψαν 1217 video σκελετού σε δυο διαστάσεις και 1217 video σκελετού σε τρεις διαστάσεις. Το σύνολο των δεδομένων THETIS χωρίζεται σε έξι υποφακέλους, όπως φαίνεται στο σχήμα 3.9.



Σχήμα 3.9 : Δομή της βάσης δεδομένων THETIS

Σε αυτό το σημείο, παρουσιάζεται μια περίληψη των περιεχομένων του κάθε φακέλου.

- .txt: περιλαμβάνει λεπτομερή περιγραφή των περιεχομένων της βάσης δεδομένων.

- Video_RGB: περιέχει 1980 αρχεία AVI, σε 12 υποφακέλους (ανά κλάση). Σε κάθε φάκελο, υπάρχουν 3 επαναλήψεις από κάθε άτομο για την κίνηση αυτή.

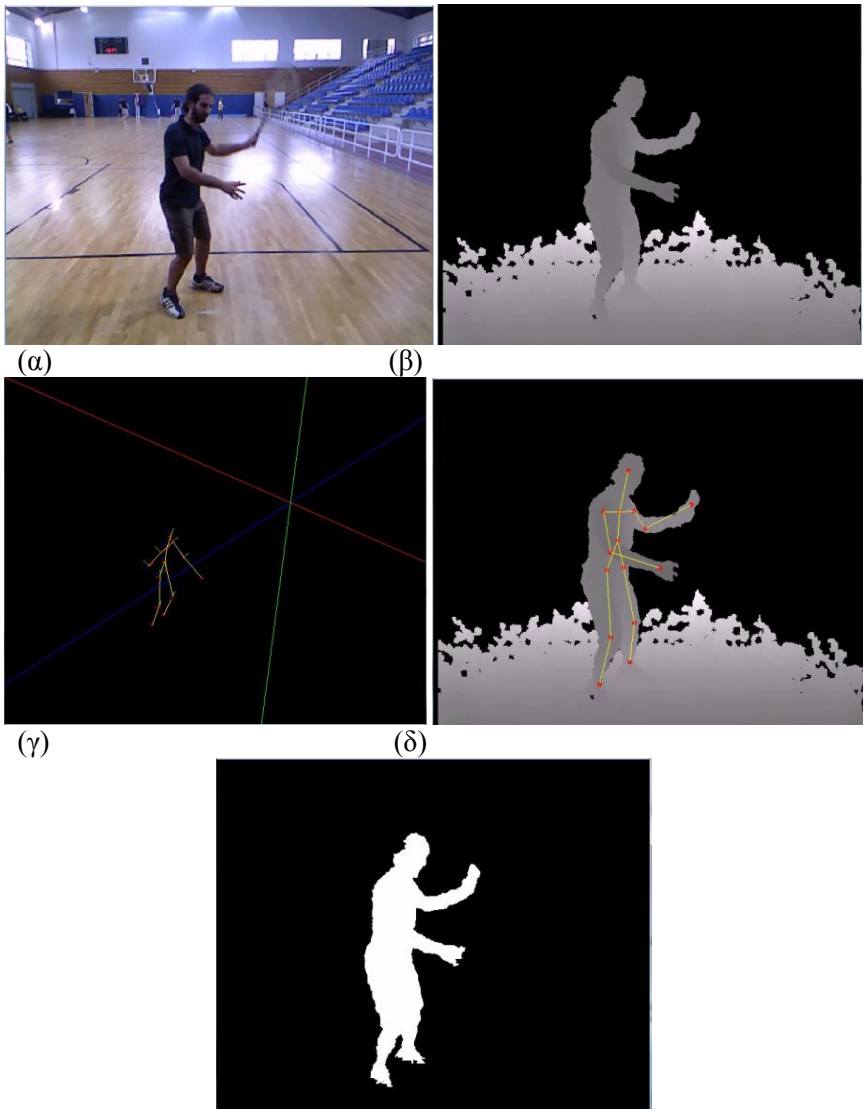
- Video_Depth: περιέχει 1980 αρχεία AVI, σε 12 υποφακέλους (ανά κλάση). Σε κάθε φάκελο, υπάρχουν 3 επαναλήψεις από κάθε άτομο για την κίνηση αυτή.

- Video_Mask: περιέχει 1980 αρχεία AVI, σε 12 υποφακέλους (ανά κλάση). Σε κάθε φάκελο, υπάρχουν 3 επαναλήψεις από κάθε άτομο για την κίνηση αυτή.

- Video_Skelet2D: περιέχει 1217 αρχεία AVI, σε 12 υποφακέλους (ανά κλάση).

- Video_Skelet3D: περιέχει 1217 αρχεία AVI, σε 12 υποφακέλους (ανά κλάση).

Στο σχήμα 3.10 απεικονίζονται στιγμιότυπα του ίδιου ατόμου, να εκτελεί την κίνηση forehand slice, από όλες τις κατηγορίες video της βάσης.



Σχήμα 3.10 : (α) RGB video , (β) Depth video, (γ) Skelet3D video, (δ) Skelet2D video, (ε) Mask video

3.3.2 RGB videos

Για την ονομασία των αρχείων σε αυτόν τον φάκελο ακολουθούμε την εξής σύμβαση: THETIS_{actor}_{action}_{sequence}.avi

Το πεδίο actor αντιστοιχεί σε ένα από τα 55 άτομα που συμμετείχαν. Κάθε άτομο αντιστοιχεί σε έναν κωδικό από p1 έως p55. Οι κωδικοί p1 έως p31 αντιστοιχούν σε αρχάριους, ενώ οι κωδικοί p32 έως p55, αντιστοιχούν σε έμπειρους στην αντισφαίριση. Το πεδίο sequence αναφέρεται στην αρίθμηση μεταξύ των επαναλήψεων και κάθε φορά παίρνει μια από τις τιμές s1, s2 και s3. Το πεδίο action αναφέρεται στο είδος της κίνησης που περιέχει το video. Το σχήμα 3.11 παρουσιάζει την αντιστοίχιση μεταξύ των πιθανών τιμών του πεδίου action και του τυπικού ονόματος της κίνησης.

Όνομα Κίνησης	Κίνηση
backhand	Backhand
backhand2h	Backhand with two hands
bslice	Backhand slice
foreflat	Forehand flat
foreopen	Forehand open stands
fslice	Forehand slice
serflat	Flat service
serkick	Kick service
serslice	Slice service
smash	Smash
fvolley	Forehand volley
bvolley	Backhand volley

Σχήμα 3.11: Αντιστοίχιση τιμών πεδίου action και πραγματικής ονομασίας της κίνησης.

Πιο αναλυτικά, ο φάκελος περιέχει 12 υποφακέλους που το περιεχόμενό τους περιγράφεται στο σχήμα 3.12.

Ονομασία Φακέλου	**Αριθμός Video**	**Άτομα**	**Επαναλήψεις ανά άτομο**
backhand	165	55	3
backhand2hands	165	55	3
backhand_slice	165	55	3
forehand_flat	165	55	3
forehand_openstands	165	55	3
forehand_slice	165	55	3
flat_service	165	55	3
kick_service	165	55	3
slice_service	165	55	3
smash	165	55	3
forehand_volley	165	55	3
backhand_volley	165	55	3

Σχήμα 3.12: Περιγραφή των περιεχομένων του φακέλου Video_RGB.

3.3.3 Depth videos

Για την ονομασία των αρχείων σε αυτόν τον φάκελο ακολουθούμε την εξής σύμβαση: THETIS_{actor}_{action}_depth_{sequence}.avi
Το πεδίο actor αντιστοιχεί σε ένα από τα 55 άτομα που συμμετείχαν. Κάθε άτομο αντιστοιχεί σε έναν κωδικό από p1 έως p55. Το πεδίο sequence αναφέρεται στην αρίθμηση μεταξύ των επαναλήψεων και κάθε φορά παίρνει μια από τις τιμές s1, s2 και s3. Πιο αναλυτικά, ο φάκελος περιέχει 12 υποφακέλους που το περιεχόμενό τους περιγράφεται στο σχήμα 3.13.

Ονομασία Φακέλου	**Αριθμός Video**	**Άτομα**	**Επαναλήψεις ανά άτομο**
backhand	165	55	3
backhand2hands	165	55	3
backhand_slice	165	55	3
forehand_flat	165	55	3
forehand_openstands	165	55	3
forehand_slice	165	55	3
flat_service	165	55	3
kick_service	165	55	3
slice_service	165	55	3
smash	165	55	3
forehand_volley	165	55	3
backhand_volley	165	55	3

Σχήμα 3.13: Περιγραφή των περιεχομένων του φακέλου Video_Depth.

3.3.4 Mask videos

Για την ονομασία των αρχείων σε αυτόν τον φάκελο ακολουθούμε την εξής σύμβαση: THETIS_{actor}_{action}_mask_{sequence}.avi
Το πεδίο actor αντιστοιχεί σε ένα από τα 55 άτομα που συμμετείχαν. Κάθε άτομο αντιστοιχεί σε έναν κωδικό από p1 έως p55. Το πεδίο sequence αναφέρεται στην αρίθμηση μεταξύ των επαναλήψεων και κάθε φορά παίρνει μια από τις τιμές s1, s2 και s3. Πιο αναλυτικά, ο φάκελος περιέχει 12 υποφακέλους που το περιεχόμενό τους περιγράφεται στο σχήμα 3.14.

Ονομασία Φακέλου	**Αριθμός Video**	**Άτομα**	**Επαναλήψεις ανά άτομο**
backhand	165	55	3
backhand2hands	165	55	3
backhand_slice	165	55	3
forehand_flat	165	55	3
forehand_openstands	165	55	3
forehand_slice	165	55	3
flat_service	165	55	3
kick_service	165	55	3
slice_service	165	55	3
smash	165	55	3
forehand_volley	165	55	3
backhand_volley	165	55	3

Σχήμα 3.14: Περιγραφή των περιεχομένων του φακέλου Video_Mask.

3.3.5 Skelet2D videos

Για την ονομασία των αρχείων σε αυτόν τον φάκελο ακολουθούμε την εξής σύμβαση :
THETIS_{actor}_{action}_skelet2D_{sequence}.avi
Το πεδίο actor αντιστοιχεί σε ένα από τα 55 άτομα που συμμετείχαν. Κάθε άτομο αντιστοιχεί σε έναν κωδικό από p1 έως p55. Το πεδίο sequence αναφέρεται στην αρίθμηση μεταξύ των επαναλήψεων και κάθε φορά παίρνει μια από τις τιμές s1, s2 και s3. Πιο αναλυτικά, ο φάκελος περιέχει 12 υποφακέλους που το περιεχόμενό τους περιγράφεται στο σχήμα 3.15.

Ονομασία Φακέλου	**Αριθμός Video**	**Άτομα**	**Παρατηρήσεις**
backhand	97	51	{p1,p5,p6,p12,p13,p17,p18, p22,p29,p32,p40,p50}x1 επαναληψη, {p7,p28,p33,p45}x0 επαναλήψεις, {p2,p3,p23,p26,p46,p51,p53,p54}x 3 επαναλήψεις, όλα τα υπόλοιπα άτομα x 2 επαναλήψεις το κάθε ένα

backhand2hands	107	54	{p11,p14,p23,p34,p55 }x1 επαναληψη, {p8}x0 επαναλήψεις, {p6,p20,p53}x 3 επαναλήψεις, όλα τα υπόλοιπα άτομα x 2 επαναλήψεις το κάθε ένα
backhand_slice	100	54	{p2,p7,p10,p12,p25,p27,p41, p43,p44,p47,p55}x1 επαναληψη, {p8}x0 επαναλήψεις, {p6,p20,p53}x 3 επαναλήψεις, όλα τα υπόλοιπα άτομα x 2 επαναλήψεις το κάθε ένα
forehand_flat	110	55	{p6,p19,p21,p32 p50,p55}x1 επαναληψη, {p3,p4,p8,p11,p31,p49}x 3 επαναλήψεις, όλα τα υπόλοιπα άτομα x 2 επαναλήψεις το κάθε ένα
forehand_openstands	101	55	{p7,p12,p15,p21,p22,p33 P35,p36,p58,p50,p52}x1 επαναληψη, {p3,p4,p9,p53}x 3 επαναλήψεις, όλα τα υπόλοιπα άτομα x 2 επαναλήψεις το κάθε ένα
forehand_slice	97	55	p6,p7,p28,p32,p33,p34,p35,p40,p44, p45,p47,p48,p49,p50}x1 επαναληψη, {p1}x 3 επαναλήψεις, όλα τα υπόλοιπα άτομα x 2 επαναλήψεις το κάθε ένα
flat_service	96	51	{p1,p3,p5,p7,p13,p33,p36,p41, p45,p51 }x1 επαναληψη, {p37,p50,p22,p55}x0 επαναλήψεις, {p2,p40,p44,54} x 3 επαναλήψεις, όλα τα υπόλοιπα άτομα x 2 επαναλήψεις το κάθε ένα
kick_service	109	54	{p3,p8,p21,p27,p28,p32,p45 }x1 επαναληψη, {p37}x0 επαναλήψεις, {p1,p7,p11,p17,p31,p44,p48,p49,p53} x 3 επαναλήψεις, όλα τα υπόλοιπα άτομα x 2 επαναλήψεις το κάθε ένα
slice_service	100	54	{p2,p7,p11,p12,p19,p22,p31,p36, p41,p45,p52,p55}x1 επαναληψη, {p37}x0 επαναλήψεις, {p21,p51} x 3 επαναλήψεις, όλα τα υπόλοιπα άτομα x 2 επαναλήψεις το κάθε ένα
smash	104	52	{p3,p22,p27,p32,p34, p43}x1 επαναληψη, {p11,p37,p45}x0 επαναλήψεις, {p5,p16,p25,p28,p42,p55} x 3 επαναλήψεις, όλα τα υπόλοιπα άτομα x 2 επαναλήψεις το κάθε ένα

forehand_volley	93	52	{p7,p12,p21,p26,p27,p33,p34,p36,p37, p43,p51,p52,p53 }x1 επαναληψη, { p32,p49,p50}x0 επαναλήψεις, {p30} x 3 επαναλήψεις, όλα τα υπόλοιπα άτομα x 2 επαναλήψεις το κάθε ένα
backhand_volley	103	53	{p5,p21,p22,p27,p35,p37,p45}x1 επαναληψη, { p32,p37}x0 επαναλήψεις, {p2,p13,p41,p48,p51} x 3 επαναλήψεις, όλα τα υπόλοιπα άτομα x 2 επαναλήψεις το κάθε ένα

Σχήμα 3.15: Περιγραφή των περιεχομένων του φακέλου Video_Skelet2D.

3.3.6 Skelet3D videos

Για την ονομασία των αρχείων σε αυτόν τον φάκελο ακολουθούμε την εξής σύμβαση:
THETIS_{actor}_{action}_skelet3D_{sequence}.avi
Το πεδίο actor αντιστοιχεί σε ένα από τα 55 άτομα που συμμετείχαν. Κάθε άτομο αντιστοιχεί σε έναν κωδικό από p1 έως p55. Το πεδίο sequence αναφέρεται στην αρίθμηση μεταξύ των επαναλήψεων και κάθε φορά παίρνει μια από τις τιμές s1, s2 και s3. Πιο αναλυτικά, ο φάκελος περιέχει 12 υποφακέλους που το περιεχόμενό τους περιγράφεται στο σχήμα 3.16.

Ονομασία Φακέλου	**Αριθμός Video**	**Άτομα**	**Παρατηρήσεις**
backhand	97	51	{p1,p5,p6,p12,p13,p17,p18, p22,p29,p32,p40,p50}x1 επαναληψη, {p7,p28,p33,p45}x0 επαναλήψεις, {p2,p3,p23,p26,p46,p51,p53,p54}x 3 επαναλήψεις, όλα τα υπόλοιπα άτομα x 2 επαναλήψεις το κάθε ένα
backhand2hands	107	54	{p11,p14,p23,p34,p55 }x1 επαναληψη, {p8}x0 επαναλήψεις, {p6,p20,p53}x 3 επαναλήψεις, όλα τα υπόλοιπα άτομα x 2 επαναλήψεις το κάθε ένα
backhand_slice	100	54	{p2,p7,p10,p12,p25,p27,p41,p43,p44,p47,p55} x1 επαναληψη, {p8}x0 επαναλήψεις, {p6,p20,p53}x 3 επαναλήψεις, όλα τα υπόλοιπα άτομα x 2 επαναλήψεις το κάθε ένα
forehand_flat	110	55	{p6,p19,p21,p32p50,p55}x1 επαναληψη, {p3,p4,p8,p11,p31,p49}x 3 επαναλήψεις, όλα τα υπόλοιπα άτομα x 2 επαναλήψεις το κάθε ένα
forehand_openstands	101	55	{p7,p12,p15,p21,p22,p33 p35,p36,p58,p50,p52}x1 επαναληψη,

			{p3,p4,p9,p53}x 3 επαναλήψεις, όλα τα υπόλοιπα άτομα x 2 επαναλήψεις το κάθε ένα
forehand_slice	97	55	p6,p7,p28,p32,p33,p34,p35,p40,p44, p45,p47,p48,p49,p50}x1 επαναληψη, {p1}x 3 επαναλήψεις, όλα τα υπόλοιπα άτομα x 2 επαναλήψεις το κάθε ένα
flat_service	96	51	{p1,p3,p5,p7,p13,p33,p36,p41, p45,p51 }x1 επαναληψη, {p37,p50,p22,p55}x0 επαναλήψεις, {p2,p40,p44,54} x 3 επαναλήψεις, όλα τα υπόλοιπα άτομα x 2 επαναλήψεις το κάθε ένα
kick_service	109	54	{p3,p8,p21,p27,p28, P32,p45 }x1 επαναληψη, {p37}x0 επαναλήψεις, {p1,p7,p11,p17,p31,p44 P48,p49,p53} x 3 επαναλήψεις, όλα τα υπόλοιπα άτομα x 2 επαναλήψεις το κάθε ένα
slice_service	100	54	{p2,p7,p11,p12,p19,p22,p31,p36, P41,p45,p52,p55}x1 επαναληψη, {p37}x0 επαναλήψεις, {p21,p51} x 3 επαναλήψεις, όλα τα υπόλοιπα άτομα x 2 επαναλήψεις το κάθε ένα
smash	104	52	{p3,p22,p27,p32,p34, p43}x1 επαναληψη, {p11,p37,p45}x0 επαναλήψεις, {p5,p16,p25,p28, p42,p55} x 3 επαναλήψεις, όλα τα υπόλοιπα άτομα x 2 επαναλήψεις το κάθε ένα
forehand_volley	93	52	{p7,p12,p21,p26,p27,p33,p34,p36,p37,p43,p51 ,p52,p53 }x1 επαναληψη, { p32,p49,p50}x0 επαναλήψεις, {p30} x 3 επαναλήψεις, όλα τα υπόλοιπα άτομα x 2 επαναλήψεις το κάθε ένα
backhand_volley	103	53	{p5,p21,p22,p27,p35,p37,p45}x1 επαναληψη, { p32,p37}x0 επαναλήψεις, {p2,p13,p41,p48,p51} x 3 επαναλήψεις, όλα τα υπόλοιπα άτομα x 2 επαναλήψεις το κάθε ένα

Σχήμα 3.16 : Περιγραφή των περιεχομένων του φακέλου Video_Skelet3D.

3.4 Εργαλεία

3.4.1 Μετατροπή αρχείων ONI σε αρχεία AVI

Σκοπός της δημιουργίας της βάσης δεδομένων THETIS είναι να χρησιμοποιηθεί ως εργαλείο αξιολόγησης και ανάλυσης των αλγορίθμων που προτείνονται για την επίλυση του προβλήματος της αναγνώρισης ανθρώπινης δραστηριότητας και πιο συγκεκριμένα για εφαρμογές gaming, αυτοματοποιημένου σχολιασμού αθλητικών event κ.α. Επομένως, κρίνοντας πως η διάθεση της σχετικής πληροφορίας σε αρχεία τύπου ONI θα ήταν περιοριστική για ευρεία χρήση, καθώς θα ήταν επιβεβλημένη η χρήση του πλαισίου εφαρμογών OpenNI κρίθηκε αναγκαία η μεταροπή των καταγεγραμμένων αρχείων τύπου ONI, σε μια ευρέως διαδεδομένη μορφή αρχείων για την αποθήκευση δεδομένων πολυμέσων. Η μετατροπή των αρχείων ONI σε AVI πραγματοποιήθηκε με τη χρήση μιας εφαρμογής, βασισμένης στο διαγλωσσικό πλαίσιο εφαρμογών OpenNI, που έχει αναπτυχθεί από τον κ. Σταύρο Αποστόλου. Η εφαρμογή παρέχει τις εξής δυνατότητες :

- Απομόνωση των δεδομένων του βάθους που έχει καταγράψει ο αισθητήρας βάθους, και η αποθήκευση σε αρχείο avi.

- Εξαγωγή της σιλουέτας του ατόμου με χρήση αλγορίθμων που υποστηρίζει το OpenNI και η αποθήκευση της πληροφορίας σε αρχείο avi.

- Εξαγωγή του σκελετού του ανθρώπινου σώματος, μέσω του εντοπισμού των αρθρώσεων που επίσης υποστηρίζεται από το OpenNI.

- Η απεικόνιση της πληροφορίας που αφορά τον σκελετό τόσο στις δυο διαστάσεις όσο και στις τρεις.

3.4.2 Περικοπή των AVI αρχείων

Με τη διαδικασία περικοπής των αρχικών αρχείων AVI, δημιουργήθηκαν για κάθε αρχικό αρχείο, περισσότερα μικρότερης διάρκειας. Έτσι από 3300 αρχεία, προέκυψαν 8374 νέα. Σε κάθε ένα από τα αρχικά αρχεία, παρουσιάζεται η εκτέλεση μερικών επαναλήψεων μιας συγκεκριμένης κίνησης αντισφαίρισης. Αντίθετα, μετά την περικοπή, κάθε αρχείο περιέχει μόνο μια επανάληψη της επιθυμητής κίνησης αποφεύγοντας την καταγραφή κινήσεων που δεν σχετίζονται με την επιθυμητή δραστηριότητα. Έτσι, τα τελικά έχουν περιεχόμενο περισσότερο σαφές και πιο σχετικό με την επιθυμητή κίνηση.

Η διαδικασία της περικοπής των αρχείων πραγματοποιήθηκε χειρωνακτικά με τη χρήση του εργαλείου περικοπής αρχείων video Virtual Dub 1.9.11.

ΚΕΦΑΛΑΙΟ 4

Διεξαγωγή Πειραμάτων

4.1 Εισαγωγή

Κατά το τελευταίο στάδιο της εργασίας πραγματοποιείται η πειραματική δοκιμή της βάσης δεδομένων THETIS με την εφαρμογή δυο μεθόδων αναγνώρισης δραστηριότητας που έχουν ήδη προταθεί. Για τον σκοπό αυτό, όλα τα video που περιέχουν την αναπαράσταση του σκελετού σε τρεις διαστάσεις (**Skelet3D video**), καθώς και τα video που περιέχουν την πληροφορία του βάθους (**Depth video**), κωδικοποιούνται με περιγραφείς τελευταίας τεχνολογίας, που περιγράφονται εκτενώς στις ενότητες 4.2.1 και 4.2.2. Στη συνέχεια, αφού κβαντοποιηθούν οι περιγραφείς των video με τη διαδικασία που περιγράφεται στην ενότητα 4.3, εισάγονται ως είσοδοι σε μη-γραμμικές μηχανές διανυσμάτων υποστήριξης (nonlinear SVM) (ενότητα 4.4) για την ταξινόμηση των video σε κλάσεις με βάση το είδος της κίνησης που περιέχουν.

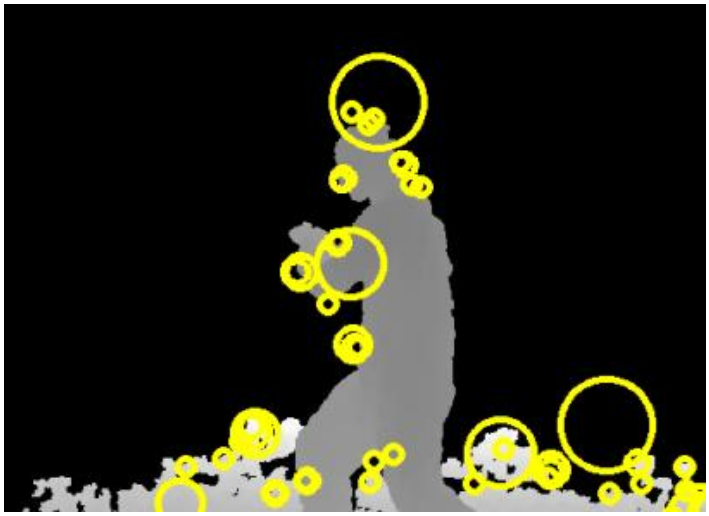
Επιπλέον, παρουσιάζονται σε αντιπαράθεση τα αποτελέσματα που προέκυψαν από την εφαρμογή των ίδιων περιγραφέων και με το ίδιο πρωτόκολλο ταξινόμησης στη βάση δεδομένων **KTH.** Σκοπός είναι να παρουσιαστούν οι προκλήσεις που προκύπτουν από το νέο σύνολο δεδομένων κινήσεων THETIS.

4.2 Μέθοδοι Εξαγωγής Περιγραφέων

4.2.1 Μέθοδος εντοπισμού σημείων ενδιαφέροντος στο χωροχρόνο

Ένας αποδεδειγμένα πολύ χρήσιμος τρόπος εντοπισμού των στοιχείων που μπορεί να μας ενδιαφέρουν μέσα σε μία εικόνα και κατ' επέκταση σε ένα βίντεο, είναι η ανεύρεση σημείων ενδιαφέροντος (interest points). Στην προσπάθεια να βρεθούν τέτοια σημεία στο χωροχρόνο που να είναι αμετάβλητα στις αλλαγές της κλίμακας των υπό εξέταση αντικειμένων ή ανθρωπίνων κινήσεων προτάθηκαν κατά καιρούς διάφορες μέθοδοι.

Στην παρούσα εργασία για τον εντοπισμό χωροχρονικών σημείων ενδιαφέροντος και την εξαγωγή των περιγραφέων τους, ακολουθήσαμε τη μέθοδο Space-Time Interest Points-STIP που χρησιμοποιήθηκε στο [18] και χρησιμοποιήσαμε τον κώδικα που παρέχουν οι συγγραφείς. Ο κώδικας προεκτείνει τον Ανιχνευτή Harris 3D (Harris 3D Detector), των Laptev και Lindeberg [15] που εντοπίζει χωροχρονικά σημεία ενδιαφέροντος (Σχήμα 4.1) και υπολογίζει τους τοπικούς χωροχρονικούς περιγραφείς Ιστογράμματα Προσανατολισμένης Κλίσης (Histograms of Oriented Gradient-HOG) και Ιστογράμματα Οπτικής Ροής (Histograms of Optical Flow - HOF). Όπως στο [18], χρησιμοποιούμε την έκδοση εκείνη του κώδικα που δεν χρησιμοποιεί επιλογή κλίμακας,αλλά αντίθετα ένα σύνολο πολλαπλών συνδυασμών από χωρικές και χρονικές κλίμακες.



Σχήμα 4.1 : Εφαρμογή του κώδικα STIP σε βίντεο της βάσης THETIS που απεικονίζει το βάθος.

4.2.1.1 Ανιχνευτής Harris 3D

Ο Ανιχνευτής Harris 3D (Harris 3D Detector) αναζητά τις χωροχρονικές τιμές της εικόνας που παρουσιάζουν μεγάλες αποκλίσεις τόσο στη διάσταση του χώρου, όσο και του χρόνου. Τα σημεία αυτά θα είναι χωρικά σημεία ενδιαφέροντος, αλλά με μια χαρακτηριστική χρονική θέση, που αντιστοιχεί σε στιγμές μη συνεχόμενης κίνησης της εικόνας σε μια χωροχρονική γειτονιά.
Συνεπώς, για τη χωροχρονική ακολουθία της εικόνας, f : $\Re^2 \times \Re \rightarrow \Re$ έχουμε την κατασκευή της γραμμικής χωροχρονικής αναπαράστασης D : $\Re^2 \times \Re \rightarrow {\Re_+}^2$, με συνέλιξη της f με έναν ανισοτροπικό Γκαουσιανό πυρήνα συγκεκριμένης χωρικής σ_l^2 και χρονικής διακύμανσης τ_l^2:

$$D(\cdot;\sigma_l^2,\tau_l^2) = g(\cdot;\sigma_l^2,\tau_l^2) * f(\cdot) \qquad (4.2.1)$$

όπου ο χωροχρονικά διαχωρίσιμος Γκαουσιανός πυρήνας ορίζεται ως :

$$g(x,y,t;\sigma_l^2,\tau_l^2) = \frac{\exp\left(-(x^2+y^2)/2\sigma_l^2 - t^2/2\tau_l^2\right)}{\sqrt{(2\pi)^3\sigma_l^4\tau_l^2}} \qquad (4.2.2)$$

Θεωρούμε τη χωροχρονική μήτρα δεύτερων ροπών (spatio-temporal second moment matrix), η οποία είναι ένας 3×3 πίνακας που αποτελείται από χωρικές και χρονικές παραγώγους πρώτης τάξης, κανονικοποιημένες από μία Γκαουσιανή συνάρτηση βάρους $g(;\sigma_l^2,\tau_l^2)$

$$\mu = g(;\sigma_l^2,\tau_l^2) * \begin{pmatrix} D_x^2 & D_xD_y & D_xD_t \\ D_xD_y & D_y^2 & D_yD_t \\ D_xD_t & D_yD_t & D_t^2 \end{pmatrix} \qquad (4.2.3)$$

,όπου οι κλίμακες ενσωμάτωσης (integration scales) είναι $\sigma_l^2 = s\sigma_l^2$ και $\tau_l^2 = s\tau_l^2$, οι παράγωγοι πρώτης τάξης ορίζονται ως $D_\xi(;\sigma_l^2,\tau_l^2) = \partial_\xi(g * f)$. Η αναζήτηση για σημεία ενδιαφέροντος γίνεται πλέον στις περιοχές της f που οι ιδιοτιμές $\lambda_1,\lambda_2,\lambda_3$ του μ παίρνουν ιδιάζουσες τιμές.

Επεκτείνοντας πάλι τον απλό Harris ανιχνευτή, δημιουργείται ένα τρισδιάστατο μέτρο Harris/ γωνιότητας

$$(4.2.4) \qquad H = det(\mu) - k \cdot trace^3(\mu) = \lambda_1\lambda_2\lambda_3 - k(\lambda_1 + \lambda_2 + \lambda_3)^3$$

Έτσι λοιπόν, τα σημεία ενδιαφέροντος που επιλέγουμε είναι αυτά που αντιστοιχούν στα θετικά τοπικά μέγιστα της H. Για την επιλογή της κλίμακας που συνοδεύει το κάθε σημείο ενδιαφέροντος εμπλέκουμε την Λαπλασιανή σε χωροχρονική μορφή και κανονικοποιημένη ως προς την κλίμακα, ώστε να ορίζεται με τον εξής τρόπο:

$$(4.2.5) \qquad \nabla^2_{norm} D = D_{xx,norm} + D_{yy,norm} + D_{tt,norm}$$

όπου $D_{xx,norm} = \sigma^{2\alpha}\tau^{2b}\mathrm{D}_{xx}$ και $D_{tt,norm} = \sigma^{2c}\tau^{2d}\mathrm{D}_{tt}$. Με την επιλογή των κατάλληλων παραμέτρων για τα a, b, c, d μπορούμε να φέρουμε τη Λαπλασιανή σε μορφή που είναι διαχειρίσιμη ώστε τελικά, αναζητούμε τα σημεία που μεγιστοποιούν τοπικά ανάμεσα σε χωρικές και χρονικές κλίμακες τη συνάρτηση:

$$(4.2.6) \qquad \nabla^2_{norm} D = \sigma^2\tau^{1/2}\left(D_{xx} + D_{yy}\right) + \sigma\tau^{3/2}D_{tt}$$

4.2.1.2 Ιστογράμματα Προσανατολισμένης Κλίσης και Ιστογράμματα Οπτικής Ροής

Για την εξαγωγή χαρακτηριστικών που αφορούν τα σημεία ενδιαφέροντος που εντοπίζονται με τον Harris 3D detector, όπως αναφέρθηκε προηγουμένως, χρησιμοποιούνται τα ιστογράμματα HOG και HOF.

Για την υλοποίηση των HOG (Σχήμα 4.2) απαιτείται η ακόλουθη διαδικασία. Αρχικά, γίνεται ο υπολογισμός της κλίσης (gradient) με το φιλτράρισμα των χρωματικών τιμών ή των τιμών έντασης φωτεινότητας (αν πρόκειται για ασπρόμαυρες εικόνες) της εικόνας με τα διακριτά φίλτρα - μάσκες παραγώγισης και για τις δύο διαστάσεις: $[-1;\ 0;\ 1] και\ [-1;\ 0;\ 1]^T$.

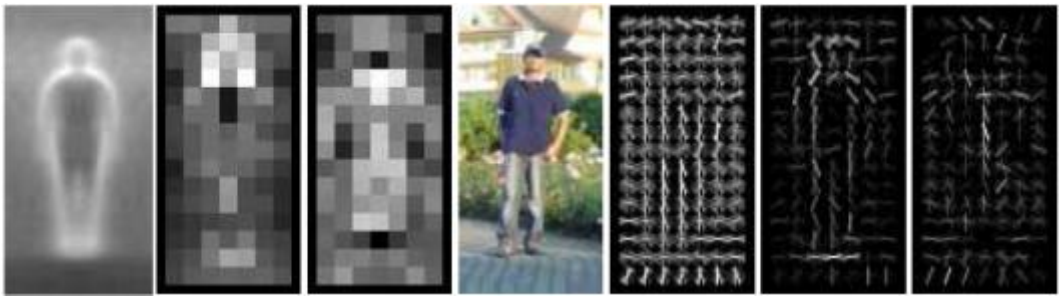
Στη συνέχεια, πρέπει να κατασκευαστούν τα ιστογράμματα για κάθε κελί. Κάθε pixel μέσα στο κελί προσθέτει στο ιστόγραμμα τις πληροφορίες για την κατεύθυνση της κλίσης του με ένα συγκεκριμένο βάρος για την κάθε ψήφο προς μία κατεύθυνση. Συνήθως για αυτό το λόγο χρησιμοποιείται απλά σαν βάρος το πλάτος της κλίσης προς μία κατεύθυνση, αφού έχει αποδειχθεί πως δίνει τα καλύτερα αποτελέσματα. Για την καλύτερη κανονικοποίηση των πλατών της κλίσης, πολλά κελιά ομαδοποιούνται σε μεγαλύτερα χωρικά τμήματα (blocks). Πιο σύνηθες είδος σχήματος που χρησιμοποιείται είναι το ορθογώνιο που χαρακτηρίζεται από 3 παραμέτρους: α, β, γ. Ένα ορθογώνιο αναπαριστάται από $\alpha\ x\ \alpha$ πλέγματα από $\beta\ x\ \beta$ κελιά pixels, που το καθένα περιέχει γ διαφορετικές κατευθύνσεις.

Για την κανονικοποίηση των τμημάτων προτείνονται επίσης διάφορα σχήματα, που περιλαμβάνουν το v, το μη κανονικοποιημένο διάνυσμα που περιέχει όλα τα ιστογράμματα ενός τμήματος, τις νόρμες του και μια μικρή σταθερά e. Ο υπολογισμός των κανονικοποιημένων τιμών μπορεί να γίνεται με έναν από τους παρακάτω τρόπους:

$$L_{2-norm}:\ f = \frac{v}{\sqrt{\|v\|_2^2 + e^2}}$$

$$L_{1-norm}:\ f = \frac{v}{\sqrt{\|v\|_2 + e}}$$

$$L_{1-sqrt}:\ f = \sqrt{\frac{v}{\|v\|_2 + e}}$$



Σχήμα 4.2: Απεικόνιση ανθρώπου και υπολογισμών της προσανατολισμένης κλίσης συνολικά, αλλά και στα διαφορετικά κελιά για την εξαγωγή των HOG όπως στο σχήμα του [71].

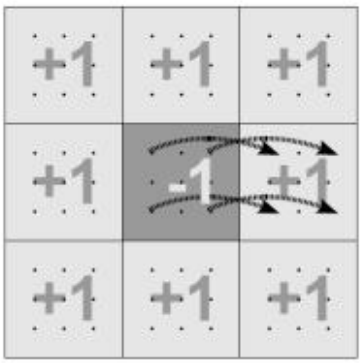
Η υλοποίηση για τον υπολογισμό των ιστογραμμάτων HOF (Σχήμα 4.3) είναι αρκετά όμοια ως προς τα στάδιά της με την μέθοδο για τον υπολογισμό των ιστογραμμάτων HOG. Για τον υπολογισμό της οπτικής ροής w συγκεκριμένα, χρησιμοποιείται για κάθε pixel ανεξάρτητα η λύση της μετριασμένης γραμμικής εξίσωσης Ελαχίστων Τετραγώνων (damped Linear Least Squares equation):

$$w = (A^T A + \beta I)^{-1} A^T b$$

πάνω σε μια μικρή $N \times N$ γειτονιά, όπου το b είναι ένα N^2διάνυσμα στήλης που κωδικοποιεί τις χρονικές διαφορές της εικόνας, ο A είναι ένας $N^2 \times 2$ πίνακας από χωρικές κλίσεις $[I_x I_y]$ και β είναι ένας μετριαστικός παράγοντας για να μειωθούν αριθμητικά θέματα από τον μοναδιαίο $A^T A$.

Στη συνέχεια, για τον υπολογισμό των διαφορών των οπτικών ροών, οι οποίες θα συμμετέχουν στα ιστογράμματα, έχουν προταθεί αρκετοί τρόποι, αλλά ένας που αναδεικνύει καλύτερα τις σχετικές κινήσεις των άκρων είναι ο IMHcd (Internal Motion Histograms cell differences). Σύμφωνα με αυτή την τεχνική, έχουμε 3 × 3 τμήματα από κελιά και σε καθένα pixel από τα 8 εξωτερικά κελιά υπολογίζονται οι διαφορές οπτικής ροής ως προς τα αντίστοιχα pixel του κεντρικού κελιού, ώστε να προκύψει ένα ιστόγραμμα προσανατολισμού. Τα 8 ιστογράμματα που προκύπτουν κανονικοποιούνται σαν ένα τμήμα (Σχήμα 4.3 (β΄)).

(α΄)



(β΄)

Σχήμα 4.3: Ιστογράμματα Οπτικής Ροής-HOF. α') Υπολογισμός οπτικής ροής και πλάτους ροής μεταξύ δύο διαδοχικών εικόνων με ανάδειξη ορίων κίνησης. β') Σχήμα κωδικοποίησης IMHcd, όπου ο υπολογισμός των διαφορών γίνεται μεταξύ αντίστοιχων pixel σε γειτονικά κελιά όπως στο σχήμα του [72].

4.2.1.3 Παράμετροι της μεθόδου εντοπισμού σημείων ενδιαφέροντος στο χωροχρόνο

Με την εισαγωγή των βίντεο στην είσοδο του συστήματός, οι ακολουθίες από καρέ επεξεργάζονται σύμφωνα με τον αλγόριθμο Harris3D. Σε κάθε σημείο του βίντεο υπολογίζεται η μήτρα δεύτερων ροπών (second moment matrix) (4.2.3) με τη χρήση ανεξάρτητων τιμών για τις χωρικές και χρονικές κλίμακες σ και τ, μιας διαχωρίσιμης συνάρτησης Γκαουσιανής ομαλοποιήσης g και των χωροχρονικών παραγώγων ∇D. Οι τελικές θέσεις των χωροχρονικών σημείων ενδιαφέροντος δίνονται από τα τοπικά μέγιστα της (4.2.4), ενώ για την επιλογή της βέλτιστης χωροχρονικής κλίμακας γίνεται χρήση της (4.2.6), αλλά σε μια απλοποιημένη μορφή όπου δοκιμάζονται συνδυασμοί χωροχρονικών κλιμάκων και επιλέγεται ο καλύτερος συνδυασμός. Με αυτόν τον τρόπο έχουμε ίδια ή και καλύτερα αποτελέσματα και σημαντική επιτάχυνση του αλγορίθμου σε σημείο να πλησιάζει την ταχύτητα του βίντεο.

Έπειτα από την επιλογή των σημείων ενδιαφέροντος για την εξαγωγή των χαρακτηριστικών μας χρησιμοποιούμε την περιγραφή των συνδυασμένων Ιστογραμμάτων Προσανατολισμένης Κλίσης και Οπτικής Ροής (HOG - HOF descriptors) στις γειτονιές των σημείων ενδιαφέροντος. Για αυτό τον συνδυασμό τα μεγέθη της περιγραφής της γειτονιάς του σημείου ενδιαφέροντος καθορίζονται ως $\Delta_x(\sigma) = \Delta_y(\sigma) = 18\sigma$, $\Delta_t(\tau) = 8\tau$, όπου $\Delta_x(\sigma), \Delta_y(\sigma)$ είναι τα χωρικά μεγέθη συναρτήσει της χωρικής κλίμακας σ και $\Delta_t(\tau)$ είναι η χρονική έκταση συναρτήσει της χρονικής κλίμακας τ. Κάθε όγκος χωρίζεται σε ένα $n_x \times n_y \times n_t$ πλέγμα από κελιά, ενώ για κάθε κελί υπολογίζονται τα Ιστογράμματα

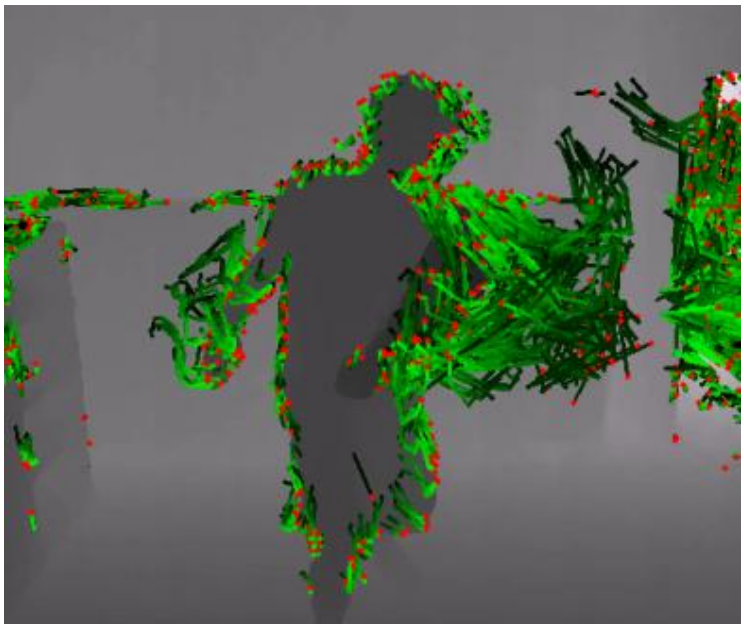
Προσανατολισμένης Κλίσης 4-κατευθύνσεων (4-bin HOG) και τα Ιστογράμματα Οπτικής Ροής 5-κατευθύνσεων (5- bin HOF).

Έπειτα από κανονικοποίηση τα στοιχεία των ιστογραμμάτων συνθέτουν το διάνυσμα των χαρακτηριστικών μας. Για την δική μας εφαρμογή χρησιμοποιούμε παραμέτρους n_x= $n_y = 3$, $n_t = 2$. Κατά συνέπεια, από κάθε σημείο ενδιαφέροντος έχουμε τον υπολογισμό ενός διανύσματος χαρακτηριστικών με 3 × 3 × 2 × 4 = 72 HOG και 3×3×2×5 = 90 HOF, δηλαδή ένα διάνυσμα διάστασης 3 (χωροχρονικη_θέση_σημείου ενδιαφέροντος - x,y,t) + 2 (χωροχρονικές_κλίμακες -σ^2, τ^2) + 72 + 90 = 167.

4.2.2 Μέθοδος εντοπισμού πυκνών τροχιών κίνησης

Η δεύτερη μέθοδος αναγνώρισης ανθρώπινης δραστηριότητας σε video, που εφαρμόστηκε στα video της βάσης δεδομένων THETIS - video που απεικονίζουν το βάθος και τον σκελετό σε τρεις διαστάσεις – είναι η μέθοδος Dense Trajectories ή πυκνές τροχιές κίνησης που προτάθηκε από τον Wang και τους συνεργάτες του [73]. Η μέθοδος Dense Trajectories στηρίζεται στην πυκνή δειγματοληψία σημείων από κάθε καρέ και παρακολουθεί την μετατόπισή τους με βάση την πληροφορία που λαμβάνει από τα πεδία οπτικής ροής. Ο αριθμός των σημείων που παρακολουθούνται μπορεί εύκολα να πολλαπλασιαστεί, εφόσον υπολογιστούν τα πεδία οπτικής ροής χωρίς κόστος. Έτσι, οι πυκνές τροχιές των σημείων περιγράφουν την κίνηση στο video (Σχήμα 4.4).

Επιπλέον για την αντιμετώπιση των προβλημάτων που προέρχονται από την κίνηση της κάμερας, στο [73] οι συγγραφείς εισήγαγαν έναν νέο τοπικό περιγραφέα που συγκεντρώνεται στην κίνηση του προσκηνίου. Ο περιγραφέας αυτός αποτελεί επέκταση του τρόπου κωδικοποίησης της κίνησης με Ιστογράμματα Ορίων Κίνησης (Motion Boundary Histograms) [72].



Σχήμα 4.4: Εφαρμογή της μεθόδου Dense Trajectories σε βίντεο της βάσης THETIS που απεικονίζει το βάθος.

4.2.2.1 Εξαγωγή Τροχιών

Οι τροχιές εξάγονται για πολλαπλές χωρικές κλίμακες (Σχήμα 4.5 αριστερά). Τα χαρακτηριστικά σημεία δειγματοληπτούνται σε ένα πλέγμα με W pixels, όπου W το βήμα δειγματοληψίας και παρακολουθούνται χωριστά για κάθε κλίμακα. Κάθε σημείο $P_t = (x_t, y_t)$ από το καρέ t, εντοπίζεται στο επόμενο καρέ t+1, μέσω ενός φίλτρου μεσαίου (median filter) σε ένα πυκνό οπτικό πεδίο οπτικής ροής $\omega = (u_t, v_t)$.

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * \omega)|_{(\bar{x}_t, \bar{y}_t)} \quad (4.2.7)$$

όπου M είναι ο πυρήνας του median filter και $(\bar{x}_t, \bar{y}_t)$ είναι η στρογγυλοποιημένη θέση του σημείου (x_t, y_t). Μόλις υπολογιστεί το πεδίο οτικής ροής, τα σημεία μπορούν να εντοπιστούν έυκολα χωρίς επιπλέον κόστος. Τα σημεία των ακολουθιών των καρέ συνενώνονται και σχηματίζουν την τροχιά $(P_t, P_{t+1}, P_{t+2}, \dots)$.

Για τον υπολογισμό της οπτικής ροής χρησιμοποιείται ο αλγόριθμος του Farneback, όπως υλοποιείται από τη βιβλιοθήκη OpenCV[7]. Ένα κοινό πρόβλημα κατά τον υπολογισμό της τροχιάς είναι η ολίσθηση από την αρχική της θέση. Για την αντιμετώπισή του, πρέπει να περιοριστεί το μήκος της τροχιάς σε L καρέ. Όταν ξεπεραστεί το όριο L, η τροχιά αφαιρείται από τη διαδικασία παρακολούθησης (Σχήμα 4.5 κέντρο). Επίσης για να διασφαλιστεί η πυκνή κάλυψη του video και για να υπάρχει σε κάθε καρέ η παρουσία ενός ίχνους στο πλέγμα, στην περίπτωση που δεν εντοπιστεί κάποιο σημείο στη γειτονιά W x W λαμβάνεται με δειγματοληψία ένα σημείο και προστίθεται στη διαδικασία για τον εντοπισμό της τροχιάς.

Το σχήμα μιας τροχιάς κωδικοποιεί τοπικά μοτίβα κίνησης. Με δεδομένη μια τροχιά μήκους L, η μέθοδος Dense Trajectories περιγράφει το σχήμα της μέσω μιας ακολουθίας $S = (\Delta P_t, \dots, \Delta P_{t+L-1})$ διανυσμάτων μετατόπισης $\Delta P_t = (P_{t+1} - P_t) = x_{t+1} - x_t, y_{t+1} - y_t)$. Το τελικό διάνυσμα κανονικοποιείται από το άθροισμα των μεγεθών των διανυσμάτων μετατόπισης :

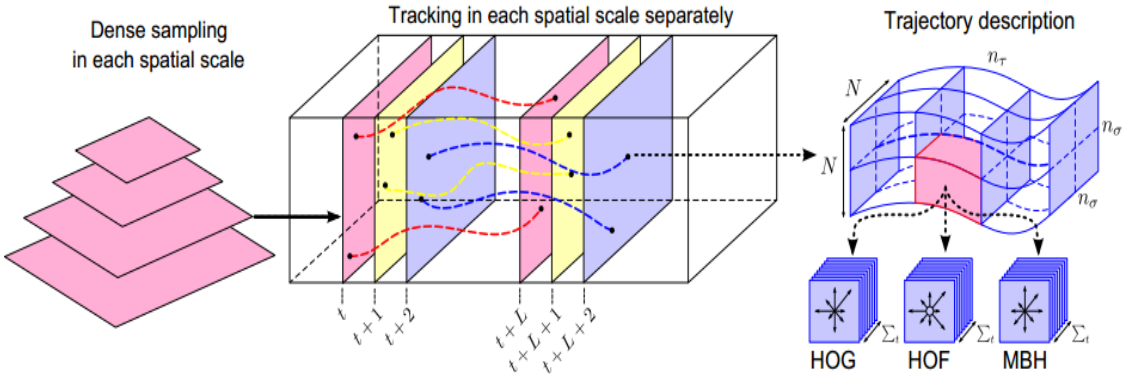
$$S' = \frac{(\Delta P_t, \dots, \Delta P_{t+L-1})}{\sum_{j=t}^{t+L-1} \|\Delta P_j\|} \quad (4.2.8)$$

Το διάνυσμα αυτό ονομάζεται περιγραφέας τροχιάς (trajectory descriptor).

4.2.2.2 Περιγραφείς Κίνησης

Προκειμένου να αποκτηθεί η πληροφορία για την κίνηση από τις τροχιές, υπολογίζονται κάποιοι περιγραφείς στον όγκο χωροχρόνου γύρω από την τροχιά (Σχήμα 4.5 δεξιά). Το μέγεθος του όγκου είναι N x N pixels και L καρέ. Για να ενσωματωθεί η πληροφορία της δομής στην παρουσίαση, ο όγκος διαιρείται σε ένα πλέγμα χωροχρόνου μεγέθους $\boldsymbol{n_\sigma}$ x $\boldsymbol{n_\sigma}$ x $\boldsymbol{n_\tau}$. Οι περιγραφείς που υπολογίζονται είναι τα Ιστογράμματα Προσανατολισμένης Κλίσης (Histograms of Oriented Gradient-HOG), τα Ιστογράμματα Οπτικής Ροής (Histograms of Optical Flow - HOF) και τα Ιστογράμματα Ορίων Κίνησης (Motion Boundary Histograms-MBH).

[7] http://opencv.willowgarage.com/wiki/



Σχήμα 4.5: Παρουσίαση της περιγραφής με πυκνές τροχιές. Αριστερά: Πυκνή δειγματοληψία χαρακτηριστικών σημείων σε πολλαπλές χωρικές κλίμακες. Κέντρο: Η παρακολούθηση τους πραγμτοποιείται στην αντίστοιχη χωρική κλίμακα για L καρέ. Δεξιά: Οι περιγραφείς τροχιάς βασίζονται στο σχήμα τους που αναπαριστάται από τις σχετικές συντεταγμένες του σημείου, και στην πληροφορία που αφορά στην εμφάνιση και την κίνηση σε μια τοπική γειτονιά από N × N pixels γύρω από την τροχιά. Για να ενσωματωθεί η πληροφορία της δομής στην παρουσίαση, ο όγκος διαιρείται σε ένα πλέγμα χωροχρόνου μεγέθους $n_\sigma \times n_\sigma \times n_\tau$.

Ο περιγραφέας MBH χωρίζει το πεδίο οπτικής ροής $I_\omega = (I_x, I_y)$ σε δυο συνιστώσες x και y. Στη συνέχεια, υπολογίζονται οι χωρικές παράγωγοι για την καθεμιά τους και η πληροφορία προσανατολισμού κβαντοποιείται σε ιστογράμματα όπως στον περιγραφέα HOG. Εφόσον οι περιγραφείς MBH αναπαριστούν την κλίση της οπτικής ροής, η πληροφορία για τις αλλαγές στο πεδίο οπτικής ροής απομονώνονται και αποθηκεύονται.

4.2.2.3 Παράμετροι της μεθόδου εντοπισμού πυκνών τροχιών κίνησης

Για τη διεξαγωγή των πειραμάτων στο πλαίσιο της εργασίας, το βήμα δειγματοληψίας που χρησιμοποιήθηκε είναι W=15. Επίσης, χρησιμοποιήθηκαν 8 χωρικές κλίμακες με παράγοντα απόστασης $1/\sqrt{2}$. Ως μήκος τροχιάς χρησιμοποιήσαμε L=15 καρέ.

Όσον αφορά στην οργάνωση του χωροχρονικού πλέγματος του όγκου σε κελιά, οι τιμές των παραμέτρων έχουν ως εξής: N=32 pixels, $n_\sigma = 2$, $n_\tau = 3$. Για κάθε σημείο ενδιαφέροντος υπολογίζονται η τροχιά Trajectory (2-bin), τα Ιστογράμματα Προσανατολισμένης Κλίσης 8-κατευθύνσεων (8-bin HOG), τα Ιστογράμματα Οπτικής Ροής 9- κατευθύνσεων (9- bin HOF), τα Ιστογράμματα Ορίων Κίνησης άξονα x 8-κατευθύνσεων (8-bin MBHx) και τα Ιστογράμματα Ορίων Κίνησης άξονα y 8-κατευθύνσεων (8-bin MBHy). Έπειτα από κανονικοποίηση τα στοιχεία των ιστογραμμάτων συνθέτουν το διάνυσμα των χαρακτηριστικών μας. Κατά συνέπεια, από κάθε σημείο ενδιαφέροντος έχουμε τον υπολογισμό ενός διανύσματος χαρακτηριστικών με 2xL=2x15=30 Trajectory, 2 × 2 × 3× 8 = 96 HOG και 2×2×3×9 = 108 HOF, 2 × 2 × 3× 8 = 96 MBHx και 2 × 2 × 3× 8 = 96 MBHy.

4.3 Κβαντοποίηση των περιγραφέων

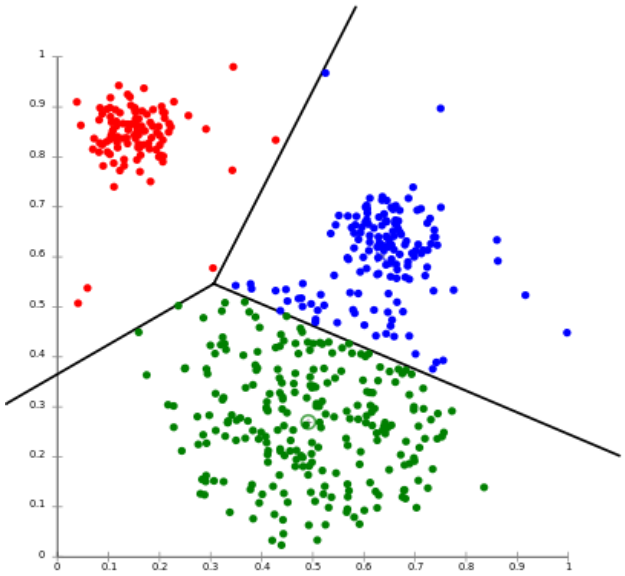
4.3.1 Δημιουργία Οπτικού Λεξικού

Είναι εύκολα κατανοητό από τα προηγούμενα κεφάλαια ότι η εξαγωγή χαρακτηριστικών κίνησης από εικονοσειρές είναι διεργασία με υψηλό υπολογιστικό κόστος, η οποία παράγει σετ δεδομένων πολύ μεγάλου μεγέθους. Πολλές φορές τα χαρακτηριστικά αυτά μπορούν να φτάσουν την τάξη των 10^5 διακριτών μεγεθών. Είναι προφανές ότι τέτοιου τύπου ακατέργαστα χαρακτηριστικά δε μπορούν να χρησιμοποιηθούν αποδοτικά σε αλγορίθμους ταξινόμησης (classification) λόγω της υπολογιστικής πολυπλοκότητας που επιβάλλουν. Ένας τρόπος να αντιμετωπιστούν τέτοιου είδους προβλήματα είναι η τεχνική **bag of words (features)**, δηλαδή η δημιουργία ενός «οπτικού λεξικού». Με τον τρόπο αυτό επιτυγχάνεται η κβαντοποίηση των περιγραφέων. Η δημιουργία του οπτικού λεξικού πραγματοποιείται με αλγορίθμους ομαδοποίησης ή αλλιώς **ομαδοποίησης (clustering)** των δεδομένων. Η ομαδοποίηση με την ευρεία έννοια είναι ένα από τα σημαντικά προβλήματα που έχει απασχολήσει την ερευνητική κοινότητα, ειδικότερα όταν έχει να κάνει με δεδομένα τα οποία μεταβάλλονται με το χρόνο, όπως συμβαίνει στις χρονοσειρές.

Με τον όρο συσταδοποίηση (clustering) αναφερόμαστε στη διαδικασία εκείνη με την οποία προσπαθούμε να εντοπίσουμε εσωτερικές δομές-σχέσεις σε ένα σύνολο δεδομένων άγνωστης κατηγορίας, οργανώνοντας τα δεδομένα αυτά σε ομογενείς ομάδες. Στόχος είναι τα δεδομένα της ίδιας ομάδας να παρουσιάζουν μεγάλη ομοιότητα μεταξύ τους και διακριτή ετερότητα με αυτά των άλλων ομάδων. Με την ομαδοποίηση, πολυπληθή σετ δεδομένων μπορούν να χρησιμοποιηθούν αποδοτικά επιλέγοντας τους χαρακτηριστικούς εκπροσώπους των κυριότερων ομάδων.

Οι αλγόριθμοι ομαδοποίησης μπορεί να είναι ιεραρχικοί (hierarchical) ή μη ιεραρχικοί/κατατμητικοί (non-hierarchical/partitional). Οι ιεραρχικοί αλγόριθμοι βρίσκουν διαδοχικές ομάδες χρησιμοποιώντας κάθε φορά ήδη καθιερωμένες ομάδες, ενώ οι μη ιεραρχικοί καθορίζουν τις ομάδες άμεσα και εξ αρχής. Οι ιεραρχικοί αλγόριθμοι χωρίζονται στους συσσωρευτικούς (agglomerative) και στους διαχωριστικούς (divisive). Οι πρώτοι αντιμετωπίζουν κάθε στοιχείο σαν μια ομάδα και στη συνέχεια συγχωνεύεται σε μεγαλύτερες ομάδες. Οι δεύτεροι ξεκινούν με ολόκληρο το σύνολο και το διασπούν σε μικρότερες ομάδες.

Στην μελέτη της εξαγωγής χαρακτηριστικών κίνησης από εικονοσειρές είδαμε ότι οι τεχνικές που χρησιμοποιήθηκαν οδηγούν σε μεγάλα σετ τέτοιων χαρακτηριστικών. Για να μπορέσουμε να περιορίσουμε τα χαρακτηριστικά αυτά σε πλήθος ικανό να χρησιμοποιηθεί αποδοτικά στη διαδικασία της αναγνώρισης ανθρώπινων κινήσεων, χρησιμοποιήθηκε ως τεχνική ομαδοποίησης η μέθοδος σαφούς διαμέρισης k μέσων (k-means). Με τη μέθοδο αυτή δοθέντος ενός συνόλου αντικειμένων, κατασκευάζονται k διαμερίσεις του συνόλου, όπου κάθε διαμέριση αναπαριστά μια ομάδα που περιέχει τουλάχιστον ένα αντικείμενο.



Σχήμα 4.6: Ομαδοποίηση Χαρακτηριστικών με αλγόριθμο k-means.

4.3.2 Ο αλγόριθμος Κ-μέσων

Η πρώτη προσέγγιση για την κατασκευή του οπτικού λεξικού έγινε από τους Sivic και Zisserman [70], οι οποίοι χρησιμοποίησαν το γνωστό επαναληπτικό αλγόριθμο συσταδοποίησης Κ-μέσων (k-means). Για την κατασκευή οπτικού λεξικού με τον αλγόριθμο k-means, ορίζουμε αρχικά τον αριθμό των κέντρων (οπτικών λέξεων) που επιθυμούμε να έχουμε, αρχικοποιούμε τα κέντρα των συστάδων με κάποια από τα σημεία του συνόλου εκπαίδευσης που θα συσταδοποιήσουμε και εκτελούμε στη συνέχεια τον επαναληπτικό αλγόριθμο. Η διαδικασία της συσταδοποίησης έχει δύο στάδια. Το στάδιο ανάθεσης και το στάδιο ανανέωσης. Ο αλγόριθμος σταματάει μέχρι να υπάρχει κάποια σύγκλιση ή μέχρι να ξεπεραστεί ένα όριο επαναλήψεων που έχουμε θέσει. Ακολούθως, παρουσιάζουμε συνοπτικά τον αλγόριθμο k-means για την κατασκευή οπτικού λεξικού.

Δοθέντος ενός συνόλου περιγραφέων $(x_1, x_2, \dots, x_n)$ διάστασης d, ο αλγόριθμος k-means στοχεύει στον διαχωρισμό των n περιγραφέων σε K σύνολα $S = \{S_1, S_2, \dots, S_K\}$ με $k \leq N$ με σκοπό την ελαχιστοποίηση του εξής αθροίσματος

$$arg \min_{S} \sum_{i=1}^{K} \sum_{x_j \in S_i} \|x_j - m_i\|^2 \qquad (4.3.1)$$

όπου m_i είναι μέσος όρος των σημείων στο S_i και στην προκειμένη περίπτωση αντιστοιχούν στα διανύσματα των λέξεων του οπτικού λεξικού.

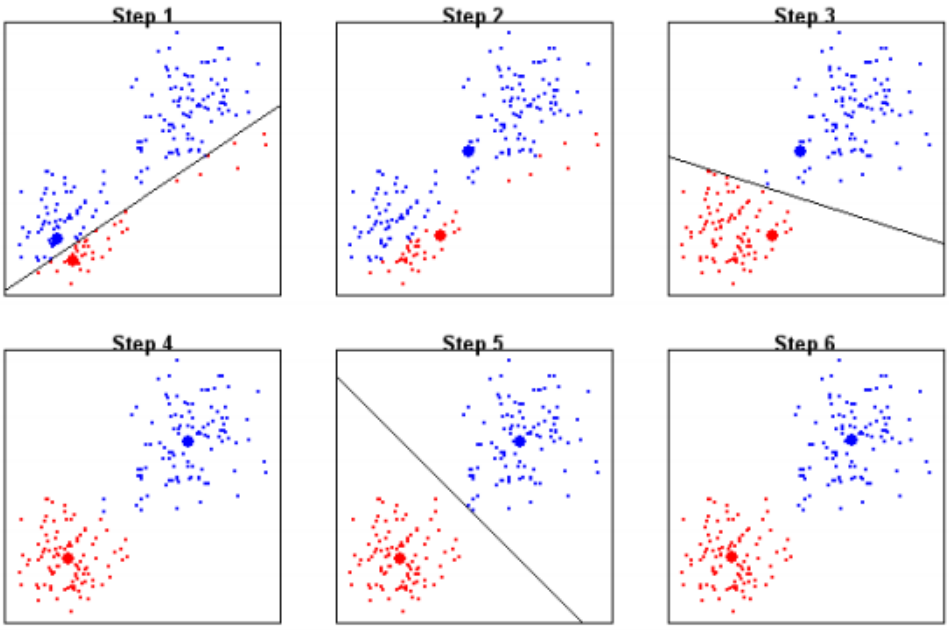
Με δεδομένο ένα αρχικό σύνολο κέντρων $m_1^{(1)}, \dots, m_K^{(1)}$, ο αλγόριθμος k-means προβαίνει στην εκτέλεση των δύο εναλλασσόμενων σταδίων:
Στάδιο Ανάθεσης: Αναθέτουμε κάθε περιγραφέα του συνόλου δεδομένων στη συστάδα με το κοντινότερο κέντρο (δηλαδή, την κοντινότερη οπτική λέξη):

$$(4.3.2)\quad S_i^{(t)} = \left\{x_p : \left\|x_p - m_i^{(t)}\right\| \leq \left\|x_p - m_j^{(t)}\right\| \forall\ 1 \leq j \leq K\right\}$$

όπου κάθε X_p ανατίθεται μόνο σε ένα $S_i^{(t)}$.

Στάδιο ανανέωσης: Υπολογίζουμε τους νέους μέσους όρους των στοιχείων κάθε συστάδας τους οποίους θεωρούμε ύστερα σαν τα νέα κέντρα της επόμενης επανάληψης $t+1$ του αλγορίθμου:

$$(4.3.3)\qquad m_i^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|} \sum_{x_j \in S_i^{(t)}} x_j$$



Σχήμα 4.7: Εφαρμογή του αλγορίθμου k-means σε ένα πρόβλημα δύο διαστάσεων. Αρχικά τα κέντρα (κύκλος) επιλέγονται τυχαία από τα δεδομένα. Στη συνέχεια ο αλγόριθμος συγκλίνει διαχωρίζοντας τα σημεία σε δύο διαφορετικές συστάδες τη μπλε και την κόκκινη.

Όπως φαίνεται από το σχήμα 4.7, στην πρώτη επανάληψη του k-means επιλέγουμε τα κέντρα τυχαία μέσα από τα δεδομένα και αναθέτουμε τα σημεία του χώρου σε αυτά (αρχικοποίηση αλγορίθμου - step 1). Στο step 2 επαναυπολογίζουμε τα κέντρα ως το μέσο όρο των στοιχείων που έχουν ανατεθεί σε αυτά. Στην επόμενη επανάληψη με δεδομένα τα νέα κέντρα, υπολογίζουμε τις νέες αναθέσεις των σημείων και επαναυπολογίζουμε τα κέντρα. Θεωρούμε ότι ο αλγόριθμος έχει συγκλίνει όταν δεν συντελείται καμία αλλαγή στο στάδιο ανάθεσης.

Ο αλγόριθμος k-means είναι εξαιρετικά διαδεδομένος καθώς είναι αρκετά γρήγορος (συγκλίνει γρήγορα σε κάποιο όριο). Σε ότι αφορά την απόδοση του, ο αλγόριθμος δεν εγγυάται ότι θα προσεγγίσει πάντα το βέλτιστο. Η ποιότητα της τελικής λύσης εξαρτάται πολύ από το αρχικό σύνολο ομάδων και μπορεί να είναι πολύ χαμηλότερη από το συνολικό βέλτιστο. Βασικό χαρακτηριστικό του αλγόριθμου που επιβάλλει και τους περιορισμούς στην απόδοσή του είναι ότι ο αριθμός των ομάδων πρέπει να οριστεί εξαρχής.

4.3.3 Ιστογράμματα Συχνότητας

Οι εικονοσειρές (video) εκπροσωπούνται από σύνολα σημείων κλειδιά που περιγράφονται από τους περιγραφείς. Επειδή όμως διαφέρουν στο πλήθος, δεν έχει κανένα νόημα η σειρά τους. Αυτό δημιουργεί δυσκολίες για τις μεθόδους ταξινόμησης που απαιτούν συνήθως διανύσματα χαρακτηριστικών με σταθερές διαστάσεις ως είσοδο. Επομένως, πριν από την διαδικασία της ταξινόμησης κάθε σημείο κλειδί κωδικοποιείται από τον δείκτη της ομάδας στην οποία ανήκει. Με αυτόν τον τρόπο, κατασκευάζεται ένα ιστόγραμμα συχνότητας των οπτικών λέξεων για κάθε εικονοσειρά (video) που θα χρησιμοποιηθεί από τον αλγόριθμο ταξινόμησης.

4.3.4 Παράμετροι αλγορίθμου Κ-μέσων

Στο πλαίσιο της διεξαγωγής των πειραμάτων, για τη δημιουργία του bag of features κατασκευάστηκε για κάθε έναν από τους περιγραφείς ένα οπτικό λεξικό με τον αλγόριθμο k-means του προγράμματος Matlab R2010b. Συγκεκριμένα, σε κάθε πείραμα πραγματοποιήθηκε k-means (k=500) για να ομαδοποιήσουμε ένα υποσύνολο 100Κ χαρακτηριστικών που επιλέχθηκαν τυχαία από το σύνολο εκπαίδευσης του πειράματος. Έτσι, πραγματοποιήθηκε k-means (k=500):

1. Στο υποσύνολο 100Κ χαρακτηριστικών του περιγραφέα HOG που προέκυψαν από την εφαρμογή του αλγορίθμου STIP στο σύνολο δεδομένων KTH.

2. Στο υποσύνολο 100Κ χαρακτηριστικών του περιγραφέα HOF που προέκυψαν από την εφαρμογή του αλγορίθμου STIP στο σύνολο δεδομένων KTH.

3. Στο υποσύνολο 100Κ χαρακτηριστικών του περιγραφέα trajectory που προέκυψαν από την εφαρμογή του αλγορίθμου Dense Trajectories στο σύνολο δεδομένων KTH.

4. Στο υποσύνολο 100Κ χαρακτηριστικών του περιγραφέα MBH που προέκυψαν από την εφαρμογή του αλγορίθμου Dense Trajectories στο σύνολο δεδομένων KTH.

5. Στο υποσύνολο 100Κ χαρακτηριστικών του περιγραφέα HOG που προέκυψαν από την εφαρμογή του αλγορίθμου Dense Trajectories στο σύνολο δεδομένων KTH.

6. Στο υποσύνολο 100Κ χαρακτηριστικών του περιγραφέα HOF που προέκυψαν από την εφαρμογή του αλγορίθμου Dense Trajectories στο σύνολο δεδομένων KTH.

7. Στο υποσύνολο 100Κ χαρακτηριστικών του περιγραφέα HOG που προέκυψαν από την εφαρμογή του αλγορίθμου STIP στο σύνολο δεδομένων THETIS_Depth.

8. Στο υποσύνολο 100Κ χαρακτηριστικών του περιγραφέα HOF που προέκυψαν από την εφαρμογή του αλγορίθμου STIP στο σύνολο δεδομένων THETIS_Depth.

9. Στο υποσύνολο 100Κ χαρακτηριστικών του περιγραφέα trajectory που προέκυψαν από την εφαρμογή του αλγορίθμου Dense Trajectories στο σύνολο δεδομένων THETIS_Depth .

10. Στο υποσύνολο 100Κ χαρακτηριστικών του περιγραφέα MBH που προέκυψαν από την εφαρμογή του αλγορίθμου Dense Trajectories στο σύνολο δεδομένων THETIS_Depth.

11. Στο υποσύνολο 100K χαρακτηριστικών του περιγραφέα HOG που προέκυψαν από την εφαρμογή του αλγορίθμου Dense Trajectories στο σύνολο δεδομένων THETIS_Depth.

12. Στο υποσύνολο 100K χαρακτηριστικών του περιγραφέα HOF που προέκυψαν από την εφαρμογή του αλγορίθμου Dense Trajectories στο σύνολο δεδομένων THETIS_Depth.

13. Στο υποσύνολο 100K χαρακτηριστικών του περιγραφέα HOG που προέκυψαν από την εφαρμογή του αλγορίθμου STIP στο σύνολο δεδομένων THETIS_Skelet3D.

14. Στο υποσύνολο 100K χαρακτηριστικών του περιγραφέα HOF που προέκυψαν από την εφαρμογή του αλγορίθμου STIP στο σύνολο δεδομένων THETIS_ Skelet3D.

15. Στο υποσύνολο 100K χαρακτηριστικών του περιγραφέα trajectory που προέκυψαν από την εφαρμογή του αλγορίθμου Dense Trajectories στο σύνολο δεδομένων THETIS_ Skelet3D.

16. Στο υποσύνολο 100K χαρακτηριστικών του περιγραφέα MBH που προέκυψαν από την εφαρμογή του αλγορίθμου Dense Trajectories στο σύνολο δεδομένων THETIS_ Skelet3D.

17. Στο υποσύνολο 100K χαρακτηριστικών του περιγραφέα HOG που προέκυψαν από την εφαρμογή του αλγορίθμου Dense Trajectories στο σύνολο δεδομένων THETIS_ Skelet3D.

18. Στο υποσύνολο 100K χαρακτηριστικών του περιγραφέα HOF που προέκυψαν από την εφαρμογή του αλγορίθμου Dense Trajectories στο σύνολο δεδομένων THETIS_ Skelet3D.

4.4 Ταξινόμηση

4.4.1 Μηχανές Διανυσμάτων Υποστήριξης

4.4.1.1 Εισαγωγή

Οι μηχανές διανυσμάτων υποστήριξης (ΜΔΥ) *(support vector machines)* (SVM) αποτελούν συστήματα εκμάθησης με χώρο υποθέσεων που περιλαμβάνει γραμμικές συναρτήσεις, δέχονται ως εισόδους δείγματα από χώρους χαρακτηριστικών πολλών διαστάσεων και εκπαιδεύονται με βάση έναν αλγόριθμο εκμάθησης στο πλαίσιο της θεωρίας βελτιστοποίησης. Τα SVM αναπτύχθηκαν ως στρατηγική εκμάθησης από τον Vapnik [77] και τους συνεργάτες του.

Συγκεκριμένα, βασική λειτουργία των SVM είναι η εύρεση υπολογιστικά αποδοτικού τρόπου εκμάθησης υπερεπιπέδων *(hyperlines)* διαχωρισμού σε χώρους χαρακτηριστικών πολλών διαστάσεων. Στην ταξινόμηση διανυσμάτων υποστήριξης *(support vector classification),* ως υπερεπίπεδα διαχωρισμού ορίζονται αυτά που βελτιστοποιούν τα όρια γενίκευσης (*generalization bounds*) και ως υπολογιστικά αποδοτικούς αλγορίθμους εκμάθησης, αυτούς που ανταποκρίνονται σε μεγέθη συνόλων δεδομένων της τάξης των 100000 δειγμάτων.

Τα SVM αντιμετωπίζουν επιτυχώς τα προβλήματα αποδοτικότητας όσον αφορά στην εκπαίδευση, στον έλεγχο νέων δεδομένων *(testing)* ,στο υπερταίριασμα *(overfitting)* των μηχανών εκμάθησης και στην αναξιοπιστία των εκάστοτε χειρισμών *(heuristics),* όπως προκύπτει από τα παρακάτω:

- Λόγω των συνθηκών του Mercer για τους πυρήνες (kernels) που χρησιμοποιούνται στην εκμάθηση, τα αντίστοιχα προβλήματα βελτιστοποίησης είναι κυρτά και συνεπώς, δεν εμφανίζονται τοπικά ελάχιστα και κάθε τοπική λύση αποτελεί και ολικό βέλτιστο. Επομένως, εξασφαλίζεται η ύπαρξη λύσεων ακόμη και για μεγάλα σύνολα εκπαίδευσης.

- Η δυνατότητα χρήσης συναρτήσεων πυρήνων *(kernel functions)* αποτελεί το κλειδί για την αποδοτική χρήση των SVM σε χώρους πολλών διαστάσεων.

- Η θεωρία γενίκευσης που εφαρμόζουν εξασφαλίζει τον έλεγχο της χωρητικότητας και συνεπώς την αποφυγή του overfitting, το οποίο είναι εγγενές σε χώρους πολλών διαστάσεων ελέγχοντας τα μέτρα του περιθωρίου *(margin)* των υπερεπιπέδων διαχωρισμού, ενώ παράλληλα η θεωρία βελτιστοποίησης παρέχει τις μαθηματικές τεχνικές που απαιτούνται για την εύρεση των υπερεπιπέδων που βελτιστοποιούν αυτά τα μέτρα.

- Παρέχουν συμπαγείς και αραιές δυικές αναπαραστάσεις της εκάστοτε υπόθεσης μειώνοντας τον αριθμό των παραμέτρων εκμάθησης, ευνοώντας την εξαγωγή αποδοτικών αλγορίθμων εκμάθησης.

Για όλους τους παραπάνω λόγους, κατά τη διεξαγωγή των πειραμάτων επιλέχθηκε η χρήση μηχανών διανυσμάτων υποστήριξης για την ταξινόμηση των κινήσεων.

4.4.1.2 Βέλτιστο υπερεπίπεδο για γραμμικά διαχωρίσιμα πρότυπα.

Θεωρούμε δεδομένο εκπαίδευσης το $\{(x_i, d_i)\}_{i=1..n}$, όπου το x_i είναι το πρότυπο εισόδου για το i-οστό παράδειγμα και d_iείναι το επιθυμητό αποτέλεσμα (target output). Για αρχή, θεωρούμε ότι το πρότυπο (κλάση) που αναπαρίσταται από το υποσύνολο d_i = +1 και το πρότυπο που αναπαρίσταται από το υποσύνολο d_i = -1 είναι «γραμμικά διαχωρίσιμα». Η εξίσωση για την επιφάνεια απόφασης στη μορφή ενός υπερεπιπέδου που κάνει το διαχωρισμό είναι:

$$w^T x + b = 0 \qquad (4.4.1)$$

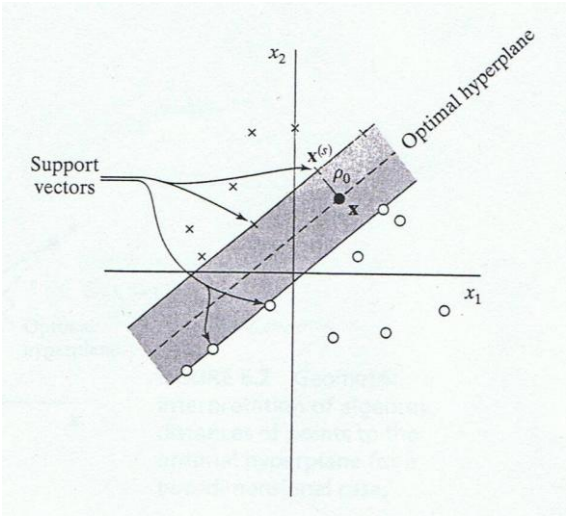
όπου x είναι ένα διάνυσμα εισόδου (input vector), το w είναι ένα ρυθμιζόμενο διάνυσμα βαρών, και το b είναι το κατώφλι. Μπορούμε επομένως να γράψουμε:

$$w^T x_i + b \geq 0 \ \ \gamma\iota\alpha \ d_i = +1 \qquad (4.4.2)$$
$$w^T x_i + b \leq 0 \ \ \gamma\iota\alpha \ d_i = -1$$

Η υπόθεση για γραμμικά διαχωρίσιμα πρότυπα γίνεται εδώ για να εξηγήσουμε τη βασική ιδέα πίσω από ένα support vector machine υπό ένα πιο απλό σκηνικό.

Για δεδομένο διάνυσμα βαρών w και κατώφλι b, ο διαχωρισμός μεταξύ της υπερεπιφάνειας που ορίστηκε στην εξίσωση (4.4.1) και του κοντινότερου data point (σημείο δεδομένων) λέγεται margin of separation ή αλλιώς περιθώριο διαχωρισμού, που δηλώνεται από το ρ. Ο στόχος ενός support vector machine είναι να βρεί την συγκεκριμένη υπερεπιφάνεια για την οποία το περιθώριο διαχωρισμού ρ να είναι το μέγιστο. Υπό αυτές τις

συνθήκες η επιφάνεια διαχωρισμού αναφέρεται ως η βέλτιστη υπερεπιφάνεια (optimal hyperplane). Το σχήμα (4.8) απεικονίζει τη γεωμετρική κατασκευή μιας βέλτιστης υπερεπιφάνειας για ένα δισδιάστατο χώρο εισόδου.



Σχήμα 4.8: Απεικόνιση μιας βέλτιστης υπερεπιφάνειας για γραμμικά διαχωρίσιμα πρότυπα.

Έστω ότι τα w_0 *και* b_0 δηλώνουν τις βέλτιστες τιμές του διανύσματος βαρών και του κατωφλίου, αντίστοιχα. Παρομοίως, η βέλτιστη υπερεπιφάνεια που αναπαριστά μια πολυδιάστατη γραμμική επιφάνεια απόφασης στο χώρο είσοδου ορίζεται ως

$$w_0^T x + b_0 = 0 \qquad (4.4.3)$$

η οποία είναι επαναδιατύπωση της εξίσωσης (4.4.1).

Η συνάρτηση:

$$g_x = w_o^T x + b_0 \qquad (4.4.4)$$

δίνει την αλγεβρική ποσότητα της απόστασης μεταξύ του x και της βέλτιστης υπερεπιφάνειας.

Ίσως ο πιό εύκολος τρόπος για να το δούμε αυτό είναι να εκφράσουμε το x ως:

$$x = x_p + r\left(\frac{w_0}{\|w_0\|}\right)$$

όπου το x_p είναι η κανονική προβολή του x πάνω στη βέλτιστη υπερεπιφάνεια, και το r είναι η επιθυμητή αλγεβρική απόσταση (το r είναι θετικό αν το x βρίσκεται στη θετική πλευρά της βέλτιστης υπερεπιφάνειας και αρνητικό αν το x βρίσκεται στην αρνητική πλευρά). Εφόσον εξ'ορισμού το g(x_p) = 0 τότε ακολούθως,

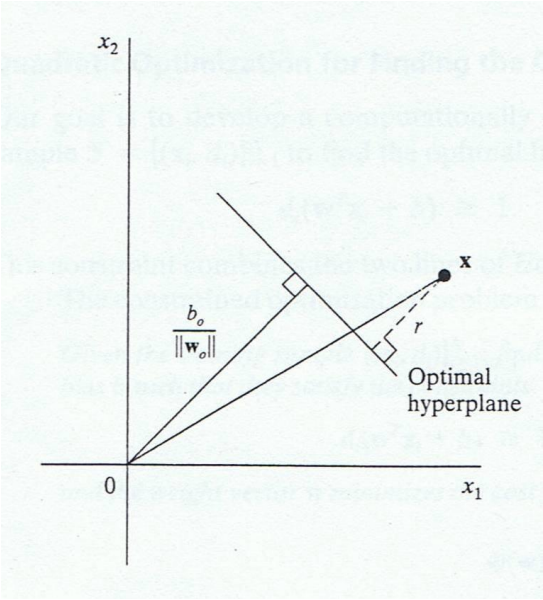
$$g_x = w_o^T x + b_0 = r\|w_0\| \qquad (4.4.5)$$

ή

$$r = \frac{g(x)}{\|w_0\|}$$

Πιο συγκεκριμένα, η απόσταση από την αρχή συντεταγμένων (δηλαδή, x = 0) μέχρι τη βέλτιστη υπερεπιφάνεια δίνεται από το $\frac{b_0}{\|w_o\|}$.

Αν $b_0 > 0$, η αρχή συντεταγμένων βρίσκεται στη θετική πλευρά της βέλτιστης υπερεπιφάνειας και αν $b_0 < 0$, βρίσκεται στην αρνητική πλευρά. Αν $b_0 = 0$, τότε η βέλτιστη υπερεπιφάνεια περνά διαμέσου της αρχής συντεταγμένων. Μια γεωμετρική αναπαράσταση αυτών των αλγεβρικών αποτελεσμάτων βρίσκεται στο σχήμα 4.9.



Σχήμα 4.9: Γεωμετρική αναπαράσταση των αλγεβρικών αποστάσεων μεταξύ των σημείων και της βέλτιστης υπερεπιφάνειας για δισδιάστατο χώρο.

Το ζητούμενο είναι να βρούμε τις παραμέτρους $w_0 και\ b_0$ της βέλτιστης υπερεπιφάνειας, δεδομένου ενός συνόλου εκπαίδευσης $J = \{(x_i, d_i)\}$. Λαμβάνοντας υπ'όψη τα αποτέλεσματα του σχήματος 4.9, βλέπουμε ότι το ζεύγος (w_0, b_0) πρέπει να τηρεί τους περιορισμούς:

$$(4.4.6) \qquad \begin{aligned} {w_0}^T x_i + b_0 &\geq 1 \ \ για\ d_i = +1 \\ w_0^T x_i + b_0 &\leq -1 \ \ για\ d_i = -1 \end{aligned}$$

Να σημειώσουμε ότι, αν η εξίσωση (4.4.2) ισχύει, δηλαδή τα πρότυπα είναι γραμμικά διαχωρίσιμα, μπορούμε πάντα να ξαναφτιάξουμε τα $w_0 και\ b_0$ σε μικρότερη κλίμακα έτσι ώστε να ισχύει και η εξίσωση (4.4.6) (αυτό αφήνει την εξίσωση (4.4.3) ανεπηρέαστη).

Τα ειδικά σημεία (x_i, d_i) για τα οποία η πρώτη ή δεύτερη γραμμή της εξίσωσης (4.4.6) ικανοποιείται με το σήμα ισότητας λέγονται διανύσματα υποστήριξης (support vectors), εξ'ού και το όνομα "support vector machine". Αυτά τα διανύσματα (vectors) παίζουν αξιοπρόσεκτο ρόλο στη λειτουργία αυτού του είδους αλγορίθμων εκμάθησης. Τα support vectors είναι εκείνα τα σημεία τα οποία βρίσκονται κοντινότερα στην επιφάνεια απόφασης και επομένως είναι και τα πιο δύσκολα για να κατηγοριοποιηθούν. Θεωρούμε ένα support vector $x^{(s)}$ για το οποίο $d^{(s)} = +1$. Τότε εξ'ορισμού, έχουμε

$$(4.4.7) \qquad g\left(x^{(s)}\right) = w_0^T x^{(s)} \pm b_0 = \pm 1 \ \ για\ \ d^{(s)} = \pm 1$$

Από την εξίσωση (4.4.5) η αλγεβρική απόσταση μεταξύ του support vector $x^{(s)}$ και της βέλτιστης υπερεπιφάνειας είναι

$$(4.4.8) \qquad r = \frac{g(x^{(s)})}{\|w_0\|} = \begin{cases} = \frac{1}{\|w_0\|} & \alpha\nu \ d^{(s)} = +1 \\ = \frac{-1}{\|w_0\|} & \alpha\nu \ d^{(s)} = -1 \end{cases}$$

όπου το θετικό πρόσημο υποδηλώνει ότι το $x^{(s)}$ βρίσκεται στη θετική πλευρά της βέλτιστης υπερεπιφάνειας και το αρνητικό πρόσημο υποδηλώνει ότι βρίσκεται στην αρνητική πλευρά. Έστω ότι το ρ δηλώνει τη μέγιστη τιμή του περιθωρίου διαχωρισμού (margin of separation) μεταξύ δύο κλάσεων που αποτελούνται από το σύνολο εκπαίδευσης J. Τότε από την εξίσωση (4.4.8) προκύπτει ότι:

$$(4.4.9) \qquad \rho = 2r = \frac{2}{\|w_0\|}$$

Η εξίσωση (4.4.9) δηλώνει ότι μεγιστοποιώντας το περιθώριο διαχωρισμού μεταξύ των κλάσεων είναι ισοδύναμο με την ελαχιστοποίηση του Ευκλείδειου μέτρου (Euclidean norm) του διανύσματος βαρών w.

Περιληπτικά, η βέλτιστη υπερεπιφάνεια που ορίζεται από την εξίσωση (4.4.3) είναι μοναδική υπό την έννοια ότι το βέλτιστο διάνυσμα βαρών w_0 παρέχει το μέγιστο δυνατό διαχωρισμό μεταξύ των θετικών και αρνητικών παραδειγμάτων. Αυτή η βέλτιστη κατάσταση επιτυγχάνεται με την ελαχιστοποίηση του Ευκλείδειου μέτρου του διανύσματος βαρών w.

4.4.1.3 Βέλτιστο υπερεπίπεδο για μη-γραμμικά διαχωρίσιμα πρότυπα

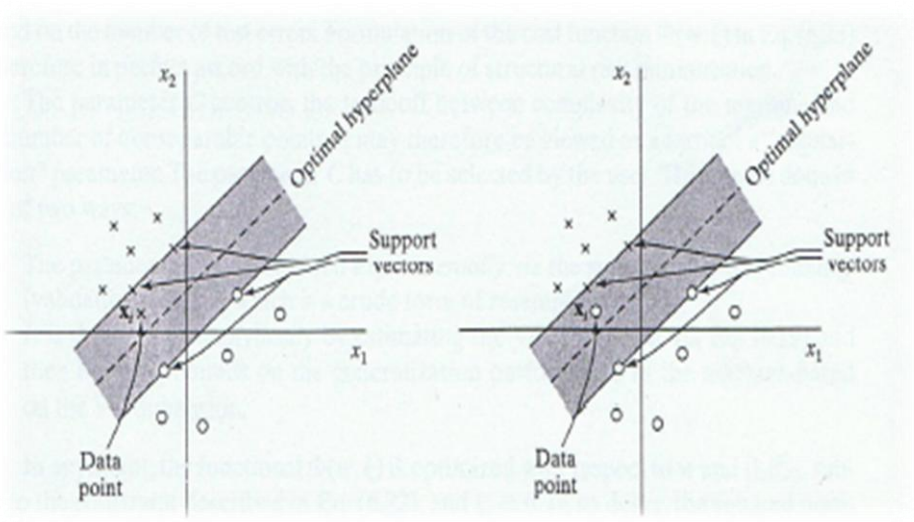
Μέχρι τώρα έχουμε αναφερθεί στα γραμμικά διαχωρίσιμα πρότυπα. Στη συνέχεια, θα ασχοληθούμε με την πιό δύσκολη περίπτωση μη γραμμικών διαχωρίσιμων προτύπων. Δεδομένου ενός τέτοιου συνόλου εκπαίδευσης, είναι δύσκολο να κατασκευάσουμε μια διαχωριστική υπερεπιφάνεια χωρίς να αντιμετωπίσουμε λάθη κατηγοριοποίησης. Παρ'όλα αυτά, θέλουμε να βρούμε μια βέλτιστη υπερεπιφάνεια η οποία ελαχιστοποιεί την πιθάνοτητα λαθών στην κατηγοριοποίηση. Το περιθώριο διαχωρισμού μεταξύ των κλάσεων λέγεται soft, δηλαδή μαλακό, στην περίπτωση όπου ένα δεδομένο (xi, di) παραβιάζει την εξής συνθήκη:

$$(4.4.10) \qquad d_i(w^T x_i + b) \geq 1 \ \ \gamma\iota\alpha \ i = 1,2,\dots,N$$

Αυτή η παραβίαση μπορεί να προκύψει για δύο λόγους:

- Το δεδομένο (x_i, d_i) πέφτει εντός της περιοχής διαχωρισμού αλλά στη σωστή πλευρά της επιφάνειας απόφασης (σχήμα 4.10 α)

- Το δεδομένο (x_i, d_i)πέφτει στη λάθος πλευρά της επιφάνειας απόφασης (σχήμα 4.10 β)



Σχήμα 4.10: α) Το σημείο x_i (που ανήκει στη κλάση β1) πέφτει εντός της περιοχής διαχωρισμού αλλά στη σωστή πλευρά της επιφάνειας απόφασης. β) Το σημείο x_i (που ανήκει στην κλάση β2) πέφτει στη λάθος πλευρά της επιφάνειας απόφασης.

Παρατηρούμε ότι στην πρώτη περίπτωση έχουμε σωστή κατηγοριοποίηση, ενώ στη δεύτερη περίπτωση έχουμε λανθασμένη κατηγοριοποίηση. Για το χειρισμό των μη γραμμικά διαχωρίσιμων δεδομένων εισάγουμε ένα καινούργιο σύνολο από μη αρνητικές βαθμωτές μεταβλητές $\{\xi_i\}_{i=1..N}$ στον ορισμό της υπερεπιφάνειας διαχωρισμού (δηλαδή επιφάνειας απόφασης) όπως φαίνεται πιο κάτω:

$$d_i(w^T x_i + b) \geq 1 - \xi_i \,, για\ i = 1,2,\dots,N \qquad (4.4.11)$$

Η μεταβλητή ξ_i λέγεται slack variable, και μετρά την απόκλιση ενός δεδομένου από την ιδανική κατάσταση στο διαχωρισμό προτύπων. Για $0 \leq \xi_i \leq 1$, το δεδομένο πέφτει εντός της περιοχής διαχωρισμού αλλά στη σωστή πλευρά της επιφάνειας απόφασης, όπως φαίνεται και στο σχήμα 4.10 α. Για $\xi_i > 1$, πέφτει στη λάθος πλευρά της υπερεπιφάνειας διαχωρισμού, όπως φαίνεται και στο σχήμα 4.10 β. Τα support vectors είναι εκείνα τα σημεία δεδομένων τα οποία ικανοποιούν ακριβώς την εξίσωση (4.4.11), ακόμα και αν $\xi_i > 0$. Σημειώνουμε ότι εάν ένα σημείο με $\xi_i > 0$ δεν συμπεριληφθεί στο σύνολο εκπαίδευσης, τότε η επιφάνεια απόφασης θα αλλάξει. Επομένως, τα support vectors ορίζονται με τον ίδιο ακριβώς τρόπο για τα διαχωρίσιμα και για τα μη γραμμικά διαχωρίσιμα δεδομένα.

Ο στόχος μας είναι να βρούμε μια διαχωριστική υπερεπιφάνεια για την οποία το λάθος της λαθασμένης κατηγοριοποίησης να είναι το ελάχιστο. Μπορούμε να το κάνουμε αυτό με την ελαχιστοποίηση της συνάρτησης:

$$\Phi(\xi) = \sum_{i=1\dots N} I(\xi_i - 1)$$

σε σχέση με το διάνυσμα βαρών w, υπό τους περιορισμούς που περιγράφηκαν στην εξίσωση (4.4.11) και τον περιορισμό στο $\|w^2\|$. Η συνάρτηση I(ξ) είναι μια συνάρτηση δείκτη *(indicator function*) που ορίζεται από το

$$I(\xi) = 0 \ \ \alpha\nu\ \xi \leq 0$$
$$I(\xi) = 1 \ \ \alpha\nu\ \xi > 0$$

Δυστυχώς, η ελαχιστοποίηση του Φ(ξ) σε σχέση με το w είναι ένα nonconvex πρόβλημα βελτιστοποίησης που είναι NP-complete. Για να κάνουμε το πρόβλημα βελτιστοποίησης μαθηματικώς υπάκουο, προσεγγίζουμε τη συνάρτηση Φ(ξ) γράφοντας :

$$\Phi(\xi) = \sum_{\iota=1\ldots N} \xi_i$$

Επιπλέον, απλοποιούμε τον υπολογισμό διατυπώνοντας τη συνάρτηση που θα ελαχιστοποιηθεί σε σχέση με το διάνυσμα βαρών w ως ακολούθως:

$$(4.4.12) \qquad \Phi(w, \xi) = \frac{1}{2} w^T w \; + C \sum_{\iota=1\ldots N} \xi_i$$

4.4.1.4 ΜΔΥ-πολλών κλάσεων

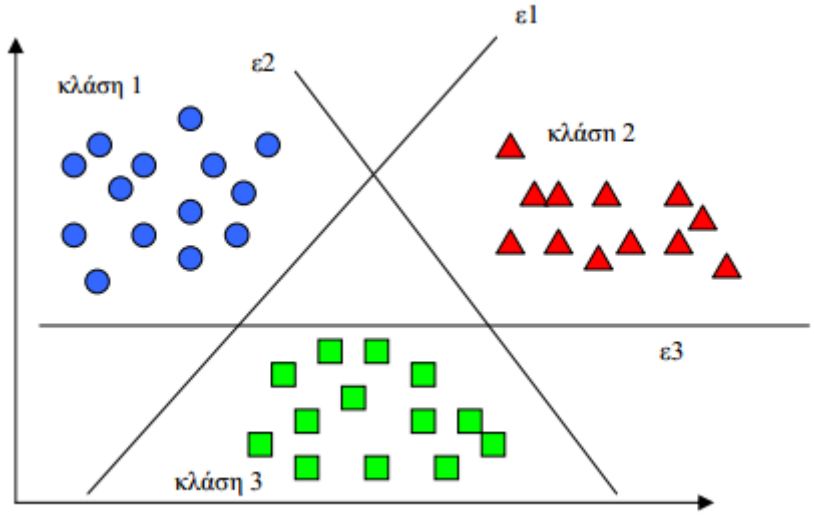
Στις προηγούμενες παραγράφους περιγράφηκε ο τρόπος διαχωρισμού δυο κλάσεων. Στην πλειοψηφία των προβλημάτων όμως, θα πρέπει να γίνει η ταξινόμηση σε περισσότερες των δύο κλάσεων. Έχουν αναπτυχθεί διάφορες μέθοδοι για το συνδυασμό πολλών δυαδικών SVM, με στόχο τη δημιουργία ΜΔΥ-πολλών κλάσεων. Οι πιο συχνά χρησιμοποιούμενες προσεγγίσεις (Vapnik, 1998) είναι οι one-against-all, all-together, one-against-one.

- *One-against-all*: Με αυτόν τον τρόπο προσπαθούμε να βρούμε τα όρια της κάθε κλάσης εναντίον όλων των άλλων μαζί και έτσι έχουμε k συναρτήσεις απόφασης αφού έχουμε k διαχωριστικά υπερεπίπεδα. Τα όρια της κάθε κλάσης προέρχονται ύστερα από συμψηφισμό των αποτελεσμάτων των k SVM. Για την διαδικασία της κατάταξης τώρα, υπολογίζουμε την τιμή των k συναρτήσεων απόφασης και κατατάσσουμε το σημείο στην κλάση που αντιστοιχεί στην συνάρτηση απόφασης με τη μεγαλύτερη τιμή. Αυτή η λογική επιφέρει λάθη κατά τη διαδικασία της κατάταξης αφού εξετάζει μόνο τη τιμή της συνάρτησης απόφασης. Στο Σχήμα 4.11 φαίνεται η αρχή της μεθόδου αυτής σε ένα δισδιάστατο παράδειγμα με τρεις κλάσεις και τα διαχωριστικά επίπεδα που προήλθαν από αυτή τη μέθοδο. Η ευθεία ε1 διαχωρίζει την κλάση 1 από όλες τις υπόλοιπες. Ομοίως, οι ευθείες ε2 και ε3, διαχωρίζουν τις κλάσεις 2 και 3 από όλες τις υπόλοιπες, αντίστοιχα.

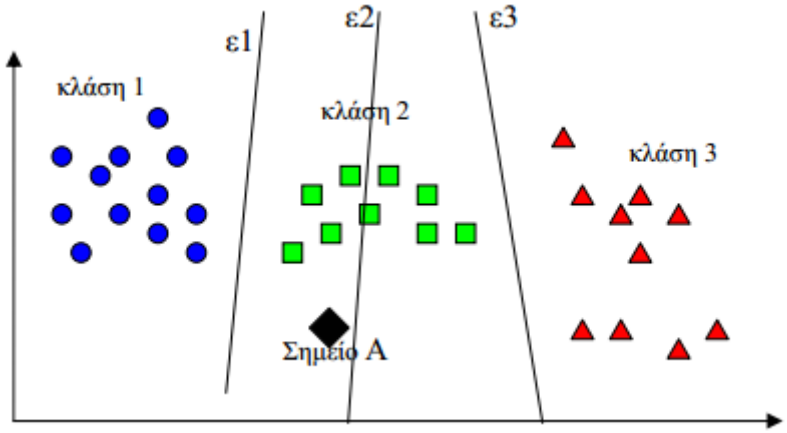
- *All-together:* Η μέθοδος αυτή είναι παραπλήσια της one-against-all. Δημιουργούνται k συναρτήσεις απόφασης για τις k κλάσεις, όπου, για παράδειγμα η n συνάρτηση απόφασης διαχωρίζει την n κλάση από όλες τις άλλες. Στην one-against-all επιλύονται k SVM για να βρεθούν τα όρια της μίας κλάσης, ενώ εδώ, συνοψίζονται στη λύση ενός. Η συνάρτηση απόφασης αυτής της μεθόδου είναι η ίδια με την one-against-all και το σημείο κατατάσσεται στην κλάση που έχει τη μέγιστη τιμή η αντίστοιχη συνάρτηση απόφασης.

- *One-against-one*: Σε αντίθεση με τη one-against-all, εδώ θεωρούνται όλα τα πιθανά ζεύγη των κλάσεων και λύνεται το πρόβλημα διαχωρισμού τους. Δηλαδή, για k κλάσεις έχουμε k!/[2(k-2)!] δυνατούς συνδυασμούς των κλάσεων. Επίσης, τόσες είναι και οι συναρτήσεις απόφασης. Πιο απλά, εφαρμόζουμε τη μεθοδολογία που αναπτύχθη στις προηγούμενες παραγράφους και βρίσκουμε το διαχωριστικό υπερεπίπεδο κάθε πιθανού ζεύγους κλάσεων. Η πιο διαδεδομένη μέθοδος για την κατάταξη ενός σημείου αν έχουμε ακολουθήσει τη διαδικασία της one-against-one είναι η max-wins. Με τον αλγόριθμο αυτό, κάθε προς κατάταξη σημείο ελέγχεται με όλες τις συναρτήσεις απόφασης και κάθε φορά κατατάσσεται σε μια κλάση. Η κλάση στην οποία έχει καταταχθεί τις περισσότερες φορές είναι και η κλάση στην οποία

τελικά τοποθετείται το σημείο. Στο Σχήμα 4.12 βλέπουμε ένα δισδιάστατο πρόβλημα τριών κλάσεων και τις διαχωριστικές ευθείες.



Σχήμα 4.11 : Διαχωρισμός τριών κλάσεων με τη μέθοδο one-against-all.



Σχήμα 4.12 : Δισδιάστατο πρόβλημα τριών κλάσεων που διαχωρίζονται με τη μέθοδο one-against-one, καθώς επίσης και η κατάταξη ενός τυχαίου σημείου.

4.4.2 Παράμετροι ΜΔΥ

Εφόσον τα σύνολα δεδομένων THETIS_Depth, THETIS_Skelet3D και KTH περιέχουν video με περισσότερες από δυο κατηγορίες κινήσεων (12 κλάσεις κινήσεων τα THETIS_Depth, THETIS_Skelet3D και 6 κλάσεις κινήσεων το KTH), για την ταξινόμηση των video χρησιμοποιήσαμε ένα μη-γραμμικό SVM με RBF kernel. Οι διαφορετικοί περιγραφείς συνδυάζονται σε μια προσέγγιση πολυκαναλική (multi-channel kernel) όπως στο [74]

$$K(x_i, x_j) = \exp\left(-\sum_c \frac{1}{A^c} D(x_i^c, x_j^c)\right),$$

όπου $D\left(x_i^c, x_j^c\right)$ είναι η ευκλείδια απόσταση μεταξύ των video x_i και x_j, λαμβάνοντας υπ'όψιν το c-ιοστό κανάλι-περιγραφέα. A^c είναι η μέση τιμή των χ^2 αποστάσεων των δειγμάτων εκπαίδευσης για τον c-ιοστό περιγραφέα [75]. Επομένως, στην πραγματικότητα αθροίζονται οι αποστάσεις $D\left(x_i^c, x_j^c\right)$ και εφαρμόζεται ένα RBF kernel στο άθροισμά τους.

Επίσης, επειδή πρόκειται για ταξινόμηση πολλών κλάσεων χρησιμοποιείται η προσέγγιση one-against-all. Για την εκπαίδευση και την επαλήθευση του SVM χρησιμοποιήσαμε το πρωτόκολλο **leave-one-out cross-validation (LOOCV)**. Η τεχνική αυτή κρατά τα δείγματα (video) που αφορούν ένα συγκεκριμένο άτομο ως σύνολο επαλήθευσης (testing set), ενώ τα υπόλοιπα δείγματα χρησιμοποιούνται ως (training set). Αυτό επαναλαμβάνεται N φορές ώστε το σύνολο δειγμάτων για κάθε άτομο να χρησιμοποιηθεί ακριβώς μια φορά ως (testing set). Επομένως, για το σύνολο δεδομένων KTH που περιλαμβάνει 24 άτομα, η διαδικασία επαναλαμβάνεται 24 φορές ενώ για τα σύνολα δεδομένων THETIS_Depth και THETIS_Skelet3D η διαδικασία επαναλαμάνεται 55 φορές. Τέλος, σε κάθε πείραμα τα N αποτελέσματα συνδυάζονται και προκύπτει το τελικό αποτέλεσμα.

4.6 Παρουσίαση Αποτελεσμάτων

4.6.1 Δείκτες Αξιολόγησης

Για την αξιολόγηση των μεθόδων χρησιμοποιούμε τέσσερις δείκτες επίδοσης, το μέσο όρο της *ορθότητας* ταξινόμησης *(average classification accuracy*), όπως αυτός διαμορφώνεται επί ενός αριθμού διαφορετικών χωρισμάτων των video σε σύνολα εκπαίδευσης και ελέγχου ,την *ακρίβεια* ταξινόμησης *(classification precision)* ανά κλάση, την *ορθότητα* ταξινόμησης (*classification accuracy*) ανά κλάση και τον *πίνακα "σύγχυσης*" *(confusion matrix*). Ο δείκτης ορθότητας ταξινόμησης (classification accuracy) ορίζεται ως

$$Μέσος\ όρος\ ορθότητας\ ταξινόμησης = \frac{ορθά\ ταξινομημένα\ video\ ελέγχου}{σύνολο\ video\ ελέγχου}$$

Ο δείκτης ακρίβειας ταξινόμησης (classification precision) για μια κλάση ορίζεται ως

$$ακρίβεια\ ταξινόμησης\ κλάσης\ j =$$

$$= \frac{video\ ελέγχου\ που\ ταξινομήθηκαν\ ορθώς\ στην\ κλάση\ j}{video\ ελέγχου\ που\ ταξινομήθηκαν\ ορθώς\ στην\ κλάση\ j\ + \\ video\ ελέγχου\ που\ ταξινομήθηκαν\ \ λανθασμένα\ στην\ κλάση\ j}$$

Ο δείκτης ορθότητας ταξινόμησης (classification accuracy) για μια κλάση ορίζεται ως

$$ορθότητα\ ταξινόμησης\ κλάσης\ j =$$

$$= \frac{video\ ελέγχου\ που\ ταξινομήθηκαν\ ορθώς\ \ στην\ κλάση\ j}{video\ ελέγχου\ που\ ταξινομήθηκαν\ ορθώς\ \ στην\ κλάση\ j\ + \\ video\ ελέγχου\ κλάσης\ j\ που\ δεν\ \ ταξινομήθηκαν\ \ \ στην\ κλάση\ j}$$

Ακολούθως, ο πίνακας "σύγχυσης" ορίζεται ως

$$M_{ij} = \frac{\left|\{V_K \in C_j : h(V_K) = i\}\right|}{C_j},$$

όπου $i, j \in \{1, \dots, N_c\}$, M_{ij} είναι η τιμή του πίνακα σύγχυσης στη θέση (j, i) και C_j το σύνολο των video ελέγχου κατηγορίας j. $h(V_K)$είναι η κατηγορία, η οποία έλαβε τη μέγιστη τιμή απόφασης ταξινομητή για το video V_K, δηλαδή η κατηγορία στην οποία τελικά ταξινομήθηκε το video V_K. Επομένως, η τιμή του M_{ij} ισούται με το πλήθος των video που ανήκουν στην κατηγορία j αλλά ταξινομήθηκαν στην κατηγορία i , δια το πλήθος των video που ανήκουν στην κατηγορία j. Συνήθως οι τιμές των accuracy, precision, accuracy και confusion matrix εκφράζονται σε ποσοστά επί τις εκατό (%).

4.6.2 Αποτελέσματα Πρώτης Μεθόδου

Στην ενότητα αυτή παρουσιάζονται τα αποτελέσματα που προέκυψαν από την εφαρμογή της μεθόδου STIP στα δεδομένα των συνόλων THETIS_Depth , THETIS_Skelet3D και Depth για τον συνδυασμό των δυο περιγραφέων HOG και HOF. Στο σχήμα 4.13 καταγράφεται ο μέσος όρος ακρίβειας / ορθότητας (accuracy) των αποτελεσμάτων ταξινόμησης και για τα τρια σύνολα δεδομένων.

Σύνολο Δεδομένων	Average Accuracy (%)
THETIS_Depth	60.23%
THETIS_Skelet3D	54.40%
KTH	92.99%

Σχήμα 4.13 : Μέσος όρος ακρίβειας ταξινόμησης για τη μέθοδο STIP.

Στα σχήματα 4.14, 4.15 και 4.16 παρουσιάζονται πιο αναλυτικά τα αποτελέσματα ταξινόμησης για κάθε κλάση κινήσεων, από την εφαρμογή της μεθόδου STIP στο σύνολο δεδομένων THETIS_Depth.

Είδος Κίνησης	Precision (%)	Accuracy (%)
Backhand with two hands	61,29%	69,09%
Backhand	63,93%	70,91%
Backhand slice	63,47%	64,24%
Backhand volley	68,79%	65,45%
Forehand flat	59,30%	62,20%
Forehand open stands	**73,17%**	**72,73%**
Forehand slice	64,97%	61,82%
Forehand volley	63,82%	58,79%
Service flat	46,95%	46,67%
Service kick	52,60%	49,09%
Service slice	53,95%	49,70%
Smash	50,29%	52,12%

Σχήμα 4.14: Ποσοστά precision και accuracy κάθε κλάσης για το σύνολο THETIS_Depth.

Κινήσεις	1	2	3	4	5	6	7	8	9	10	11	12
1	114	8	4	4	15	8	3	0	5	1	1	2
2	11	117	4	5	6	6	4	5	0	2	2	3
3	4	13	106	19	6	1	4	5	3	2	0	2
4	1	7	21	108	8	2	6	11	0	0	0	1
5	17	5	5	0	102	14	5	6	3	1	3	3
6	8	14	3	2	6	120	5	4	0	1	0	2
7	9	5	11	2	9	3	102	17	1	1	2	3
8	5	3	5	15	9	1	25	97	2	0	1	2
9	5	2	2	0	4	4	0	2	77	21	26	22
10	7	3	2	1	5	1	0	1	23	81	17	24
11	3	1	4	0	1	1	1	0	31	20	82	21
12	2	5	0	1	1	3	2	4	19	24	18	86

Σχήμα 4.15: Πίνακας σύγχυσης (confusion marix) σε απόλυτες τιμές για το σύνολο THETIS_Depth. Η αρίθμηση των κινήσεων συμφωνεί με την αρίθμηση που παρουσιάστηκε στην ενότητα 3.3.1.

Κινήσεις	1	2	3	4	5	6	7	8	9	10	11	12
1	69,1	4,8	2,4	2,4	9,1	4,8	1,8	0,0	3,0	0,6	0,6	1,2
2	6,7	70,9	2,4	3,0	3,6	3,6	2,4	3,0	0,0	1,2	1,2	1,8
3	2,4	7,9	64,2	11,5	3,6	0,6	2,4	3,0	1,8	1,2	0,0	1,2
4	0,6	4,2	12,7	65,5	4,8	1,2	3,6	6,7	0,0	0,0	0,0	0,6
5	10,3	3,0	3,0	0,0	61,8	8,5	3,0	3,6	1,8	0,6	1,8	1,8
6	4,8	8,5	1,8	1,2	3,6	72,7	3,0	2,4	0,0	0,6	0,0	1,2
7	5,5	3,0	6,7	1,2	5,5	1,8	61,8	10,3	0,6	0,6	1,2	1,8
8	3,0	1,8	3,0	9,1	5,5	0,6	15,2	58,8	1,2	0,0	0,6	1,2
9	3,0	1,2	1,2	0,0	2,4	2,4	0,0	1,2	46,7	12,7	15,8	13,3
10	4,2	1,8	1,2	0,6	3,0	0,6	0,0	0,6	13,9	49,1	10,3	14,5
11	1,8	0,6	2,4	0,0	0,6	0,6	0,6	0,0	18,8	12,1	49,7	12,7
12	1,2	3,0	0,0	0,6	0,6	1,8	1,2	2,4	11,5	14,5	10,9	52,1

Σχήμα 4.16: Πίνακας σύγχυσης (confusion marix) σε ποσοστά % για το σύνολο THETIS_Depth.

Στα σχήματα 4.17, 4.18 και 4.19 παρουσιάζονται πιο αναλυτικά τα αποτελέσματα ταξινόμησης για κάθε κλάση κινήσεων, από την εφαρμογή της μεθόδου STIP στο σύνολο δεδομένων THETIS_Skelet3D.

Είδος Κίνησης	Precision (%)	Accuracy (%)
Backhand with two hands	65,66%	60,75%
Backhand	60,64%	58,76%
Backhand slice	61,22%	60,00%
Backhand volley	62,14%	50,52%
Forehand flat	52,07%	57,27%

Forehand open stands	61,48%	82,18%
Forehand slice	51,04%	62,14%
Forehand volley	58,51%	59,14%
Service flat	46,51%	41,67%
Service kick	37,29%	40,37%
Service slice	44,32%	39,00%
Smash	50,59%	41,35%

Σχήμα 4.17 : Ποσοστά precision και accuracy κάθε κλάσης για το σύνολο THETIS_ Skelet3D.

Κινήσεις	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	65	5	3	9	4	8	4	1	2	3	2	1
2	9	57	2	3	1	11	1	2	2	2	4	3
3	4	2	60	13	6	3	7	1	0	1	0	3
4	8	2	14	64	2	1	6	5	0	0	0	1
5	3	5	2	5	63	5	11	5	0	5	4	2
6	0	5	0	1	1	83	0	1	0	3	3	4
7	3	5	4	2	15	0	49	16	0	1	1	1
8	4	2	6	4	10	1	9	55	1	1	0	0
9	1	2	1	0	5	5	1	2	40	21	9	9
10	1	4	1	0	2	4	2	0	23	44	16	12
11	1	1	2	0	6	8	5	0	8	24	39	6
12	0	4	3	2	6	6	1	6	10	13	10	43

Σχήμα 4.18 : Πίνακας σύγχυσης (confusion marix) σε απόλυτες τιμές για το σύνολο THETIS_Skelet3D. Η αρίθμηση των κινήσεων συμφωνεί με την αρίθμηση που παρουσιάστηκε στην ενότητα 3.3.1.

Κινήσεις	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	60,7	4,7	2,8	8,4	3,7	7,5	3,7	0,9	1,9	2,8	1,9	0,9
2	9,3	58,8	2,1	3,1	1,0	11,3	1,0	2,1	2,1	2,1	4,1	3,1
3	4,0	2,0	60,0	13,0	6,0	3,0	7,0	1,0	0,0	1,0	0,0	3,0
4	7,8	1,9	13,6	62,1	1,9	1,0	5,8	4,9	0,0	0,0	0,0	1,0
5	2,7	4,5	1,8	4,5	57,3	4,5	10,0	4,5	0,0	4,5	3,6	1,8
6	0,0	5,0	0,0	1,0	1,0	82,2	0,0	1,0	0,0	3,0	3,0	4,0
7	3,1	5,2	4,1	2,1	15,5	0,0	50,5	16,5	0,0	1,0	1,0	1,0
8	4,3	2,2	6,5	4,3	10,8	1,1	9,7	59,1	1,1	1,1	0,0	0,0
9	1,0	2,1	1,0	0,0	5,2	5,2	1,0	2,1	41,7	21,9	9,4	9,4
10	0,9	3,7	0,9	0,0	1,8	3,7	1,8	0,0	21,1	40,4	14,7	11,0
11	1,0	1,0	2,0	0,0	6,0	8,0	5,0	0,0	8,0	24,0	39,0	6,0
12	0,0	3,8	2,9	1,9	5,8	5,8	1,0	5,8	9,6	12,5	9,6	41,3

Σχήμα 4.19 : Πίνακας σύγχυσης (confusion marix) σε ποσοστά % για το σύνολο THETIS_Skelet3D.

Στα σχήματα 4.20, 4.21 και 4.22 παρουσιάζονται πιο αναλυτικά τα αποτελέσματα ταξινόμησης για κάθε κλάση κινήσεων, από την εφαρμογή της μεθόδου STIP στο σύνολο δεδομένων KTH.

Είδος Κίνησης	Precision (%)	Accuracy (%)
Boxing	95,19%	99,00%
Handclapping	97,98%	97,98%
Handwaving	100,00%	96,00%
Jogging	81,90%	86,00%
Running	88,89%	80,00%
Walking	94,29%	99,00%

Σχήμα 4.20: Ποσοστά precision και accuracy κάθε κλάσης για το σύνολο KTH.

Κινήσεις	1	2	3	4	5	6
1	99	1	0	0	0	0
2	2	97	0	0	0	0
3	3	1	96	0	0	0
4	0	0	0	86	10	4
5	0	0	0	18	80	2
6	0	0	0	1	0	99

Σχήμα 4.21: Πίνακας σύγχυσης (confusion marix) σε απόλυτες τιμές για το σύνολο KTH. Η αρίθμηση των κινήσεων συμφωνεί με τον πίνακα 4.20.

Κινήσεις	1	2	3	4	5	6
1	99,0	1,0	0,0	0,0	0,0	0,0
2	2,0	98,0	0,0	0,0	0,0	0,0
3	3,0	1,0	96,0	0,0	0,0	0,0
4	0,0	0,0	0,0	86,0	10,0	4,0
5	0,0	0,0	0,0	18,0	80,0	2,0
6	0,0	0,0	0,0	1,0	0,0	99,0

Σχήμα 4.22: Πίνακας σύγχυσης (confusion marix) σε ποσοστά % για το σύνολο KTH.

Όπως παρατηρούμε, τα ποσοστά ακρίβειας ταξινόμησης για το σύνολο THETIS_Depth είναι σταθερά υψηλότερα από τα αντίστοιχα ποσοστά του συνόλου THETIS_Skelet3D. Ωστόσο, σε σύγκριση με τα αποτελέσματα του συνόλου κινήσεων KTH είναι χαμηλότερα. Αξίζει να σημειωθεί ότι η κλάση που εμφανίζει τα υψηλότερα ποσοστά ακρίβειας είναι η forehand open stands (73,17% στο THETIS_Depth και 61,48% THETIS_Skelet3D). Αντιθέτως, τόσο για το σύνολο THETIS_Depth, όσο και για το σύνολο THETIS_Skelet3D, τα χαμηλότερα ποσοστά ακρίβειας προκύπτουν στις κινήσεις, service flat, service kick, service slice και smash, όπως φαίνεται στους πίνακες 4.14 και 4.17. Παραδείγματος χάριν, για την κίνηση service flat τα ποσοστά ακρίβειας είναι 46,95% και 46,51% αντίστοιχα, ενώ

το χαμηλό ποσοστό ακρίβειας 37,29% εμφανίζει η κίνηση service kick του συνόλου THETIS_Skelet3D.

Τα αποτελέσματα που αφορούν τις κινήσεις service flat, service kick, service slice και smash είναι αναμενόμενα καθώς η ομοιότητα των κινήσεων αυτών είναι υψηλή. Μάλιστα, οι τρεις πρώτες αποτελούν παραλλαγές της ίδιας κίνησης που είναι το service. Επιπροσθέτως, η εκτέλεσή τους από αρχάριους στην αντισφαίριση καθιστά την αναγνώρισή τους ακόμη και από τον ίδιο τον άνθρωπο ένα δύσκολο πρόβλημα.

4.6.3 Αποτελέσματα Δεύτερης Μεθόδου

Στην ενότητα αυτή παρουσιάζονται τα αποτελέσματα που προέκυψαν από την εφαρμογή της μεθόδου Dense Trajectories στα δεδομένα των συνόλων THETIS_Depth , THETIS_Skelet3D και Depth. Για κάθε σύνολο πραγματοποιήθηκαν τρια πειράματα , ένα με βάση τον περιγραφέα Trajectory, ένα με βάση τον περιγραφέα MBH και ένα με βάση τον συνδυασμό των τεσσάρων περιγραφέων Trajectory, HOG, HOF και MBH. Στο σχήμα 4.23 καταγράφεται ο μέσος όρος ακρίβειας / ορθότητας (accuracy) των αποτελεσμάτων ταξινόμησης και για τα τρία σύνολα δεδομένων, για κάθε διαφορετικό περιγραφέα.

Σύνολο Δεδομένων	**Trajectory**	**MBH**	**TRAJECTORY,HOG, HOF,MBH**
THETIS_Depth	51,59 %	54,32 %	**57,50** %
THETIS_Skelet3D	46,84 %	50,78 %	**53,08 %**
KTH	86,98 %	**92,32 %**	90,65 %

Σχήμα 4.23 : Μέσος όρος ακρίβειας ταξινόμησης (average accuracy) με τη μέθοδο Dense Trajectories, για διαφορετικούς συνδυασμούς περιγραφέων.

Στα σχήματα 4.24, 4.25, 4.26, 4.27 παρουσιάζονται πιο αναλυτικά τα αποτελέσματα ταξινόμησης για κάθε κλάση κινήσεων για την εφαρμογή των διαφόρων περιγραφέων της μεθόδου Dense Trajectories στο σύνολο δεδομένων THETIS_Depth.

	Precision			**Accuracy**		
Είδος Κίνησης	**Trajectory**	**MBH**	**Trajectory, HOG, HOF, MBH**	**Trajectory**	**MBH**	**Trajectory, HOG, HOF, MBH**
Backhand with two hands	54,79%	57,07%	59,04%	62,42%	66,06%	67,27%
Backhand	56,83%	59,24%	61,29%	63,03%	66,06%	69,09%
Backhand slice	50,93%	55,78%	60,38%	49,70%	49,70%	58,18%
Backhand volley	58,29%	60,23%	65,03%	61,82%	64,24%	64,24%
Forehand flat	50,30%	53,05%	56,29%	50,61%	53,05%	57,32%

Forehand open stands	**67,90%**	69,41%	**71,43%**	66,67%	71,52%	72,73%
Forehand slice	60,42%	62,82%	64,56%	52,73%	59,39%	61,82%
Forehand volley	53,94%	58,71%	62,58%	53,94%	55,15%	58,79%
Service flat	**35,50%**	38,69%	**43,35%**	36,36%	39,39%	45,45%
Service kick	41,98%	45,16%	48,03%	41,21%	42,42%	44,24%
Service slice	40,44%	41,35%	47,06%	33,33%	33,33%	38,79%
Smash	46,15%	47,22%	49,43%	47,27%	51,52%	52,12%

Σχήμα 4.24: Ποσοστά precision και accuracy κάθε κλάσης για το σύνολο THETIS_Depth, για τους διάφορους περιγραφείς.

Κινήσεις	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	103	8	6	6	18	9	3	0	5	2	3	2
2	15	104	6	6	6	9	5	5	0	2	2	5
3	7	15	82	32	7	1	5	6	3	3	1	3
4	2	4	25	102	6	0	6	15	2	1	1	1
5	18	8	7	4	83	19	3	10	4	1	4	3
6	8	17	3	2	10	110	6	4	0	2	1	2
7	10	7	14	2	9	3	87	25	1	1	2	4
8	5	4	7	16	11	4	23	89	1	1	1	3
9	5	2	4	1	5	1	0	3	60	28	31	25
10	8	6	2	1	5	0	1	2	31	68	18	23
11	5	3	5	1	4	1	1	0	41	29	55	20
12	2	5	0	2	1	5	4	6	21	24	17	78

Σχήμα 4.25: Πίνακας σύγχυσης (confusion marix) σε απόλυτες τιμές για το σύνολο THETIS_Depth, με περιγραφέα Trajectory.

Κινήσεις	1	2	3	4	5	6	7	8	9	10	11	12
1	109	8	4	5	18	9	3	0	5	1	1	2
2	14	109	4	6	6	7	5	5	0	2	2	5
3	7	15	82	35	7	1	4	6	3	3	0	2
4	1	7	24	106	4	2	6	14	0	0	0	1
5	18	6	6	3	87	19	5	9	3	1	4	3
6	8	15	3	2	6	118	5	4	0	2	0	2
7	10	5	11	2	9	3	98	19	1	1	2	4
8	5	4	5	15	11	2	27	91	2	0	1	2
9	5	2	2	0	5	4	0	2	65	25	30	25
10	8	6	2	1	6	1	0	1	26	70	19	25
11	4	2	4	0	4	1	1	0	44	26	55	24
12	2	5	0	1	1	3	2	4	19	24	19	85

Σχήμα 4.26: Πίνακας σύγχυσης (confusion marix) σε απόλυτες τιμές για το σύνολο THETIS_Depth, με περιγραφέα MBH.

Κινήσεις	1	2	3	4	5	6	7	8	9	10	11	12
1	111	8	4	4	18	8	3	0	5	1	1	2
2	12	114	4	5	6	7	5	5	0	2	2	3
3	6	15	96	24	7	1	4	5	3	2	0	2
4	1	7	23	106	5	2	6	14	0	0	0	1
5	17	6	5	3	94	17	5	6	3	1	4	3
6	8	14	3	2	6	120	5	4	0	1	0	2
7	9	5	11	2	9	3	102	17	1	1	2	3
8	5	3	5	15	9	1	25	97	2	0	1	2
9	5	2	2	0	4	4	0	2	75	23	26	22
10	8	6	2	1	6	1	0	1	24	73	18	25
11	4	1	4	0	2	1	1	0	41	24	64	23
12	2	5	0	1	1	3	2	4	19	24	18	86

Σχήμα 4.27: Πίνακας σύγχυσης (confusion marix) σε απόλυτες τιμές για το σύνολο THETIS_Depth, με περιγραφείς Trajectory, HOG, HOF και MBH.

Στα σχήματα 4.26, 4.27, 4.28 και 4.29 παρουσιάζονται πιο αναλυτικά τα αποτελέσματα ταξινόμησης για κάθε κλάση κινήσεων για την εφαρμογή των διαφόρων περιγραφέων της μεθόδου Dense Trajectories στο σύνολο δεδομένων THETIS_Skelet3D.

	Precision			**Accuracy**		
Είδος Κίνησης	**Trajectory**	**MBH**	**Trajectory, HOG, HOF, MBH**	**Trajectory**	**MBH**	**Trajectory, HOG, HOF, MBH**
Backhand with two hands	49,41%	57,14%	**63,73%**	39,25%	52,34%	60,75%
Backhand	44,58%	52,75%	59,34%	38,14%	49,48%	55,67%
Backhand slice	45,65%	54,17%	59,18%	42,00%	52,00%	58,00%
Backhand volley	50,94%	53,85%	59,80%	52,43%	54,37%	59,22%
Forehand flat	49,59%	50,00%	51,67%	53,55%	54,55%	56,36%
Forehand open stands	**60,14%**	**62,41%**	61,03%	82,18%	82,18%	82,18%
Forehand slice	44,34%	47,00%	48,96%	48,45%	48,39%	48,45%
Forehand volley	54,08%	55,21%	56,84%	56,99%	56,99%	58,06%
Service flat	45,35%	47,56%	45,45%	40,53%	40,63%	41,67%
Service kick	**34,88%**	**37,19%**	**36,13%**	41,28%	41,28%	39,45%
Service slice	39,18%	42,22%	43,68%	38,00%	38,00%	38,00%
Smash	39,47%	47,67%	49,40%	28,85%	39,42%	39,42%

Σχήμα 4.28: Ποσοστά precision και accuracy κάθε κλάσης για το σύνολο THETIS_ Skelet3D, για τους διάφορους περιγραφείς.

Κινήσεις	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	42	9	8	13	6	8	6	4	2	5	2	2
2	13	37	2	4	2	14	3	2	3	4	9	4
3	5	4	42	19	7	2	10	4	1	3	0	3
4	9	3	18	54	2	1	7	4	0	1	1	3

5	3	5	3	5	60	4	14	5	0	5	4	2
6	0	5	0	1	0	83	0	1	0	4	3	4
7	4	5	6	2	14	0	47	16	0	1	1	1
8	5	2	6	4	11	1	9	53	1	1	0	0
9	2	2	1	0	5	5	1	2	39	21	9	9
10	1	5	1	0	2	5	2	0	20	45	16	12
11	1	2	2	1	6	8	5	0	7	24	38	6
12	0	4	3	3	6	7	2	7	13	15	14	30

Σχήμα 4.29: Πίνακας σύγχυσης (confusion marix) σε απόλυτες τιμές για το σύνολο THETIS_ Skelet3D, με περιγραφέα Trajectory.

Κινήσεις	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	56	6	4	13	6	8	4	2	2	3	2	1
2	12	48	2	4	2	10	3	2	2	2	6	4
3	5	4	52	15	6	2	7	4	0	2	0	3
4	9	3	16	56	2	1	7	4	0	1	1	3
5	3	5	3	5	60	4	14	5	0	5	4	2
6	0	5	0	1	0	83	0	1	0	4	3	4
7	4	5	6	2	14	0	47	16	0	1	1	1
8	5	2	6	4	11	1	9	53	1	1	0	0
9	2	2	1	0	5	5	1	2	39	21	9	9
10	1	5	1	0	2	5	2	0	20	45	16	12
11	1	2	2	1	6	8	5	0	7	24	38	6
12	0	4	3	3	6	6	1	7	11	12	10	41

Σχήμα 4.30: Πίνακας σύγχυσης (confusion marix) σε απόλυτες τιμές για το σύνολο THETIS_ Skelet3D, με περιγραφέα MBH.

Κινήσεις	1	2	3	4	5	6	7	8	9	10	11	12
1	65	5	3	9	4	8	4	1	2	3	2	1
2	10	54	2	4	1	12	1	2	2	2	4	3
3	6	2	58	13	6	3	7	1	0	1	0	3
4	8	2	16	61	2	1	6	6	0	0	0	1
5	3	5	2	5	62	5	12	5	0	5	4	2
6	0	5	0	1	1	83	0	1	0	3	3	4
7	3	5	4	3	15	0	47	17	0	1	1	1
8	4	2	6	4	10	1	10	54	1	1	0	0
9	1	2	1	0	5	5	1	2	40	21	9	9
10	1	4	1	0	2	4	2	0	24	43	16	12
11	1	1	2	0	6	8	5	0	9	24	38	6
12	0	4	3	2	6	6	1	6	10	15	10	41

Σχήμα 4.31: Πίνακας σύγχυσης (confusion marix) σε απόλυτες τιμές για το σύνολο THETIS_ Skelet3D, με περιγραφείς Trajectory, HOG, HOF και MBH.

Στα σχήματα 4.30, 4.31, 4.32 και 4.33 παρουσιάζονται πιο αναλυτικά τα αποτελέσματα ταξινόμησης για κάθε κλάση κινήσεων για την εφαρμογή των διαφόρων περιγραφέων της μεθόδου Dense Trajectories στο σύνολο δεδομένων KTH.

	Precision			**Accuracy**		
Είδος Κίνησης	**Traje-ctory**	**MBH**	**Trajectory, HOG, HOF, MBH**	**Traje-ctory**	**MBH**	**Trajectory, HOG, HOF, MBH**
Boxing	84,85%	95,05%	91,26%	84,00%	96,00 %	94,00%
Handclapping	79,21%	92,00%	91,00%	80,81%	92,93 %	91,92%
Handwaving	**96,94%**	**97,00%**	**98,94%**	95,00%	97,00 %	93,00%
Jogging	80,37%	89,01%	84,00%	86,00%	81,00 %	84,00%
Running	84,44%	84,62%	82,83%	76,00%	88,00 %	82,00%
Walking	96,15%	96,12%	96,12%	100,00%	99,00 %	99,00%

Σχήμα 4.32: Ποσοστά precision και accuracy κάθε κλάσης για το σύνολο KTH, για τους διάφορους περιγραφείς.

Κινήσεις	1	2	3	4	5	6
1	84	16	0	0	0	0
2	15	80	3	0	0	1
3	0	5	95	0	0	0
4	0	0	0	86	14	0
5	0	0	0	21	76	3
6	0	0	0	0	0	100

Σχήμα 4.33: Πίνακας σύγχυσης (confusion marix) σε απόλυτες τιμές για το σύνολο KTH, με περιγραφέα Trajectory.

Κινήσεις	1	2	3	4	5	6
1	96	4	0	0	0	0
2	4	92	3	0	0	0
3	1	2	97	0	0	0
4	0	1	0	81	16	2
5	0	0	0	10	88	2
6	0	1	0	0	0	99

Σχήμα 4.34: Πίνακας σύγχυσης (confusion marix) σε απόλυτες τιμές για το σύνολο KTH, με περιγραφέα MBH.

Κινήσεις	1	2	3	4	5	6
1	94	4	0	0	1	1
2	7	91	1	0	0	0
3	2	4	93	0	1	0
4	0	0	0	84	14	2
5	0	1	0	16	82	1
6	0	0	0	0	1	99

Σχήμα 4.35: Πίνακας σύγχυσης (confusion marix) σε απόλυτες τιμές για το σύνολο KTH, με περιγραφείς Trajectory, HOG, HOF και MBH.

Όπως παρατηρούμε, τα ποσοστά ακρίβειας ταξινόμησης για το σύνολο THETIS_Depth είναι σταθερά υψηλότερα από τα αντίστοιχα ποσοστά του συνόλου THETIS_Skelet3D. Ωστόσο, σε σύγκριση με τα αποτελέσματα του συνόλου κινήσεων KTH είναι χαμηλότερα. Επιπροσθέτως, είναι φανερό ότι ο συνδυασμός περιγραφέων που βελτιστοποιεί τα αποτελέσματα ταξινόμησης για τα σύνολα δεδομένων THETIS_Depth και THETIS_Skelet3D είναι ο συνδυασμός όλων των περιγραφέων που υπολογίζει ο κώδικας

Dense Trajectories, δηλαδή ο συνδυασμός των περιγραφέων Trajectory, HOG, HOF και MBH. Επίσης, παρατηρούμε πως για το σύνολο δεδομένων KTH, τα αποτελέσματα βελτιστοποιούνται με την χρήση του περιγραφέα MBH αποκλειστικά.

Αξίζει να σημειωθεί ότι στα σύνολα THETIS_Depth και THETIS_Skelet3D, τα χαμηλότερα ποσοστά ακρίβειας προκύπτουν στις κινήσεις service flat (35,50%) και service kick (34,88%) αντίστοιχα, όπως φαίνεται στους πίνακες 4.24 και 4.26. Αντιθέτως, τα υψηλότερα ποσοστά παρουσιάζει η κίνηση foreflat open stands (71,43%) για το σύνολο THETIS_Depth, και η κίνηση backhand with 2 hands (63,73%) για το σύνολο THETIS_Skelet3D.

4.6.4 Συγκριτικά Αποτελέσματα

Στην ενότητα αυτή, παρουσιάζονται τα αποτελέσματα της μεθόδου STIP σε αντιπαράθεση με τα αποτελέσματα της μεθόδου Dense Trajectories. Όπως προκύπτει και από το σχήμα 4.32, η εφαρμογή της μεθόδου STIP οδήγησε σε καλύτερα αποτελέσματα τόσο για τα video της βάσης KTH, όσο και για τα video της παρουσιαζόμενης βάσης δεδομένων THETIS.

Σύνολο Δεδομένων	**Dense Trajectories**			**STIP**
Περιγραφείς	Trajectory	MBH	Trajectory, HOG, HOF, MBH	HOG,HOF
THETIS_Depth	51,59 %	54,32 %	57,50 %	**60.23%**
THETIS_Skelet3D	46,84 %	50,78 %	53,08 %	**54.40%**
KTH	86,98 %	92,32 %	90,65 %	**92.99%**

Σχήμα 4.36: Σύγκριση των αποτελεσμάτων των μεθόδων Dense Trajectories και STIP, για όλα τα σύνολα δεδομένων που χρησιμοποιήθηκαν στην πειραματική διαδικασία.

Ειδικότερα, παρατηρούμε πως η καλύτερη επίδοση του συνόλου THETIS_Depth σημειώνεται για τη μέθοδο STIP (60.23%). Τα πειράματα με την χρήση του περιγραφέα MBH της μεθόδου Dense Trajectories στο σύνολο KTH, έδωσαν πολύ καλά αποτελέσματα (92,32 %) σχεδόν τόσο καλά όσο η μέθοδος STIP (92,99 %). Η χρήση του περιγραφέα MBH, άλλωστε, σκοπό έχει την απομόνωση του θορύβου που προκύπτει από την κίνηση της κάμερας. Στην παρουσιαζόμενη βάση, που η κάμερα είναι στατική, φαίνεται πως δεν δίνει αντίστοιχα καλά αποτελέσματα σε σύγκριση με τα αποτελέσματα της μεθόδου STIP. Επιπροσθέτως, και στα τρια σύνολα δεδομένων κίνησης, η εφαρμογή του περιγραφέα τροχιάς Trajectories ως αποκλειστικό περιγραφέα, δίνει τα λιγότερο καλά αποτελέσματα.

Γενικότερα, είναι σαφές πως η εφαρμογή των μεθόδων STIP και Dense Trajectories δεν οδήγησε σε τόσο υψηλή ακρίβεια στην αναγνώριση των κινήσεων της προτεινόμενης βάσης, όσο στην αναγνώριση των κινήσεων της KTH. Στο σημείο αυτό πρέπει να σημειωθεί η διαφορετικότητα της βάσης δεδομένων κίνησης THETIS, σε σχέση με τη βάση KTH. Κατ'αρχάς, οι κινήσεις της βάσης KTH είναι πιο απλές από τις κινήσεις της βάσης THETIS, που περιλαμβάνει κινήσεις αντισφαίρισης, δηλαδή πιο σύνθετες και λιγότερο διαχωρίσιμες μεταξύ τους. Παραδείγματος χάριν, το σύνολο THETIS, περιέχει τρια είδη service που είναι διαχωρίσιμα μόνο από έμπειρους παίκτες, ενώ είναι πολύ δύσκολο να γίνει διάκριση μεταξύ τους από ένα μέσο θεατή.

Επιπροσθέτως, κρίνεται απαραίτητο να σημειωθεί πως στα video των συνόλων THETIS_Depth και THETIS_Skelet3D καταγράφεται η κίνηση άλλων ατόμων στο πίσω μέρος της σκηνής (background). Το περιβάλλον καταγραφής των κινήσεων δεν είναι πλήρως ελεγχόμενο, καθώς στο πλάνο εισχωρούν συχνά άλλα άτομα τα οποία επιδίδονται σε ποικίλες δραστηριότητες άσχετες με την κίνηση που καταγράφεται τη δεδομένη χρονική στιγμή. Μπορεί η χρήση του Kinect για την παραγωγή 3D δεδομένων να παρέχει ένα σημαντικό πλεονέκτημα για την εξαγωγή του background, ο περιβαλλοντικός θόρυβος όμως δεν παύει να αποτελεί μια επιπλέον πρόκληση. Τέλος, υπάρχει συχνά θόρυβος στην απεικόνιση του βάθους, λόγω της σκέδασης του υπέρυθρου φωτός από διάφορες επιφάνειες, όπως καθρέπτης, ξύλινο πάτωμα που αντανακλά το φως, κ.α.

Επιπλέον, το σύνολο δεδομένων THETIS_Skelet3D που καταγράφει την κίνηση του σκελετού του ανθρώπινου σώματος στις τρεις διαστάσεις του χώρου είναι ένα νέο είδος συνόλου δεδομένων κίνησης. Συνδυάζει την πληροφορία του βάθους που κατασκευάζεται με τη βοήθεια της κάμερας υπερύθρων και την ανακατασκευή του σκελετού του ανθρώπινου σώματος, που υποστηρίζεται από το διαπλατφορμικό πλαίσιο εφαρμογών OpenNI Framework. Επομένως, είναι δίκαιο να αναφέρουμε πως δεν μπορούν να συγκριθούν άμεσα τα αποτελέσματα των μεθόδων STIP και Dense Trajectories του συνόλου δεδομένων KTH με εκείνα του συνόλου THETIS_Skelet3D, καθώς το είδος της πληροφορίας που έχει καταγραφεί στα video κάθε περίπτωσης είναι εντελώς διαφορετικό. Σε κάθε περίπτωση, η σύγκριση γίνεται για να αναδείξουμε τη δυναμική της προτεινόμενης βάσης και τις προκλήσεις που αυτή παρουσιάζει όταν αποτελεί αντικείμενο αξιολόγησης αλγορίθμων κατηγοριοποίησης τελευταίας τεχνολογίας.

ΚΕΦΑΛΑΙΟ 5

5.1 Συμπεράσματα

Στο κεφάλαιο αυτό γίνεται μια σύνοψη του αποτελέσματος της εργασίας και επιχειρείται η αξιολόγησή του και η διερεύνηση πιθανών προοπτικών για περαιτέρω μελέτη και έρευνα.

Στο πλαίσιο της παρούσας διπλωματικής εργασίας δημιουργήθηκε ένα νέο μεγάλο σύνολο από video που περιέχουν κινήσεις του αθλήματος της αντισφαίρισης. Η νέα βάση δεδομένων THETIS συνδυάζει έναν αριθμό πλεονεκτήμάτων που μπορούν να αξιοποιηθούν από μελλοντικές εφαρμογές στα πλαίσια της αναγνώρισης της ανθρώπινης δραστηριότητας.

Πρώτα απ΄όλα η βάση THETIS είναι πρωτότυπη ως προς το είδος των κινήσεων καθώς σε κανένα άλλο σύνολο δεδομένων από τα ήδη υπάρχοντα, δεν περιλαμβάνεται όλο το φάσμα των βασικών κινήσεων της αντισφαίρισης. Επιπλέον, περιλαμβάνει video όχι μόνο εικόνας RGB, αλλά και video που αναπαριστούν την κάθε κίνηση στις τρεις διαστάσεις του χώρου. Το πλεονέκτημα αυτό, που πηγάζει από τη χρήση της συσκευής Kinect ως μέσο καταγραφής, παρέχει σε ένα μέρος των δεδομένων της βάσης THETIS, ανεξαρτησία από την γωνία λήψης, η οποία σε άλλα σύνολα δεν υπάρχει.

Επιπροσθέτως τα video σκελετού 3D, σκελετού 2D και περιγράμματος παρέχουν σε γενικές γραμμές το πλεονέκτημα της ανεξαρτησίας από το background (φόντο), εκτός ελαχίστων εξαιρέσεων όπου δεν κατέστη δυνατή η πλήρης απομόνωση της κίνησης από το φόντο.

Ένα ακόμη πλεονέκτημα του συνόλου δεδομένων THETIS είναι αδιαμφισβήτητα τα video που απεικονίζουν τις κινήσεις του σκελετού των συμμετεχόντων σε σύστημα τριών χωρικών διαστάσεων. Με την αναπαράσταση αυτή προτείνουμε την προσέγγιση του προβλήματος της αναγνώρισης της ανθρώπινης δραστηριότητας από video, με έναν τρόπο λιγότερο σύνθετο σε σύγκριση με τις προσεγγίσεις που χρησιμοποιούν εικονοακολουθίες RGB για την αξιολόγηση των διαφόρων μεθόδων αναγνώρισης κινήσεων. Μέσω της αναπαράστασης αυτής δίνεται η δυνατότητα για χρήση της πληροφορίας που αφορά στη μετατόπιση των αρθρώσεων του σώματος, απομονωμένη από την υπόλοιπη πληροφορία ενός video.

Ως προς την πειραματική διαδικασία, δημιουργήθηκε ένα σύστημα αναγνώρισης κινήσεων από video, που βασίζεται σε δυο αλγόριθμους ανάλυσης της κίνησης τελευταίας τεχνολογίας, τον Space-Time Interest Points και τον Dense trajectories. Επίσης, κατασκευάστηκε το πρωτόκολλο για την κβαντοποίηση των χαρακτηριστικών διανυσμάτων που εξάγουν αυτοί οι αλγόριθμοι και για την ταξινόμηση των video με βάση το περιεχόμενο, με χρήση SVM. Τα πειράματα ταξινόμησης εφαρμόστηκαν στα video βάθους και σκελετού 3D του συνόλου μας και στα δεδομένα της βάσης KTH. Τα αποτελέσματα των πειραμάτων στα δεδομένα της βάσης THETIS, είναι ενθαρρυντικά αλλά υπάρχει σίγουρα χώρος για την βελτίωσή τους. Οι λόγοι που δεν προέκυψαν υψηλά ποσοστά ακρίβειας μπορούν να εξηγηθούν.

Κατ'αρχάς, στη δημιουργία της βάσης συμμετείχαν τόσο έμπειροι όσο και αρχάριοι στο άθλημα της αντισφαίρισης, συγκεκριμένα συμμετείχαν 31 αρχάριοι και 24 έμπειροι αντισφαιριστές. Μάλιστα, ο λόγος που η συμμετοχή των αρχαρίων κρίθηκε αναγκαία θα εξηγηθεί στη συνέχεια, όπου γίνεται λόγος για τις προοπτικές επέκτασης της εργασίας.

Παρ'όλα αυτά, το γεγονός ότι στη διαδικασία εκπαίδευσης του SVM χρησιμοποιήθηκαν και τα δείγματα των αρχαρίων στην αντισφαίριση, απέτρεψε την κατασκευή ένος τέλειου λεξιλογίου για το κάθε είδος της κίνησης. Είμαστε βέβαιοι, πως αν στη διαδικασία εκπαίδευσης συμμετείχαν μόνο δείγματα έμπειρων, τα αποτελέσματα ακρίβειας θα ήταν υψηλότερα.

Επιπροσθέτως, η βάση δεδομένων THETIS αποτελείται από αρκετά εξειδικευμένες κινήσεις της αντισφαίρισης. Ειδικότερα, τρεις από τις δώδεκα κινήσεις αποτελούν παραλλαγές της ίδιας κίνησης που είναι το service. Στην περίπτωσή τους, οι διαφορές τους είναι πολύ δυσδιάκριτες, ώστε μόνο οι πολύ έμπειροι στο άθλημα της αντισφαίρισης να είναι σε θέση να τις διαχωρίσουν. Γι'αυτό το λόγο, παρατηρούμε πως τα αποτελέσματα ακρίβειας που αφορούν στις συγκεκριμένες κλάσεις, είναι έως και 20% χαμηλότερα από τις υπόλοιπες κινήσεις.

Μάλιστα, είμαστε πεπεισμένοι πως τα ποσοστά ακρίβειας μπορούν να αυξηθούν αρκετά, αν διερευνηθούν ακόμα περισσότερο κάποιες παράμετροι που επηρεάζουν άμεσα την επίδοση του συστήματος, όπως οι παράμετροι που αφορούν στον πυρήνα kernel του SVM της ταξινόμησης, που διερευνήθηκαν εποπτικά, ενώ μπορεί να γίνει εξαντλητική διερεύνησή τους με τα κατάλληλα υπολογιστικά μέσα.

Όσον αφορά στις μελλοντικές προεκτάσεις της εργασίας, μπορούμε να πούμε πως υπάρχουν προοπτικές για περαιτέρω μελέτη και έρευνα. Ξεκινώντας από εκείνα που μπορούν να γίνουν εύκολα με τα δεδομένα που έχουν ήδη συγκεντρωθεί, προτείνεται η χρήση των video που απεικονίζουν την κίνηση της σιλουέτας των ατόμων σε ανάλογο σύστημα ανάλυσης και κατηγοριοποίησης. Ήδη στην ενότητα 2.2, αναφέρθηκαν μέθοδοι αναγνώρισης κινήσεων που ανέδειξαν την αξία της εξαγωγής της σιλουέτας από ένα video.

Επιπροσθέτως, μελλοντικές ερευνητικές προσπάθειες μπορούν να επιχειρήσουν την επέκταση του συστήματος σε ταξινομητή expert-non expert , με βάση τα video του συνόλου μας, που όπως αναφέρθηκε και πιο πάνω περιέχουν δείγματα και έμπειρων και αρχαρίων. Πιο συγκεκριμένα, θα μπορούσε να κατασκευαστεί ένα σύστημα που να είναι σε θέση να ξεχωρίζει αν το άτομο που εκτελεί μια κίνηση αντισφαίρισης σε ένα video είναι αρχάριο ή έμπειρο στο άθλημα. Μάλιστα, μια πιθανή δυνατότητα ενός τέτοιου συστήματος, θα ήταν να βαθμολογεί την επίδοση κάθε εκτέλεσης σε μια κλίμακα από το 1 έως το 10.

Αν επεκτείνει κανείς αυτή την ιδέα, θα μπορούσε να δώσει εκπαιδευτικό χαρακτήρα στο σύστημα και να αποτελέσει εργαλείο εκμάθησης για το άθλημα της αντισφαίρισης. Δηλαδή, με την ανάπτυξη ενός διαδραστικού περιβάλλοντος, θα μπορούσε το σύστημα να χρησιμοποιείται κατά τη διαδικασία προπόνησης και στο τέλος, να παρουσιάζει στατιστικά στοιχεία για την επίδοση κατά την εκτέλεση κάθε κίνησης.

Τέλος, το σύνολο δεδομένων THETIS, θα μπορούσε να αποτελέσει χρήσιμο εργαλείο για την ανάπτυξη εφαρμογών αυτόματης ανάλυσης αγώνων αντισφαίρισης (sports play analysis), που αποτελεί ένα από τα πιο δημοφιλή αθλήματα. Ένα τέτοιο σύστημα που θα μπορούσε να αναγνωρίζει σε πραγματικό χρόνο την κίνηση που εκτελεί ο εκάστοτε παίκτης, θα ήταν σε θέση να επιτύχει την συλλογή στατιστικών στοιχείων, την ανάλυση της τακτικής του παιχνιδιού και την αυτόματη περιγραφή ενός αγώνα αντισφαίρισης.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Aggarwal, J. K. and Ryoo, M. S. 2011. Human activity analysis: A review. ACM Comput. Surv. 43, 3, Article 16 (April 2011), 43 pages.

[2] Bobick, A. and Davis, J. 2001. The recognition of human movement using temporal templates. *IEEE Trans. Patt. Anal. Mach. Intel. 23,* 3, 257–267.

[3] Shechtman, E. and Irani, M. 2005. Space-time behavior based correlation. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* Vol. 1, IEEE, Los Alamitos, CA,405–412.

[4] Ke, Y., Sukthankar, R., and Hebert, M. 2007. Spatio-temporal shape and flow correlation for action recognition. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, Los Alamitos, CA.

[5] Schuldt, C., Laptev, I., and Caputo, B. 2004. Recognizing human actions: A local SVM approach. *In Proceedings of the International Conference on Pattern Recognition (ICPR).* Vol. 3, 32–36.

[6] Rodriguez, M. D., Ahmed, J., and Shah, M. 2008. Action MACH: A spatio-temporal maximum average correlation height filter for action recognition. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, Los Alamitos, CA.

[7] Blank, M., Gorelick, L., Shechtman, E., Irani, M., and Basri, R. 2005. Actions as space-time shapes. *In Proceedings of the IEEE International Conference on Computer Vision (ICCV).* IEEE, Los Alamitos, CA, 1395–1402.

[8] Sheikh, Y., Sheikh, M., and Shah, M. 2005. Exploring the space of a human action. *In Proceedings of the IEEE International Conference on Computer Vision (ICCV).* Vol. 1, IEEE, Los Alamitos, CA, 144–149.

[9] Yilmaz, A. and Shah, M. 2005. *Recognizing human actions in videos acquired by uncalibrated moving cameras (ICCV).* IEEE, Los Alamitos, CA.

[10] Campbell, L. W. and Bobick, A. F. 1995. Recognition of human body motion using phase space constraints. *In Proceedings of the IEEE International Conference on Computer Vision (ICCV).* IEEE Los Alamitos, CA, 624–630.

[11] Rao, C. and Shah, M. 2001. View-invariance in action recognition. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 2, IEEE, Los Alamitos, CA, 316–322.

[12] Chomat, O. and Crowley, J. 1999. Probabilistic recognition of activity using local appearance. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* Vol. 2, IEEE Los Alamitos, CA.

[13] Zelnik-Manor, L. and Irani, M. 2001. Event-based analysis of video. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, Los Alamitos, CA.

[14] Blank, M., Gorelick, L., Shechtman, E., Irani, M., and Basri, R. 2005. Actions as space-time shapes. *In Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, Los Alamitos, CA, 1395–1402
[15] Laptev, I. and Lindeberg, T. 2003. Space-time interest points. *In Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, Los Alamitos, CA, 432.

[16] Lowe, D. G. 1999. Object recognition from local scale-invariant features. *In Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, Los Alamitos, CA, 1150–1157.

[17] Dollar, P., Rabaud, V., Cottrell, G., and Belongie, S. 2005. Behavior recognition via sparse spatiotemporal features. *In Proceedings of the 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance* (VS-PETS). IEEE, Los Alamitos, CA. 65–72.

[18] Savarese, S., Delpozo, A., Niebles, J., and Fei-Fei, L. 2008. Spatial-temporal correlatons for unsupervised action classification. *In Proceedings of the IEEE Workshop on Motion and Video Computing (WMVC)*. IEEE, Los Alamitos, CA.

[19] Laptev, I.,Marszalek, M., Schmid, C. and Rozenfeld, B. 2008. Learning realistic human actions from movies. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Los Alamitos, CA.

[20] Ryoo, M. S. and Aggarwal, J. K. 2009. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. *In Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, Los Alamitos, CA.

[21] Darrell, T. and Pentland, A. 1993. Space-time gestures. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Los Alamitos, CA. 335–340.

[22] Gavrila, D. and Davis, L. 1995. Towards 3-D model-based tracking and recognition of human movement. *In Proceedings of the International Workshop on Face and Gesture Recognition*. 272–277.

[23] Yacoob, Y. and Black, M. 1998. Parameterized modeling and recognition of activities. *In Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, Los Alamitos, CA, 120–127.

[24] Efros, A.,Berg, A., Mori, G., and Malik, J. 2003. Recognizing action at a distance. *In Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Vol. 2, IEEE, Los Alamitos, CA, 726–733.

[25] Lublinerman, R.,Ozay, N., Zarpalas, D., and Camps, O. 2006. Activity recognition from silhouettes using linear systems and model (in)validation techniques. *In Proceedings of the International Conference on Pattern Recognition (ICPR)*. 347–350.

[26] Veeraraghavan, A., Chellappa, R., and Roy-Chowdhury, A. 2006. The function space of an activity. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 1, IEEE, Los Alamitos, CA, 959–968.

[27] Y. Rui, T. S. Huang, and S. F. Chang. 1999. Image retrieval: current techniques, promising directions and open issues. Journal of Visual Communication and Image Representation 10, 4, 39–62.

[28] C. Stauffer and W. E. L. Grimson. 2002. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*, 8, 747–757.
[29] W. Hu, D. Xie, T. Tan, and S. Maybank. 2004. Learning activity patterns using fuzzy self-organizing neural network. *IEEE Transactions on Systems, Man and Cybernetics 34*, 3, 1618–1626.

[30] Yamato, J., Ohya, J., and Ishii, K. 1992. Recognizing human action in time-sequential images using hidden Markov models. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Los Alamitos, CA, 379–385.

[31] Starner, T. and Pentland, A. 1995. Real-time American Sign Language recognition from video using hidden Markov models. *In Proceedings of the International Symposium on Computer Vision.* 265.

[32] Bobick, A. F. andwilson, A. D. 1997. A state-based approach to the representation and recognition of gesture. *IEEE Trans. Patt. Anal. Mach. Intell. 19*, 12, 1325–1337.

[33] Oliver, N. M.,Rosario, B., and Pentland, A. P. 2000. A Bayesian computer vision system for modeling human interactions. *IEEE Trans. Patt. Anal. Mach. Intell. 22*, 8, 831–843.

[34] Park, S. and Aggarwal, J. K. 2004. A hierarchical Bayesian network for event recognition of human actions and interactions. *Multimedia Syst. 10*, 2, 164–179.

[35] Natarajan, P. and Nevatia, R. 2007. Coupled hidden semi-Markov models for activity recognition. *In Proceedings of the IEEE Workshop on Motion and Video Computing (WMVC)*. IEEE, Los Alamitos, CA.

[36] Oliver, N., Horvitz, E., and Garg, A. 2002. Layered representations for human activity recognition. *In Proceedings of the IEEE International Conference on Multimodal Interfaces (ICMI)*. IEEE, Los Alamitos, CA, 3–8.

[37] Nguyen, N. T., Phung, D. Q., Venkatesh, S., and Bui, H. H. 2005. Learning and detecting activities from movement trajectories using the hierarchical hidden Markov models. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 2, IEEE, Los Alamitos, CA, 955–960.

[38] Zhang, D., Gatica-Perez, D., Bengio, S., and Mccowan, I. 2006. Modeling individual and group actions in meetings with layered hmms. *IEEE Trans. Multimedia 8*, 3, 509–520.

[39] Gong, S. and Xiang, T. 2003. Recognition of group activities using dynamic probabilistic networks. *In Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, Los Alamitos, CA, 742.

[40] Dai, P.,Di, H.,Dong, L., Tao, L., and Xu, G. 2008. Group interaction analysis in dynamic context. *IEEE Trans. Syst. Man Cybern. Part B 38*, 1, 275–282.

[41] Damen, D. and Hogg, D. 2009. Recognizing linked events: Searching the space of feasible explanations. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Los Alamitos, CA.

[42] Shi, Y., Huang, Y., Minnen, D., Bobick, A. F., and Essa, I. A. 2004. Propagation networks for recognition of partially ordered sequential action. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 2, IEEE, Los Alamitos, CA, 862–869.

[43] Ivanov, Y. A. and Bobick, A. F. 2000. Recognition of visual activities and interactions by stochastic parsing. *IEEE Trans. Patt. Anal. Mach. Intell. 22*, 8, 852–872.

[44] Moore, D. J. and Essa, I. A. 2002. Recognizing multitasked activities from video using stochastic context-free grammar. *In Proceedings of the AAAI Conference on Innovative Applications of Artificial Intelligence*. 770–776.

[45] Minnen, D., Essa, I. A., and Starner, T. 2003. Expectation grammars: Leveraging high-level expectations for activity recognition. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 2, IEEE, Los Alamitos, CA, 626–632.

[46] Joo, S.-W. and Chellappa, R. 2006. Attribute grammar-based event recognition and anomaly detection. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*. IEEE, Los Alamitos, CA, 107.

[47] Allen, J. F. 1983. Allen, J. F. 1983. Maintaining knowledge about temporal intervals. *Comm. ACM 26*, 11, 832–843.

[48] Pinhanez, C. S. and Bobick, A. F. 1998. Human action detection using PNF propagation of temporal constraints. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Los Alamitos, CA, 898.

[49] Intille, S. S. and Bobick, A. F. 1999. A framework for recognizing multi-agent action from visual evidence. *In Proceedings of the AAAI Conference on Innovative Applications of Artificial Intelligence. AAAI/IAAI*. 518–525.

[50] Nevatia, R., Zhao, T., and Hongeng, S. 2003. Hierarchical language-based representation of events in video Streams. In *Proceedings of the IEEE Workshop on Event Mining*. IEEE, Los Alamitos, CA.

[51] Ryoo, M. S. and Aggarwal, J. K. 2006a. Recognition of composite human activities through context-free grammar based representation. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Los Alamitos, CA, 1709–1718.

[52] Tran, S. D. and Davis, L. S. 2008. Event modeling and recognition using Markov logic networks. *In Proceedings of European Conference on Computer Vision (ECCV)*. 610–623.

[53] Schuldt, C., Laptev, I., and Caputo, B. 2004. Recognizing human actions: A local SVM approach. *In Proceedings of the International Conference on Pattern Recognition (ICPR)*. Vol. 3, 32–36.

[54] Blank, M., Gorelick, L., Shechtman, E., Irani, M., and Basri, R. 2005. Actions as space-time shapes. *In Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, Los Alamitos, CA, 1395–1402.

[55] Ke, Y., Sukthankar, R., and Hebert, M. 2007. Spatio-temporal shape and flow correlation for action recognition. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Los Alamitos, CA.

[56] Laptev, I.,Marszalek, M., Schmid, C., and Rozenfeld, B. 2008. Learning realistic human actions from movies. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Los Alamitos, CA.

[57] Rodriguez, M. D., Ahmed, J., and Shah, M. 2008. Action MACH: A spatio-temporal maximum average correlation height filter for action recognition. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Los Alamitos, CA.

[58] M. Marszałek, I. Laptev, and C. Schmid. 2009. Actions in context. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Los Alamitos, CA.

[59] M. Rodriguez, J. Ahmed, and M. Shah. 2008. Action MACH: A spatio-temporal maximum average correlation height filter for action recognition. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Los Alamitos, CA.

[60] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos in the wild. 2009. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Los Alamitos, CA.

[61] D. Weinland, R. Ronfard, and E. Boyer. Free Viewpoint Action Recognition Using Motion History Volumes. *Computer Vision and Image Understanding*, 104, nos. 2/3 ,249-257.

[62] K. Mikolajczyk and H. Uemura. 2008. Action Recognition with MotionAppearance Vocabulary Forest. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Los Alamitos, CA, 1–8.

[63] L. Yeffet and L. Wolf. 2009. Local Trinary Patterns for Human Action Recognition. *In Proceedings of the IEEE 12thConference on Computer Vision*. 492-497.

[64] R. Messing, C. Pal, and H. Kautz. 2009. Activity Recognition Using the Velocitym Histories of Tracked Keypoints. *In Proceedings of the IEEE 12th Conference on Computer Vision*. 104-111.

[65] A. Patron-Perez, M. Marszalek, A. Zisserman, and I. Reid. 2010. High Five: Recognising Human Interactions in TV Shows. *In Proceedings of the British Machine Vision Conference.*

[66] J.C. Nieblesand, C.-W. Chen, and L. Fei-Fei. 2010. Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification. *In Proceedings of the IEEE 11th European Conference on Computer Vision*. 392-405.

[67] G. Yo, J. Yuan, and Z. Liu. 2011. Unsupervised Random Forest Indexing for Fast Action Search. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Los Alamitos, CA , 865-872.

[68] B. Georis, M. Maziere, F. Bremond, and M. Thonnat. 2004. A Video Interpretation Platform Applied to Bank Agency Monitoring. *In Proceedings of the* 2nd *Workshop on Intelligent Distributed Surveillance Systems (IDSS).*

[69] Janez Pers, Matej Kristan, Matej Perse, Stanislav Kovacic. 2008. Analysis of Player Motion in Sport Matches. In: Arnold Baca , Martin Lames, Keith Lyons, Bernhard Nebel, Josef Wiemeyer (eds.), *Computer Science in Sport - Mission and Methods*, (Dagstuhl Seminar Proceedings, no.08372), Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Germany.

[70] J. Sivic and A. Zisserman. 2003. Video Google: A text retrieval approach to object matching in videos. *In Proceedings of the International Conference on Computer Vision.* Vol 2, 1470–1477.

[71] N. Dalal and B. Triggs.2005. Histograms of oriented gradients for human detection. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. I:886--893, IEEE Computer Society.

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

[73] H. Wang, A. Kläser, C. Schmid, L. Cheng-Lin. 2011. Action Recognition by Dense Trajectories. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE ,Colorado Springs, US. 3169-3176.

[74] M.M. Ullah, S.N. Parizi and I. Laptev. 2010. *In Proceedings of BMVC'10.* Aberystwyth, UK.

[75] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. 2007. *International Journal of Computer Vision,* 73, 2, 213-238.

[76] V. Bloom, D. Makris and V. Argyriou. 2012. G3D : A Gaming Action Dataset and Real Time Action Recognition Evaluation Framework. *In Proceedings of the IEEE 3rd International Workshop on Computer Vision for Computer Games (CVCG)*.

[77] C. Cortes and V. Vapnik. 1995. Support-vector network. *Machine Learning, 20,* 273-297.

.