Εθνικό Μετσόβιο Πολυτεχνείο
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΥΠΟΛΟΓΙΣΤΩΝ

Εύρωστη αναγνώριση ανθρώπινης δραστηριότητας σε ρεαλιστικά περιβάλλοντα

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

του

ΓΕΩΡΓΙΟΥ Σ. ΓΟΥΔΕΛΗ

Διπλωματούχου Ηλεκτρονικού Μηχανικού
Πανεπιστημίου του Κεντ Η.Β. (2004)

Αθήνα, Αύγουστος 2015

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
& ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΥΠΟΛΟΓΙΣΤΩΝ

Εύρωστη αναγνώριση ανθρώπινης δραστηριότητας σε ρεαλιστικά περιβάλλοντα

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

του

ΓΕΩΡΓΙΟΥ Σ. ΓΟΥΔΕΛΗ

Διπλωματούχου Ηλεκτρονικού Μηχανικού
Πανεπιστημίου του Κεντ Η.Β. (2004)

Συμβουλευτική Επιτροπή: Στέφανος Κόλλιας
Ανδρέας Σταφυλοπάτης
Κωνσταντίνος Καρπούζης

Εγκρίθηκε από την επταμελή επιτροπή την 31η Αυγούστου 2015.

. . .	. . .	. . .
Στέφανος Κόλλιας	Ανδρέας Σταφυλοπάτης	Κωνσταντίνος Καρπούζης
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Ερευνητής Α΄ ΕΠΙΣΕΥ
. . .	. . .	. . .
Στάμου Γεώργιος	Μαγκλογιάννης Ηλίας	Τσανάκας Παναγιώτης
Επ. Καθηγητής Ε.Μ.Π.	Επ. Καθηγητής ΠΑ.ΠΕΙ.	Καθηγητής Ε.Μ.Π.

. . .
Στέφανος Ζαφειρίου
Επ. Καθηγητής Imperial College of London

Αθήνα, Αύγουστος 2015

Περιεχόμενα

Η ανάλυση εικόνας κέρδισε το ενδιαφέρον μου από νωρίς στα προπτυχιακά μου έτη και στην συνέχεια ακόμα περισσότερο με την εκπόνηση της διπλωματικής μου εργασίας πάνω στον εντοπισμό και την παρακολούθηση οχημάτων σε βίντεο. Στο μεταπτυχιακό μου, επέλεξα να παραμείνω στο πεδίο και να ασχοληθώ αυτή τη φορά με τη χρήση της εικόνας στη βιομετρική αναγνώριση προσώπου. Κατά την ενασχόλησή μου με το αντικείμενο, το ενδιαφέρον μου κέντρισε ιδιαίτερα, η χρήση της εικόνας στην αλληλεπίδραση ανθρώπου-μηχανής. Οι αμέτρητες πιθανές εφαρμογές και το διαρκώς αυξανόμενο ενδιαφέρον που παρουσιάζει σε παγκόσμιο επίπεδο, με οδήγησαν στο να ασχοληθώ εκτενέστερα στη διδακτορική μου διατριβή. Το πως οι μηχανές έχουν γίνει αναπόσπαστο κομμάτι της καθημερινότητάς μας και το πώς ο ευφυής σχεδιασμός τους μπορεί να προσδώσει ποιότητα στη ζωή μας, νομίζω ότι είναι κάτι που πάντα θα με γοητεύει.

Στο σημείο αυτό, θα ήθελα να ευχαριστήσω τον καθηγητή Στέφανο Κόλλια ο οποίος με εμπιστεύθηκε δίνοντας μου την ευκαιρία να ασχοληθώ εκτενέστερα με το παραπάνω ενδιαφέρον κομμάτι της επεξεργασίας εικόνας και ο οποίος ήταν πάντα στο πλευρό μου. Καθ' όλη την διάρκεια της εκπόνησης της διατριβής, στενή συνεργασία είχα με τον ερευνητή Α', Κωνσταντίνο Καρπούζη. Τον ευχαριστώ ιδιαιτέρως για τη στήριξη και την εξαιρετική συνεργασία.

Επίσης, θα ήθελα να ευχαριστήσω τη Σοφία Γούργαρη και τον Γεώργιο Τσατίρη, για την πολύ καλή συνεργασία που είχαμε σε επί μέρους τμήματα της ερευνητικής μου εργασίας. Θα ήθελα επίσης να ευχαριστήσω την ευρύτερη ερευνητική ομάδα του εργαστηρίου, τους περισσότερους εκ των οποίων έχω την τιμή να αποκαλώ φίλους, για το πολύ καλό και ευχάριστο κλίμα στο οποίο συνέβαλαν όλα αυτά τα χρόνια και το οποίο συνετέλεσε στη δημιουργική ενασχόλησή μου με το ερευνητικό έργο.

Η διατριβή αυτή, δεν θα είχε ποτέ υλοποιηθεί χωρίς τη στήριξη και την αφοσίωση των γονιών μου Σταύρου και Βασιλικής. Τους ευχαριστώ γιατί με την αγάπη τους, ήταν αυτοί που στήριξαν από μικρό τα όνειρά μου και φρόντιζαν να έχω πάντα τις απαραίτητες προϋποθέσεις για να τα πραγματοποιώ.

Ένα μεγάλο ευχαριστώ θα ήθελα επίσης να απευθύνω στη σύζυγό μου Δήμητρα, για την υπομονή και την κατανόηση που επέδειξε όλα αυτά τα χρόνια, καθώς και για τη στήριξή της σε αποφάσεις που έπρεπε να λάβω κατά τη διάρκεια αυτών των χρόνων και που μας αφορούσαν εξίσου. Τέλος, ειδική αναφορά θέλω να κάνω στους μικρούς Βασιλική κι Αλέξη. Μπορεί να μην βοήθησαν ιδιαίτερα κατά την ερευνητική ενασχόληση και το γράψιμο της παρούσας διατριβής, είναι αυτοί όμως που με γεμίζουν χαρά και ενέργεια να θέλω να κάνω ακόμα περισσότερα πράγματα στο μέλλον. Εύχομαι να τους κάνω πάντα υπερήφανους.

ΠΕΡΙΛΗΨΗ

Η μηχανική αναγνώριση ανθρωπίνων δράσεων έχει γίνει ιδιαίτερα δημοφιλής την τελευταία δεκαετία. Αυτοματοποιημένα μη επανδρωμένα συστήματα παρακολούθησης, διαδραστικά βιντεοπαιχνίδια, μηχανική μάθηση και ρομποτική, είναι μόνο μερικές από τις περιοχές οι οποίες εμπλέκουν την αναγνώριση της ανθρώπινης δραστηριότητας. Η παρούσα διατριβή εξετάζει την ικανότητα του γνωστού μετασχηματισμού Trace, στη μηχανική αναγνώριση ανθρώπινης δραστηριότητας και προτείνει αρχικά, δύο νέες μεθόδους για την εξαγωγή χαρακτηριστικών βασισμένες στο συγκεκριμένο μετασχηματισμό. Η πρώτη μέθοδος εξάγει μετασχηματισμούς από δυαδικές σιλουέτες που αναπαριστούν διάφορα στάδια μιας απλής περιόδου της κίνησης. Μια τελική αναπαράσταση, αποτελούμενη από τους παραπάνω μετασχηματισμούς, αναπαριστά ολόκληρη την ακολουθία της κίνησης εμπεριέχοντας πολλή από την πολύτιμη χωρο-χρονική πληροφορία της ανθρώπινης κίνησης. Η δεύτερη μέθοδος, βασίζεται στον ίδιο μετασχηματισμό για την κατασκευή ενός συνόλου αμετάβλητων χαρακτηριστικών τα οποία αναπαριστούν την ανθρώπινη κίνηση και μπορούν να αντεπεξέλθουν στις συνήθεις παραμορφώσεις που προκαλούνται κατά την λήψη ενός βίντεο. Η συγκεκριμένη μέθοδος, εκμεταλλεύεται τις φυσικές ιδιότητες του μετασχηματισμού Trace για να παράγει εύρωστα στο θόρυβο χαρακτηριστικά τα οποία είναι αμετάβλητα στην μετατόπιση, στην περιστροφή στην κλιμάκωση και είναι αποτελεσματικά, απλά και γρήγορα στην κατασκευή τους. Σε συνέχεια των παραπάνω, δημιουργήθηκε μια νέα τεχνική η οποία επεκτείνει την τελευταία μέθοδο στον τρισδιάστατο χώρο με τη δημιουργία για πρώτη φορά στη βιβλιογραφία μιας τρισδιάστατης μορφής του Trace, του ονομαζόμενου 3D Cylindrical Trace transform. Σε συνδυασμό με τα χωρο-χρονικά σημεία ενδιαφέροντος (STIPs), εφαρμόστηκε για την εξαγωγή εύρωστων χαρακτηριστικών από βίντεο τόσο για την αναγνώριση ανθρώπινης δραστηριότητας όσο και για τον εντοπισμό ανθρωπίνων πτώσεων. Πειράματα κατηγοριοποίησης που πραγματοποιήθηκαν σε πέντε δημοφιλείς και απαιτητικές βάσεις με τη χρήση SVM πυρήνα Radial Basis Function, παρείχαν εντυπωσιακά αποτελέσματα αναδεικνύοντας τις δυνατότητες των προτεινόμενων τεχνικών. Τέλος, στην προσπάθεια ανάδειξης νέων προκλήσεων στο πεδίο της αναγνώρισης της ανθρώπινης δραστηριότητας, δημιουργήθηκε και προτάθηκε μια νέα μεγάλη βάση δεδομένων (THETIS) η οποίο κατά την πειραματική της αξιολόγηση αναδείχθηκε

σε ένα ιδιαίτερα απαιτητικό σύνολο δεδομένων, το οποίο έχει αρχίσει ήδη να προσελκύει το ενδιαφέρον των ερευνητών.

Λέξεις κλειδιά. Ανάλυση εικόνας, αναγνώριση ανθρωπίνων κινήσεων, εντοπισμός ανθρωπίνων πτώσεων, αλληλεπίδραση ανθρώπου υπολογιστή.

ABSTRACT

Machine based human action recognition has become very popular in the last decade. Automatic unattended surveillance systems, interactive video games, machine learning and robotics are only few of the areas that involve human action recognition. This thesis examines the capability of a known transform, the so-called Trace, for the task of human action recognition and proposes two new feature extraction methods based on the specific transform. The first method, extracts Trace transforms from binarized silhouettes representing different stages of a single action period. A final history template composed from the above transforms, represents the whole sequence containing much of the valuable spatio-temporal information contained in a human action. The second, involves Trace for the construction of a set of invariant features that represents the action sequence and can cope with variations usually appeared in video capturing. The specific method takes advantage of the natural specifications of the Trace transform, to produce noise robust features that are invariant to translation, rotation, scaling and are effective, simple and fast to create. As a follow-up to the above developments, a new technique has been developed, which extends the last referred method to the 3D domain, creating for the first time in bibliography a 3D form of the Trace, named 3D Cylindrical Trace transform. Combined with the spatio-temporal interest points (STIPs), it was applied for the extraction of robust features from videos, both for the scenarios of human action recognition and human fall detection. Classification experiments on five popular and demanding datasets, using SVMs of RBF kernel, provided impressive results indicating the potentials of the proposed techniques. Finally, trying to give prominence to the challenges in the action recognition field, a new dataset named THETIS was created and proposed. The experimental evaluation of THETIS indicated a very challenging dataset which has already attracted researcher's interest.

Keywords. Image processing, human action recognition, human fall detection, human computer interaction.

Κεφάλαιο 1

Εισαγωγή – Αναγνώριση ανθρωπίνων κινήσεων

1.1 Εισαγωγή

Η αναγνώριση ανθρωπίνων κινήσεων έχει εξελιχθεί σε μια ιδιαίτερα σημαντική περιοχή της υπολογιστικής όρασης ενώ έχει φτάσει να αφορά μεγάλο κομμάτι της έρευνας στο συγκεκριμένο πεδίο. Σκοπός της αναγνώρισης της ανθρώπινης δραστηριότητας είναι, η αυτόματη αναγνώριση μιας κίνησης ενός ατόμου η οποία εμπεριέχεται σε μια άγνωστη εικονοσειρά (βίντεο). Σε μια πιο απλοποιημένη προσέγγιση, όπου το βίντεο παρέχεται σε τμήματα τα οποία περιγράφουν την περίοδο μιας κίνησης, η αναγνώριση αφορά στην κατηγοριοποίηση του συγκεκριμένου τμήματος σε μια συγκεκριμένη κατηγορία κινήσεων. Σε αυτή την περίπτωση, πρέπει το σύστημα να κάνει συνεχή παρακολούθηση και εντοπισμό των χρονικών πλαισίων (καρέ έναρξης και καρέ τέλους) όλων των κινήσεων που λαμβάνουν χώρα στο εξεταζόμενο βίντεο.

Η αυτοματοποιημένη αναγνώριση της ανθρώπινης δραστηριότητας βρίσκει χρήση σε μια πληθώρα εφαρμογών. Παραδείγματα αποτελούν: η παρακολούθηση δημόσιων χώρων για τον εντοπισμό ύποπτων ή επικίνδυνων κινήσεων όπως σε καταστήματα, αεροδρόμια κτλ., η επιτήρηση ατόμων με ιδιαιτερότητες (παιδιά, άτομα με ειδικές ανάγκες, ηλικιωμένοι), με στόχο την προστασία και την υποβοήθησή τους, η αυτοματοποιημένη ανάλυση αθλητικών δρώμενων για ποικίλους σκοπούς (αυτοματοποιημένη αναμετάδοση, δημιουργία στατιστικών), η εφαρμογή σε έξυπνα σπίτια με διαφόρους τρόπους, ενώ δεν θα πρέπει να παραβλεφθεί και η κατηγορία του παιχνιδιού (gaming) η οποία έχει φανατικούς και διαρκώς αυξανόμενο σε αριθμό φίλους.

Το ενδιαφέρον που έχει προκαλέσει η αναγνώριση της ανθρώπινης δραστηριότητας ιδιαίτερα στο πλαίσιο της αλληλεπίδρασης ανθρώπου-μηχανής, έχει αυξήσει κατά πολύ την πολυπλοκότητα του προβλήματος. Παραμορφώσεις όπως μετατοπίσεις του υποκει-

μένου, κλιμακώσεις, διαφοροποιήσεις στο φωτισμό και μετατοπίσεις, είναι μερικά μόνο από τα βασικά προβλήματα. Δεν μπορούμε φυσικά να αγνοήσουμε την ιδιαιτερότητα του κάθε ατόμου και του τρόπου με τον οποίο μπορεί να εκτελεί μια κίνηση και να τη διαφοροποιεί σε σχέση με την ίδια κίνηση εκτελεσμένη από κάποιον άλλο. Όλα τα παραπάνω επηρεάζονται δραματικά και εξαρτώνται από το είδος της ζητούμενης εφαρμογής, η οποία θέτει ουσιαστικά και τις επί μέρους απαιτήσεις.

1.2 Σύνοψη συνεισφοράς

Στην παρούσα διατριβή θα εξετάσουμε την ικανότητα δυο ευρέως διαδεδομένων μετασχηματισμών, του Radon και ειδικότερα του Trace στην αναγνώριση ανθρωπίνων κινήσεων. Θα προτείνουμε επίσης μεθόδους βασισμένες στους συγκεκριμένους μετασχηματισμούς για την εξαγωγή εύρωστων χαρακτηριστικών, τα οποία δίνουν λύση σε πολλά από τα βασικά προβλήματα της αναγνώρισης ανθρώπινης δραστηριότητας.

Στις επόμενες ενότητες του τρέχοντος κεφαλαίου, θα παρουσιαστούν οι βασικές τεχνικές για την αυτοματοποιημένη αναγνώριση της ανθρώπινης δραστηριότητας. Οι τεχνικές που παρουσιάζονται είναι κατηγοριοποιημένες βάσει των χαρακτηριστικών που εξάγουν και αποτελούν τμήμα της σημαντικότερης βιβλιογραφίας στο συγκεκριμένο πεδίο.

Στο Κεφάλαιο 2 παρουσιάζουμε μια καινοτόμο μέθοδο για την εξαγωγή χαρακτηριστικών με τη χρήση του μετασχηματισμού Radon. Εκμεταλλευόμενοι τις φυσικές ιδιότητες του μετασχηματισμού, δημιουργούμε εύρωστα στο θόρυβο χαρακτηριστικά και προτείνουμε μια νέα αναπαράσταση η οποία περιγράφει ολόκληρη την κινησιοακολουθία. Η μέθοδος εφαρμόστηκε σε ιδιαίτερα θορυβώδεις σιλουέτες, εξαχθείσες από τη βάση δεδομένων KTH και έδειξε να ανταποκρίνεται πολύ ικανοποιητικά ενώ παρουσιάζεται να είναι ιδιαίτερα γρήγορη. Η συγκεκριμένη εργασία μέρος της οποίας το κεφάλαιο αποτελεί, παρουσιάστηκε στα πρακτικά του συνεδρίου *WIAMIS11* [33] και τιμήθηκε από την επιτροπή του συνεδρίου με το βραβείο καλύτερης δημοσίευσης (*best paper award*).

Στο Κεφάλαιο 3, βασιζόμενοι στην προαναφερθείσα μέθοδο, εξετάσαμε την περίπτωση ενός νεότερου και πιο ευμετάβλητου μετασχηματισμού (Trace) και παρουσιάζουμε τη χρήση του στην αναγνώριση της ανθρώπινης δραστηριότητας. Αρχικά, εξετάζουμε την ικανότητα του μετασχηματισμού για τον διαχωρισμό αντίστοιχων κλάσεων και δημιουργούμε μια αναπαράσταση για την διαισθητική κατανόηση της ικανότητας αυτής. Στη συνέχεια, εφαρμόζουμε τη μέθοδο εξαγωγής χαρακτηριστικών που προτάθηκε αρχικά για τον μετασχηματισμό Radon, εφαρμόζοντας μια σειρά από συναρτησιακά που προσδίδουν διαφορετικές ιδιότητες στην τελική αναπαράσταση, έχοντας σαν αποτέλεσμα ιδιαίτερα εύρωστα στο θόρυβο χαρακτηριστικά. Η εφαρμογή των διαφορετικών συναρτησιακών για των υπολογισμό της τελική προτεινόμενης αναπαράστασης δείχνει να βελτιώνει σημαντικά

τις επιδώσεις συγκριτικά με τη μέθοδο που βασίζεται αποκλειστικά στον μετασχηματισμό Radon.

Στη συνέχεια, λαμβάνοντας υπόψιν μερικά από τα πιο βασικά προβλήματα που παρουσιάζονται στην αναγνώριση της ανθρώπινης δραστηριότητας, εισαγάγαμε μια νέα μέθοδο αναπαράστασης των δεδομένων. Πιο συγκεκριμένα, προτείνουμε μια μέθοδο δημιουργίας χαρακτηριστικών (HTFs) που παρουσιάζονται αμετάβλητα σε μια σειρά στρεβλώσεων όπως η κλιμάκωση/αποκλιμάκωση (zoom-in zoom-out), η περιστροφή (rotation) και η μετατόπιση (translation). Αξίζει να σημειωθεί ότι η τελική αναπαράσταση μια κίνησης, γίνεται με τη χρήση ενός μόνο διανύσματος πολύ μικρής διάστασης. Τα χαρακτηριστικά που προκύπτουν παρουσιάζονται ιδιαιτέρως εύρωστα στο θόρυβο ενώ μια σειρά πειραμάτων σε δυο ευρέως διαδεδομένες βάσεις δεδομένων, αναδεικνύουν την αποτελεσματικότητα της προτεινόμενης μεθόδου. Το κεφάλαιο αυτό βασίζεται κατά κύριο λόγο στο κείμενο της δημοσίευσης [35] που δημοσιεύτηκε στο περιοδικό *Pattern Recognition* της *Elsevier*.

Θέλοντας να εξετάσουμε περαιτέρω τις δυνατότητες της παραπάνω μεθόδου, την εφαρμόσαμε και για το σενάριο του εντοπισμού ανθρωπίνων πτώσεων. Η εφαρμογή αυτή αν και αρκετά συγγενής, επιδέχεται διαφορετικής προσέγγισης και τις περισσότερες φορές αποτελεί ειδική κατηγορία. Στο Κεφάλαιο 4 παρουσιάζεται η εφαρμογή του αλγορίθμου HTFs στο σενάριο του εντοπισμού πτώσεων και η επίδοσή του σε δυο νέες και απαιτητικές βάσεις δεδομένων. Ο αλγόριθμος έδειξε να ανταποκρίνεται άριστα επιτυγχάνοντας κατηγοριοποίηση της τάξεως του 100% και στις δύο βάσεις, ξεπερνώντας τον ανταγωνισμό αλλά και τις προτεινόμενες μεθόδους στις δημοσιεύσεις που συνοδεύουν τα εν λόγω σύνολα δεδομένων.

Δουλεύοντας στο πεδίο της αναγνώρισης της ανθρώπινης δραστηριότητας, θελήσαμε να αναδείξουμε κάποια από τα προβλήματα που υπάρχουν, αλλά και να δώσουμε τόσο στον εαυτό μας, όσο και στη λοιπή ερευνητική κοινότητα, τη δυνατότητα να δουλέψουμε με σύγχρονα δεδομένα που αφορούν ολοένα και περισσότερες εφαρμογές. Στο Κεφάλαιο 5, παρουσιάζουμε μια νέα βάση δεδομένων που δημιουργήσαμε ακριβώς για το σκοπό αυτό. Η ονομαζόμενη βάση THETIS, αποτελεί ένα μεγάλο σύνολο δεδομένων το οποίο περιλαμβάνει 8374 βίντεο, 5 διαφορετικών τύπων, των οποίων η λήψη έχει γίνει με τη χρήση του αισθητήρα Kinect.

Τα βίντεο που απαρτίζουν το σύνολο THETIS, περιέχουν δράσεις αθλητικού περιεχομένου και έχουν κατατμηθεί χειροκίνητα ούτως ώστε να αναπαριστούν την περίοδο της κάθε περιλαμβανόμενης κίνησης. Οι τύποι των βίντεο που παρουσιάζονται είναι εκτός από το κλασσικό RGB, πληροφορία βάθους, πληροφορία σιλουέτας και σκελετός σε δυο και σε τρεις διαστάσεις. Πειράματα που διεξήχθησαν για την αξιολόγηση της βάσης με αλγορίθμους τελευταίας γενιάς, έδειξαν ότι πρόκειται για ένα ιδιαίτερα απαιτητικό σύνολο το οποίο δίνει τη δυνατότητα για πολύπλευρο πειραματισμό. Αξίζει να σημειωθεί πως την

ώρα που γράφεται η παρούσα διατριβή, έχει προσελκύσει ήδη το ενδιαφέρον ερευνητικών ομάδων οι οποίες έχουν ήδη δημοσιεύσει αποτελέσματα επάνω στη THETIS. Η βάση δεδομένων THETIS παρουσιάστηκε στα πρακτικά του συνεδρίου *CVPR13* [37].

Έχοντας κατά νου πως οι ανθρώπινες δράσεις είναι στην ουσία χωρο-χρονικοί όγκοι και στη διαρκή προσπάθειά μας να αποδώσουμε όσο το δυνατόν περισσότερη από τη χωρο-χρονική πληροφορία σε μια διαχειρίσιμη αναπαράσταση, παρουσιάζουμε μια νέα μέθοδο στο Κεφάλαιο 6. Αρχικά στο συγκεκριμένο κεφάλαιο, θα παρουσιάσουμε τον προτεινόμενο μετασχηματισμό 3D Cylindrical Trace (CTT), ο οποίος αποτελεί την πρώτη προτεινόμενη τρισδιάστατη μορφή του μετασχηματισμού Trace και ο οποίος επεκτείνει τις δυνατότητές του, στο τρισδιάστατο χώρο. Ο μετασχηματισμός έγινε με γνώμονα την εφαρμογή του για την εξαγωγή χαρακτηριστικών από χωρο-χρονικούς όγκους και είναι σχεδιασμένος να εκτελείται σημαντικά πιο γρήγορα σε σχέση με άλλους τρισδιάστατους μετασχηματισμούς όπως ο 3D Radon.

Στη συνέχεια του ίδιου κεφαλαίου, παρουσιάζεται μια νέα μεθοδολογία η οποία συνδυάζει τα χαρακτηριστικά του CTT, με τον αλγόριθμο τελευταίας γενιάς Selective Spatio-Temporal Interest Points (SSTIPs) και τα HTFs που παρουσιάζονται σε προηγούμενη ενότητα. Η τεχνική χάρη στον προτεινόμενο μετασχηματισμό, καταφέρνει να ”συλλάβει” τους διακριτούς σε κάθε κίνηση γεωμετρικούς συσχετισμούς των χωρο-χρονικών σημείων ενδιαφέροντος τους οποίους χάρη στα HTFs, αποδίδει σε ένα πολύ μικρού μεγέθους διάνυσμα. Έχοντας προεκτείνει τις δυνατότητες του Trace στον τρισδιάστατο χώρο, η τεχνική είναι παραμετροποιήσιμη και προσδίδει στην τελική αναπαράσταση μεγάλη ευρωστία στο θόρυβο και σε γνωστές στρεβλώσεις όπως η μετατόπιση η περιστροφή και η κλιμάκωση. Η μέθοδος δοκιμάστηκε σε πέντε γνωστές βάσεις δεδομένων τόσο για το σενάριο της αναγνώρισης ανθρώπινης δραστηριότητας, όσο και για τον εντοπισμό ανθρωπίνων πτώσεων. Τα πειραματικά αποτελέσματα έδειξαν μια εξαιρετική συμπεριφορά και για τις δύο εφαρμογές, ξεπερνώντας στις περισσότερες τον περιπτώσεων τις μεθόδους τελευταίας τεχνολογίας. Η παραπάνω έρευνα έχει υποβληθεί και είναι υπό κρίση στο περιοδικό *Image and Video Computing* της *Elsevier* [34].

Η διατριβή καταλήγει με το Κεφάλαιο 7, στο οποίο θα παρουσιαστούν τα συμπεράσματα της μέχρι τώρα έρευνας, θα αναφερθούν οι τρέχουσες προσπάθειες και θα προταθούν μελλοντικές κατευθύνσεις.

1.3 Βασικές τεχνικές αναγνώρισης ανθρωπίνων κινήσεων

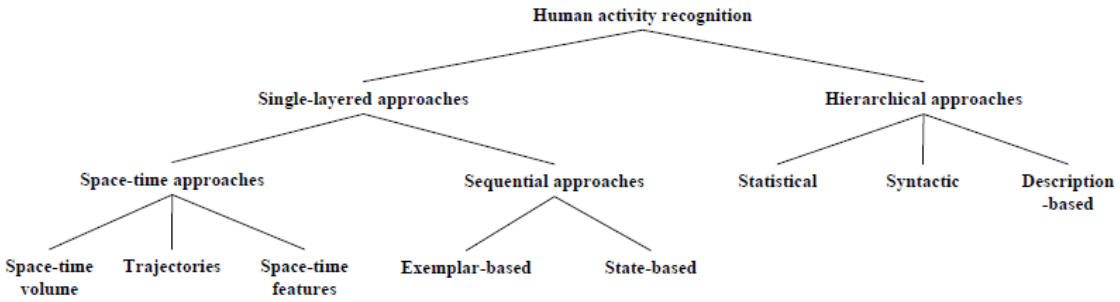
Σε μια πρόσφατα δημοσιευμένη επισκόπηση του πεδίου της ανάλυσης ανθρωπίνων κινήσεων [1], η ανθρώπινη δραστηριότητα κατηγοριοποιείται σε τέσσερα διαφορετικά επίπεδα με άξονα την πολυπλοκότητα της κάθε κίνησης: χειρονομίες, δράσεις, αλληλεπιδρά-

σεις, και δραστηριότητες ομάδων. Ως χειρονομίες ορίζονται οι στοιχειώδεις κινήσεις ενός μέλους ενός ατόμου και είναι τα ατομικά στοιχεία εκείνα που περιγράφουν μια σημασιολογική κίνησή του. Τέτοια παραδείγματα είναι το "σήκωμα του ποδιού" ή το "κούνημα του χεριού". Οι δράσεις είναι δραστηριότητες ενός μόνο ατόμου οι οποίες μπορεί να αποτελούνται από πολλαπλές χειρονομίες οργανωμένες χρονικά, όπως το "βάδισμα" το "τρέξιμο" ή το "άλμα". Οι αλληλεπιδράσεις είναι ανθρώπινες δραστηριότητες οι οποίες απαιτούν την συμμετοχή δύο ή περισσοτέρων ατόμων, ή ανθρώπων και αντικειμένων. Για παράδειγμα, η "χειραψία" είναι χαρακτηριστικό παράδειγμα αλληλεπίδρασης δύο ατόμων ενώ το "σήκωμα μιας καρέκλας" είναι αλληλεπίδραση ατόμου/αντικειμένου. Τέλος, δραστηριότητα ομάδας είναι η δραστηριότητα που πραγματοποιείται από ένα σύνολο ανθρώπων, όπως για παράδειγμα μια "επαγγελματική σύσκεψη" ή μια "διαδήλωση".

Σύμφωνα με το προαναφερθέν άρθρο, οι μέθοδοι αναγνώρισης δραστηριότητας κατατάσσονται σε δυο κύριες κατηγορίες: τις προσεγγίσεις μονής στιβάδας (single-layered approaches) και τις ιεραρχικές προσεγγίσεις (hierarchical approaches). Οι προσεγγίσεις μονής στιβάδας είναι αυτές οι οποίες αναπαριστούν και αναγνωρίζουν τις δραστηριότητες απευθείας από μια δοσμένη ακολουθία εικόνων. Εξ αιτίας της φύσης του, ο συγκεκριμένος τύπος μεθόδων ενδείκνυται για την αναγνώριση χειρονομιών και δράσεων με χαρακτηριστικά αλληλουχίας. Στον αντίποδα, οι ιεραρχικές προσεγγίσεις αναπαριστούν σύνθετες, υψηλού επιπέδου κινήσεις περιγράφοντάς τες σε αντιστοιχία με άλλες απλούστερες κινήσεις, οι οποίες γενικά αποκαλούνται υπο-συμβάντα (subevents). Τα αντίστοιχα συστήματα αναγνώρισης, αποτελούνται από πολλαπλά επίπεδα κατασκευασμένα με τέτοιο τρόπο που να επιτρέπουν την ανάλυση σύνθετων κινήσεων.

Με τη σειρά τους, οι προσεγγίσεις μονής στιβάδας κατηγοριοποιούνται σε δυο τύπους ανάλογα με το πως μοντελοποιούν την κίνηση: αυτοί είναι οι χωροχρονικές προσεγγίσεις και οι προσεγγίσεις ακολουθίας. Οι χωροχρονικές προσεγγίσεις εξετάζουν το βίντεο της κίνησης σαν ένα τρισδιάστατο όγκο (XYZ), ενώ οι προσεγγίσεις ακολουθίας το ερμηνεύουν σαν μια ακολουθία παρατηρήσεων. Ανάλογα με τα χαρακτηριστικά που χρησιμοποιούν, οι χωροχρονικές μέθοδοι χωρίζονται με τη σειρά τους σε τρεις νέες κατηγορίες ανάλογα με τα χαρακτηριστικά που εξάγουν από τον τρισδιάστατο όγκο: τους ίδιους τους όγκους, τροχιές ή τοπικά σημεία ενδιαφέροντος. Οι μέθοδοι ακολουθίας κατηγοριοποιούνται ανάλογα με το αν χρησιμοποιούν παραδειγματικού τύπου διαδικασίες αναγνώρισης ή βασισμένες σε μοντέλα. Μια σχηματική ταξινομία της παραπάνω κατηγοριοποίησης δίνεται στο σχήμα 1.1.

Οι ιεραρχικές προσεγγίσεις κατηγοριοποιούνται βάσει των μεθοδολογιών αναγνώρισης που χρησιμοποιούν: στατιστικές προσεγγίσεις, συντακτικές και προσεγγίσεις βασισμένες σε περιγραφείς. Οι στατιστικές μέθοδοι κατασκευάζουν στατιστικά μοντέλα βασισμένα σε στάδια συνδεόμενα ιεραρχικά (π.χ. Hidden Marcov Models) με σκοπό την ανα-



Σχήμα 1.1: *Ιεραρχική ταξινομία μεθόδων αναγνώρισης ανθρώπινης δραστηριότητας βάσει προσέγγισης [1].*

παράσταση και την αναγνώριση υψηλού επιπέδου κινήσεων. Αντίστοιχα, οι συντακτικές προσεγγίσεις, χρησιμοποιούν ένα γραμματικό συντακτικό όπως μια στοχαστική γραμματική ανοιχτού γλωσσικού περιβάλλοντος, για να μοντελοποιήσουν δραστηριότητες ακολουθίας. Ουσιαστικά μοντελοποιούν μια κίνηση υψηλού επιπέδου σε μια σειρά από απλές, ατομικού επιπέδου κινήσεις. Οι περιγραφικές μέθοδοι, αναπαριστούν τις ανθρώπινες κινήσεις περιγράφοντας υπο-συμβάντα των κινήσεων και τη χωρική, τη χρονική και τη λογική δομή τους.

1.3.1 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 και άλλα. 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 δραστηριό-

τητες 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.

1.3.2 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 βασικές 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 και επιτήρησης, 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, οι σιδηροδρομικοί σταθμοί και οι τράπεζες, 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 βίντεο [109], [39].

* *Διαδραστικά περιβάλλοντα και εφαρμογές*. Η κατανόηση της αλληλεπίδρασης μεταξύ ανθρώπου και υπολογιστή παραμένει μια διαρκής πρόκληση στο πρόβλημα του σχεδιασμού σχετικών διαπροσωπειών. 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 ανθρώ-

πινης 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.

* *Ανάλυση βίντεο με βάση το περιεχόμενο*. Τα βίντεο αποτελούν μέρος της καθημερινότητας των ανθρώπων και με την συνεχή εξάπλωση των ηλεκτρονικών κοινωνικών δικτύων που διαμοιράζουν πάσης φύσεως βίντεο, κρίνεται αναγκαία η αποτελεσματική δημιουργία ευρετηρίου και αποθήκευση τους για την διευκόλυνση του χρήστη. Αυτή η διαδικασία απαιτεί την εκμάθηση προτύπων από βίντεο 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 βίντεο αυξήθηκε σημαντικά [95]. 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.

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

1.3.3 Μέθοδοι αναγνώρισης ανθρωπίνων 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 Aggarwal και M.S. Ryoo [1]. Η ταξινόμηση αυτή απεικονίζεται στο Σχήμα 1.1 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.

Μέθοδοι Single-layered ή μονής στιβάδας

Ως 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).

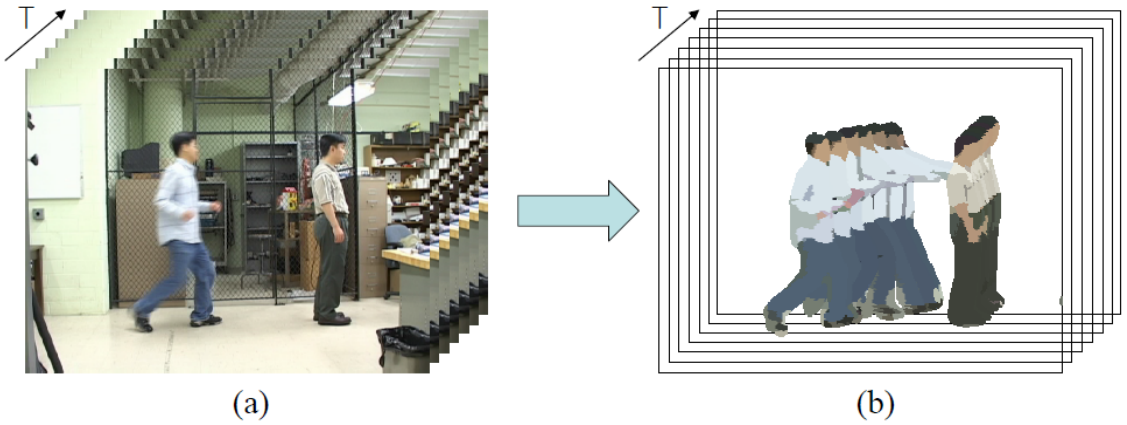
Space-time

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

Μια τυπική μεθοδολογία αναγνώρισης ανθρώπινης δραστηριότητας που βασίζεται στον τρισδιάστατο αυτό όγκο ενός βίντεο 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 χωροχρονικός όγκος που την αντιπροσωπεύει. Τέλος, χρησιμοποιώντας έναν αλγόριθμο ταιριασμάτων προτύπων, ο νέος όγκος χωροχρόνου συγκρίνεται με τα υπάρχοντα πρότυπα και επιλέγεται η δραστηριότητα εκείνη που το πρότυπό της ταιριάζει περισσότερο (Σχήμα 1.2).

Εκτός από την τυπική αναπαράσταση των βίντεο 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 αναπα-



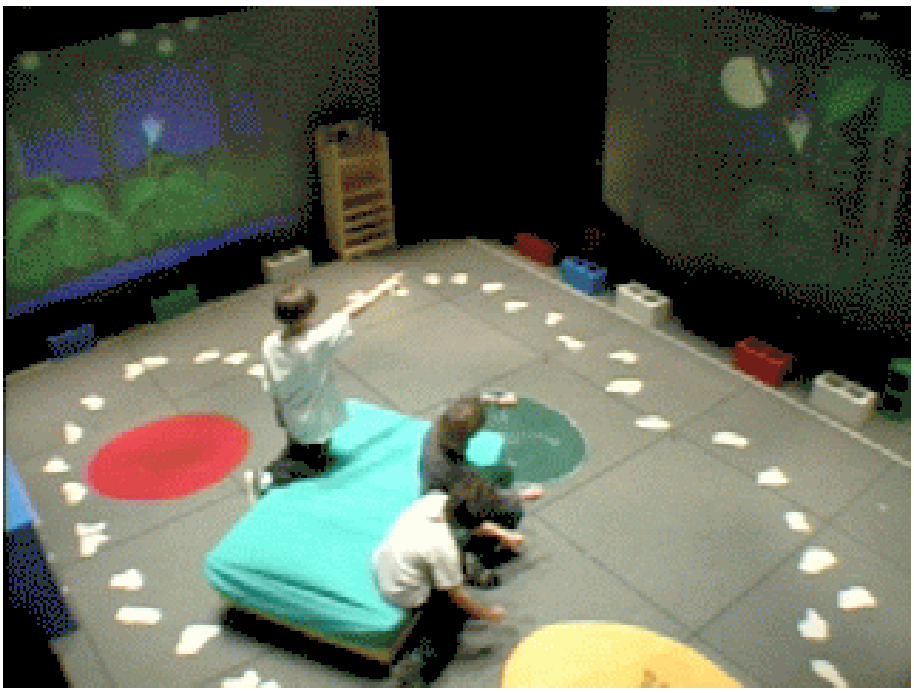
Σχήμα 1.2: *Παραδείγματα τρισδιάστατων όγκων XYT κατακευασμένα από: (a) ολόκληρες εικόνες και (b) blob εικόνων από εικονοσειρά που αναπαριστά την κίνηση «γρονθοκοπώ». [1].*

ράστασης της ακολουθίας εικόνων, καθώς και στους διάφορους αλγορίθμους που χρησιμοποιούνται στην αναγνώριση της ανθρώπινης δραστηριότητας.

Space-time volume Η αναγνώριση κινήσεων μέσω της αναπαράστασης του όγκου στο χωροχρόνο 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 [8] 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). Παραδείγματα χωροχρονικών αναπαραστάσεων δίνονται στο Σχήμα 1.4. Οι δύο εικόνες κατασκευάζονται από μια ακολουθία εικόνων στο μπροστινό πλάνο, οι οποίες αποτελούν ουσιαστικά δισδιάστατες προβολές (XY) του αρχικού τρισδιάστατου όγκου XYT στο 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 1.3).

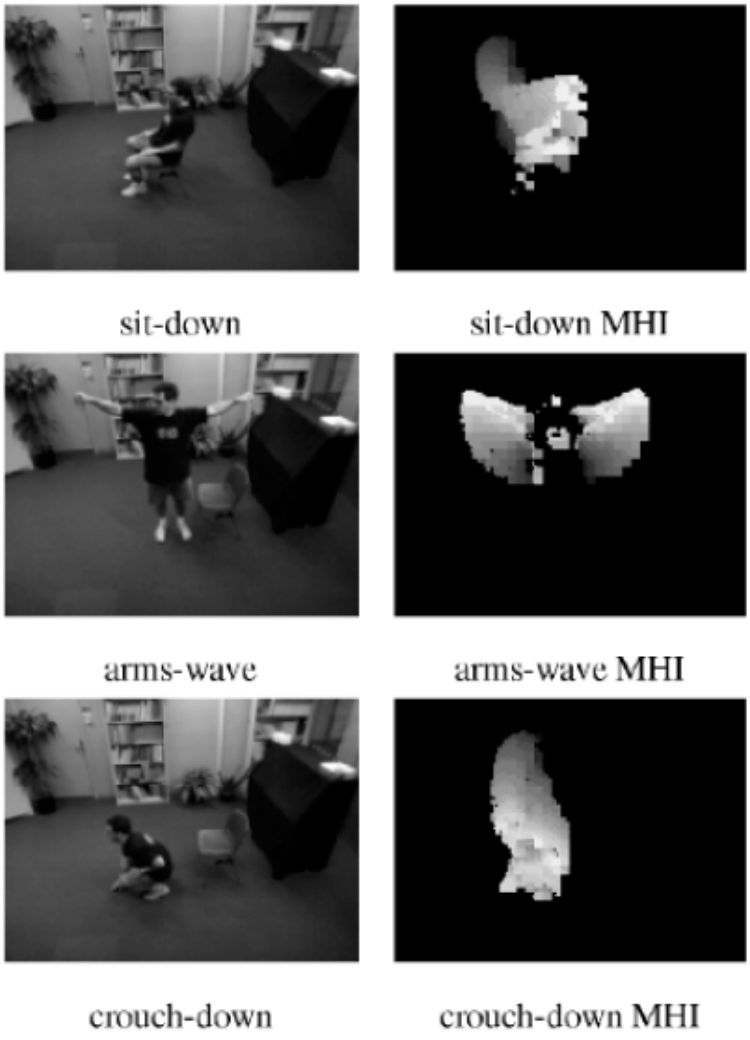
Οι Shechtman και Irani [101] για την επίτευξη της αναγνώρισης της ανθρώπινης δραστηριότητας χρησιμοποιούν την οπτική ροή (optical flow) του τρισδιάστατου χωροχρο-



Σχήμα 1.3: *Διαδραστικός αφηγηματικός χώρος παιχνιδιού (Kids Room) [9].*

νικού όγκου. Επιπλέον, μετρώντας την ομοιότητα που υπάρχει μεταξύ του εξαγόμενου όγκου ενός νέου βίντεο και των πρότυπων όγκων που έχουν στη διάθεσή τους, κατασκευάζουν μια συσχέτιση με τα πρότυπα βίντεο. 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 βίντεο 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 βίντεο που εξετάζει το σύστημα κάθε φορά. Έτσι, όταν δοθεί ένα άγνωστο βίντεο, 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..

Οι Ke et al. [50] αξιοποίησαν την κατάτμηση του όγκου στο χωροχρόνο για να μοντελοποιήσουν ανθρώπινες δραστηριότητες. Το σύστημά τους εφαρμόζει έναν ιεραρχικό αλγόριθμο meanshift για να κατηγοριοποιήσει τα voxels ανάλογα με το χρώμα τους, αποκτώντας έτσι κατατμημένους όγκους. Η αναγνώριση της κίνησης επιτυγχάνεται ψάχνοντας για ένα υποσύνολο κατατμημένων όγκων που ταιριάζουν περισσότερο με το πρότυπο της κίνησης. Το σύστημα εφαρμόστηκε στην αναγνώριση κινήσεων της βάσης KTH [100], καθώς επίσης και σε αγώνες αντισφαίρισης σε βίντεο με πιο πολύπλοκο φόντο (background).



Σχήμα 1.4: *Παραδείγματα χωροχρονικών αναπαραστάσεων: motion-history images [8]. Αυτή η αναπαράσταση μπορεί να γίνει αντιληπτή σαν μια ζυγισμένη προβολή των 3D XYT όγκων σε 2D XY διαστάσεις [9].*

Μια διαφορετική τεχνική χρησιμοποίησαν οι Rodriguez et al. [92] 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 βίντεο και αναλύοντας την απόκρισή του. Πειράματα με χρήση της μεθόδου αυτής πραγματοποιήθηκαν πάνω στις βάσεις KTH και Weizmann [6].

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.

Space-time trajectories Για την αναγνώριση της ανθρώπινης δραστηριότητας υπάρχουν προσεγγίσεις που αντιλαμβάνονται την δραστηριότητα ως ένα σύνολο από τροχιές στο χωροχρόνο. 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. [103] 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 [129] κάνουν χρήση όμοιας αναπαράστασης για τη σύγκριση βίντεο από κάμερες που κινούνται.

Μια διαφορετική προσέγγιση εισήγαγαν οι Campbell και Bobick [12], 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 βίντεο 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 [90] 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 που βασίζονται στην ανάλυση τροχιών είναι ανεξάρτητες από την οπτική γωνία. Παρόλα 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 τρισδιάστατης ανίχνευσης και εντοπισμού των μελών του ανθρώπινου σώματος.

Space-time features 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 δραστηριότητα. Για να περιγραφεί επαρκώς μια μέθοδος τύπου space-time features, είναι απαραίτητο να απαντηθούν τρία ερωτήματα που την αφορούν. Πρώτον, ποιά τοπικά χαρακτηριστικά εξάγει, δεύτερον, με ποιόν τρόπο τα αξιοποιεί για να αναπαραστήσει μια κίνηση και τέλος, ποια μεθοδολογία χρησιμοποιεί για την ταξινόμηση των κινήσεων.

Οι Chomat και Crowley [15] 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.

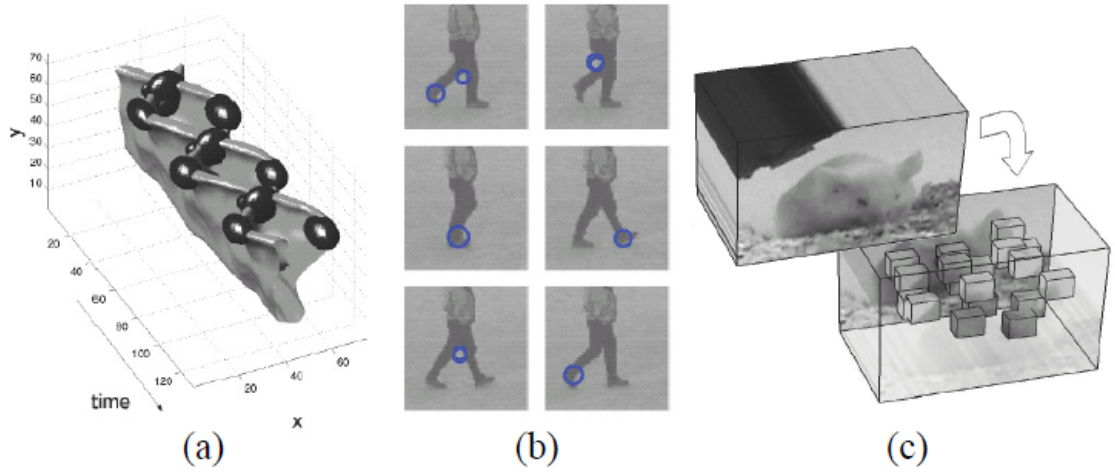
Μια διαφορετική προσέγγιση που βασίζεται στη χρήση τοπικών χαρακτηριστικών σε πολλαπλές βαθμίδες πρότειναν οι Zelnic-Manor και Irani [132]. Η ανάλυση του όγκου των βίντεο σε χρονικές βαθμίδες αποσκοπεί στην αντιμετώπιση της διαφοροποίησης που μπορεί να υπάρξει στην ταχύτητα εκτέλεσης μιας δραστηριότητας. Για κάθε σημείο ΧΥΤ 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. [6], με τη διαφορά ότι η εξαγωγή πραγματοποιήθηκε με χρήση της συνάρτησης Poisson. Ακριβέστερα, για κάθε pixel κατασκευάζεται ένας τρισδιάστατος όγκος ΧΥΤ 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 σύστημά 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 [56], οι οποίοι εξάγουν αραιά σημεία ενδιαφέροντος (interest points) από τις ακολουθίες εικόνων. Συγκεκριμένα, ο ανιχνευτής τους εντοπίζει γωνίες στον τρισδιάστατο χώρο ΧΥΤ, οι οποίες συλλαμβάνουν ποικίλες μεταβολές στην κίνηση, όπως αλλαγή κατεύθυνσης ενός αντικειμένου, συγχώνευση ή διαχωρισμό της δομής μιας εικόνας. Παραδείγματα χωροχρονικών τοπικών χαρακτηριστικών δίνονται στο Σχήμα 1.5.

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 περιγραφέας του Lowe [62].

Οι Dollar et al. [21] 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 1.5 c)). Στη συνέχεια, εφαρμόζει κατάλληλους μετασχηματισμούς για να προκύψουν τα τελικά τοπικά χαρακτηριστικά. Τέλος, το



Σχήμα 1.5: *Παραδείγματα χωροχρονικών τοπικών χαρακτηριστικών εξαχθέντα από βίντεο της κίνησης "περπάτημα" και αυτών από ένα βίντεο κίνησης ποντικιού. Το a) δείχνει αλυσιδωτές XYT επιφάνειες των ποδιών ενός ατόμου και τα σημεία ενδιαφέροντος που έχουν προκύψει [56]. Το b) δείχνει τα ίδια σημεία τοποθετημένα στις αρχικές εικόνες. Το c) δείχνει κυβοειδή εξαχθέντα χαρακτηριστικά [21].*

σύστημα μοντελοποιεί κάθε κίνηση με ένα ιστόγραμμα που χρησιμοποιείται στην τελική αναγνώριση εκφράσεων προσώπου, συμπεριφορών ποντικιών και ανθρώπινων δραστηριοτήτων.

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. [98] για την εξαγωγή της πληροφορίας περί συσχέτισης των χαρακτηριστικών. Ομοίως, οι Laptev et al. [58] 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 βίντεο περισσότερο ρεαλιστικά.

Στην ίδια κατεύθυνση βρίσκεται και η προσέγγιση των Ryoo και Aggarwal [97], 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 απλές 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.

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 τρισδιάστατη 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.

Sequential

Σε αυτήν την κατηγορία ανήκουν οι μέθοδοι αναγνώρισης μέσω της ανάλυσης ακολουθιών από χαρακτηριστικά. Σε αυτά τα συστήματα, ένα βίντεο 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 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.

Exemplar-based Τα συστήματα που χρησιμοποιούν πρότυπα, 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 βίντεο, συγκρίνεται η ακολουθία διανυσμάτων με τα χαρακτηριστικά από το βίντεο αυτό με την ακολουθία πρότυπο. Εάν υπάρχει αρκετά μεγάλη ομοιότητα, το σύστημα συμπεραίνει ότι το βίντεο 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 [19] 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 βίντεο, 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 είσοδοι.

Οι Gavrila και Davis [28] 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 [125], 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.

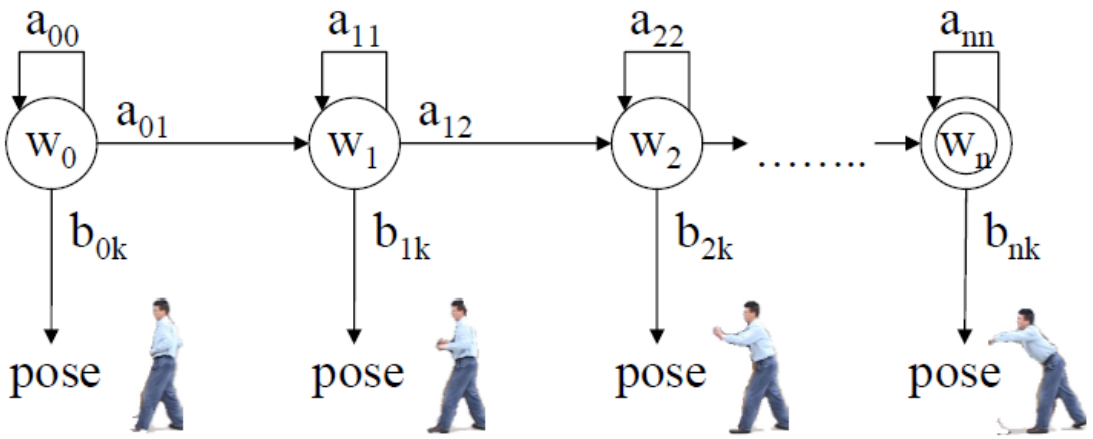
Η αναγνώριση ανθρώπινων κινήσεων σε μεγάλη απόσταση όπου οι κινήσεις είναι δυσδιάκριτες, απασχόλησε τους Efros et al. [24] οι οποίοι χρησιμοποίησαν περιγραφείς κινή-

σεων 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 βίντεο 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. [64], 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 [119] στη σαφή μοντελοποίηση των ενδοατομικών και διατομικών διαφοροποιήσεων 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.

State model-based 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 εικό-



Σχήμα 1.6: *Δείγμα απλού, αυστηρά ακολουθιακού κρυφού Μαρκοβιανού μοντέλου, για την κίνηση "τέντωμα βραχίωνα". Κάθε εικόνα ηθοποιού στο σχήμα, αναπαριστά μια πόζα με την μέγιστη πιθανότητα παρατήρησης b_{jk} για το στάδιό του w_j [1].*

νων 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 υπόψιν 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 βίντεο που έχει δοθεί ως είσοδος. Στο Σχήμα 1.6 παρουσιάζεται ένα παράδειγμα ακολουθιακού HMM.

Η εργασία των Yamato et al. [126] 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 χα-

ρακτηριστικό 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 αλλαγές στα χαρακτηριστικά των μοντέλων κατά την εκτέλεση ανθρώπινων δραστηριοτήτων.

Ομοίως, συμβατικά HMMs χρησιμοποιήθηκαν και από τους Starner και Pentland [108] με σκοπό την αναγνώριση της Αμερικανικής 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. [126]. 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 [10] έκαναν χρήση των μοντέλων κατάστασης. Στο σύστημά τους, μια χειρονομία παρουσιάζεται ως μια δισδιάστατη ΧΥ τροχιά που περιγράφει τις αλλαγές στη θέση των χεριών. Κάθε καμπύλη αναλύεται σε ακολουθίες διανυσμάτων και κατ’ επέκταση σε ακολουθίες καταστάσεων που υπολογίζονται από ένα παράδειγμα εκπαίδευσης. Επιπροσθέτως, κάθε κατάσταση είναι ασαφής ώστε να λαμβάνονται υπόψιν 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. [83] 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 [85] 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 [74] 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.

Σύγκριση Γενικά, οι ακολουθιακές προσεγγίσεις λαμβάνουν υπόψιν 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 υπόψιν 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 (βίντεο) αφού οι δραστηριότητες που σκοπεύουν να αναγνωρίσουν είναι περισσότερο σύνθετες.

1.3.4 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 αναγνώρι-

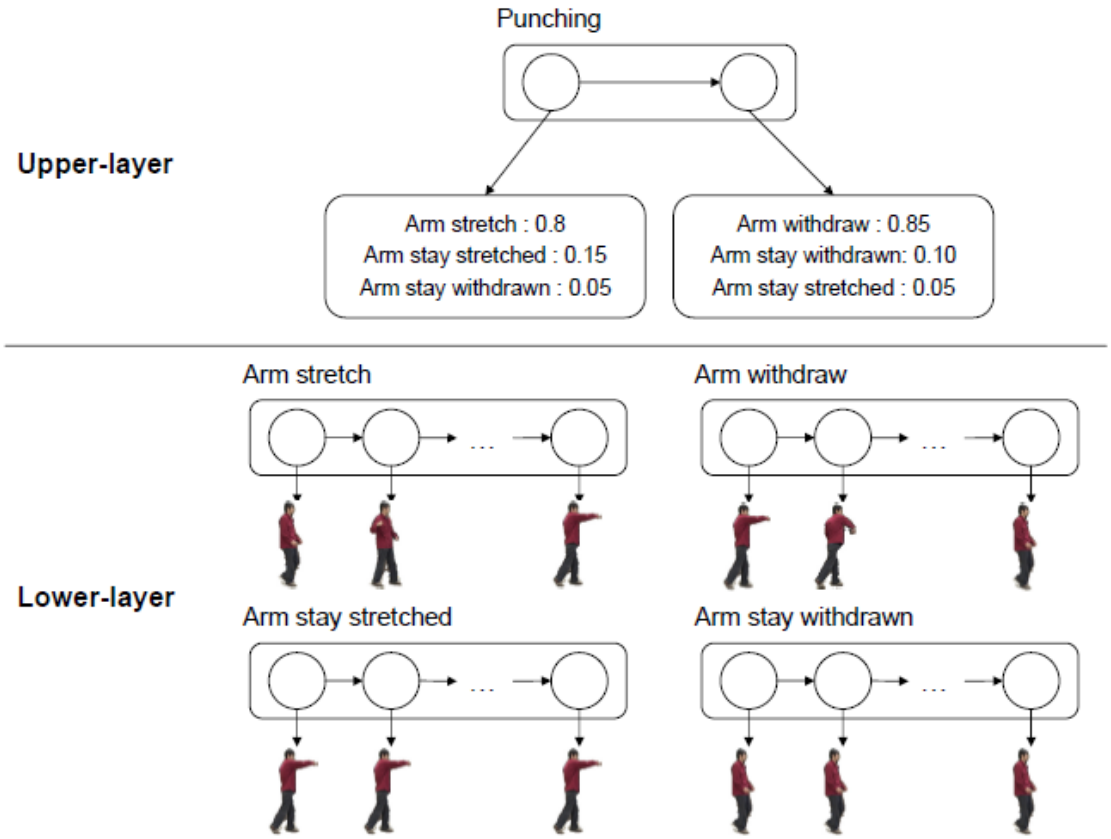
σης διευκολύνεται από την ιεραρχική μοντελοποίηση των δραστηριοτήτων υψηλού επιπέδου. Η ανθρώπινη γνώση μπορεί να συμπεριληφθεί στο σύστημα απαριθμώντας σημαντικές σημασιολογικά υποδραστηριότητες, που συνθέτουν μια δραστηριότητα υψηλού επιπέδου και/ή προσδιορίζοντας τις σχέσεις τους. Κατά την μοντελοποίηση 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.

Statistical

Η αναγνώριση ανθρώπινης δραστηριότητας μέσω των στατιστικών μεθοδολογιών, πραγματοποιείται 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 είσοδος στο σύστημα. Στο Σχήμα 1.7 παρουσιάζεται ένα παράδειγμα ενός τέτοιου μοντέλου για την αναγνώριση της δραστηριότητας ”γρονθοκοπώ”.

Οι Oliver et al. [82] εισήγαγαν μια από τις θεμελιώδεις 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 1.7: *Δείγμα ιεραρχικού μοντέλου για την αναγνώριση της κίνησης "γρονθοκοπώ". Το μοντέλο αποτελείται από δύο επίπεδα. Στο χαμηλότερο επίπεδο το HMM χρησιμοποιείται για την αναγνώριση διαφόρων ατομικού επιπέδου δραστηριοτήτων όπως το "τέντωμα" και η "ανάκληση". Το ανώτερο επίπεδο HMM λαμβάνει τα αποτελέσματα του κατώτερου HMM σαν είσοδο, αναγνωρίζοντας ότι το "γρονθοκοπώ" είναι "τέντωμα" και ότι η "ανάκληση" είναι μέρος μιας ακολουθίας [1].*

Το παραπάνω παράδειγμα το μιμήθηκαν πολλοί ερευνητές, όπως οι ερευνητές στο [77], οι οποίοι κατασκεύασαν ένα HMM δυο επιπέδων για την αναγνώριση πολύπλοκων ακολουθιακών δραστηριοτήτων, (π.χ., "ένα άτομο παίρνει ένα γεύμα" και "ένα άτομο τρώει ένα snack"). Ομοίως, οι Zhang et al. [133] 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 [31] 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. [16] 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 [18], 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. [104]. 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 προϋπόθεση 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.

Syntactic

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 κι εδώ, 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 [71] επέκτειναν την παραπάνω μέθοδο, αυξάνοντας την ακρίβεια και βελτιώνοντας την ανίχνευση λάθους. Αντιθέτως, το σύστημα ΑΑΔ των Minnen et. al. [70] 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 [47], 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 τελευ-

ταίων πρέπει να είναι αυστηρώς ακολουθιακή.

Description-based

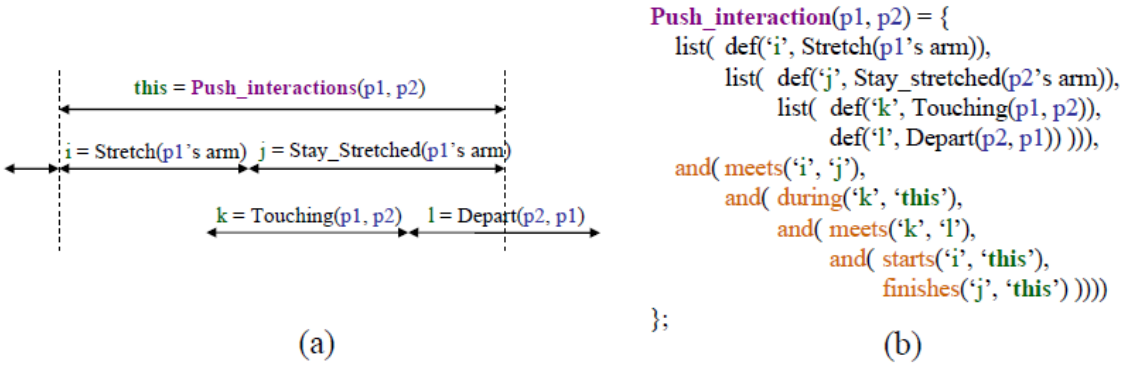
Οι περιγραφικές ιεραρχικές μεθοδολογίες αναπαριστούν την ανθρώπινη δραστηριότητα υψηλού επιπέδου ως ένα σύνολο από πιο απλές κινήσεις (π.χ. υπο-γεγονότα), περιγράφοντας τις χωρικές, χρονικές και λογικές τους συσχετίσεις. Επομένως, η αναγνώριση μιας δραστηριότητας επιτυγχάνεται ψάχνοντας όλα τα υπο-γεγονότα 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 [3], τα οποία είναι τα εξής: before, meets, overlaps, during, starts, finishes και equals. Αξίζει να σημειωθεί πως τα κατηγορήματα before και meets περιγράφουν διαδοχικές σχέσεις, ενώ τα υπόλοιπα ορίζουν σύγχρονες σχέσεις. Στο Σχήμα 1.8 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 δραστηριότητας. 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 1.8 b)). Επομένως, η CFG απλώς διαβεβαιώνει πως η αναπαράσταση της δραστηριότητας πληροί τους κανόνες της γραμματικής.

Οι Pinhanez και Bobick [88] 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 [42] χρησιμοποίησαν μια περιγραφική μέθοδο για την ανάλυση αγώνων Αμερικανικού ποδοσφαίρου, αναπαριστώντας την ανθρώπινη δραστηριότητα σε τρία επίπεδα: atomic, individual και team-level.

Αξίζει να αναφερθεί η γλώσσα προγραμματισμού VERL που σχεδιάστηκε από τους Nevatia et al. [75], με σκοπό την περιγραφή της ανθρώπινης δραστηριότητας. Αρχικά όρισαν τρία επίπεδα ιεραρχίας για τις δραστηριότητες: ”απλό γεγονός”, ”σύνθετο γεγονός”



Σχήμα 1.8: *a)Χρονικά διαστήματα μιας αλληλεπίδρασης και των υπο-συμβάντων της b)ο σχετικός ψευδοκώδικας [96].*

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 [96] 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 [116], 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.

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).

Κεφάλαιο 2

Μετασχηματισμός Radon για την αναγνώριση ανθρωπίνων κινήσεων

2.1 Εισαγωγή

Η παρατήρηση και η ανάλυση της ανθρώπινης συμπεριφοράς είναι ένα ανοικτό ερευνητικό θέμα την τελευταία δεκαετία. Η αναγνώριση της ανθρώπινης δραστηριότητας στην καθημερινή ζωή, είναι ένα αρκετά πολύπλοκο εγχείρημα το οποίο βρίσκει εφαρμογές σε διάφορα πεδία όπως αυτοματοποιημένη παρακολούθηση πλήθους, ανάλυση συμπεριφοράς εντός καταστημάτων, αυτόματη ανάλυση αθλητικών δραστηριοτήτων και άλλα. Κάποιος θα μπορούσε να ορίσει το πρόβλημα ως την ικανότητα ενός συστήματος να κατηγοριοποιήσει αυτόματα την δραστηριότητα που εκτελείται από έναν άνθρωπο, δοσμένης της εικονοσειράς (video) που την περιέχει.

Παρότι το πρόβλημα μπορεί εύκολα να περιγραφεί, η λύση ενός τέτοιου προβλήματος είναι ένα εξαιρετικά δύσκολο έργο το οποίο απαιτεί διαφορετικές προσεγγίσεις σε μια σειρά υπό-προβλημάτων. Η πρόκληση του εγχειρήματος προέρχεται από μια πληθώρα παραγόντων οι οποίοι επηρεάζουν το ποσοστό αναγνώρισης. Η ιδιοσύστατη του κάθε ατόμου είναι ένα μεγάλο θέμα καθώς η ίδια κίνηση μπορεί να εκτελεστεί διαφορετικά από τον καθένα. Πολύπλοκα φόντα, μερικές αποκρύψεις-επικαλύψεις του υποκειμένου, διακυμάνσεις φωτισμού, σταθερότητα κάμερας και γωνία λήψης είναι μόνο μερικά από τα προβλήματα που αυξάνουν την πολυπλοκότητα και δημιουργούν μια σειρά από αναγκαίες προϋποθέσεις.

Η μέθοδος που παρουσιάζουμε στο συγκεκριμένο κεφάλαιο, προτείνει την εξαγωγή απλών χαρακτηριστικών για την αναγνώριση ανθρωπίνων κινήσεων βασισμένη στις γνωστές ιδιότητες του μετασχηματισμού Trace. Πιο συγκεκριμένα στην παρούσα μελέτη, κάνουμε μια πρώτη αποτίμηση του μετασχηματισμού Trace για την αναγνώριση κινήσεων ξεκινώντας από την πιο γνωστή μορφή του, τον μετασχηματισμού Radon. Στην προτεινό-

μενη μέθοδο δημιουργούμε ”όγκους προτύπων” τα επονομαζόμενα και ”History Traces”, τα οποία με τη σειρά τους αναπαριστούν μια περίοδο για κάθε εξεταζόμενη κίνηση. Μηχανές διανυσμάτων υποστήριξης (Support Vector Machines) με πυρήνες Radial Basis Function (RBF) χρησιμοποιούνται για την αξιολόγηση του συστήματος, αναδεικνύοντας μια ιδιαιτέρως ανταγωνιστική επίδοση της τάξεως του 87,7%.

2.1.1 Σχετική βιβλιογραφία

Αν θα έπρεπε να κατηγοριοποιήσουμε την αναγνώριση ανθρώπινης δραστηριότητας σε διαφορετικές υποκατηγορίες, κάποιος θα μπορούσε να το κάνει εξετάζοντας τα χαρακτηριστικά που χρησιμοποιούνται για να αναπαρασταθεί η κάθε κίνηση. Όπως επισημαίνουν οι συγγραφείς στο [113], βάσει των υποκείμενων χαρακτηριστικών διακρίνονται δύο βασικές κατηγορίες. Η πιο επιτυχημένη βασίζεται στα *δυναμικά χαρακτηριστικά* και αποτελεί το ερευνητικό αντικείμενο της πλειοψηφίας των μελετών. Η δεύτερη, βασίζεται στα *χαρακτηριστικά στατικής πόζας* και παρέχει το πλεονέκτημα της εξαγωγής χαρακτηριστικών από ακίνητες εικόνες.

Ένα σύστημα εμπνευσμένο από τα Local Binary Patterns (LBP) που παρουσιάζεται στη [128], δείχνει να είναι ”ελαστικό” στις μεταβολές της υφής συγκρίνοντας τις γειτονικές κηλίδες, ενώ ταυτόχρονα διατηρεί την ικανότητά του στη παρουσία κίνησης. Τεχνικές βασισμένες στα LBP έχουν επίσης προταθεί στο [51] όπου ο χωρο-χρονικός όγκος διαμερίζεται κατά μήκος των τριών αξόνων ((x, y, t) για τη δημιουργία LBP ιστογραμμάτων των xt και yt επιπέδων. Μια άλλη προσέγγιση στο [127] προκειμένου να συλλάβει τοπικά χαρακτηριστικά στην οπτική ροή, υπολογίζει μια μεταβλητή του LBP και αναπαριστά τις κινήσεις ως *τοπικά άτομα* (local atoms). Στο [44] μια άλλη εργασία εμπνευσμένη από τη βιολογία χρησιμοποιεί ”ιεραρχικά ταξινομημένους χώρο-χρονικούς ανιχνευτές χαρακτηριστικών”. Χωρο-χρονικά σημεία ενδιαφέροντος χρησιμοποιούνται για να αναπαραστήσουν και να μάθουν της κλάσεις ανθρωπίνων κινήσεων στο [78]. Βελτίωση αποτελεσμάτων αναφέρεται στα [99], [58] όπου πληροφορία βασισμένη στην οπτική ροή συνδυάζεται με πληροφορία εμφάνισης. Σε μια πιο πρόσφατη μελέτη στο [91], προτείνεται ένας ανιχνευτής χωρο-χρονικών χαρακτηριστικών ο οποίος βασίζεται στον υπολογισμό μοντέλων προεκβολών.

Όπως αναφέρθηκε νωρίτερα, τα απαιτούμενα χαρακτηριστικά για την αναγνώριση ανθρωπίνων κινήσεων, μπορούν να εξαχθούν είτε από βίντεο, είτε από απλές εικόνες που περιγράφουν διαφορετικές στατικές πόζες. Οι μέθοδοι που βασίζονται στις στατικές πόζες, βασίζονται κατά κύριο λόγο σε σιλουέτες και παρόλο που δεν έχουν την ακρίβεια των τεχνικών που βασίζονται σε ακολουθίες, παρουσιάζουν το πλεονέκτημα της εξαγωγής απόφασης από ένα απλό καρέ. Αντιπροσωπευτικά παραδείγματα αυτής της κατηγορίας παρουσιάζονται στις εργασίες [30] και [41]. Πιο συγκεκριμένα στην [41] η κατηγο-

ριοποίηση της συμπεριφοράς επιτυγχάνεται εξάγοντας ιδιο-σχήματα (eigenshapes) από απλές σιλουέτες με την εφαρμογή "ανάλυσης κυρίων συνιστωσών" (Principal Component Analysis). Στο [30] αντίστοιχα, η μοντελοποίηση της πόζας των ατόμων από ένα αποσπασμένο καρέ γίνεται με τη χρήση της μεθόδου bag of rectangles.

Μια άλλη τεχνική στο [134] εμπλέκει υπέρυθρες εικόνες προκειμένου να εξάγει πιο καθαρές πόζες. Ακολούθως, ταξινόμηση επιτυγχάνεται με τη χρήση απλών ποζών βασισμένες σε περιγραφείς ιστογραμμάτων προσανατολισμένης κλίσης (histogram of oriented gradient hog). Ένας τύπος HOG περιγραφέα χρησιμοποιείται επίσης στο [63] σε ένα σύνολο από ήδη εξαχθείσες πόζες που αναπαριστούσαν δραστηριότητες παικτών χόκει. Προκειμένου να διαχειριστούν με καλύτερο τρόπο τις ευκρινείς πόζες και τα ανεπιθύμητα φόντα, οι συγγραφείς στο [113], επεκτείνανε τους περιγραφείς βασισμένους σε HOG και αναπαριστούν τις κλάσεις των κινήσεων με ιστογράμματα "χονδροειδών" ποζών.

Επίσης, σε αντίθεση με άλλες τεχνικές που χρησιμοποιούν πολύπλοκες αναπαραστάσεις δράσεων, οι συγγραφείς του [5] προτείνουν μια μέθοδο η οποία βασίζεται στην εξαγωγή μιας "πόζας κλειδί" από την κάθε ακολουθία. Η μέθοδος επιλέγει την πιο αντιπροσωπευτική και την πιο χαρακτηριστική πόζα από ένα σύνολο υποψηφίων ούτως ώστε να μπορέσει να διαχωρίσει μια κίνηση από μια άλλη.

Μια άλλη κατηγοριοποίηση των προσεγγίσεων που σχετίζονται με την αναγνώριση κινήσεων επιχειρείται από τους συγγραφείς στο [48]. Η διαφορά εδώ μεταξύ των μεθόδων έγκειται στην αναπαράσταση που χρησιμοποιείται από τους συγγραφείς. Η χρονική εξέλιξη των ανθρώπινων σιλουετών έχει συχνά χρησιμοποιηθεί για την περιγραφή των κινήσεων. Για παράδειγμα, στο [8] οι συγγραφείς πρότειναν την αναπαράσταση των κινήσεων με χρονικά πατρόν (tempaltes) με το όνομα "Motion History" και "Motion Energy" αντίστοιχα. Μια επέκταση της δουλειάς αυτής στο [123], εμπνευσμένη από τα MH templates, εισάγει τους Motion History όγκους ως μια αναπαράσταση "ελεύθερης οπτικής γωνίας". Δουλεύοντας προς την ίδια κατεύθυνση, οι συγγραφείς του [111] πρότειναν τους "κυλίνδρους δράσης" (action cylinders), αναπαριστώντας την ακολουθία μιας κίνησης ως ένα γενικευμένο κύλινδρο ενώ στο [129], χωρο-χρονικοί όγκοι παράγονταν βασισμένοι σε μια ακολουθία από δισδιάστατα περιγράμματα σε συνάρτηση με το χρόνο. Αυτά τα περιγράμματα είναι οι δισδιάστατες προβολές των σημείων που βρίσκονται στην εξωτερική πλευρά των ορίων ενός αντικειμένου που εκτελεί μια δράση σε 3 διαστάσεις. Χωρο-χρονικοί όγκοι χρησιμοποιούνται επίσης στα [38], [32] βασισμένοι σε σιλουέτες που εξάγονται προς την κατεύθυνση του χρόνου.

Μια άλλη πρόσφατη κατηγορία τεχνικών, έχοντας επίσης προσανατολισμό στο χωροχρόνο, βασίζεται στην ανάλυση της δομής τοπικών τρισδιάστατων "επιρραμάτων" (3D patches) στο βίντεο που περιέχει την κίνηση [102], [21], [57], [29]. Διάφορα τοπικά χαρακτηριστικά (χωρο-χρονικά ή όχι) έχουν συνδυαστεί με διάφορες τεχνικές της μηχανι-

κής μάθησης. "Κρυφά Μαρκοβιανά Μοντέλα" (Hidden Markov Models) [2], [45], [122], "Εξαρτημένα Τυχαία Πεδία" (Conditional Random Fields) [121], [106], [72] και οι "Μηχανές Διανυσμάτων Υποστήριξης" Support Vector Machines, [100], [58], [33] είναι μόνο μερικές από τις τεχνικές αυτές.

2.2 Σύνοψη του προτεινόμενου συστήματος

Ο πιο κοινός τρόπος για να συλλάβουμε μια ανθρώπινη κίνηση είναι με μια δισδιάστατη (2D) κάμερα. Κατά αυτό τον τρόπο, η κίνηση περιλαμβάνεται σε μια ακολουθία αποτελούμενη από μια σειρά διαφορετικών καρέ. Στη δική μας περίπτωση, δουλεύουμε στη βάση δεδομένων KTH η οποία αποτελείται από ένα μεγάλο αριθμό ακολουθιών βίντεο. Καθώς το φόντο σε όλα τα βίντεο είναι ομαλό, το αφαιρούμε εφαρμόζοντας έναν αλγόριθμο grassfire [36]. Όπως αναφέρθηκε νωρίτερα, η εξαγωγή σιλουετών είναι μια κοινή τεχνική σε αρκετές μελέτες οι οποίες αναφέρονται στην παρακολούθηση της ανθρώπινης δυναμικής [11], [54]. Όπως συνηθίζεται στις περισσότερες προσεγγίσεις αναγνώρισης ανθρώπινης κίνησης, έχουμε δημιουργήσει τα δείγματα ελέγχου και εκπαίδευσης χειροκίνητα (κατατμίζοντας τόσο σε χώρο όσο και σε χρόνο). Έχουμε επίσης στοιχίσει τις παρεχόμενες ακολουθίες ούτως ώστε, κάθε δείγμα μιας κίνησης να αντιπροσωπεύεται από ένα χρονικά διαβαθμισμένο βίντεο το οποίο περιέχει μια περίοδο. Παρότι το φόντο είναι ομοιόμορφο, οι εξαχθείσες σιλουέτες παρουσιάζονται ιδιαίτερα θορυβώδεις καθώς μια σειρά εξωτερικών παραγόντων (π.χ. συνθήκες φωτισμού κ.α.) εξακολουθούν να επηρεάζουν το τελικό αποτέλεσμα. Ωστόσο, χάρη στις ιδιότητες του μετασχηματισμού που χρησιμοποιείται, τα νέα χαρακτηριστικά που προκύπτουν, παρουσιάζονται να είναι εύρωστα στο θόρυβο και δεν απαιτούν προγενέστερο φιλτράρισμα. Κατ' αυτόν τον τρόπο, ένα τελικό πατρόν (template) το οποίο αναπαριστά ολόκληρη την κίνηση δημιουργείται ως αποτέλεσμα της ενσωμάτωσης πολλών δυαδικών μετασχηματισμών. Στη συνέχεια, τα τελικά templates αποτελούν το διάνυσμα το οποίο θα εκπαιδεύσει ισάριθμα με τον αριθμό των κλάσεων RBF kernel SVMs. Η κατηγοριοποίηση αυτής της νέας διάταξης, επιτυγχάνεται μετρώντας την απόσταση των εξεταζόμενων διανυσμάτων από τα διανύσματα υποστήριξης της κάθε κλάσης. Ωστόσο, καθώς το ζητούμενο είναι να αξιολογήσουμε την συνολική απόδοση του συστήματος, υπολογίζουμε τον συνολικό αριθμό των επιτυχημένων κατηγοριοποιήσεων για κάθε διάνυσμα που περνάει από το κάθε εκπαιδευμένο SVM αντίστοιχα. Για την εξέταση, ακολουθήσαμε το πρωτόκολλο «leave-one-person-out». Περισσότερες πληροφορίες για την πειραματική διαδικασία παρέχονται στη σχετική ενότητα 2.3

2.2.1 Constructing History Trace Templates

Έχει αποδειχθεί [49] ότι τα ολοκληρώματα κατά μήκος ευθειών γραμμών μιας 2D μεταβλητής μπορεί να την αναπαραστήσει πλήρως. Η αναπαράσταση που χρησιμοποιούμε στη συγκεκριμένη εργασία, είναι ουσιαστικά μια υπο-περίπτωση του Trace, ο γνωστός και ως μετασχηματισμός Radon ο οποίος έχει βρει χρησιμότητα σε μια πληθώρα σημαντικών εφαρμογών από τη μηχανοργανωμένη τομογραφία ως και την αναγνώριση βηματισμού [11].

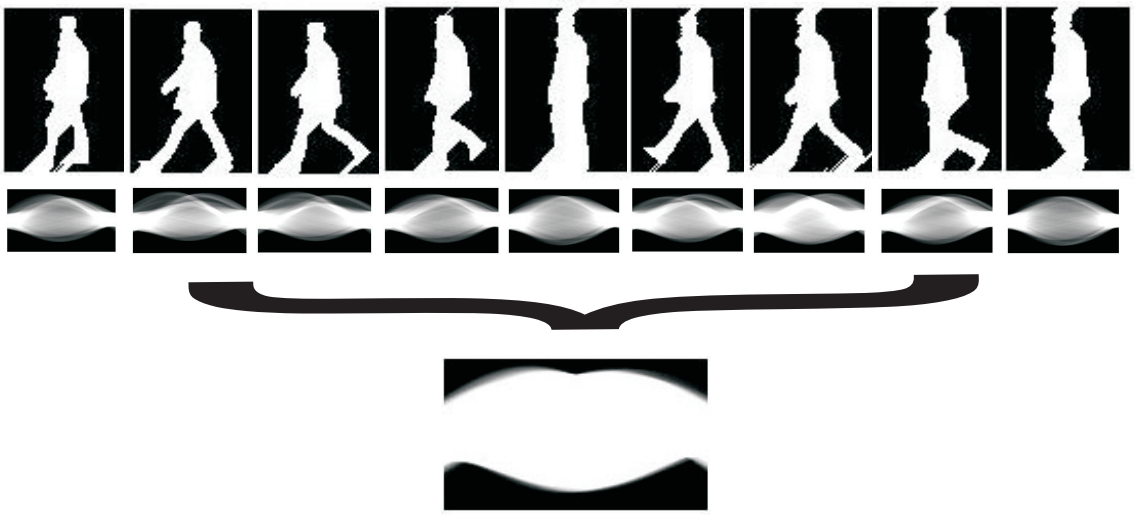
Έστω $f(x, y)$ είναι μια 2D συνάρτηση στον Ευκλείδειο χώρο $\Re^2$ η οποία έχει αποσπαστεί από μια βίντεο-ακολουθία και περιέχει μια σιλουέτα δυαδικής μορφής. Ο μετασχηματισμός Radon R_f είναι μια συνάρτηση που ορίζεται επάνω στο χώρο των ευθειών L στο $\Re^2$ από το ολοκλήρωμα κατά μήκος κάθε τέτοιας ευθείας και δίνεται από:

$$R_f(p, \theta) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\delta(p - x\cos\theta - y\sin\theta)dxdy \tag{2.1}$$

όπου $R_f(p, \theta)$ είναι το ολοκλήρωμα μιας γραμμής μιας εικόνας από $-\infty$ στο ∞. p και θ είναι οι παράμετροι που ορίζουν τη θέση της γραμμής. Έτσι, $R_f(p, \theta)$ είναι το αποτέλεσμα του ολοκληρώματος του f επάνω στη γραμμή $p = x\cos\theta + y\sin\theta$. Στην περίπτωσή μας ως σημείο αναφοράς, ορίζεται το κέντρο της σιλουέτας. Έτσι, καθώς οι ανθρώπινες δραστηριότητες είναι ουσιαστικά χωρο-χρονικοί όγκοι, στόχος είναι η αναπαράσταση όσο το δυνατόν περισσότερης δυναμικής και δομικής πληροφορίας της κίνησης είναι δυνατόν. Σε αυτό το σημείο, ο Radon παρουσιάζει μεγάλη καταλληλότητα για το συγκεκριμένο σκοπό. Μετασχηματίζει εικόνες δύο διαστάσεων σε ένα πεδίο πιθανών παραμέτρων γραμμών, όπου κάθε γραμμή μέσα στην εικόνα θα είναι μια κορυφή τοποθετημένη στις αντίστοιχες παραμέτρους των ευθειών.

Όταν ο Radon υπολογίζεται έχοντας ως αναφορά το κέντρο της σιλουέτας, συγκεκριμένοι συντελεστές θα έχουν "αιχμαλωτίσει" πολλή από την ενέργεια της σιλουέτας. Αυτοί οι συντελεστές θα διαφέρουν κατά τη χρονική διάρκεια της κίνησης και θα παρέχουν σημαντικές διαφορές από μια κίνηση στην άλλη για το ίδιο χρονικό πλαίσιο.

Εκτός της δομικής πληροφορίας, προκειμένου να συλλάβουμε και τη χρονική πληροφορία που εμπεριέχεται σε μια κίνηση, προτείνουμε τη δημιουργία ενός ($History\ Trace\ Template$.) Αυτό το πατρόν είναι ουσιαστικά ένας συνεχής μετασχηματισμός στη χρονική κατεύθυνση μιας ακολουθίας. Έστω $f(p, \vartheta, t)$ είναι η ακολουθία μιας ανθρώπινης κίνησης. Αν $\breve{g}_n(p, \theta)$ είναι ο μετασχηματισμός Radon μιας σιλουέτας $s_n(p, \theta)$, για το n καρέ όπου $n = 1...N$, τότε το History Trace Template για τη συγκεκριμένη ακολουθία θα δίνεται από:



Σχήμα 2.1: *Εξαχθείσες σιλουέτες και μετασχηματισμοί Radon για μια περίοδο περπατήματος. Το History Trace template φαίνεται στο κάτω μέρος.*

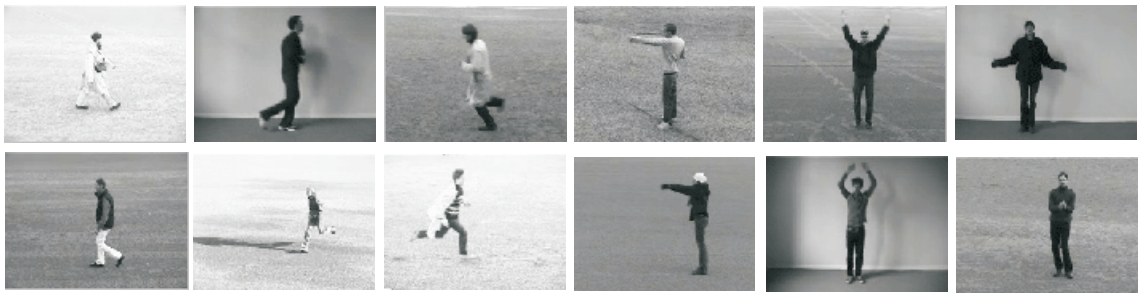
$$T_N(p,\theta) = \sum_{n=1}^{N} \check{g}_n(p,\theta). \tag{2.2}$$

Κατ' αυτό τον τρόπο, τα εξαχθέντα χαρακτηριστικά θα είναι μια συνάρτηση πολλαπλών σημαντικών διακρίσεων που περιέχονται σε μια σειρά μετασχηματισμών που έχουν παραχθεί για μια περίοδο της κίνησης αντίστοιχα. Όπως προαναφέρθηκε, στην εργασία μας όλες οι περίοδοι των κινήσεων έχουν χρονικά στοιχηθεί στον ίδιο αριθμό καρέ N. Το Σχήμα 2.1 δείχνει τους μετασχηματισμούς που έχουν εξαχθεί για μια περίοδο της κίνησης "περπάτημα". Το τελικό History Trace template φαίνεται στο κάτω μέρος της εικόνας.

2.3 Πειραματικά αποτελέσματα

Σε αυτή την ενότητα, θα παρουσιάσουμε τα πειραματικά αποτελέσματα με σκοπό να παρουσιάσουμε την αποτελεσματικότητα της προτεινόμενης μεθόδου για την αναγνώριση ανθρωπίνων κινήσεων. Για την αξιολόγηση της απόδοσης χρησιμοποιούμε την προσέγγιση "leave-one-person-out cross-validation". Τα πειράματα διεξάχθηκαν σε έναν υπολογιστή με επεξεργαστή Intel Core i5 (650@3,2 GHz) με 4GB RAM. Για την αξιολόγηση χρησιμοποιήθηκε η βάση δράσεων KTH [100]. Δείγματα από το συγκεκριμένο σύνολο δεδομένων για διαφορετικούς τύπους κινήσεων παρουσιάζονται στο Σχήμα 2.2.

Η συγκεκριμένη βάση δεδομένων περιέχει 6 τύπους ανθρώπινων δράσεων (walking, jogging, running, boxing, hand waving and hand clapping) εκτελεσμένες αρκετές φορές από 25 διαφορετικά άτομα σε τέσσερα διαφορετικά σενάρια, σε διαφορετικές συνθήκες



Σχήμα 2.2: *Δείγματα δράσεων από το σύνολο δεδομένων KTH για τις κινήσεις walking, jogging, running, boxing, hand waving και hand clapping αντίστοιχα.*

φωτισμού: εξωτερικοί χώροι, εξωτερικοί χώροι με διαβαθμίσεις μεγέθους, (εστίαση και αποεστίαση του υποκειμένου), εξωτερικοί χώροι με διαφορετικό ρουχισμό και εσωτερικοί χώροι. Η βάση περιέχει 600 ακολουθίες. Όλα τα βίντεο έχουν ληφθεί έχοντας ομοιογενές φόντο με στατική κάμερα και ταχύτητα λήψης 25 καρέ το δευτερόλεπτο.

Στα πειράματα που ακολουθούν, τα βίντεο έχουν περικοπεί στη ανάλυση των 160*120 εικονοστοιχείων και έχουν διάρκεια κατά μέσο όρο τέσσερα δεύτερα. Τα δείγματα εκπαίδευσης δημιουργήθηκαν κατατμίζοντας χειροκίνητα (τόσο χρονικά όσο και χωρικά) και στοιχίζοντας τις παρεχόμενες ακολουθίες. Το φόντο αφαιρέθηκε με τη χρήση ενός αλγορίθμου ”grassfire” [36]. Η προσέγγιση ”leave-one-person-out cross-validation” χρησιμοποιήθηκε για να αξιολογηθεί η γενική απόδοση των κατηγοριοποιητών για το πρόβλημα της αναγνώρισης κίνησης.

Σε αυτό το σημείο, θα θέλαμε να τονίσουμε ότι το πρόβλημα της αναγνώρισης δράσεων είναι πρόβλημα πολλών κλάσεων. Για να διαχειριστούμε αυτή την πολυπλοκότητα, ανασκευάζουμε το πρόβλημα σε μια γενίκευση δυαδικών κατηγοριοποιήσεων. Πιο συγκεκριμένα, εκπαιδεύσαμε 6 διαφορετικά SVMs (ένα για κάθε κλάση) χρησιμοποιώντας ένα πρωτόκολλο ”ένας-εναντίον-όλων ”. Η τελική απόφαση λαμβάνεται κατηγοριοποιώντας κάθε εξεταζόμενο δείγμα σε μια κλάση $\mathcal{C}_a$, βάσει της απόστασης d του εξεταζόμενου διανύσματος από τα διανύσματα υποστήριξης. Όπου $\mathcal{C}_a$ είναι ένα σύνολο από templates καταχωρημένα σε μια κλάση δράσης (π.χ. boxing). Ωστόσο, καθώς θέλουμε να επιτύχουμε μια γενικευμένη αξιολόγηση του συστήματος, μετρήσαμε τις επιτυχείς δυαδικές κατηγοριοποιήσεις για κάθε δείγμα, που εξετάζεται σε καθένα από τα 6 διαφορετικά εκπαιδευμένα SVMs. Με αυτόν τον τρόπο καταφέραμε να επιτύχουμε 25*6*24=3600 κατηγοριοποιήσεις αντί 600 (άτομα*δράσεις*δείγματα για κάθε άτομο). Τα αποτελέσματα ανέδειξαν μια πολύ ανταγωνιστική απόδοση της τάξης του 87.7%. Αυτό καθίσταται ακόμη πιο ενδιαφέρον κοιτάζοντας τους χρόνους εκτέλεσης της μεθόδου. Για 25 επαναλήψεις (εξέταση όλων των ατόμων), συμπεριλαμβανομένης της εκπαίδευσης, απαιτήθηκαν 6 λεπτά ενώ ο καθαρός χρόνος εξέτασης για κάθε δείγμα ήταν 0.01 δευτερόλεπτα. Τα ποσοστά κα-

Πίνακας 2.1: *Ποσοστά κατηγοριοποίησης (%) για τους διαφορετικούς τύπους δράσεων της βάσης KTH.*

Τύπος Δράσης	Boxing	Handclapping	Handwaving	Jogging	Walking	Running	Overall
Κατηγοριοποίηση	92.2	90.0	85.6	84.3	88.1	86.0	87.7

τηγοριοποίησης για την κάθε κλάση που χρησιμοποιήθηκε στην πειραματική διαδικασία δίνεται στον Πίνακα 2.1.

Κεφάλαιο 3

Ερευνώντας τον μετασχηματισμό Trace για την εύρωστη αναγνώριση ανθρωπίνων κινήσεων

3.1 Εισαγωγή

Η δουλειά που εισάγεται σε αυτή την εργασία είναι ουσιαστικά η επέκταση της εργασίας που παρουσιάστηκε στο [33] και στο Κεφάλαιο 2. Στη συγκεκριμένη μελέτη, ο μετασχηματισμός Radon προτάθηκε για την εξαγωγή χαρακτηριστικών ικανών να αναπαραστήσουν μια κίνηση με τη μορφή ενός template. Ο Radon που στην πράξη είναι μια υπο-περίπτωση του μετασχηματισμού Trace, έχει βρει μια πληθώρα από σημαντικές εφαρμογές από την μηχανοποιημένη τομογραφία ως την αναγνώριση βάδισης [11].

Σε αυτή τη μελέτη, δημιουργούμε χαρακτηριστικά εξετάζοντας τη δυνατότητα του μετασχηματισμού Trace για την αναγνώριση ανθρώπινης δραστηριότητας. Στο πρώτο στάδιο, χρησιμοποιούμε διάφορα συναρτησιακά για τη δημιουργία του Trace τα οποία προσδίδουν στο τελικό template διαφορετικές ιδιότητες και τα οποία ονομάσαμε *History Trace Template* (HTT). Πιο αναλυτικά, εξετάζουμε διάφορα συναρτησιακά του Trace με σκοπό να δημιουργήσουμε template όγκους που ο καθένας θα αναπαριστά μια περίοδο μιας κίνησης. Radial Basis Function (RBF) SVMs που χρησιμοποιήθηκαν για την αξιολόγηση της μεθόδου, έδειξαν μια πολύ ανταγωνιστική απόδοση του συστήματος επιτυγχάνοντας 90.22% για τη βάση δεδομένων KTH και 93.4% για τη Weizmann αντίστοιχα.

Στο δεύτερο στάδιο, επεκτείνουμε περαιτέρω τη μέθοδο εισάγοντας μια νέα τεχνική για την εξαγωγή χαρακτηριστικών για την παραγωγή ανεπηρέαστων σε διάφορες μεταβολές χαρακτηριστικών, όπως η περιστροφή, η μετατόπιση και η διαβάθμιση. Πιο συγκεκριμένα, για κάθε καρέ της κινησιο-ακολουθίας, υπολογίζεται ένα σύνολο από μετασχηματισμούς Trace. Υπολογίζοντας διάφορα συναρτησιακά με συγκεκριμένες ιδιότητες στους

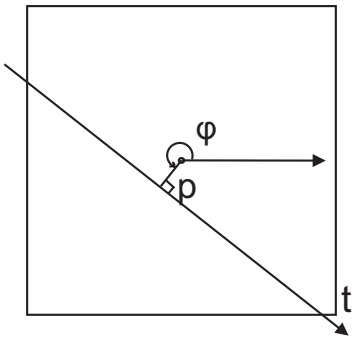
συγκεκριμένους μετασχηματισμούς, εξάγεται ένα σύνολο από ανεπηρέαστα "τριπλά χαρακτηριστικά" (Triple Features). Η κίνηση τελικά αναπαρίσταται από ένα διάνυσμα χαμηλής διάστασης που ονομάζουμε "Ιστορικό Τριπλό Χαρακτηριστικό" *History Triple Features* (HTFs) και περιέχει τα πιο διακριτά από ένα σύνολο εξαχθέντων χαρακτηριστικών σε κάθε καρέ. Πειράματα κατηγοριοποίησης με τη χρήση των προαναφερθέντων βάσεων δεδομένων, παρουσίασαν μια ακόμη πιο βελτιωμένη απόδοση της τάξης του 93.14% και 95.42% αντίστοιχα, αναδεικνύοντας τις δυνατότητες της προτεινόμενης μεθόδου.

Με την καλύτερη γνώση που μπορεί να έχει ο συγγραφέας, αυτή είναι η πρώτη φορά που ο μετασχηματισμός Trace, κάποια από της εκφάνσεις του ή τα παράγωγά του χρησιμοποιείται για εξαγωγή χαρακτηριστικών για την αναγνώριση ανθρωπίνων κινήσεων.

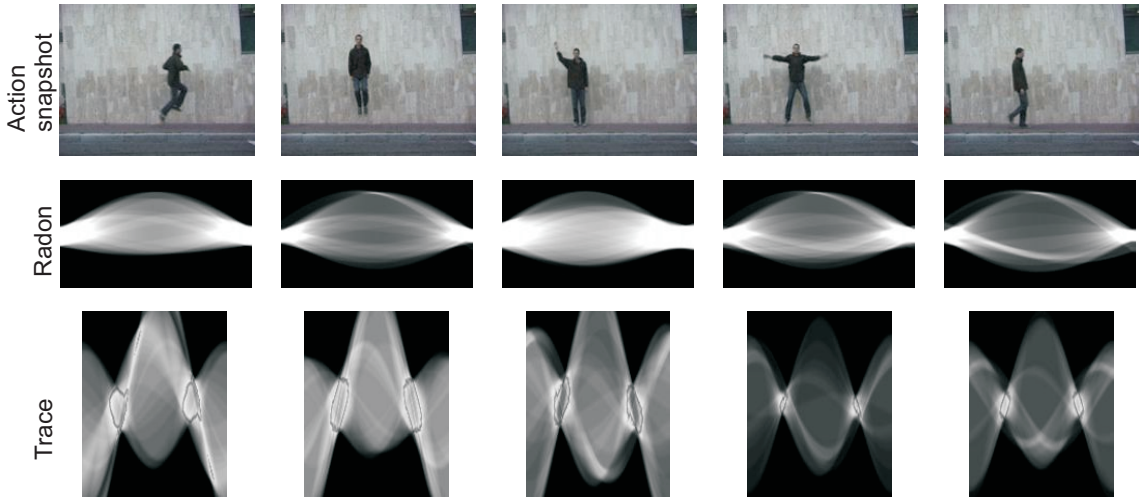
3.2 Μετασχηματισμός Trace

Όπως αναφέρθηκε και στο Κεφάλαιο 2, ο μετασχηματισμός Trace είναι μια γενίκευση του μετασχηματισμού Radon [20] ενώ την ίδια στιγμή, ο Radon αποτελεί υπο-περίπτωσή του. Ενώ ο μετασχηματισμός Radon μιας εικόνας είναι η δισδιάστατη αναπαράσταση της εικόνας με συντεταγμένες ϕ και p με την τιμή του ολοκληρώματος της εικόνας να υπολογίζεται κατά μήκος της αντίστοιχης γραμμής, τοποθετημένο σε κελί (ϕ, p), ο Trace υπολογίζει το συναρτησιακό T επάνω στην παράμετρο t κατά μήκος της γραμμής, το οποίο δεν είναι απαραίτητα το ολοκλήρωμα. Ο μετασχηματισμός Trace δημιουργείται διατρέχοντας μια εικόνα με ευθείες γραμμές όπου συγκεκριμένα συναρτησιακά της συνάρτησης (εικόνας) υπολογίζονται. Από την ίδια εικόνα μπορούν να δημιουργηθούν διάφοροι μετασχηματισμοί με διαφορετικές ιδιότητες. Ο μετασχηματισμός που δημιουργείται, είναι στην πράξη μια δισδιάστατη συνάρτηση των παραμέτρων της κάθε γραμμής ίχνους. Ορισμός των παραπάνω παραμέτρων για μια διατρέχουσα γραμμή δίνεται στο Σχήμα 3.1. Παραδείγματα των Radon και Trace μετασχηματισμών για στιγμιότυπα διαφόρων κινήσεων δίνονται στο Σχήμα 3.2. Στη συνέχεια, θα δώσουμε την περιγραφή της διαδικασίας εξαγωγής χαρακτηριστικών για το μετασχηματισμό Trace βασισμένο στη θεωρία που παρέχεται στο [49].

Για την καλύτερη κατανόηση του συγκεκριμένου μετασχηματισμού, ας υποθέσουμε ένα γραμμικά παραμορφωμένο αντικείμενο (π.χ. περιστροφή, μετατόπιση, διαβάθμιση). Θα μπορούσαμε να πούμε ότι το αντικείμενο απλά παρατηρείται σε ένα άλλο σύστημα συντεταγμένων επίσης γραμμικά παραμορφωμένο. Αυτό μπορεί να εξηγηθεί ευκολότερα αν ονομάσουμε το αρχικό σύστημα συντεταγμένων της εικόνας C_1 και το νέο παραμορφωμένο C_2. Έστω ότι το παραμορφωμένο σύστημα μπορεί να αποκτηθεί περιστρέφοντας το C_1 κατά γωνία $-\theta$, μεταβάλλοντας τους άξονες βάσει της παραμέτρου v και μετατοπίζοντας με διάνυσμα $(-s_0 \cos\psi_0, -s_0 \sin\psi_0)$. Έστω ότι υπάρχει ένα 2D αντικείμενο F το



Σχήμα 3.1: *Ορισμός των παραμέτρων μιας ιχνο-γραμμής.*



Σχήμα 3.2: *Δείγματα των Radon και Trace μετασχηματισμών που έχουν δημιουργηθεί από σιλουέτες διαφόρων στιγμιοτύπων διαφόρων κινήσεων από τη βάση δεδομένων Weizmann.*

οποίο παρακολουθείται από το C_1 με $F_1(x, y)$ και από το C_2 με $F_2(\tilde{x}, \tilde{y})$. Το $F_2(\tilde{x}, \tilde{y})$ μπορεί να θεωρηθεί ως μια εικόνα πού έχει δημιουργηθεί από την $F_1(x, y)$ περιστρέφοντάς την κατά θ, κλιμακώνοντάς τη κατά v^{-1} και μετατοπίζοντάς την κατά $(s_0 \cos\psi_0, s_0 \sin\psi_0)$.

Μια γραμμικά μετασχηματισμένη εικόνα στην ουσία μεταφέρεται κατά μήκος των γραμμών ενός άλλου συστήματος συντεταγμένων, καθώς οι ευθείες γραμμές στο νέο σύστημα συντεταγμένων εμφανίζονται επίσης ως ευθείες γραμμές. Οι παράμετροι μια ευθείας στο C_2 παραμετροποιημένοι κατά (ϕ, p, t) στο παλιό σύστημα C_1, είναι:

$$\phi_{old} = \phi - \theta \tag{3.1}$$

$$p_{old} = v[p - s_0 \cos(\psi_0 - \phi)] \tag{3.2}$$

$$t_{old} = v[t - s_0 \sin(\psi_0 - \phi)]. \tag{3.3}$$

Ας ορίσουμε ως Λ ένα σύνολο από γραμμές που σαρώνουν μια εικόνα προς όλες τις κατευθύνσεις. Ο μετασχηματισμός Trace είναι μια συνάρτηση g ορισμένη στο συγκεκριμένο σύνολο με τη βοήθεια ενός συναρτησιακού T της συνάρτησης (εικόνας), όταν αυτή εξετάζεται ως συνάρτηση της μεταβλητής t. Το συναρτησιακό T καλείται *Trace functional*. Εάν $L(C_1; \phi, p, t)$ είναι μια γραμμή στο σύστημα συντεταγμένων C_1, τότε

$$g(F; C_1; \phi, p) = T(F(C_1; \phi, p, t)), \tag{3.4}$$

όπου $F(C_1; \phi, p, t)$ δηλώνει τις τιμές της συνάρτησης κατά μήκος μια επιλεγμένης γραμμής. Παίρνοντας αυτό το συναρτησιακό, η μεταβλητή t εξαλείφεται. Αυτό έχει σαν αποτέλεσμα μια συνάρτηση δύο διαστάσεων με μεταβλητές ϕ και p. Η καινούργια συνάρτηση είναι επίσης εικόνα ορισμένη στο Λ.

Όπως περιγράφεται και στο [49], χρησιμοποιώντας δύο επιπλέον συναρτησιακά δηλωμένα με τα γράμματα P and Φ, μπορεί να οριστεί ένα "Τριπλό χαρακτηριστικό" (*triple feature*). Όπου το P καλείται διαμετρικό και το Φ καλείται κυκλικό συναρτησιακό αντίστοιχα. Το P είναι ένα συναρτησιακό της συνάρτησης του Trace, όταν εκλαμβάνεται ως συναρτησιακό που λειτουργεί στη μεταβλητή της κατεύθυνσης αφού οι δύο προηγούμενες λειτουργίες έχουν ολοκληρωθεί. Έτσι, το τριπλό χαρακτηριστικό Π ορίζεται ως:

$$\Pi(F, C_1) = \Phi(P(T(F(C_1; \phi, p, t)))). \tag{3.5}$$

Σε αυτό το σημείο, πρέπει να επιλεχθούν τα τρία συναρτησιακά. Στη συνέχεια τα συναρτησιακά *αμετάβλητο* και *ευαίσθητο* στη μετατόπιση που θα χρησιμοποιούνται, για λόγους συντομίας θα αναφέρονται "αμετάβλητο" και "ευαίσθητο" αντίστοιχα. Ένα συναρτησιακό Ξ μιας συνάρτησης $\xi(x)$ είναι *αμετάβλητο* εάν

$$\Xi(\xi(x+b)) = \Xi(\xi(x)), \qquad \forall b \in \Re \quad (I_1).$$

Ένα αμετάβλητο συναρτησιακό θα πρέπει να χαρακτηρίζεται από τις ακόλουθες ιδιότητες

• Κλιμακώνοντας την ανεξάρτητη μεταβλητή κατά α, κλιμακώνει το αποτέλεσμα κατά ένα συντελεστή, $a(\alpha)$:

$$\Xi(\xi(\alpha x)) = a(\alpha)\Xi(\xi(x)), \qquad \forall \alpha > 0 \quad (i_i).$$

• Κλιμακώνοντας τη συνάρτηση κατά c κλιμακώνει το αποτέλεσμα κατά ένα συντελεστή, $\gamma(c)$:

$$\Xi(c\xi(x)) = \gamma(c)\Xi(\xi(x)), \qquad \forall c > 0 \quad (i_2).$$

Έχει αποδειχθεί ότι μπορούμε να γράψουμε:

$$a(\alpha) = \alpha^{k_\Xi} \quad and \quad \gamma(c) = c^{\lambda_\Xi}, \tag{3.6}$$

όπου οι παράμετροι k_Ξ και λ_Ξ χαρακτηρίζουν το συναρτησιακό Ξ.

Απαιτούνται συναρτησιακά με τις ακόλουθες ιδιότητες: εφαρμοζόμενο σε μια 2π περιοδική συνάρτηση u, το παραγόμενο αποτέλεσμα θα πρέπει να είναι το ίδιο με αυτό που θα παραγόταν εάν το συναρτησιακό εφαρμοζόταν στην αρχική συνάρτηση μείον την πρώτη αρμονική της $u^{(1)}$, που ορίζεται από $u^{\perp} \equiv u - u^{(1)}$:

$$Z(u) = Z(u^{(\perp)}) \quad (si_1).$$

Ένα συναρτησιακό Z καλείται *ευαίσθητο* εάν

$$Z(\zeta(x+b)) = Z(\zeta(x)) - b, \qquad \forall b \in \Re \quad (S_1).$$

Ένα ευαίσθητο συναρτησιακό μιας περιοδικής συνάρτησης ορίζεται ως εξής: Έστω r είναι η περίοδος της συνάρτησης στην οποία ορίζεται το Z. Μια συνάρτηση αποκαλείται *r-ευαίσθητη* εάν:

$$Z(\zeta(x+b)) = Z(\zeta(x)) - b_{(mod\ r)}, \quad \forall b \in \Re \quad (S_2).$$

Οι επόμενες ιδιότητες πρέπει επίσης να ισχύουν σε ένα ευαίσθητο συναρτησιακό:

• Κλιμακώνοντας την ανεξάρτητη μεταβλητή κλιμακώνεται το αποτέλεσμα ανάστροφα.

$$Z(\zeta(\alpha x)) = \frac{1}{\alpha} Z(\zeta(x)), \qquad \forall \alpha > 0 \quad (s_1)$$

Συνδυασμός των παραπάνω με το (S_1), έχει ως αποτέλεσμα:

$$Z(\zeta(\alpha(x+b))) = \frac{1}{\alpha} Z(\zeta(x)) - b \quad (s_{11})$$

και

$$Z(\zeta(\alpha x + b)) = \frac{1}{\alpha} Z(\zeta(x)) - \frac{b}{\alpha} \quad (s_{12}).$$

• Κλιμάκωση της συνάρτησης δεν επηρεάζει το αποτέλεσμα:

$$Z(c\zeta(x)) = Z(\zeta(x)), \qquad \forall c > 0 \quad (s_2).$$

3.2.1 Κατασκευή αμετάβλητων χαρακτηριστικών

Έστω ότι τα συναρτησιακά T, P και Φ έχουν επιλεγεί ούτως ώστε να είναι αμετάβλητα με το T να υπακούει στην ιδιότητα (i_1), το P στις ιδιότητες (i_1) και (i_2) και το Φ να υπακούει στην ιδιότητα (i_2).

Ο τρόπος με το οποίον η γραμμική διαστρέβλωση της εικόνας επηρεάζει την τιμή του τριπλού χαρακτηριστικού παρουσιάζεται στη συνέχεια. Μπορεί να παρατηρηθεί ότι το τριπλό χαρακτηριστικό της διαστρεβλωμένης εικόνας δίνεται από:

$$\Pi(F, C_2) = \Phi(P(T(F(C_1; \phi_{old}, p_{old}, t_{old})))). \tag{3.7}$$

Αν από την (3.1) αντικαταστήσουμε το (3.2) και το (3.3), προκύπτει

$$\Pi(F, C_2) = \Phi(P(T(F(C_1; \phi - \theta, v[p - s_0 \cos(\psi_0 - \phi)], v[t - s_0 \sin(\psi_0 \sin(\psi_0 - \phi)]))))). \tag{3.8}$$

Εξαιτίας της αμεταβλητότητας του T και της ιδιότητας του (i_1) αυτό μπορεί να γραφεί ως:

$$\Pi(F, C_2) = \Phi(P(\alpha_T(v)T(F(C_1; \phi - \theta, v[p - s_0 \cos(\psi_0 - \phi)], t)))). \tag{3.9}$$

Εξαιτίας της ιδιότητας (i_2) που ακολουθείται από το P, αυτό γράφεται:

$$\Pi(F, C_2) = \Phi(\gamma_p(\alpha_T(v))P(T(F(C_1; \phi - \theta, v[p - s_0 \cos(\psi_0 - \phi)], t)))). \tag{3.10}$$

Από την ιδιότητα (i_1) που ακολουθείται από το το P και της αμεταβλητότητάς του, προκύπτει:

$$\Pi(F, C_2) = \Phi(\gamma_p(\alpha_T(v))\alpha_p(v)P(T(F(C_1; \phi - \theta, p, t)))). \tag{3.11}$$

Εάν το Φ είναι αμετάβλητο και υπακούει στην ιδιότητα του P τότε:

$$\Pi(F, C_2) = \gamma_\Phi(\gamma_P(\alpha_T(v)\alpha_p(v)))\Phi(P(T(F(C_1; \phi, p, t))))). \tag{3.12}$$

Αυτή η συνθήκη μπορεί να εκφραστεί σε αντιστοιχία με τους εκθέτες των συναρτησιακών κ και λ όπου προκύπτει:

$$\Pi(F, C_2) = v^{\lambda_\phi(\kappa_T\lambda_p + \kappa_p)}\Pi(F, C_1). \tag{3.13}$$

Έτσι, η αμεταβλητότητα θα πρέπει να ακολουθείται από την προφανή συνθήκη:

$$\lambda_\phi(\kappa_T\lambda_p + \kappa_p) = 0 \tag{3.14}$$

Αυτή η συνθήκη δεν είναι απαραίτητη εάν δεν υπάρχει διαφορά στην κλιμάκωση μεταξύ των αντικειμένων τα οποία συγκρίνονται, ενώ μπορούν να χρησιμοποιηθούν οποιαδήποτε αμετάβλητα συναρτησιακά τα οποία υπακούουν τις απαραίτητες ιδιότητες.

Επιλέγοντας το συναρτησιακό T, το Φ να είναι αμετάβλητο και το συναρτησιακό P να είναι ευαίσθητο και να υπακούει στην ιδιότητα (s_{11}), το Φ θα υπακούει επίσης και στην ιδιότητα si_1. Έτσι το (3.10) δεν προκύπτει από το (3.9). Αντιθέτως, θα μπορούσαμε να εφαρμόσουμε την ιδιότητα (s_{11}) του P, η οποία έχει ως αποτέλεσμα:

$$\Pi(F, C_2) = \Phi(\frac{1}{v}P(T(F(C_1; \phi, p, t))) + s_0 \cos(\psi_0 - \phi)). \tag{3.15}$$

Εξαιτίας της ιδιότητας si_1 του Φ, παίρνουμε ότι:

$$\Pi(F, C_2) = \gamma_\Phi(\frac{1}{v})\Phi(P(T(F(C_1; \phi, p, t)))) \tag{3.16}$$

ή ισοδύναμα,

$$\Pi(F, C_2) = v^{-\lambda_\Phi}\Pi(F, C_1). \tag{3.17}$$

Επιλέγοντας το Φ έτσι ώστε

$$\lambda_\Phi = 0, \tag{3.18}$$

μπορεί να γίνει αντιληπτό ότι το υπολογιζόμενο τριπλό χαρακτηριστικό είναι και πάλι αμετάβλητο στην περιστροφή, τη μετατόπιση και την κλιμάκωση.

Εν τούτοις, οι συνθήκες (3.14) και (3.18) είναι πολύ περιοριστικές. Η σχέση μεταξύ των τριπλών χαρακτηριστικών που υπολογίστηκαν στις δυο περιπτώσεις μπορεί να γενικευθεί από

$$\Pi(F, C_2) = v^{-\omega}\Pi(F, C_1), \tag{3.19}$$

για τη (3.13), $\omega \equiv -\lambda_\Phi(\kappa_T\lambda_P + \kappa_P)$, ενώ για τη (3.17), $\omega \equiv \lambda_\Phi$. Καθώς μπορούμε να επιλέξουμε τον τύπο του συναρτησιακού που θα δημιουργήσουμε, επιλέγουμε το ω να είναι γνωστό. Έτσι, κάθε τριπλό χαρακτηριστικό που υπολογίζεται μπορεί να κανονικοποιηθεί.

$$\Pi_{norm}(F, C_1) = \sqrt[\omega]{|\Pi(F, C_1)|sign(\Pi(F, C_1))}, \tag{3.20}$$

Ενώ η (3.19) μπορεί να απλοποιηθεί στο:

$$\Pi(F, C_2) = v^{-1}\Pi_{norm}(F, C_1), \tag{3.21}$$

Διαιρώντας δύο τριπλά χαρακτηριστικά τα οποία έχουν δημιουργηθεί κατά αυτόν τον τρόπο, μπορεί να παραχθεί ένα αμετάβλητο.

3.3 Επισκόπηση του προτεινόμενου συστήματος

Ο πιο συνηθισμένος τρόπος να συλλάβουμε μια ανθρώπινη κίνηση είναι με τη χρήση μιας κοινής 2D κάμερας. Με αυτό τον τρόπο η κίνηση εμπεριέχεται σε μια βιντεοακολουθία αποτελούμενη από μια σειρά διαφορετικών καρέ. Στο δικό μας σχήμα, έχουμε δουλέψει επάνω στις βάσεις δεδομένων KTH [100] και Weizmann [6]. Και οι δύο περιέχουν

ένα μεγάλο αριθμό από βίντεο με ανθρώπινες δραστηριότητες, ενώ παράλληλα έχουνε ευρέως χρησιμοποιηθεί για την αξιολόγηση σχετικών αλγορίθμων. Όπως συνηθίζεται στις περισσότερες μεθόδους που αφορούν στην αναγνώριση ανθρώπινης κίνησης, κατασκευάσαμε τα δείγματα εκπαίδευσης και εξέτασης χειροκίνητα κατατμίζοντάς τα (τόσο χωρικά όσο και χρονικά). Έχουμε επίσης στοιχίσει τις παρεχόμενες ακολουθίες ούτως ώστε, κάθε δείγμα να αντιπροσωπεύεται από ένα χρονο-κλιμακούμενο βίντεο το οποίο περιέχει μια περίοδο.

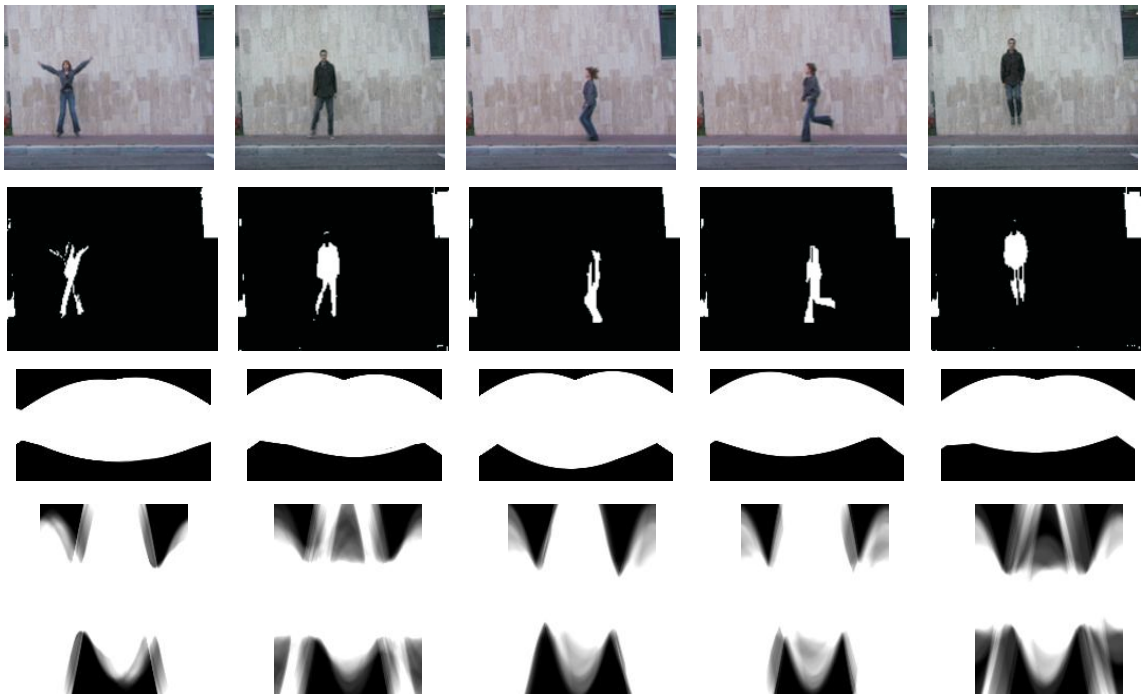
3.3.1 History Trace Templates (HTTs)

Παρόλο που το φόντο είναι ομοιόμορφο, οι εξαχθείσες σιλουέτες παρουσιάζονται θορυβώδεις καθώς ένα σύνολο εξωτερικών παραγόντων (όπως συνθήκες φωτισμού, επικαλύψεις κ.α.), εξακολουθεί να επηρεάζει σημαντικά το αποτέλεσμα. Προκειμένου να αναδείξουμε τις δυνατότητες της προτεινόμενης μεθόδου, δεν χρησιμοποιούμε κάποιο εξελιγμένο αλγόριθμο για την εξαγωγή των σιλουετών όπως επίσης δεν εφαρμόζουμε κάποιου είδους προκαταρκτικό φίλτρο. Παρ' όλα αυτά, εξαιτίας των ιδιοτήτων του μετασχηματισμού Trace, τα νέα χαρακτηριστικά τα οποία δημιουργούνται παρουσιάζονται ιδιαιτέρως εύρωστα στο θόρυβο. Έτσι, για κάθε σιλουέτα, δημιουργείται ένας Trace μετασχηματισμός. Ένα τελικό template το οποίο ονομάστηκε History Trace Template (HTT) και το οποίο τελικά αναπαριστά ολόκληρη την κίνηση, δημιουργείται ως αποτέλεσμα πολλαπλών ενσωματώσεων σε αυτό.

Στη συνέχεια, τα τελικά templates αποτελούν τα διανύσματα τα οποία θα εκπαιδεύσουν ίδιο αριθμό κλάσεων, SVMs RBF πυρήνα. Παραδείγματα από εξαχθείσες σιλουέτες από καρέ διαφόρων κινήσεων και τα παραγόμενα HTTs για τα συγκεκριμένα βίντεο, παρουσιάζονται στο Σχήμα 3.3. Η κατηγοριοποίηση επιτυγχάνεται μετρώντας την απόσταση των εξεταζόμενων διανυσμάτων από τα διανύσματα υποστήριξης της κάθε κλάσης. Ωστόσο, καθώς σκοπός είναι να αξιολογήσουμε τη συνολική απόδοση του νέου συστήματος, μετρήσαμε το συνολικό αριθμό σωστών κατηγοριοποιήσεων για κάθε διάνυσμα το οποίο "διέρχεται" από κάθε εκπαιδευμένο SVM αντίστοιχα. Για την εξέταση, ακολουθήσαμε ένα πρωτόκολλο τύπου "leave-one-person-out". Περισσότερες πληροφορίες σχετικά με την πειραματική διαδικασία παρέχονται στη σχετική ενότητα 3.5.

3.3.2 History Triple Features (HTFs)

Διερευνώντας τις δυνατότητες του μετασχηματισμού Trace επεκτείναμε τη μέθοδο που βασίζεται στα HTTs δημιουργώντας ακόμη πιο αποτελεσματικά χαρακτηριστικά για την αναγνώριση ανθρωπίνων κινήσεων. Τα νέα χαρακτηριστικά αποτελούνται από ένα σύνολο διαιρέσεων τριπλών χαρακτηριστικών και παρουσιάζουν αμεταβλητότητα σε διά-



Σχήμα 3.3: *Παραδείγματα από History Trace Templates που έχουν παραχθεί για τις κινήσεις jack, side, skip, run και pjump της βάσης Weizmann. Η δεύτερη σειρά απεικονίζει τις σιλουέτες για τα παραπάνω στιγμιότυπα ενώ η τρίτη και η τέταρτη σειρά παρουσιάζει δύο διαφορετικούς τύπους HTT που έχουν παραχθεί για καθένα από τα βίντεο δράσεων.*

φορες παραμορφώσεις.

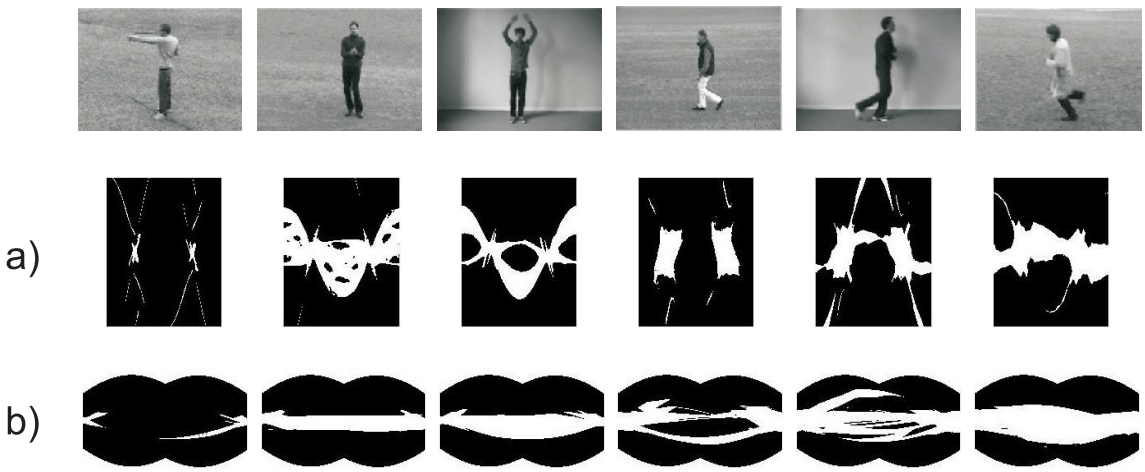
Για κάθε βίντεο-ακολουθία, το φόντο και οι σιλουέτες εξάγονται με τον τρόπο που περιγράφηκε παραπάνω για τις ίδιες βάσεις δεδομένων. Σε αυτή την περίπτωση, χρησιμοποιώντας διαφορετικά συναρτησιακά, υπολογίζεται ένας αριθμός διαφορετικών μετασχηματισμών για κάθε καρέ. Από αυτούς τους μετασχηματισμούς, υπολογίζεται ένα διάνυσμα το οποίο αποτελείται από μια σειρά αμετάβλητων χαρακτηριστικών που έχουν υπολογιστεί για το κάθε καρέ μιας περιόδου. Κάνοντας χρήση Γραμμικής Διακριτικής Ανάλυσης (Linear Discriminant Analysis) [23] για να μειώσουμε τη διάσταση, η όλη ακολουθία αναπαριστάται από ένα νέο διάνυσμα που ονομάζουμε History Triple Feature (HTF) και είναι ένα σύνολο από πραγματικούς αριθμούς που περιέχει πολύτιμη πληροφορία για την κατηγοριοποίηση ανθρωπίνων κινήσεων. Μια πιο κατανοητή περιγραφή της συγκεκριμένης τεχνικής εξαγωγής χαρακτηριστικών, δίνεται στην ενότητα 3.4.2.

3.3.3 Ικανότητα του Trace να διαχωρίζει κλάσεις δράσεων (μια διαισθητική απεικόνιση)

Τα χαρακτηριστικά τα οποία προκύπτουν από το μετασχηματισμό Trace μπορεί να έχουν ή και όχι φυσική σημασία στην ανθρώπινη αντίληψη. Ωστόσο, έχουν τις σωστές μαθηματικές ιδιότητες οι οποίες επιτρέπουν την κατηγοριοποίηση των κινήσεων από μια ορισμένη ομάδα μετασχηματισμών. Για να απεικονίσουμε την ικανότητα αυτή του Trace να παρέχει ικανά για την κατηγοριοποίηση κινήσεων χαρακτηριστικά και προκειμένου να παρέχουμε μια διαισθητική κατανόηση αυτής της ικανότητας, κατασκευάσαμε τα ονομαζόμενα Weighted Trace Transforms (WTT) η βασική μέθοδος των οποίων, είχε αρχικά προταθεί στο [107] για την αναγνώριση προσώπου. Εφαρμόσαμε την ίδια τεχνική στα HTTs, υπολογίζοντας τα Weighted History Trace Transforms (WHTTs) για κάθε μια από τις κλάσης της βάσης KTH.

Κάθε γραμμή ίχνους, αναπαρίσταται από ένα σημείο στην Trace αναπαράσταση μιας εικόνας. Το WHTT είναι μια αναπαράσταση των ιχνο-γραμμών ζυγισμένες ανάλογα με το ρόλο που παίζουν στην αναγνώριση των διαφόρων κλάσεων. Στην πράξη, βρίσκει τα χαρακτηριστικά εκείνα τα οποία παραμένουν στο τελικό template (HTT) για την κάθε κλάση, ακόμα και αν η κίνηση εκτελείται από διαφορετικά άτομα ή η λήψη έχει γίνει από διαφορετική οπτική γωνία. Το WHTT υπολογίζεται ως ακολούθως:

Έστω D_1, D_2, D_3 είναι 3 HTTs εκπαίδευσης. Η διαφορά μεταξύ των HTTs τριών κινήσεων υπολογίζεται.



Σχήμα 3.4: *Weighted History Trace templates, για όλες τις διαφορετικές κλάσεις δράσεων της βάσης KTH με τη χρήση δύο διαφορετικών συναρτησιακών (σειρά a) και b)). Τα σημαντικά σημεία διαφέρουν ξεκάθαρα από κλάση σε κλάση.*

$$\begin{aligned} D_1(p,\theta) &\equiv |T_1(p,\theta) - T_2(p,\theta)|, \\ D_2(p,\theta) &\equiv |T_1(p,\theta) - T_3(p,\theta)|, \\ D_3(p,\theta) &\equiv |T_2(p,\theta) - T_3(p,\theta)|, \end{aligned} \tag{3.22}$$

όπου T_i είναι το HTT της i^{th} κίνησης εκπαίδευσης και κ είναι ένα κατώφλι. Ο πίνακας βαρών ορίζεται ως ακολούθως:

$$f(u) = \begin{cases} 1, & \text{εάν } D_1(p,\theta) \preceq \kappa \text{ και } D_2(p,\theta) \preceq \kappa \text{ και } D_3(p,\theta) \preceq \kappa \\ 0, & \text{διαφορετικά} \end{cases} \tag{3.23}$$

Το αποτέλεσμα είναι τελικά ένα νέο template το οποίο περιέχει αυτές τις τονισμένες ιχνο-γραμμές οι οποίες παρήγαγαν τιμές για τα HTTs που διέφεραν η μια από την άλλη έως ένα ορισμένο επίπεδο κ. Τα αποτελέσματα από τους παραπάνω υπολογισμούς και την εφαρμογή τους σε διαφορετικές κλάσεις της βάσης KTH, απεικονίζονται στο Σχήμα 3.4. Τα WHTTs έχουνε υπολογιστεί λαμβάνοντας ως σύνολο εκπαίδευσης κάθε φορά, το σύνολο των HTT δειγμάτων που αποτελούν την αντίστοιχη κλάση της κίνησης. Για να αποδείξουμε ότι διαφορετικά συναρτησιακά μπορούν να εισάγουν διαφορετικά χαρακτηριστικά μιας κίνησης, έχουν χρησιμοποιηθεί δύο διαφορετικά συναρτησιακά. Η διαφορά των τονισμένων σημείων μεταξύ των τελικών templates ανάμεσα στις κλάσεις των δράσεων, φαίνεται ξεκάθαρα.

3.4 Κατασκευάζοντας χαρακτηριστικά βασισμένα στον Trace για ανθρώπινες κινησιο-ακολουθίες

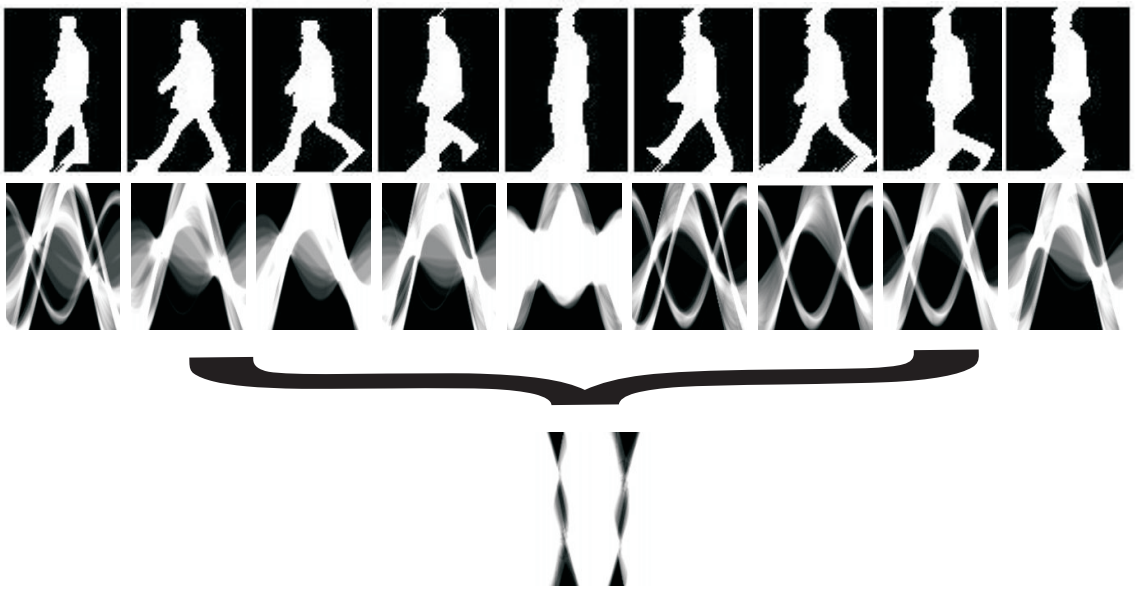
Έχει δειχθεί [49] ότι το ολοκλήρωμα κατά μήκος ευθειών ορισμένων στο πεδίο μιας 2D συνάρτησης, μπορούν να την αναπαραστήσουν πλήρως. Όπως περιγράφηκε πιο πάνω, ο μετασχηματισμός Trace παράγεται διατρέχοντας μια εικόνα με ευθείες γραμμές στις οποίες ορισμένα συναρτησιακά της συγκεκριμένης συνάρτησης έχουν υπολογιστεί. Το αποτέλεσμα είναι μια άλλη εικόνα δυο διαστάσεων η οποία αποτελεί μια καινούργια συνάρτηση η οποία εξαρτάται από τις παραμέτρους (ϕ, p) που χαρακτηρίζουν την κάθε γραμμή. Διαφορετικοί Trace μετασχηματισμοί μπορούν να δημιουργηθούν χρησιμοποιώντας διαφορετικά συναρτησιακά. Σε αυτή την εργασία, επιλέγουμε τους κατάλληλους υπολογισμούς για τα αντίστοιχα Trace συναρτησιακά ούτως ώστε να εκμεταλλευτούμε την ευρωστία του Trace στο θόρυβο και την αμεταβλητότητα στην μετατόπιση και την κλιμάκωση.

Έστω ότι $f(x, y)$ είναι μια 2D συνάρτηση στο Ευκλείδειο επίπεδο $\Re^2$ η οποία έχει εξαχθεί από μια βίντεο-ακολουθία που περιέχει δυαδικής μορφής σιλουέτες. Ο μετασχηματισμός Trace $\breve{g}_f$, είναι μια συνάρτηση ορισμένη στο χώρο των ευθειών L μέσα στο $\Re^2$ από ένα συναρτησιακό κατά μήκος κάθε τέτοιας γραμμής. Εάν για παράδειγμα αυτό το συναρτησιακό περιορίζει τη λειτουργία του στο ολοκλήρωμα της κάθε γραμμής, τότε υπάγεται στην περίπτωση του συνεχούς μετασχηματισμού Radon μιας εικόνας και δίνεται από:

$$R_f(p, \theta) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\delta(p - x\cos\theta - y\sin\theta)dxdy \tag{3.24}$$

όπου $R(p, \theta)$ είναι το ολοκλήρωμα της ευθείας μιας εικόνας κατά μήκος της γραμμής από $-\infty$ έως ∞. Το p και το θ είναι οι παράμετροι που ορίζουν τη θέση της γραμμής. Έτσι, το $R_f(p, \theta)$ είναι το αποτέλεσμα του ολοκληρώματος f επάνω στη γραμμή $p = x\cos\theta + y\sin\theta$. Ως σημείο αναφοράς ορίζεται το κέντρο της σιλουέτας.

Καθώς οι ανθρώπινες κινήσεις είναι χωρο-χρονικοί όγκοι, σκοπός είναι να αναπαραστήσουμε όσο το δυνατόν περισσότερη από τη δυναμική και τη δομική πληροφορία της κίνησης. Σε αυτό το εγχείρημα, ο μετασχηματισμός Trace παρουσιάζει εξαιρετική καταλληλότητα. Μετασχηματίζει εικόνες δύο διαστάσεων με γραμμές σε ένα πεδίο από πιθανές παραμέτρους ευθειών, όπου κάθε γραμμή στην εικόνα θα δώσει μια ακμή τοποθετημένη στην αντίστοιχη παράμετρο. Όταν ο Trace υπολογίζεται με βάση το κέντρο της σιλουέτας, συγκεκριμένοι συντελεστές θα έχουν συλλάβει αρκετή από την ενέργεια της σιλουέτας. Αυτοί οι συντελεστές θα ποικίλουν στη διάρκεια του χρόνου και θα παρέχουν σημαντικές διαφορές από μια κίνηση στην άλλη για το ίδιο χρονικό πλαίσιο.



Σχήμα 3.5: *Εξαχθείσες σιλουέτες και μετασχηματισμοί Trace για μια περίοδο της κίνησης walking. Το History Trace Template φαίνεται στο κάτω μέρος της εικόνας.*

3.4.1 Κατασκευάζοντας τα HTTs

Εκτός από τη δομική πληροφορία, προκειμένου να συλλάβουμε και τη χρονική πληροφορία που εμπεριέχεται σε μια κίνηση, προτείνουμε την κατασκευή των $History\,Trace\,Templates$. Ένα τέτοιο template είναι ουσιαστικά ένας συνεχής μετασχηματισμός στη διεύθυνση του χρόνου μιας ακολουθίας. Έστω ότι $f(p, \vartheta, t)$ είναι μια ακολουθία που αναπαριστά μια ανθρώπινη κίνηση. Εάν $\breve{g}_n(p, \theta)$ είναι ο μετασχηματισμός Trace μιας σιλουέτας $s_n(p, \theta)$ για το n καρέ, όπου $n = 1...N$, τότε το History Trace Template για την ακολουθία αυτή θα δίνεται από:

$$T_N(p, \theta) = \sum_{n=1}^{N} \breve{g}_n(p, \theta). \tag{3.25}$$

Με αυτό τον τρόπο τα χαρακτηριστικά που προκύπτουν ως αποτέλεσμα, θα είναι μια συνάρτηση πολλαπλών διακρίσεων που περιέχονται σε πολλαπλούς μετασχηματισμούς παραγόμενους για κάθε περίοδο μιας κίνησης αντίστοιχα. Όπως αναφέρθηκε νωρίτερα, όλες οι περίοδοι έχουν κλιμακωθεί χρονικά στον ίδιο αριθμό καρέ N. Το Σχήμα 3.5 δείχνει τους μετασχηματισμούς που έχουν προκύψει για κάθε σιλουέτα εξαχθείσα από μια περίοδο περπατήματος. Το τελικό HTT φαίνεται στο κάτω μέρος της εικόνας. Για την πειραματική διαδικασία, έχουμε υπολογίσει και εξετάσει έναν αριθμό από μετασχηματισμούς Trace χρησιμοποιώντας διάφορα συναρτησιακά. Η ακριβείς μορφές των παραπάνω μετασχηματισμών παρέχονται στον Πίνακα 3.1.

Πίνακας 3.1: *Διάφορα συναρτησιακά που υπολογίστηκαν για την πειραματική διαδικασία.*

Trace Transform	Functional
1	$T(f(x)) = \int_{[0,\infty]} rf(r)dr$ where $r = x - c$ and $c = median_x \{x, f(x)\}$
2	$T(f(x)) = \int_{[0,\infty]} r^2 f(r)dr$ where $r = x - c$ and $c = median_x \{x, f(x)\}$
3	$T(f(x)) = median_{r\geq 0}\{f(r), (f(r))^{\frac{1}{2}}\}$ where $r = x - c$ and $c = median_x \{x, f(x)\}$
4	$T(f(x)) = median_{r\geq 0}\{rf(r), (f(r))^{\frac{1}{2}}\}$ where $r = x - c$ and $c = median_x \{x, f(x)\}$
5	$T(f(x)) = \int_{[0,\infty]} e^{iklogr} r^p f(r)dr, (p = 0.5, k = 4)$ where $r = x - c$ and $c = median_x \{x, (f(x))^{\frac{1}{2}}\}$
6	$T(f(x)) = \int_{[0,\infty]} e^{iklogr} r^p f(r)dr, (p = 0, k = 3)$ where $r = x - c$ and $c = median_x \{x, (f(x))^{\frac{1}{2}}\}$
7	$T(f(x)) = \int_{[0,\infty]} e^{iklogr} r^p f(r)dr, (p = 1, k = 5)$ where $r = x - c$ and $c = median_x \{x, (f(x))^{\frac{1}{2}}\}$

3.4.2 Κατασκευάζοντας τα HTFs

Σε αυτό το κεφάλαιο εισάγουμε μια καινοτόμο αναπαράσταση της ανθρώπινης κίνησης χρησιμοποιώντας χαρακτηριστικά που προκύπτουν από το μετασχηματισμό Trace τα οποία από δω και στο εξής θα αποκαλούνται *History Triple Features* (HTFs). Ο μετασχηματισμός Trace είναι ένας γενικός μετασχηματισμός ο οποίος μπορεί να εφαρμοστεί σε ολόκληρες εικόνες. Είναι γνωστό ότι μπορεί να επιλέξει τόσο χαρακτηριστικά σχήματος όσο και υφής ενός αντικειμένου την περιγραφή του οποίου χρησιμοποιείται και προσφέρει μια εναλλακτική αναπαράσταση μιας εικόνας [107].

Παρουσιάσαμε παραπάνω, πως ο Trace μπορεί να χρησιμοποιηθεί για να αναπαραστήσει μια ολόκληρη κίνηση. Ωστόσο, η συγκεκριμένη αναπαράσταση μπορεί να χρησιμοποιηθεί μόνο στην περίπτωση που οι κινήσεις έχουν βιντεοσκοπηθεί κάτω από τις ίδιες συνθήκες (γωνία λήψης, κλίμακα, περιστροφή κάμερας κτλ.). Καθώς αυτό δεν είναι κάτι σύνηθες στις περισσότερες από τις εφαρμογές που αφορούν στην αναγνώριση της ανθρώπινης δραστηριότητας, προτείνουμε μια πιο εξελιγμένη τεχνική η οποία ξεπερνά πολλούς από τους παραπάνω περιορισμούς. Ο Trace είναι μια πολύ πλούσια αναπαράσταση μιας εικόνας. Για να την χρησιμοποιήσουμε απευθείας για αναγνώριση, θα μπορούσαμε να δημιουργήσουμε μια πιο απλουστευμένη εκδοχή του.

Οι συγγραφείς του [49] έχουν αποδείξει ότι χρησιμοποιώντας εξαχθέντα triple features, μπορούν να παραχθούν πολύ εύρωστα χαρακτηριστικά για την κατηγοριοποίηση διαφορετικών αλλά πολύ συγγενών κλάσεων (πχ. διαφορετικά είδη ψαριών). Στην ενότητα 3.2 παρουσιάσαμε την θεωρία πίσω από την κατασκευή των triple features. Στη συνέχεια, θα παρουσιάσουμε την κατασκευή των προτεινόμενων HTTs.

Έχοντας εξάγει τις σιλουέτες, μετασχηματίζουμε το χώρο που περιέχει τις σιλουέτες,

στο χώρο του μετασχηματισμού Trace. Ακολουθώντας την διαδικασία που περιγράφηκε στο 3.2.1 για την εξαγωγή των triple features, δημιουργείται ένα σύνολο από τέτοια χαρακτηριστικά . Ο λόγος ενός ζεύγους τέτοιων χαρακτηριστικών όπως έχει δειχθεί, μπορεί να είναι αμετάβλητο σε διάφορα είδη διαστρεβλώσεων, κατά αντιστοιχία με τα συναρτησιακά που έχουν χρησιμοποιηθεί. Αυτά τα συναρτησιακά μπορούν να έχουν επιλεχθεί ούτως ώστε να είναι ευαίσθητα ή σχετικά ανεπηρέαστα από τις πιθανές μεταβολές που συμβαίνουν στα βίντεο κινήσεων, ενώ την ίδια στιγμή διατηρούν την διακριτότητά τους.

Έστω ότι $f(p, \vartheta, t)$ είναι μια ακολουθία που περιγράφει μια ανθρώπινη κίνηση. Εφαρμόζοντας ένα Trace συναρτησιακό T, κατά μήκος γραμμών διατρέχοντας το n καρέ που αναφέρεται στην $s_n(p, \theta)$ σιλουέτα, όπου $n = 1...N$ και N είναι ο αριθμός των καρέ, παράγεται ένας μετασχηματισμός Trace $\breve{g}_n(p, \theta)$. Εφαρμόζοντας διαφορετικά T σε κάθε σιλουέτα $s_n(p, \theta)$, παράγεται ένα σύνολο από μετασχηματισμούς $\breve{g}_n(p, \theta)$. Όπου $i = 1...L$ και L είναι ο αριθμός των μετασχηματισμών που κάποιος επιλέγει να υπολογίσει. Για κάθε $\breve{g}_{n_i}(p, \theta)$ υπολογίζεται ένα σύνολο από $\Pi_{norm}(F, C)$ κανονικοποιημένων τριπλών χαρακτηριστικών.

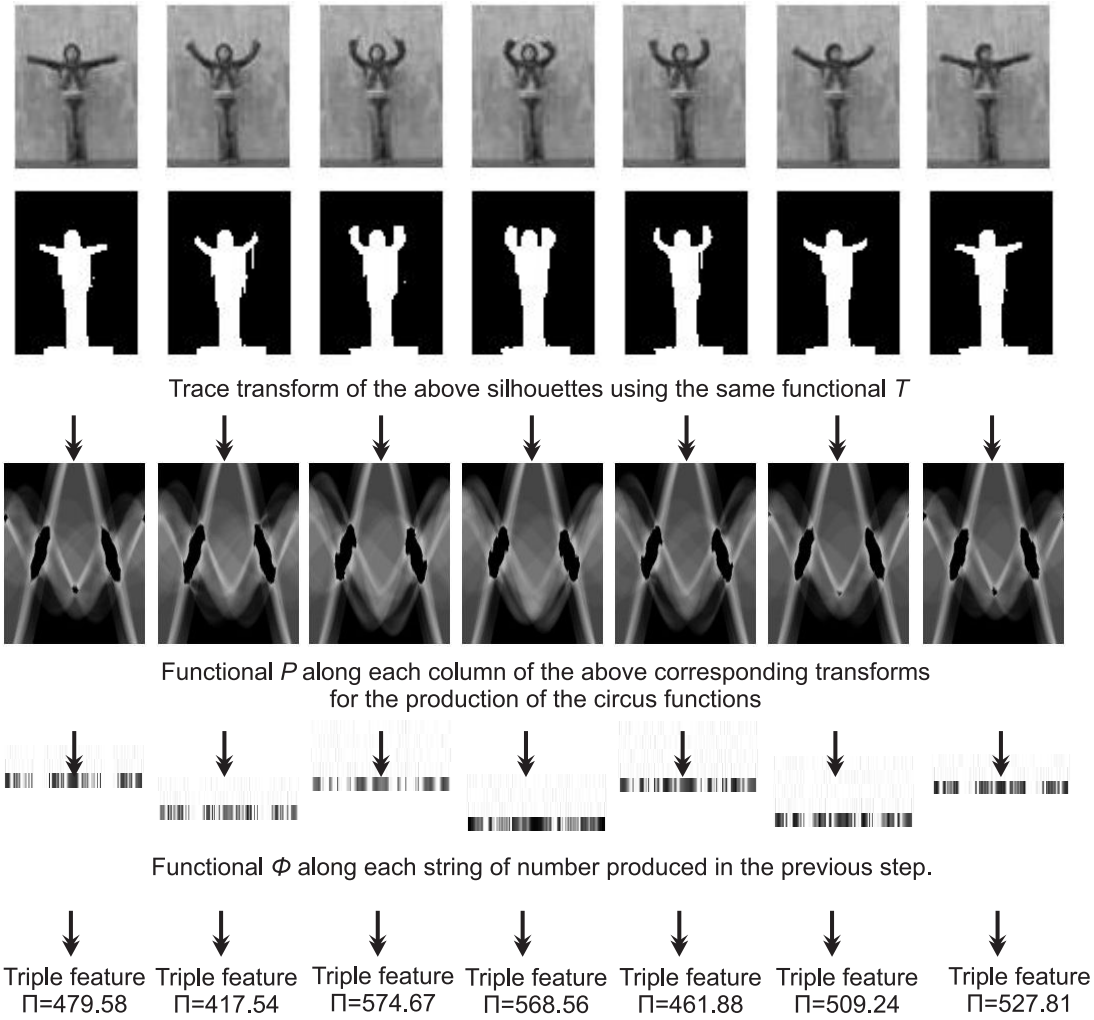
Με απλό τρόπο το triple feature κατασκευάζεται ως εξής:
α) Ο μετασχηματισμός Trace παράγεται εφαρμόζοντας ένα Trace συναρτησιακό T κατά μήκος γραμμών που διατρέχουν την εικόνα.
β) Η κυκλική (circus) συνάρτηση μιας εικόνας παράγεται εφαρμόζοντας ένα διαμετρικό συναρτησιακού P κατά μήκος των στηλών του μετασχηματισμού Trace.
γ) Το τριπλό χαρακτηριστικό παράγεται τελικά εφαρμόζοντας ένα circus συναρτησιακό Φ κατά μήκος μιας σειράς αριθμών που παράγεται από το βήμα β.
Η διαδικασία εικονίζεται στο Σχήμα 3.6.

Διαιρώντας όλα τα $\Pi_{norm}(F, C)$ μεταξύ τους, παράγεται ένα σύνολο από ανεξάρτητα χαρακτηριστικά. Έτσι, όλη η ακολουθία της κίνησης τελικά αναπαρίσταται από ένα διάνυσμα $\mathbf{v}$ το οποίο αποτελείται όλους τους λόγους τριπλών χαρακτηριστικών που έχουν υπολογιστεί για κάθε καρέ της ακολουθίας.

$$\mathbf{v} = (\Pi_{rat_1}, \Pi_{rat_2}, ..., \Pi_{rat_{g-1}}, \Pi_{rat_g}) \tag{3.26}$$

όπου Π_{rat} είναι ο λόγος των δύο κανονικοποιημένων τριπλών χαρακτηριστικών και g ο αριθμός των υπολογισμένων λόγων.

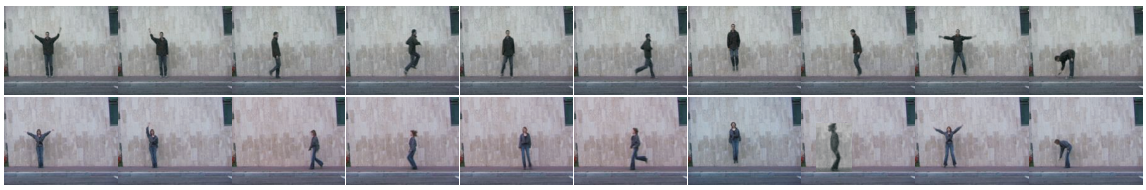
Αυτή η μέθοδος επιτρέπει την κατασκευή πολλών χαρακτηριστικών εύκολα. Αν υποθέσουμε ότι κάποιος κάνει χρήση 10 συναρτησιακών για κάθε ένα από τα στάδια κατασκευής (πχ. 10 συναρτησιακά T, 10 συναρτησιακά P και 10 συναρτησιακά Φ) σε ένα βίντεο αποτελούμενο από 10 καρέ, μπορεί να κατασκευάσει 10x10x10x10=10000 χαρακτηριστικά για μια ακολουθία. Όπως αναφέρθηκε πιο πάνω, αυτοί οι αριθμοί μπορεί να μην έχουν κά-



Σχήμα 3.6: *Εξαγωγή τριπλού χαρακτηριστικού για μια περίοδο της κίνησης wave, της βάσης δεδομένων Weizmann.*

ποια φυσική σημασία σύμφωνα με την ανθρώπινη αντίληψη, μπορούν όμως να έχουν τις απαιτούμενες μαθηματικές ιδιότητες για τεχνικές κατηγοριοποίησης.

Καθώς η διακριτή ισχύς των κατασκευασμένων χαρακτηριστικών θα ποικίλει αναμφισβήτητα, η εφαρμογή μιας τεχνικής μείωσης της διάστασης θα μπορούσε να παρέχει μια επιλογή από τα πιο διακριτά χαρακτηριστικά ενώ την ίδια στιγμή θα έκανε το πρόβλημα πιο εύκολα διαχειρίσιμο. Στο προτεινόμενο σχήμα, τα διανύσματα HTF τα οποία έχουν παραχθεί με τον τρόπο που αναφέρθηκε, υφίστανται γραμμική διακριτή ανάλυση (LDA) με το σκοπό να προσδιοριστεί ένας υπο-χώρος ο οποίος να είναι κατάλληλος για κατηγοριοποίηση. Στην πράξη, κρατάμε μόνο ένα υποσύνολο των αρχικών HTF διανυσμάτων που περιέχει τα πιο διακριτά από τα υπολογιζόμενα χαρακτηριστικά ικανά να περιγράψουν ολόκληρη την ακολουθία της κίνησης.



Σχήμα 3.7: *Δείγματα δράσεων για τις κινήσεις wave1, wave2, walk, pjump, side, run, skip, jack, jump και bend της βάσης δεδομένων Weizmann.*

3.5 Πειραματικά αποτελέσματα

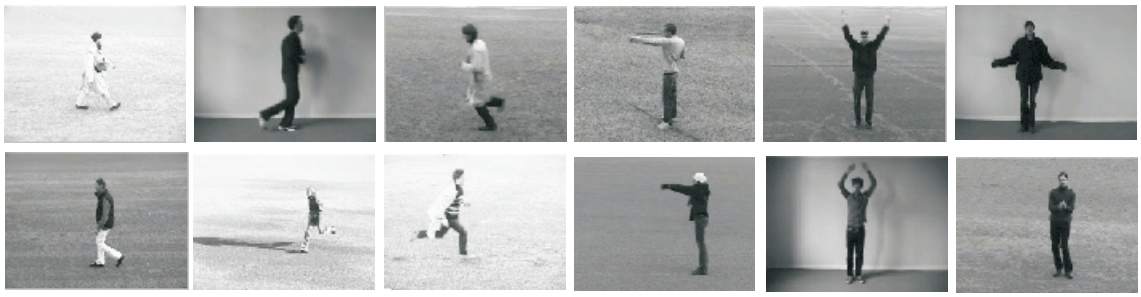
Σε αυτή την ενότητα, θα παρουσιάσουμε τα πειραματικά αποτελέσματα με σκοπό να αναδείξουμε την αποτελεσματικότητα της προτεινόμενης μεθόδου για την αναγνώριση ανθρωπίνων κινήσεων ενώ την ίδια στιγμή, θα παρέχουμε την αξιολόγηση των διαφορετικών αμετάβλητων συναρτησιακών που υπολογίστηκαν για την κατασκευή των HTTS.

Διάφορες δημοσιευμένες προσεγγίσεις έχουν χρησιμοποιήσει διαφορετικές εκδοχές αξιολόγησης. Όπως διατυπώνεται στο [27], οι περισσότεροι από τους ερευνητές που δουλεύουν στο πεδίο της αναγνώρισης ανθρώπινης δραστηριότητας, έχουν αξιολογήσει τις μεθόδους τους επάνω στην KTH [100] και την Weizmann [6]. Παρ' όλα αυτά, δεν υπάρχει ένα καθολικά αποδεκτό πρότυπο αξιολόγησης που να ακολουθείται από την πλειοψηφία των ερευνητών. Οι συγγραφείς του παραπάνω άρθρου αναφέρουν επίσης, ότι παρατηρούνται μεγάλες αποκλείσεις (έως και 10.67%) στα αποτελέσματα που δημοσιεύονται όταν διαφορετικές προσεγγίσεις αξιολόγησης εφαρμόζονται στα ίδια δεδομένα.

Στα δικά μας πειράματα, εφαρμόζουμε το πρωτόκολλο leave-one-person-out cross validation για την αξιολόγηση της μεθόδου. Η συγκεκριμένη προσέγγιση επιλέχθηκε εξαιτίας του ότι είναι πολύ δημοφιλής μεταξύ των ερευνητών. Επίσης ανακατασκευάζει τις ανάγκες των εφαρμογών πραγματικών συνθηκών με τον πιο συγγενή τρόπο. Το παραπάνω πρωτόκολλο χρησιμοποιεί ένα δείγμα ενός ατόμου για την εξέταση και το υπόλοιπο σύνολο δεδομένων χρησιμοποιείται για την εκπαίδευση. Η διαδικασία επαναλαμβάνεται N φορές όπου N είναι ο αριθμός των προσώπων στη βάση δεδομένων. Η απόδοση αναφέρεται ως η μέση ακρίβεια N επαναλήψεων.

Συνοπτικά, η διαδικασία που ακολουθείται από ένα σύστημα αναγνώρισης έχει ως εξής: η φυσική δυναμική συμπεριφορά ενός αγνώστου υποκειμένου συλλαμβάνεται από ένα σύστημα αναγνώρισης κίνησης και στη συνέχεια γίνεται επεξεργασία και σύγκριση με ένα σύνολο από προ-εγγεγραμμένα δεδομένα τα οποία έχουν προηγουμένως χρησιμοποιηθεί για την εκπαίδευση του συστήματος. Η τελική απόφαση δίνεται βάσει της σχετικότητας της εξεταζόμενης κίνησης, με κάποιο από τα δείγματα εκπαίδευσης ανάλογα με τον κανόνα συσχέτισης που ακολουθείται από το σύστημα.

Τα πειράματα εκτελέστηκαν σε έναν επεξεργαστή Intel Core i5 (650@3,2 GHz) με



Σχήμα 3.8: *Δείγματα δράσεων για τις κινήσεις walking, jogging, running, boxing, hand waving και hand clapping της βάσης δεδομένων KTH.*

4GB μνήμης RAM. Για τα πειράματα χρησιμοποιήθηκαν οι βάσεις δεδομένων KTH και Weizmann. Δείγματα των βάσεων για διάφορους τύπους κινήσεων απεικονίζονται στις Εικόνες 3.7 και 3.8. Η βάση δεδομένων KTH περιέχει 6 τύπους κινήσεων (walking, jogging, running, boxing, hand waving και hand clapping) εκτελεσμένες πολλές φορές από 25 άτομα σε 4 διαφορετικά σενάρια, κάτω από διάφορες συνθήκες φωτισμού: εξωτερικοί χώροι, εξωτερικοί χώροι με μεταβαλλόμενη κλιμάκωση (zoom-in, zoom-out), εξωτερικοί χώροι με διαφορετικά ρούχα και εσωτερικοί χώροι. Η βάση αποτελείται από 600 ακολουθίες. Όλες οι ακολουθίες έχουν κινηματογραφηθεί σε ομοιόμορφο φόντο, με στατική κάμερα και ταχύτητα λήψης 25 καρέ το δευτερόλεπτο.

Η βάση βίντεο δεδομένων Weizmann αποτελείται από 90 βίντεο χαμηλής ανάλυσης (180 x 144, ταχύτητα κλείστρου 50 καρέ το δευτερόλεπτο) που παρουσιάζουν 9 διαφορετικούς ανθρώπους. Κάθε άτομο έχει εκτελέσει 10 φυσικές δραστηριότητες όπως τρέξιμο βάδισμα και άλλα. Στη βάση οι κινήσεις που εμπεριέχονται αναφέρονται με τις ακόλουθες ονομασίες: run, walk, skip, jumping-jack (ή shortly jack), jump-forward-on-two-legs (ή jump), jump-in-place-on-two-legs (ή pjump), gallopsideways (ή side), wave-two-hands (ή wave2), waveone- hand (ή wave1), ή bend.

Στα δικά μας πειράματα, οι ακολουθίες έχουν υποστεί σμίκρυνση στην ανάλυση των 160*120 pixels και έχουνε διάρκεια 4 δευτερόλεπτα κατά μέσο όρο. Τα δείγματα εκπαίδευσης έχουν κατασκευαστεί κατατμίζοντας χειροκίνητα (τόσο σε χρόνο όσο και χώρο) και στοιχίζοντας τις παρεχόμενες ακολουθίες. Το φόντο έχει αφαιρεθεί με τη χρήση ενός απλού αλγορίθμου grassfire [36]. Για τη γενικευμένη αξιολόγηση της απόδοσης του συστήματος για το πρόβλημα της αναγνώρισης κινήσεων, χρησιμοποιήθηκε το πρωτόκολλο leave-one-person-out.

Σε αυτό το σημείο, θα έπρεπε να τονίσουμε ότι η αναγνώριση της ανθρώπινης κίνησης είναι ένα πρόβλημα πολλαπλών κλάσεων. Αντιμετωπίζουμε την ιδιαιτερότητα αυτή ανακατασκευάζοντας το πρόβλημα σαν τη γενίκευση δυαδικών κατηγοριοποιήσεων. Πιο συγκεκριμένα, για κάθε βάση δεδομένων εκπαιδεύσαμε 6 και 10 διαφορετικά SVMs (ένα

για κάθε κλάση της KTH και της Weizmann αντίστοιχα) εφαρμόζοντας ένα πρωτόκολλο ένας-εναντίων-όλων. Η τελική απόφαση λαμβανόταν κατηγοριοποιώντας κάθε δείγμα υπό εξέταση σε μια κλάση $\mathcal{C}_a$, ανάλογα με την απόσταση d του εξεταζόμενου διανύσματος από τα διανύσματα υποστήριξης. Όπου $\mathcal{C}_a$ είναι το σύνολο των templates που έχουν κατηγοριοποιηθεί σε μια κλάση (π.χ. boxing). Ωστόσο, καθώς θέλουμε να αξιολογήσουμε τη γενίκευση του αλγορίθμου με έναν πιο ευρύ τρόπο, μετρήσαμε τις επιτυχείς δυαδικές κατηγοριοποιήσεις για το κάθε δείγμα, το οποίο εξετάζεται σε καθένα από τα διαφορετικά SVMs. Με αυτό τον τρόπο καταφέραμε να επιτύχουμε 25*6*4*6=3600
(άτομα*κινήσεις*δείγματα για κάθε άτομο) κατηγοριοποιήσεις αντί για 600 για την KTH και 10*9*10=900 κατηγοριοποιήσεις αντί για 90 για την Weizmann αντίστοιχα. Η ίδια διαδικασία ακολουθήθηκε και για τις δύο μεθόδους εξαγωγής χαρακτηριστικών (HTT και HTF αντίστοιχα).

Τα αποτελέσματα ανέδειξαν μια πολύ ανταγωνιστική απόδοση και για τις δύο τεχνικές. Παρ' όλα αυτά, όπως αναμενόταν, τα HTFs απέδωσαν πολύ καλύτερα καθώς ήταν σχεδιασμένα να είναι αμετάβλητα σε διάφορες διακυμάνσεις. Πιο συγκεκριμένα, η απόδοση αναγνώρισης με την χρήση των HTTs ήταν 90.22% και 93.4% για κάθε μια από τις δύο βάσεις δεδομένων ενώ για τα HTFs η απόδοση ήταν 93.14% and 95.42% για την KTH και την Weizmann αντίστοιχα. Πρέπει επίσης να επισημάνουμε ότι οι εξαχθείσες σιλουέτες που χρησιμοποιήθηκαν, ήταν ιδιαίτερα θορυβώδεις, ενώ κανένα προκαταρκτικό φιλτράρισμα δεν εφαρμόστηκε για την βελτίωσή τους. Τα συνολικά ποσοστά απόδοσης (για όλες τις κλάσεις) που παράχθηκαν για τα διάφορα συναρτησιακά του Πίνακα 3.1 που χρησιμοποιήθηκαν για την κατασκευή των HTTs, παρέχονται στον Πίνακα 3.2.

Είναι επίσης ενδιαφέρον να αναφέρουμε ότι και οι δύο μέθοδοι εξαγωγής χαρακτηριστικών διενεργούνται αρκετά γρήγορα. Για 25 επαναλήψεις, (εξέταση όλων των δειγμάτων της KTH), συμπεριλαμβανομένης και της εκπαίδευσης, η τεχνική των HTTs απαίτησε 6 λεπτά ενώ κάθε δείγμα εξεταζόταν σε 0.01 δευτερόλεπτα. Η ίδια διαδικασία εξέτασης για τη μέθοδο των HTFs, χρειάστηκε 2.5 λεπτά ενώ κάθε δείγμα εξεταζόταν σε 0.005 δευτερόλεπτα. Ωστόσο, για τα HTFs ο χρόνος εξέτασης είναι ανάλογος και αντιστρόφως ανάλογος αντίστοιχα, του αριθμού των Trace συναρτησιακών T που κάποιος θα επιλέξει να υπολογίσει. Στη δική μας πειραματική διαδικασία, υπολογίστηκαν 9 T συναρτησιακά για κάθε καρέ του κάθε βίντεο που περιέχει μια κίνηση. Αυτό είχε σαν αποτέλεσμα την παραγωγή 40 τριπλών χαρακτηριστικών για κάθε καρέ. Με αυτό τον τρόπο, κάθε ακολουθία (όλες αποτελούνταν από 7 καρέ), αρχικά περιγράφεται από ένα διάνυσμα από 320 χαρακτηριστικά . Η εφαρμογή LDA για την επιλογή των πιο διακριτικών από αυτά, έχει σαν αποτέλεσμα ένα διάνυσμα από 31 χαρακτηριστικά $\mathbf{v}$. Ο χρόνος που απαιτήθηκε για των υπολογισμό των HTFs για μια κίνηση ήταν $\simeq 2$ δευτερόλεπτα.

Μια σύγκριση των προτεινόμενων τεχνικών με άλλες δημοσιευμένες δουλειές για τις

Πίνακας 3.2: *Αποτελέσματα που έχουν παραχθεί υπολογίζοντας HTT, εξετάζοντας διάφορα συναρτησιακά στις δυο διαφορετικές βάσεις δεδομένων.*

Trace Transform	Αποτελέσματα (%) στη KTH	Αποτελέσματα (%) στη Weizmann
Radon	87.7	91.11
1	89.82	92.20
2	88.41	92.00
3	86.66	90.52
4	88.00	93.11
5	89.82	92.20
6	89.82	92.20
7	90.22	93.41

ίδιες βάσεις δεδομένων, παρουσιάζεται στους Πίνακες 3.3 και 3.4 για την KTH και την Weizmann βάση δεδομένων αντίστοιχα. Σε αυτό το σημείο θα πρέπει να τονίσουμε ότι τα αποτελέσματα που παραθέτονται, δεν είναι τα βέλτιστα για την τεχνική HTF. Υπολογίζοντας περισσότερα χαρακτηριστικά και/ή προσαρμόζοντας καταλληλότερα συναρτησιακά για τον υπολογισμό του τελικού HTF διανύσματος το οποίο αναπαριστά μια ακολουθία, μπορεί να αυξήσει δραματικά την απόδοση. Σκοπός αυτής της εργασίας δεν είναι να παρουσιάσει ένα άριστο σύστημα αναγνώρισης ανθρωπίνων κινήσεων το οποίο να βρεθεί στην αιχμή του ανταγωνισμού. Στόχος μας ήταν να εξετάσουμε κατά κύριο λόγο τις δυνατότητες του μετασχηματισμού Trace για τη συγκεκριμένη ενέργεια και να προτείνουμε καινοτόμες τεχνικές εξαγωγής χαρακτηριστικών βασισμένες στον Trace και οι οποίες θα είναι κατάλληλες για αναγνώριση ανθρωπίνων κινήσεων, ενώ παράλληλα αντιπαρέρχονται συνήθη προβλήματα όπως η οπτική κλιμάκωση και οι ασταθείς λήψεις.

Παρά το γεγονός ότι η μέθοδος HTT έδειξε ότι μπορεί αποτελεσματικά να διακρίνει κλάσεις κινήσεων, φανερώνει κάποιους περιορισμούς όσον αφορά τις διακυμάνσεις στην λήψη του βίντεο και θα ήταν μάλλον πιο κατάλληλη για εφαρμογές ελεγχόμενου περιβάλλοντος. Από την άλλη, η μέθοδος HTF έδειξε να έχει ιδιαίτερη δυναμική για το συγκεκριμένο εγχείρημα, καθώς όχι μόνο απέδωσε ικανοποιητικά, αλλά σε θεωρητικό επίπεδο δεν υπάρχει κανένας περιορισμός στον υπολογισμό κατάλληλων συναρτησιακών. Για εφαρμογές πραγματικών συνθηκών, τα συναρτησιακά θα μπορούσαν να υπολογιστούν έτσι ώστε να εξυπηρετούν καθορισμένες απαιτήσεις αυξάνοντας την απόδοση ενός συστήματος που έχει ως μοναδικό σκοπό την αναγνώριση ανθρωπίνων κινήσεων. Εφαρμογές όπως αλληλεπίδραση ανθρώπου μηχανής και διάφορα σχετικά παιχνίδια, θα μπορούσαν επίσης να ωφεληθούν από ένα πιο γενικευμένο σχεδιασμό ο οποίος θα καλύπτει συγκεκριμένες ανάγκες ενώ την ίδια ώρα θα αντιπαρέρχεται πολλούς από τους περιορισμούς που προκύπτουν από συνήθη προβλήματα λήψεων και μεταβολών τους.

Πίνακας 3.3: *Ποσοστά κατηγοριοποίησης (%) που έχουν επιτευχθεί από διάφορες δημοσιευμένες μεθόδους, επάνω στη βάση KTH.*

Μέθοδος	Μέση Ακρίβεια	Κατηγοριοποιητής
Wong end Cipolla [124]	86.50%	SVM
Sun et al. [110]	94.00%	SVM
Liu end Shah [61]	94.16%	VWCcorrel
Dollar et al.[21]	81.20%	NNC
Schuldt et al. [100] (reported in [91])	50.33%	NNC
Rapantzikos et al. [91]	88.30%	NNC
Oikonomopoulos et al. [81] (reported in [124])	74.79%	NNC
Ke et al. [50]	80.90%	SVM
Schuldt et al. [100]	71.70%	SVM
Niebles et al. [78]	81.50%	pLSA
Jiang et all [46]	84.40%	LPBOOST
Laptev et all. [58]	91.80%	SVM
HTTs	90.22%	SVM
HTFs	93.14%	SVM

Πίνακας 3.4: *Ποσοστά κατηγοριοποίησης (%) που έχουν επιτευχθεί από διάφορες δημοσιευμένες μεθόδους, επάνω στη βάση Weizmann.*

Μέθοδος	Μέση Ακρίβεια	Κατηγοριοποιητής
Sun et al. [110]	97.80%	SVM
Klasser et al. [53]	84.3%	SVM
Jhuang et al. [44]	96.3%	SVM
Thurau [112]	86.66%	MOH
Thurau et al. [113]	94.40%	1-NN
Niebles et al. [78]	72.8%	pLSA
HTTs	93.4%	SVM
HTFs	95.42%	SVM

Κεφάλαιο 4

Εντοπισμός πτώσης με χρήση χαρακτηριστικών HTFs

4.1 Εισαγωγή

Οι δημογραφικές αλλαγές στην κοινωνία, με τη διαρκή αύξηση του αριθμού των ηλικιωμένων ανθρώπων, κρούουν τον κώδωνα του κινδύνου, ιδίως στις δυτικές χώρες. Ο όρος "eldercare" (φροντίδα των ηλικιωμένων) είναι ένας ευρύς όρος που αναφέρεται στην κάλυψη των ειδικών αναγκών και απαιτήσεων που έχουν τα άτομα τρίτης ηλικίας. Γηροκομεία, νοσηλεία εκτός νοσοκομείου, κατ' οίκον φροντίδα, υποβοηθούμενη διαβίωση κλπ. είναι κάποιες από τις διάφορες υπηρεσίες που ανάγονται σε αυτόν τον όρο. Ένα από τα κύρια σημεία της φροντίδας για ηλικιωμένους που αποτελεί κίνητρο για έρευνα, είναι η ανάγκη για ανεξαρτησία και για αξιοπρεπή διαβίωση.

Παραδοσιακά, η φροντίδα των γηραιότερων ήταν ευθύνη των μελών της οικογένειας και παρεχόταν εντός των ορίων της οικογενειακής κατοικίας. Στις σύγχρονες κοινωνίες, ο ρόλος αυτός περνάει ολοένα και περισσότερο στα χέρια κρατικών οργανισμών ή φιλανθρωπικών ιδρυμάτων [114]. Στους λόγους για την αλλαγή αυτή περιλαμβάνονται η μείωση του μέσου μεγέθους μιας οικογένειας, η αύξηση του προσδόκιμου ζωής, η γεωγραφική διασπορά των μελών μιας οικογένειας και το ότι οι γυναίκες πλέον εκπαιδεύονται και εργάζονται εκτός σπιτιού περισσότερο [114]. Παρά το ότι αυτές οι αλλαγές επηρέασαν πρώτα χώρες της Ευρώπης και της Βόρειας Αμερικής, εντοπίζονται πλέον και σε ασιατικά κράτη [105]. Σε αυτή την πραγματικότητα, η επιστημονική μελέτη έχει εστιάσει στην αυτονομία των γηραιότερων που μένουν κυρίως μόνοι ή δεν έχουν τη δυνατότητα να πληρώσουν ένα άτομο να τους προσέχει. Σύμφωνα με τη αναφορά του Κέντρου Έρευνας και Πρόληψης Τραυματισμών, όσον αφορά ηλικιωμένα άτομα, βλάβες που προκλήθηκαν από πτώσεις είναι πέντε φορές πιο συχνές από άλλους τραυματισμούς και μειώνουν σημαντικά την ευκινησία και την ανεξαρτησία τους [14].

Από τη στιγμή λοιπόν, που οι πτώσεις είναι ένα σημαντικό πρόβλημα για τη υγεία των ατόμων τρίτης ηλικίας, τα συστήματα που στοχεύουν στον εντοπισμό μιας πτώσης έχουν πληθύνει τα τελευταία χρόνια. Σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας [84], περίπου το 28-35% των ατόμων ηλικίας από 65 και πάνω έχουν τουλάχιστον μια πτώση κάθε χρόνο, με το ποσοστό να αυξάνεται στο 32-42% για όσους ξεπερνούν τα 70. Η συχνότητα των πτώσεων αυξάνεται με την ηλικία και την αδυναμία. Πράγματι, οι πτώσεις αυξάνονται εκθετικά από βιολογικές αλλαγές, σχετιζόμενες με την ηλικία, κάτι που οδηγεί σε υψηλή εμφάνιση τέτοιων συμβάντων και συναφών τραυματισμών σε γηράσκουσες κοινωνίες. Εφόσον δε ληφθούν μέτρα πρόληψης στο άμεσο μέλλον, ο αριθμός των τραυματισμών που προκαλούνται από πτώσεις εκτιμάται ότι θα είναι 100% υψηλότερος το 2030 [40]. Σε αυτό το πλαίσιο, η έρευνα στα συστήματα ανίχνευσης πτώσης έχει ενταθεί.

Τα συστήματα αυτά μπορούν να χωρισθούν σε δύο κύριες κατηγορίες: μεθόδους βασισμένες σε αισθητήρες που μπορούν να φορεθούν και τεχνικές βασισμένες σε οπτική πληροφορία. Η πρώτη κατηγορία βασίζεται συνήθως σε φορητές συσκευές όπως επιταχυνσιόμετρα και γυροσκόπια, ή σε smartphones που συνήθως περιέχουν τέτοιους αισθητήρες και βρίσκονται πάντα πάνω σε ένα άτομο (για παράδειγμα, σε τσέπη ή θήκη). Ένα χαρακτηριστικό παράδειγμα μιας πολυμεθοδικής προσέγγισης η οποία εμπλέκει φορητούς αισθητήρες, με σκοπό να εντοπίσει επείγοντα περιστατικά πτώσης, δίνεται στο [22]. Η δεύτερη κατηγορία είναι βασισμένη σε κάμερες 2D ή 3D και περιλαμβάνουν τεχνικές επεξεργασίας εικόνας και αναγνώρισης προτύπων υψηλής υπολογιστικής πολυπλοκότητας, αλλά δεν απαιτούν τη συνεχή μεταφορά μιας συσκευής ή ενός αισθητήρα. Οι ερευνητές στο [73] διαχωρίζουν τους ανιχνευτές πτώσης σε τρεις κατηγορίες: τις βασισμένες σε συσκευές που φοριούνται, τις βασισμένες σε ατμοσφαιρικούς αισθητήρες και αυτές που βασίζονται σε κάμερες (και οπτική πληροφορία). Από μια διαφορετική σκοπιά, οι ερευνητές στο [87] διακρίνουν πάλι τρεις κατηγορίες που περιλαμβάνουν μεθόδους που μετρούν επιτάχυνση, μεθόδους που, μαζί με την μέτρηση επιτάχυνσης, συνδυάζουν και άλλης φύσης μεθόδους και μεθόδους που δεν ασχολούνται καθόλου με τον παράγοντα επιτάχυνση.

Δύο από τις πρώτες προσπάθειες για μια γενική επισκόπηση του τομέα της ανίχνευσης πτώσεων δίνονται στο [80] και στο [87]. Ωστόσο, με τη συνεχή τεχνολογική πρόοδο στην περιοχή αυτή, οι ανασκοπήσεις είναι γενικά παρωχημένες. Μια πιο πρόσφατη και εκτενής ανασκόπηση παρέχεται στο [40] και περιλαμβάνει συγκριτικά στοιχεία από διάφορες μελέτες. Το άρθρο στοχεύει στο να χρησιμεύσει ως αναφορά τόσο για κλινικούς μηχανικούς όσο και για μηχανικούς βιοϊατρικής που σχεδιάζουν ή πραγματοποιούν έρευνα στο πεδίο αυτό. Οι συγγραφείς προσπαθούν κυρίως να καταγράψουν τις προκλήσεις όσον αφορά τις επιδόσεις κάτω από πραγματικές συνθήκες, καθώς και τις τρέχουσες τάσεις του χώρου. Μια πιο λεπτομερής αναφορά πάνω σε όλα τα σχετικά άρθρα παρέχεται στο [73] αλλά στερείται αναφορών σε νέες τάσεις του πεδίου (όπως τεχνικές βασισμένες σε smartphones,

κλπ.).

Η δική μας έρευνα εστιάζει σε λύσεις μη φορητών αισθητήρων και πιο συγκεκριμένα, σε μια προσέγγιση βασισμένη στην υπολογιστική όραση. Σε παρόμοια κατεύθυνση, ερευνητές στο [59] τοποθέτησαν κάμερα στην οροφή ενός δωματίου και ανέλυσαν την κατατμημένη σιλουέτα και την ταχύτητα ενός ηλικιωμένου μέσα στην εικόνα. Η αναγνώριση μιας πτώσης επιτυγχάνεται με τη χρήση ενός εμπειρικού κατωφλίου. Οι συγγραφείς του [115], για να ξεχωρίσουν το κάθισμα από την πτώση, προσέθεσαν και ηχητική πληροφορία (θόρυβος πτώσης). Ωστόσο, το εν λόγω σύστημα δε μπορεί να είναι εύρωστο, από τη στιγμή που τα περισσότερα περιβάλλοντα, όπου τέτοια συστήματα τίθενται σε χρήση, είναι θορυβώδη. Στο [93], η προτεινόμενη προσέγγιση βασίζεται σε έναν συνδυασμό ιστορικού κίνησης και στις αλλαγές του σχήματος που καταλαμβάνει ο άνθρωπος στην εικόνα. Για την κάλυψη μεγάλων περιοχών, τοποθετήθηκαν κάμερες στους τοίχους και η τελική απόφαση λαμβάνεται με την κατωφλίωση των χαρακτηριστικών που εξάγονται. Στο [26] παρουσιάζεται μια τεχνική βασισμένη σε έναν συνδυασμό ιδιόχωρων και ακολουθιών εικόνων που αναπαριστούν κίνηση. Η κατηγοριοποίηση καθημερινών κινήσεων και πτώσεων επιτυγχάνεται με την εξαγωγή ιδιο-κινήσεων και την εφαρμογή μηχανών διανυσμάτων υποστήριξης πολλών κλάσεων.

Άλλες τεχνικές [93], [117] προτείνουν ανιχνευτές πτώσης βασισμένους σε πληροφορία σχήματος. Διαχωρίζουν την ανθρώπινη σιλουέτα με τη χρήση πλαισίου οριοθέτησης (bounding box) ή έλλειψης και εξάγουν γεωμετρικές ιδιότητες όπως αναλογία διαστάσεων, προσανατολισμός ή σημεία ακμών [94]. Ωστόσο, τέτοιες προσεγγίσεις στερούνται ευρωστίας και δυνατότητας γενίκευσης στην εφαρμογή τους, καθώς εξαρτώνται από την ακριβή εξαγωγή της ανθρώπινης σιλουέτας και τους συγκεκριμένους γεωμετρικούς μετασχηματισμούς που μπορεί να υπεισέρχονται, εξ αιτίας της απόστασης και της σχετικής θέσης του ατόμου από την κάμερα. Σε μια πιο πρόσφατη μελέτη [65] παρουσιάζεται μια μέθοδος που συνδυάζει δύο τεχνικές υπολογιστικής όρασης: χαρακτηρισμό πτώσεων με βάση σχηματική πληροφορία και έναν ταξινομητή βασισμένο σε μεθόδους μάθησης για το διαχωρισμό πτώσεων από καθημερινές ενέργειες. Παράλληλα, σε άλλη μελέτη [25], το ανθρώπινο σώμα περιγράφεται με ταίριασμα έλλειψης και η κίνηση της σιλουέτας μοντελοποιείται με τη χρήση μιας κανονικοποιημένης εικόνας ενέργειας κίνησης (integrated normalized motion energy image). Τα απαραίτητα χαρακτηριστικά για την κατηγοριοποίηση των διάφορων κινήσεων του ατόμου εξάγονται από την ποσοτικοποιημένη παραμόρφωση του σχήματος της εξαγμένης σιλουέτας.

Εσχάτως, ορμώμενες από την αποδοτικότητα της πληροφορίας τριών διαστάσεων σε θέματα οπτικής γωνίας και μερικής οπτικής παρεμπόδισης, έχουν δημοσιευθεί αρκετές εργασίες που την εκμεταλλεύονται. Στο [67], η προτεινόμενη μέθοδος είναι βασισμένη σε δεδομένα ταχύτητας και λαμβάνει υπόψιν τη σύμπτυξη ή επέκταση ενός τρισδιάστατου

πλαισίου οριοθέτησης. Η μέθοδος δεν προϋποθέτει πρότερη γνώση για τη σκηνή, καθώς το σύνολο των ενεργειών που εντοπίζονται αρκεί για να ολοκληρώσει τη διαδικασία αναγνώρισης πτώσης. Σε μια άλλη προσέγγιση [89], ο προτεινόμενος αλγόριθμος χρησιμοποιεί την κάμερα βάθους Kinect και δημιουργεί δύο παραμέτρους-χαρακτηριστικά, τον προσανατολισμό του σώματος και την πληροφορία ύψους της σπονδυλικής στήλης, χρησιμοποιώντας το σύστημα συντεταγμένων της εικόνας, ή του κόσμου. Η κάμερα Kinect χρησιμοποιείται επίσης στο [76]. Ο συγκεκριμένος αλγόριθμος ανίχνευσης πτώσης, βασίζεται στην ταχύτητα κίνησης του κεφαλιού, του κεντροϊδούς του σώματος και στην απόστασή τους από το έδαφος. Με το να λαμβάνει υπόψιν τόσο τη θέση του κεφαλιού όσο και του κεντροϊδούς, ο αλγόριθμος φαίνεται να επηρεάζεται λιγότερο από τις ταλαντεύσεις του κεντροϊδούς.

Πάλι κάνοντας χρήση του Kinect, οι συγγραφείς του [135] προτείνουν μια στατιστική μέθοδο που λαμβάνει αποφάσεις βασισμένη σε πληροφορία σχετική με το πώς κινήθηκε ο άνθρωπος κατά τα πιο πρόσφατα πλαίσια του βίντεο. Ο αλγόριθμος βασίζεται στο συνδυασμό των προτεινόμενων χαρακτηριστικών μέσα σε ένα bayesian πλαίσιο. Ο στόχος είναι η δημιουργία μιας τεχνικής που, ενώ έχει εκπαιδευτεί με δεδομένα από μια συγκεκριμένη οπτική γωνία, μπορεί να χαρακτηρίσει πτώσεις που έχουν ληφθεί από διαφορετικές οπτικές γωνίες.

Σε αυτή τη μελέτη, προτείνουμε μια νέα οπτική τεχνική για αυτοματοποιημένη ανίχνευση πτώσεων. Η μέθοδος αυτή είναι βασισμένη στην καινοτόμα τεχνική εξαγωγής χαρακτηρτιστικών που έχει προταθεί στο [35] για αναγνώριση ανθρώπινης ενέργειας και την δοκιμάζουμε σε δύο καινούρια και απαιτητικά σύνολα δεδομένων. Η μέθοδος περιλαμβάνει τη χρήση του μετασχηματισμού Trace [49] για την κατασκευή ενός συνόλου ανθεκτικών χαρακτηριστικών, τα οποία αναπαριστούν την κάθε χρονοσειρά ενέργειας και μπορούν να ανταπεξέλθουν σε θόρυβο ή αλλοιώσεις που παρουσιάζονται συχνά σε ακολουθίες βίντεο. Η προτεινόμενη τεχνική εκμεταλλεύεται τις φυσικές ιδιότητες του μετασχηματισμού Trace για να παράξει εύρωστα χαρακτηριστικά τα οποία είναι ανεξάρτητα από παραμορφώσεις (μεταφορά, περιστροφή, κλιμάκωση) και είναι απλά, αποδοτικά και κατασκευάζονται γρήγορα. Η μέθοδος αυτή προτείνεται για ένα σύστημα οπτικής αναγνώρισης πτώσεων μίας κάμερας.

Μια σύντομη περιγραφή του μετασχηματισμού Trace και η θεωρία γύρω από αυτόν έχουν περιγραφεί και δίνονται στην ενότητα 3.2. Το υπόλοιπο του κεφαλαίου δομείται ως εξής: Η επισκόπηση του προτεινόμενου συστήματος δίνεται στην ενότητα 4.2.1. Η τεχνική των History Triple Features περιγράφεται στην παράγραφο 4.2.2. Η διαδικασία πειραματισμού και τα αποτελέσματα δίνονται στην ενότητα 4.3.

4.2 Προτεινόμενη Προσέγγιση

4.2.1 Γενική επισκόπηση του προτεινόμενου συστήματος

Εξερευνώντας τις δυνατότητες του μετασχηματισμού Trace επεκτείναμε τη μέθοδο η οποία βασίζεται στα History Trace Templates και είχε αρχικά προταθεί στο [33] δημιουργώντας ακόμη πιο αποτελεσματικά χαρακτηριστικά για την αναγνώριση της ανθρώπινης δραστηριότητας [35]. Τα νέα χαρακτηριστικά, όπως είδαμε στην ενότητα 3.4.2 αποτελούνται από ένα σύνολο από διαιρέσεις τριπλών χαρακτηριστικών και είναι αμετάβλητα σε διάφορες στρεβλώσεις.

Για κάθε βιντεοακολουθία, εξάγεται το υπόβαθρο και οι σιλουέτες των ανθρώπων όπως περιγράφεται σε επόμενη ενότητα. Σε αυτή την περίπτωση, με τη χρήση διαφορετικών συναρτησιακών, για κάθε καρέ υπολογίζεται ένας αριθμός διαφορετικών μετασχηματισμών. Από τους συγκεκριμένους μετασχηματισμούς, δημιουργείται ένα διάνυσμα το οποίο αποτελείται από μια σειρά αμετάβλητων χαρακτηριστικών υπολογισμένα για κάθε καρέ μιας περιόδου της εξεταζόμενης κίνησης. Ολόκληρη η ακολουθία αναπαρίσταται από ένα νέο διάνυσμα με το όνομα *History Triple Features* (HTFs) και είναι ένα σύνολο πραγματικών αριθμών το οποίο έχει πολύτιμη διακριτική πληροφορία για τον εντοπισμό πτώσης. Σε αυτό το σημείο για τη μείωση της διάστασης του διανύσματος μπορεί να χρησιμοποιηθεί η μέθοδος Principal Component Analysis (PCA). Πιο αναλυτική περιγραφή της συγκεκριμένης τεχνικής εξαγωγής χαρακτηριστικών δίνεται στην ενότητα 4.2.2.

4.2.2 Κατασκευάζοντας τα HTFs

Σε αυτή την ενότητα θα περιγράψουμε τον τρόπο με τον οποίο αναπαριστάται μια πτώση με τη χρήση χαρακτηριστικών που εξάγονται από τον μετασχηματισμό Trace και τα οποία ονομάζονται History Triple Features (HTFs). Ο μετασχηματισμός Trace είναι ένας γενικός μετασχηματισμός ο οποίος μπορεί να εφαρμοστεί σε ολόκληρες εικόνες. Είναι γνωστό ότι μπορεί να επιλέξει τόσο χαρακτηριστικά σχήματος όσο και υφής ενός αντικειμένου για την περιγραφή του οποίου χρησιμοποιείται και προσφέρει μια εναλλακτική αναπαράσταση μιας εικόνας [107].

Οι συγγραφείς του [49] έχουν αποδείξει ότι με τη χρήση εξαχθέντων τριπλών χαρακτηριστικών μπορούν να παραχθούν ιδιαιτέρως εύρωστα χαρακτηριστικά για την κατηγοριοποίηση διαφορετικών αλλά πολύ συγγενών κλάσεων (όπως για παράδειγμα διαφορετικά ήδη ψαριών). Η θεωρία πίσω από την κατασκευή των παραπάνω χαρακτηριστικών περιγράφεται στο [49]. Στη συνέχεια, παρουσιάζουμε την κατασκευή των (HTFs) τα οποία προτάθηκαν στο [35].

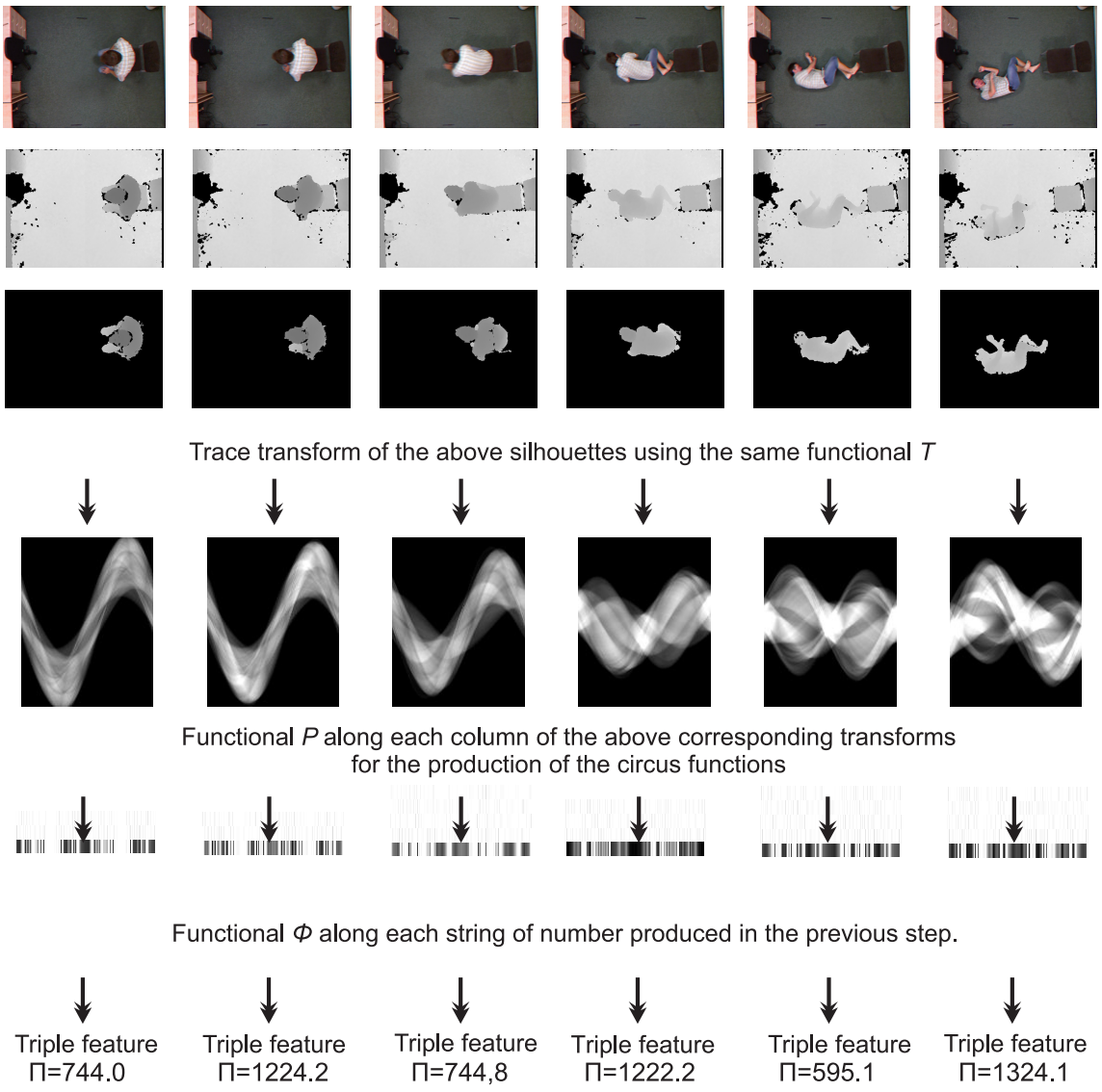
Έχοντας εξάγει τις σιλουέτες, μετασχηματίζουμε το χώρο που περιέχει τις σιλουέτες,

στο χώρο του μετασχηματισμού Trace. Ακολουθώντας την διαδικασία που περιγράφηκε στο [35] για την εξαγωγή των triple features, δημιουργείται ένα σύνολο από τέτοια χαρακτηριστικά . Ο λόγος ενός ζεύγους τέτοιων χαρακτηριστικών όπως έχει δειχθεί, μπορεί να είναι αμετάβλητο σε διάφορα είδη στρεβλώσεων, κατά αντιστοιχία με τα συναρτησιακά που έχουν χρησιμοποιηθεί. Αυτά τα συναρτησιακά μπορούν να έχουν επιλεχθεί ούτως ώστε να είναι ευαίσθητα ή σχετικά ανεπηρέαστα από τις πιθανές μεταβολές που συμβαίνουν στα βίντεο κινήσεων, ενώ την ίδια στιγμή διατηρούν την διακριτικότητά τους.

Έστω ότι $f(p, \vartheta, t)$ είναι μια ακολουθία που περιγράφει μια ανθρώπινη κίνηση. Εφαρμόζοντας ένα Trace συναρτησιακό T, κατά μήκος γραμμών διατρέχοντας το n καρέ που αναφέρεται στην $s_n(p, \theta)$ σιλουέτα, όπου $n = 1...N$ και N είναι ο αριθμός των καρέ, παράγεται ένας μετασχηματισμός Trace $\breve{g}_n(p, \theta)$. Εφαρμόζοντας διαφορετικά T σε κάθε σιλουέτα $s_n(p, \theta)$, παράγεται ένα σύνολο από μετασχηματισμούς $\breve{g}_{n_i}(p, \theta)$. Όπου $i = 1...L$ και L είναι ο αριθμός των μετασχηματισμών που κάποιος επιλέγει να υπολογίσει. Για κάθε $\breve{g}_{n_i}(p, \theta)$ υπολογίζεται ένα σύνολο από $\Pi_{norm}(F, C)$ κανονικοποιημένων τριπλών χαρακτηριστικών. Η διαδικασία για μια πτώση, απεικονίζεται στο Σχήμα 4.1.

Αυτή η μέθοδος επιτρέπει την κατασκευή πολλών χαρακτηριστικών εύκολα. Αν υποθέσουμε ότι κάποιος κάνει χρήση 10 συναρτησιακών για κάθε ένα από τα στάδια κατασκευής(π.χ. 10 συναρτησιακά T, 10 συναρτησιακά P και 10 συναρτησιακά) σε ένα βίντεο αποτελούμενο από 10 καρέ, μπορεί να κατασκευάσει 10x10x10x10=10000 χαρακτηριστικά για μια ακολουθία. Όπως αναφέρθηκε πιο πάνω, αυτοί οι αριθμοί μπορεί να μην έχουν κάποια φυσική σημασία σύμφωνα με την ανθρώπινη αντίληψη, μπορούν όμως να έχουν τις απαιτούμενες μαθηματικές ιδιότητες για τεχνικές κατηγοριοποίησης.

Καθώς η διακριτή ισχύς των κατασκευασμένων χαρακτηριστικών θα ποικίλει αναμφισβήτητα, η εφαρμογή μιας τεχνικής μείωσης της διάστασης θα μπορούσε να παρέχει μια επιλογή από τα πιο διακριτά χαρακτηριστικά ενώ την ίδια στιγμή θα έκανε το πρόβλημα πιο εύκολα διαχειρίσιμο. Στο προτεινόμενο σχήμα, τα διανύσματα HTF τα οποία έχουν παραχθεί με τον τρόπο που αναφέρθηκε, υφίστανται Ανάλυση Κυρίων Συνιστωσών (PCA) με σκοπό να προσδιοριστεί ένας υπο-χώρος ο οποίος να είναι κατάλληλος για κατηγοριοποίηση. Στην πράξη, κρατάμε μόνο ένα υποσύνολο των αρχικών HTF διανυσμάτων που περιέχει τα πιο διακριτά από τα υπολογιζόμενα χαρακτηριστικά ικανά να περιγράψουν ολόκληρη την ακολουθία της κίνησης. Η απόδοση του αλγορίθμου με τη χρήση αλλά και χωρίς Ανάλυσης Κυρίων Συνιστωσών αναφέρεται στην ενότητα 4.3.2.



Σχήμα 4.1: *Εξαγωγή τριπλών χαρακτηριστικών για μια πτώση από τη βάση δεδομένων URFall.*

4.3 Πειραματική διαδικασία και αποτελέσματα

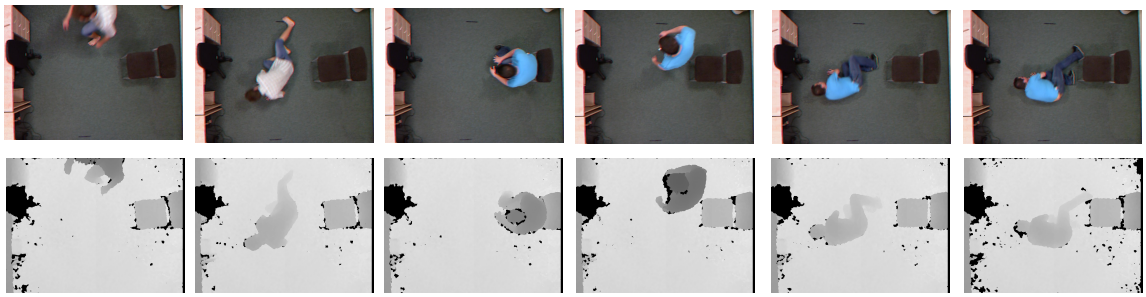
4.3.1 Βάσεις δεδομένων και πρωτόκολλα αξιολόγησης

Σε αυτή την ενότητα θα παρουσιάσουμε τα πειραματικά αποτελέσματα και τη διαδικασία που ακολουθήθηκε για την αξιολόγηση της προτεινόμενης μεθόδου. Γενικά υπάρχει μόνο ένας μικρός αριθμός διαθέσιμων βάσεων δεδομένων ο οποίος να είναι αποκλειστικά αφιερωμένος στις πτώσεις καθώς οι περισσότεροι από τους ερευνητές οι οποίοι έχουν ασχοληθεί με το αντικείμενο έχουν ελέγξει τις τεχνικές τους σε δικά τους δεδομένα. Παρόλα αυτά, προκειμένου να έχουμε ένα σημείο αναφοράς, αξιολογήσαμε την τεχνική μας σε δύο δημοσίως διαθέσιμα σύνολα δεδομένων. Τις βάσεις UR Fall Detection [52] και Le2i Fall detection [14].

Η βάση δεδομένων UR Fall περιέχει 60 ακολουθίες βιντεοσκοπημένες με τη χρήση 2 καμερών Kinect και τις σχετικές μετρικές επιτάχυνσης. Τα δεδομένα αισθητήρα συλλέχθηκαν με τη βοήθεια των συσκευών PS Move (60Hz) και x-IMU (256Hz). Η βάση περιέχει ακολουθίες από δεδομένα βάθους και RGB δεδομένα από δυο κάμερες τοποθετημένες (η μια παράλληλα στο δάπεδο και η άλλη στην οροφή αντίστοιχα) και μετρικά δεδομένα επιτάχυνσης. Κάθε βιντεο-ακολουθία είναι αποθηκευμένη σε διαφορετικούς φακέλους με τη μορφή png αρχείων. Από τη συγκεκριμένη βάση δεδομένων έχουμε χρησιμοποιήσει τα δεδομένα βάθους που προέρχονται από την κάμερα που είναι τοποθετημένη στην οροφή και ακολουθήσαμε το πειραματικό πρωτόκολλο που δίνεται από τους συγγραφείς του [52]. Δείγματα από τη βάση UR Fall δίνονται στο Σχήμα 4.2.

Η βάση Le2i Fall detection έχει ληφθεί σε ρεαλιστικά περιβάλλοντα με τη χρήση μιας απλής RGB κάμερας. Η ταχύτητα λήψης είναι 25 καρέ/δευτερόλεπτο και η ανάλυση είναι 320x240 εικονοστοιχεία (pixels). Τα δεδομένα βίντεο που έχουν ληφθεί απεικονίζουν τις κύριες δυσκολίες που μπορούν να παρουσιαστούν στο χώρο ενός ηλικιωμένου ή σε ένα συνηθισμένο περιβάλλον εργασίας (γραφείο). Τα βίντεο παρουσιάζουν διαφοροποιήσεις στη φωτεινότητα και τυπικές δυσκολίες όπως επικαλύψεις ή ανεπιθύμητο και ιδιαίτερης υφής φόντο. Οι ηθοποιοί εκτέλεσαν ποικίλες δραστηριότητες καθημερινής φύσης καθώς και πτώσεις. Το σύνολο δεδομένων περιέχει 130 υποσημειωμένα βίντεο, με επιπλέον πληροφορία που αναπαριστά τη στάθμη αληθείας (groundtruth) του σημείου πτώσης στην ακολουθία εικόνων. Η βάση δεδομένων παρουσιάζει διαφορετικές τοποθεσίες για έλεγχο και για εκπαίδευση ενώ οι συγγραφείς [14] έχουν ορίσει διάφορα πρωτόκολλα για τον έλεγχο της μεθόδου τους. Δουλεύοντας με τη συγκεκριμένη βάση, ακολουθήσαμε το πρωτόκολλο P1 όπως δίνεται στην προαναφερθείσα δημοσίευση, όπου τα σύνολα για έλεγχο και εκπαίδευση έχουν δημιουργηθεί με βίντεο από τα υποσύνολα ”Home” και ”Coffee room”. Δείγματα από τη βάση Le2i παρέχονται στο Σχήμα 4.3.

Όπως αναφέρθηκε νωρίτερα, για την εξαγωγή χαρακτηριστικών με τη χρήση της προ-



Σχήμα 4.2: *Δείγματα καρέ από τη βάση δεδομένων UR Fall για δύο διαφορετικές πτώσεις. Η επάνω σειρά παρέχει δείγματα RGB ενώ η κάτω δίνει τις αντίστοιχες εικόνες βάθους.*

Σχήμα 4.3: *Δείγματα καρέ από τη βάση Le2i. Η επάνω σειρά αναπαριστά δείγματα από καθημερινές δραστηριότητες στο περιβάλλον "CoffeeRoom" ενώ η κάτω σειρά παρέχει δείγματα από μια πτώση που έχει λάβει χώρα στο περιβάλλον "Home".*

τεινόμενης μεθόδου, έχει πρωτίστως πραγματοποιηθεί εξαγωγή ανθρωπίνων σιλουετών. Για τη βάση δεδομένων UR Fall η εξαγωγή των σιλουετών έχει γίνει υπολογίζοντας τη διαφορά βάθους ανάμεσα σε ένα εικονοστοιχείο του τρέχοντος καρέ και του αντίστοιχου εικονοστοιχείου σε ένα προϋπολογισμένο καρέ αναφοράς. Το καρέ αναφοράς έχει υπολογιστεί υπολογίζοντας τη διάμεση (median) κάθε εικονοστοιχείου βάθους μέσα σε ένα ολισθαίνον παράθυρο των 9 καρέ, σε ένα σύνολο από 80 καρέ που παρουσιάζουν μια άδεια σκηνή (χωρίς την παρουσία ανθρώπου). Έτσι, η μέση τιμή κάθε διάμεσου εικονοστοιχείου υπολογίζεται δημιουργώντας έτσι το τελικό καρέ αναφοράς και εξαλείφοντας μια σημαντική ποσότητα θορύβου που δημιουργείται από την κάμερα βάθους.

Η παρουσία ανθρώπου σε ένα συγκεκριμένο καρέ μπορεί να εντοπιστεί όταν η διαφορά τιμών μεταξύ των εικονοστοιχείων βάθους του τρέχοντος καρέ και του καρέ αναφοράς ξεπεράσει ένα προκαθορισμένο κατώφλι. Προκειμένου να προσθέσουμε ευρωστία στη μέθοδο, χρησιμοποιήσαμε ένα σύνολο τεσσάρων κατωφλίων. Αρχικά, προκειμένου ένα εικονοστοιχείο να είναι έγκυρο, απαιτήσαμε να βρίσκεται στο εύρος τιμών από 1100 έως 3620mm. Στο σημείο αυτό θα πρέπει να θυμίσουμε ότι αυτή η απόσταση αντιστοιχεί στην απόσταση που προκύπτει σε σχέση με την κάμερα βάθους που είναι τοποθετημένη στην οροφή. Στη συνέχεια, προκειμένου η διαφορά μεταξύ των καρέ υπό εξέταση και του καρέ αναφοράς να θεωρηθεί αρκετά σημαντική για τον εντοπισμό ανθρώπινης κίνησης μέσα σε αυτό, προαπαιτείται να βρίσκεται στο εύρος τιμών από 50 έως 2200mm. Αυτές οι τιμές βρέθηκαν να παρουσιάζουν μέγιστη ανοχή στον τυχαίο θόρυβο.

Ο χειρισμός της βάσης δεδομένων Le2i ήταν διαφορετικός δεδομένου ότι αποτελείται από αρχεία βίντεο RGB χαμηλής ανάλυσης. Επιπλέον, οι συνθήκες φωτισμού στις περισσότερες περιπτώσεις (ειδικά στο περιβάλλον "Home") καθιστούν τη χρήση διαφοράς μεταξύ καρέ αναξιόπιστη. Για την εύρεση ανθρώπινης παρουσίας, αξιοποιήσαμε τη μέθοδο διαχωρισμού προσκηνίου-παρασκηνίου που προτάθηκε στο [136] και στο [137] το οποίο βασίζεται στην ιδέα αφαίρεσης μεταξύ του τρέχοντος καρέ και ενός μοντελοποιημένου παρασκηνίου. Αυτό το μοντέλο ανανεώνεται διαρκώς σε επίπεδο εικονοστοιχείου, με τη χρήση μιας μεθόδου βασισμένη σε Γκαουσιανή μίξη (Gaussian mixture). Με τον τρόπο αυτό αποκρίνεται καλύτερα στις αλλαγές σκηνής και περιέχει ότι θεωρείται τελικά ως το στατικό μέρος της σκηνής.

Παρά το γεγονός ότι και οι δύο μέθοδοι που αναφέρθηκαν πιο πάνω δούλεψαν αρκετά καλά εξάγοντας ικανοποιητικά την ανθρώπινη σιλουέτα από την υπόλοιπη εικόνα, η παρουσία δεδομένων με έντονο θόρυβο δεν μπορεί να εξαλειφθεί. Παρόλα αυτά, η μέθοδος εξαγωγής χαρακτηριστικών που παρουσιάζεται στην παρούσα εργασία απέδειξε την ευρωστία της σε αυτού του τύπου τον θόρυβο. Στα πειράματα που διεξήχθησαν οι ακολουθίες είχαν υποστεί υποκλιμάκωση στη χωρική ανάλυση των 320*240 εικονοστοιχείων και έχουν χρονικό μήκος 26 και 12 καρέ για τη βάση UR και τη βάση Le2i αντίστοιχα. Τα

δείγματα εκπαίδευσης κατασκευάστηκαν κατατμίζοντας και στοιχίζοντας χειροκίνητα τις διαθέσιμες ακολουθίες.

Σε αυτό το σημείο, καλό είναι να αναφέρουμε ότι δεν ακολουθείται κάποιο ενιαίο πρότυπο για την αξιολόγηση αλγορίθμων αναγνώρισης πτώσης. Στα πειράματά μας, για να εκτιμηθεί η απόδοση του συστήματος, χρησιμοποιήθηκε η προσέγγιση ενδοπιστοποίησης "leave-one-person-out". Αυτό το συγκεκριμένο πρωτόκολλο επιλέχθηκε επειδή είναι δημοφιλές μεταξύ των ερευνητών. Επίσης, ανακατασκευάζει τις ανάγκες μιας πραγματικής εφαρμογής όσο το δυνατόν πιο πειστικά. Με αυτή την πρακτική, καταγράφεται η φυσική δυναμική συμπεριφορά ενός ατόμου, το σύστημα την επεξεργάζεται και τη συγκρίνει με ένα σύνολο καταγεγραμμένων δεδομένων, με τα οποία έχει προηγουμένως εκπαιδευτεί. Η τελική απόφαση λαμβάνεται με βάση το πόσο σχετική είναι η υπό εξέταση ακολουθία με κάποιο από τα δεδομένα που αποτελούν το σύνολο εκπαίδευσης. Επομένως, το παραπάνω πρωτόκολλο χρησιμοποιεί το δείγμα μιας κίνησης ως δεδομένο ελέγχου και το υπόλοιπο σύνολο ως δεδομένα εκπαίδευσης. Η διαδικασία επαναλαμβάνεται N φορές, όπου N είναι το πλήθος των ακολουθιών κίνησης μέσα στο σύνολο. Η απόδοση υπολογίζεται από τη μέση ακρίβεια των N επαναλήψεων.

Για την κατασκευή των διανυσμάτων χαρακτηριστικών για κάθε σειρά κίνησης χρησιμοποιήθηκαν επτά συναρτησιακά, με τη μεθοδολογία που παρουσιάζεται στο [35]. Η ίδια διαδικασία ακολουθήθηκε τόσο για το σύνολο UR, όσο και για το Le2i. Έπειτα, τα διανύσματα χρησιμοποιήθηκαν για την εκπαίδευση ενός SVM, με τη χρήση του προαναφερθέντος πρωτοκόλλου. Για την αντιστοίχηση των δεδομένων εκπαίδευσης στο χώρο, χρησιμοποιήσαμε συνάρτηση-πυρήνα Γκαουσιανής ακτινικής βάσης (Gaussian Radial Basis Function kernel), με πλήθος τιμών παράγοντα κλιμάκωσης *sigma*. Επιλέξαμε μαλακά περιθώρια και πειραματιστήκαμε με διάφορες τιμές του περιορισμού **C**.

Αρχικά, η πλήρης διαδικασία εκπαίδευσης και ελέγχου του πρωτοκόλλου εκτελέστηκε χρησιμοποιώντας ως είσοδο τα πλήρη διανύσματα. Στη συνέχεια, εφαρμόσαμε Ανάλυση Κυρίων Συνιστωσών (PCA) ξεχωριστά σε κάθε κλάση και επαναλάβαμε τη διαδικασία. Η εφαρμογή της PCA επανελήφθη επίσης για διάφορα μήκη διανύσματος.

4.3.2 Αποτελέσματα

Δείχνεται ότι η μέθοδός μας επιτυγχάνει 100% ακρίβεια και στα δύο σύνολα δεδομένων, αντίστοιχη με το ποσοστό σφάλματος 0% που παρουσιάζει η μέθοδος των Kepski και Kwolek [52] στο σύνολο δεδομένων UR και με το ποσοστό ακρίβειας 99.6% που δίνεται από τη μέθοδο της μελέτης από τους Charfi και συν. [14] που συνοδεύει τη βάση Le2i. Ωστόσο, οι μέθοδοι αυτές δείχνουν να είναι αρκετά εξειδικευμένες πάνω σε συγκεκριμένα χαρακτηριστικά του σχήματος και του πλαισίου οριοθέτησης της ανθρώπινης σιλουέτας, με έντονη εξάρτηση από τη σχετική θέση της κάμερας. Η μέθοδος που προτείνεται σε αυτό

το άρθρο είναι πιο γενικευμένη και μπορεί να παράξει χαρακτηριστικά από το σύνολο των μεταβάσεων του σχήματος του ανθρώπου ανάμεσα στα πλαίσια της ροής του βίντεο. Επίσης, παραβλέπει χαρακτηριστικά της σκηνής που δεν έχουν τίποτα να προσφέρουν στη διαχωρισιμότητα των κλάσεων. Τα αποτελέσματα των πειραμάτων μας φαίνονται στους Πίνακες 4.1, 4.2 και 4.3.

Πίνακας 4.1: *Η απόδοση του SVM στο σύνολο δεδομένων πτώσης UR (δυαδικές σιλουέτες).*

	using PCA	w/o PCA
Best score	100%	87.76%
Sigma	2	10
C	0.0001	0.0001
Vector length	8	1014

Πίνακας 4.2: *Η απόδοση του SVM στο σύνολο δεδομένων πτώσης UR (σιλουέτες βάθους).*

	using PCA	w/o PCA
Best score	100%	87.76%
Sigma	3	10
C	0.0001	0.0001
Vector length	7	1014

Πίνακας 4.3: *Η απόδοση του SVM στο σύνολο δεδομένων πτώσης Le2i*

	using PCA	w/o PCA
Best score	100%	96.6%
Sigma	3	10
C	0.0001	10
Vector length	47	468

Κεφάλαιο 5

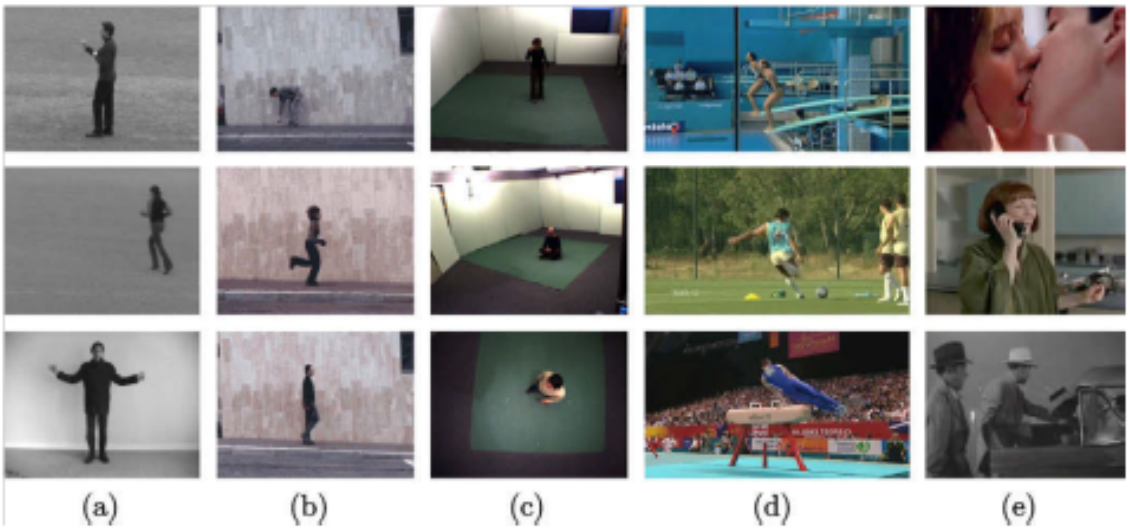
THETIS: THree Dimensional Tennis Shots. Βάση δεδομένων ανθρωπίνων κινήσεων

5.1 Εισαγωγή

Τα τελευταία χρόνια μεγάλες συλλογές από απλές αλλά και πιο σύνθετες καταγεγραμμένες κινήσεις 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.

Η μεγάλη αξία των βάσεων δεδομένων ή αλλιώς datasets που περιέχουν καταγεγραμμένες ανθρώπινες δραστηριότητες σε βίντεο, 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 [100] και η βάση Weizmann [6] που σχεδιάστηκαν για την ακαδημαϊκή αξιολόγηση συστημάτων αναγνώρισης κινήσεων γενικού σκοπού. Περιλαμβάνουν βίντεο διαφόρων ατόμων που πραγματοποιούν απλές κινήσεις, όπως ”περπατώ” και ”γνέφω” σε ένα ελεγχόμενο περιβάλλον. Τα σύνολα δεδομένων 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”



Σχήμα 5.1: *(a) KTH dataset, (b) Weizmann dataset, (c) Inria XMAS dataset, (d) UCF sports action dataset και (e) Hollywood human action dataset.*

και στοχεύει σε εφαρμογές επιτήρησης. Τέλος, έχουν κατασκευαστεί και παρουσιαστεί βάσεις που προέκυψαν από τη συλλογή πραγματικών βίντεο από μέσα ενημέρωσης, όπως τηλεοπτικές εκπομπές και ταινίες. Στο Σχήμα 5.1 απεικονίζονται στιγμιότυπα από δημοφιλείς βάσεις δεδομένων.

5.1.1 Σημαντικά σύνολα δεδομένων

Ένα σημαντικό πλήθος ερευνητών έχουν δοκιμάσει το σύστημά τους στη βάση KTH. Η βάση KTH περιέχει 2391 βίντεο, έξι δραστηριοτήτων εκτελεσμένων από 25 άτομα. Οι δραστηριότητες που περιλαμβάνονται είναι: Walking, jogging, running, boxing, hand-waving και hand-clapping. Τα βίντεο έχουν καταγραφεί σε ελαφρώς διαφορετικές κλίμακες τόσο σε εξωτερικό όσο και σε εσωτερικό περιβάλλον, με μόνο ένα άτομο στο σκηνικό. Κάθε βίντεο 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 βίντεο. Εδώ το σκηνικό είναι απλό και παραμένει ίδιο για όλα τα βίντεο. 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 βίντεο 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 βίντεο περιέχονται εκτελέσεις μιας απλής κίνησης και σκοπός είναι να αναγνωριστεί η κλάση της κίνησης του βίντεο, δεδομένου ότι αυτό ανήκει σε ένα περιο-

ρισμένο αριθμό γνωστών κλάσεων. Η δοκιμή μεθοδολογιών που χρησιμοποιούν χωροχρονικά τοπικά χαρακτηριστικά 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 χωρο-χρονικά 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 βίντεο 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 βίντεο, καταλήγοντας σε ένα σύνολο από 28 βίντεο με ανάλυση 384x288, 25fps. Τέλος, τα βίντεο 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 βίντεο 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 βίντεο καταγράφηκαν σε ένα υπόγειο σταθμό τρένου στο Λονδίνο από μια οπτική γωνία, σε συνωστισμένο περιβάλλον. Η ανάλυση των βίντεο είναι 720x576, 25fps και δεν περιέχουν μόνο άτομα και αντικείμενα, αλλά και διερ-

χόμενα τρένα όπου κόσμος επιβιβάζεται και αποβιβάζεται. Για σκοπούς εκπαίδευσης και επαλήθευσης, δημιουργήθηκαν τρία βίντεο ενώ ένα μεγαλύτερης χρονικής διάρκειας βίντεο με έξι δραστηριότητες εγκατάλειψης αποσκευών χρησιμοποιήθηκε για δοκιμή.

Τέλος, υπάρχουν βάσεις που προέκυψαν από τη συλλογή βίντεο τηλεοπτικών εκπομπών και ταινιών. Οι διαφορές τους με τις υπόλοιπες βάσεις είναι πως τα βίντεο 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 [50], [58], [92], επικεντρώθηκαν σε απλές κινήσεις όπως για παράδειγμα "φιλώ" και "κτυπώ". Παρ' όλο που οι κινήσεις είναι απλές, κάθε βίντεο 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 [66] προέκυψε από την συλλογή βίντεο 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 βίντεο με διάρκεια 20.1 ώρες περίπου. Στόχος της συγκεκριμένης 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 βίντεο είναι η βάση UCF sport [92]. Περιέχει 10 κλάσεις κινήσεων: swinging (on the pommel horse and on the floor), diving, kicking (a ball), weightlifting, horseriding, running, skateboarding, swinging, golf swinging και walking και αποτελείται συνολικά από 200 βίντεο με ανάλυση 720x480. Οι κινήσεις προέρχονται από ποικίλες αθλητικές δραστηριότητες που έχουν παρουσιαστεί σε τηλεοπτικές εκπομπές των δικτύων BBC και ECPN.

Τέλος, το σύνολο κινήσεων YouTube [60] 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 βίντεο. 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 στον Πίνακα 5.1 μαζί με μια συνοπτική περιγραφή τους.

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 βασι-

Πίνακας 5.1: *Δημοφιλή σύνολα δεδομένων κίνησης*

Ονομασία Συνόλου Δεδομένων	**Κλάσεις**	**Βίντεο**	**Συνθήκες Καταγραφής**	**Περιγραφή δεδομένων**
UMD [119]	10	100	Εργαστήριο: 1 άτομο, πολλές επαναλήψεις	Ανάλυση 300px
IXMAS [123]	11	110	Εργαστήριο: 10 άτομα, 5 γωνίες λήψεις με πολλές κάμερες	Ανάλυση 100-200 px, πολύ μικρής διάρκειας λήψεις
Olympic games [69]	17	166	Βίντεο από τους Ολυμπιακούς Αγώνες	5065 frames: υψηλή διαφοροποίηση στην ίδια κλάση, σημαντική κίνηση της κάμερας, θολή εικόνα λόγω κίνησης, διαφοροποιήσεις στην εμφάνιση
UFC [128]	2	20min	Βίντεο από τηλεοπτικές εκπομπές	Αλλαγές σε εμφάνιση, γωνία λήψης, κίνηση κάμερας, ταυτόχρονη εκτέλεση από πολλά άτομα
ADL [68]	10	150	Εργαστήριο: 5 άτομα, 3 επαναλήψεις	Σύνθετες, δραστηριότητες, στατικό background, υψηλή ανάλυση, 240x450px, 24fps
High Five [86]	4	300	23 διαφορετικά TV shows	30-600 frames, ρεαλιστική αλληλεπίδραση ατόμων
MSR II [79]	3	54	Σε περιβάλλον συνωστισμού	Πολλά άτομα, 203 στιγμιότυπα, 320x240px, 15fps
Youtube Olympic Sports [130]	16	800	Βίντεο από το Youtube	Σύνθετες δραστηριότητες, 50 σενάρια για κάθε κλάση.

κές κινήσεις του αθλήματος της αντισφαίρισης (tennis). Ακολουθώντας μια συγκεκριμένη διαδικασία που περιγράφεται στη συνέχεια, με τη χρήση της κάμερας τρισδιάστατης λήψης Kinect καταγράφηκαν συστηματικά βίντεο μερικών ωρών, που περιέχουν συγκεκριμένες κινήσεις αντισφαίρισης εκτελεσμένες από 55 διαφορετικά άτομα. Η βάση στην τελική της μορφή αποτελείται από 8374 βίντεο 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 γεγονότων 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 η οποία περιλαμβάνει κινήσεις της καθημερινότητας (π.χ. ”τρώω”, ”μιλώ στο κινητό τηλέφωνο”, ”διαβάζω ένα βιβλίο” κ.α.) και τη βάση δεδομένων G3D [7] 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 5.2 παρουσιάζονται οι λεπτομέρειες που αφορούν στις συνθήκες και στα μέσα καταγραφής των δεδομένων κίνησης. Στη συνέχεια, στην ενότητα 5.3 δίνεται μια λεπτομερής περιγραφή όλων των δεδομένων κίνησης που περιλαμβάνονται στη βάση. Η ενότητα 5.4 παρέχει πληροφορίες για τα εργαλεία που χρησιμοποιήθηκαν για την μετατροπή των αρχείων της βάσης στην τελική τους μορφή. Τέλος στην ενότητα 5.5 περιγράφεται η πειραματική διαδικασία που διεξήχθη για την αξιολόγηση της βάσης.

5.2 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 παρουσιάζει τα δικά της πλεονεκτήματα και μειονεκτήματα. Για τη βάση THETIS, χρησιμοποιήθηκε η συσκευή ανίχνευσης-καταγραφής κίνησης KINECT, της MICROSOFT που δημιουργήθηκε για την παιχνιδομηχανή XBOX 360 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 5.2.1.

5.2.1 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 βίντεο τριών χωρικών διαστάσεων. Το εύρος του αισθητήρα προσαρμόζεται αυτόνομα από το λογισμικό που τον ρυθμίζει κατάλληλα με βάση το φυσικό περιβάλλον.

Ως προς τα τεχνικά χαρακτηριστικά και την ακρίβεια του Kinect πρέπει να αναφερθεί ότι καταγράφει βίντεο 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 βίντεο σε ανάλυση 680x480 pixels ή 1280x1024 σε μικρό fps.

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 57 οριζόντια, 43 κάθετα και 70 διαγώνια. Το σύστημα στήριξης μπορεί να προσφέρει μετατόπιση στη γωνία λήψης 27 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 καλώδιο ρεύματος).

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 εικονοστοιχείο 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 γωνία μπορεί να υπολογιστεί από την τριγωνομετρία ως εξής:

$$\theta = arctan\frac{D}{L} \tag{5.1}$$

όπου η απόσταση μεταξύ του πομπού υπερύθρων και δέκτη είναι $D = 0.075\text{m}$ 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: $N = -4636,3\theta + 1092,5$. Εισάγοντας την παραπάνω εξίσωση για το θ, σε αυτήν έχουμε:

$$N = -4636.3arctan\frac{0.075}{L} + 1092.5 \tag{5.2}$$

$$L = -\frac{0.075}{tan(0.0002157N - 0.2356)}[m] \tag{5.3}$$

Η εξίσωση αυτή υπολογίζει την πραγματική απόσταση για μια δεδομένη τιμή . Ακόμα καλύτερα αποτελέσματα μπορούν να προκύψουν με τον επανυπολογισμό των N και θ, μετά από βαθμονόμηση.

5.2.2 OpenNI framework

Για την καταγραφή, σε συνδυασμό με τη συσκευή Kinect χρησιμοποιήθηκε το πλαίσιο 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.

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:

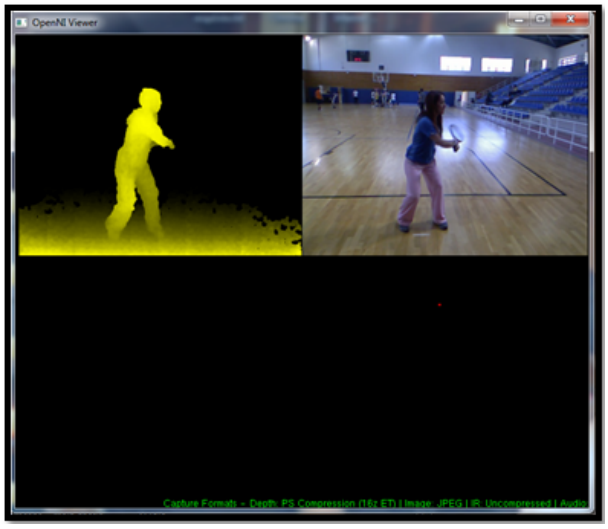
- Το διαχωρισμό του προσκηνίου (foreground) από το παρασκήνιο (background).
- Την κάτοψη του χώρου.
- Την αναγνώριση μεμονωμένων ατόμων στη σκηνή.

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

Η καταγραφή των δεδομένων κίνησης πραγματοποιήθηκε μέσω της εφαρμογής NiViewer του OpenNI ως περιγράφεται ακολούθως (Σχήμα 5.2). 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 5.4.

5.2.3 Συνθήκες Καταγραφής

Η καταγραφή των κινήσεων αντισφαίρισης που αποτελούν την βάση THETIS πραγματοποιήθηκε σε δυο διαφορετικούς εσωτερικούς χώρους στο Αθλητικό Κέντρο Εθνικού



Σχήμα 5.2: *Απεικόνιση του depth map και του image map από το NiViewer κατά τη διαδικασία καταγραφής.*

Μετσόβιου Πολυτεχνείου και στον Όμιλο Αντισφαίρισης Γλυφάδας. Η επιλογή δυο εσωτερικών χώρων για την καταγραφή των βίντεο, οφείλεται στην ευαισθησία της συσκευής Kinect στο ηλιακό φως. Συγκεκριμένα, εξαιτίας της κάμερας IR υπέρυρθων ακτίνων 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 5.3.

Όσον αφορά στο παρασκήνιο (background) των βίντεο, 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 5.3: *Επίδειξη της κίνησης backhand από την εκπαιδεύτρια αντισφαίρισης.*

Σχήμα 5.4: *Διαφοροποιήσεις στο background κατά τη διάρκεια διαφορετικών λήψεων.*

Επιπροσθέτως, στα βίντεο 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 λήψης 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 φόντο φαίνεται στο Σχήμα 5.4.

5.3 Δομή της βάσης δεδομένων THETIS

Η βάση δεδομένων κίνησης THETIS περιλαμβάνει 8374 βίντεο 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 δίνονται παρακάτω. Η χρήση των αγγλικών ονομάτων προτιμάται καθώς συμπίπτει με τη διεθνώς αποδεκτή ορολογία για το άθλημα.

1. Backhand with two hands

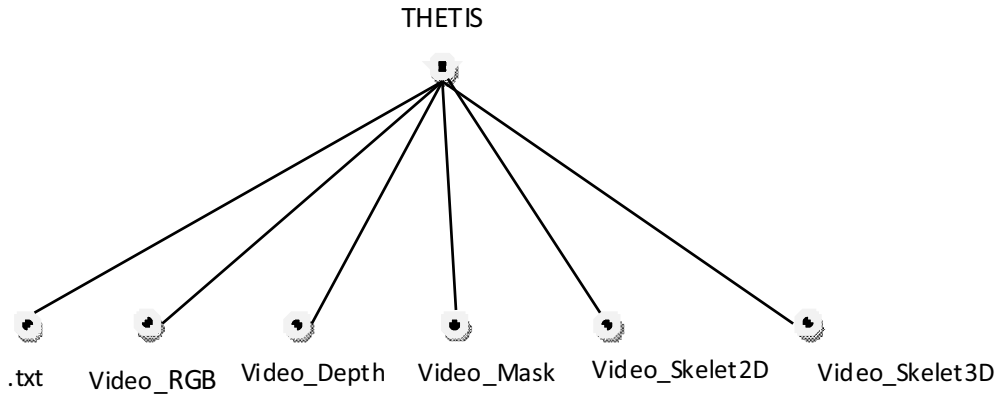
2. Backhand

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 σε αρχεία AVI με τη χρήση εφαρμογής που περιγράφεται στην ενότητα 5.4, βασισμένης στο πλαίσιο της OpenNI. Για κάθε αρχείο ONI δημιουργούνται ταυτόχρονα πέντε AVI αρχεία, ίσης διάρκειας. Συγκεκριμένα, δημιουργούνται:

- 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 διότι υφίστανται και άλλη επεξεργασία αφού κόπτονται σε επιμέρους βίντεο χειρωνακτικά. Στόχος της διαδικασίας κοψίματος είναι η δημιουργία από κάθε βίντεο τριών νέων, που το κάθε ένα θα περιέχει μόνο μια πλήρη επανάληψη της εκάστοτε κίνησης. Έτσι, προκύπτουν 1980 βίντεο RGB, 1980 βίντεο depth και 1980 βίντεο mask (silhouette). Όσον αφορά τα βίντεο που απεικονίζουν το σκελετό είτε σε δυο είτε σε τρεις διαστάσεις, δεν

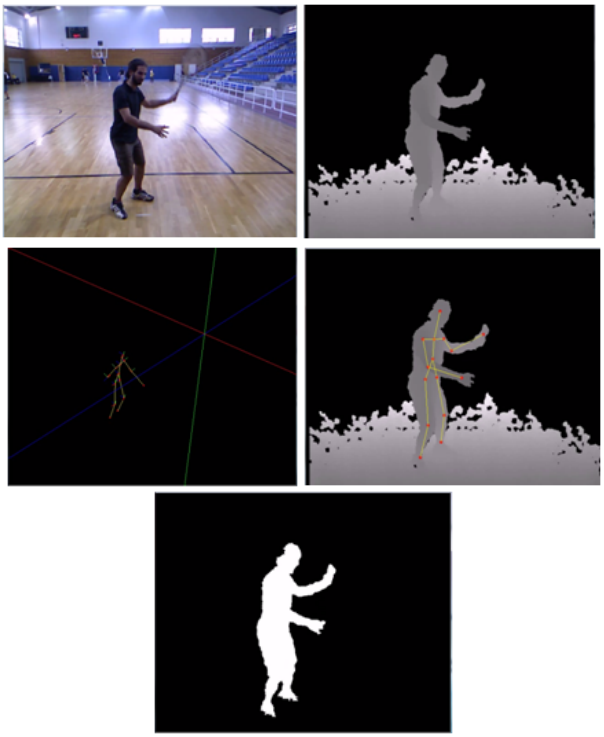


Σχήμα 5.5: *Δομή της βάσης δεδομένων THETIS.*

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 βίντεο σκελετού σε δυο διαστάσεις και 1217 βίντεο σκελετού σε τρεις διαστάσεις. Το σύνολο των δεδομένων THETIS χωρίζεται σε έξι υποφακέλους, όπως φαίνεται στο Σχήμα 5.5.

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_Mas: 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).



Σχήμα 5.6: *Στιγμιότυπα από την κίνηση forehand slice, εκτελεσμένη από το ίδιο άτομο για όλους τους τύπους βίντεο στη βάση THETIS. RGB, Depth, Skelet3D, Skelet2D και Mask αντίστοιχα.*

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

Στο Σχήμα 5.6 απεικονίζονται στιγμιότυπα του ίδιου ατόμου, να εκτελεί την κίνηση forehand slice, από όλες τις κατηγορίες βίντεο της βάσης.

5.4 Εργαλεία

5.4.1 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 μετατροπή 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. 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.

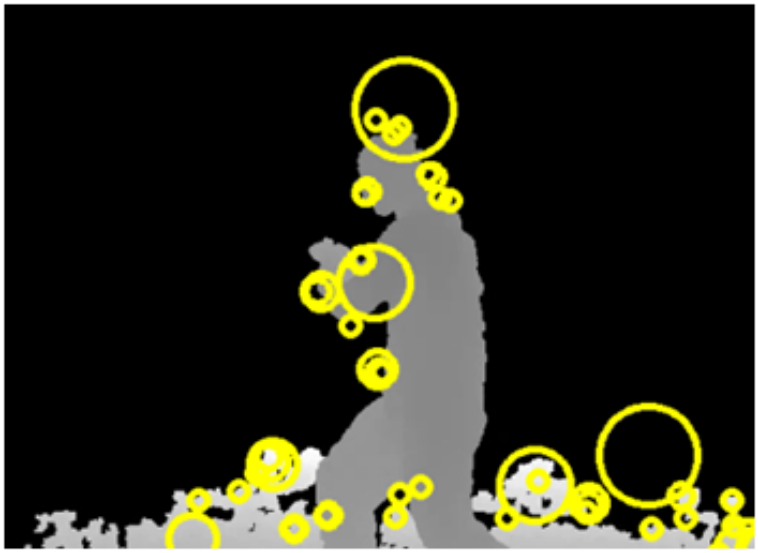
5.4.2 Περικοπή των AVI αρχείων

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.

5.5 Διεξαγωγή πειραμάτων

Κατά το τελευταίο στάδιο της δημιουργίας της βάσης δεδομένων THETIS, πραγματοποιήθηκε η πειραματική δοκιμή της με την εφαρμογή δυο μεθόδων αναγνώρισης δραστηριότητας που έχουν ήδη προταθεί. Για τον σκοπό αυτό, όλα τα βίντεο που περιέχουν την αναπαράσταση του σκελετού σε τρεις διαστάσεις (Skelet3D βίντεο), καθώς και τα βίντεο που περιέχουν την πληροφορία του βάθους (Depth βίντεο), κωδικοποιούνται με περιγραφείς τελευταίας τεχνολογίας, που θα αναφερθούν ακολούθως. Στη συνέχεια, αφού κβαντοποιηθούν οι περιγραφείς των βίντεο, εισάγονται ως είσοδοι σε μη-γραμμικές μηχανές διανυσμάτων υποστήριξης (nonlinear SVM) για την ταξινόμηση των βίντεο 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.



Σχήμα 5.7: *Εφαρμογή του αλγορίθμου STIPs σε βίντεο της βάσης THETIS τύπου δεδομένων βάθους.*

5.5.1 Μέθοδοι εξαγωγής περιγραφέων

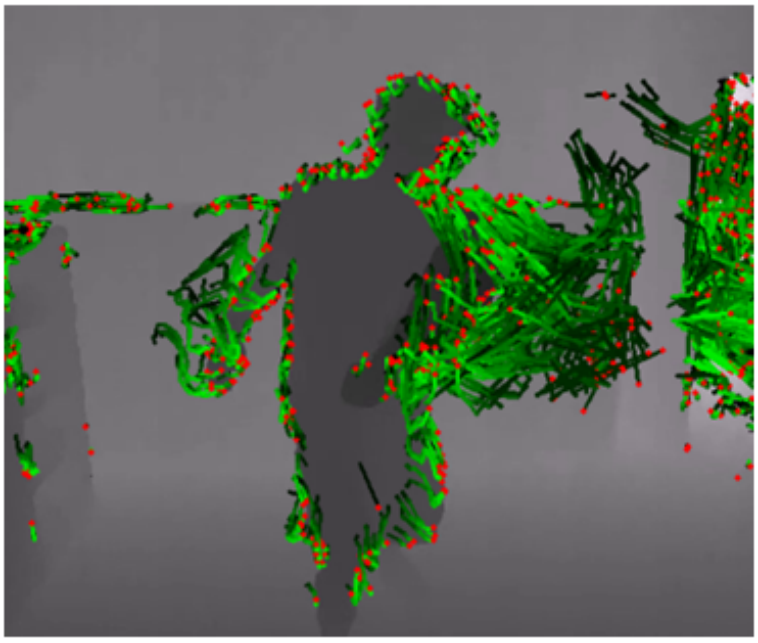
Μέθοδος Space-Time Interest Points (STIPs)

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 χωρο-χρονικών σημείων ενδιαφέροντος και την εξαγωγή των περιγραφέων τους, ακολουθήσαμε τη μέθοδο που χρησιμοποιήθηκε στο [98] και χρησιμοποιήσαμε τον κώδικα που παρέχουν οι συγγραφείς. Ο κώδικας προεκτείνει τον ανιχνευτή Harris 3D (Harris 3D Detector), των Laptev και Lindeberg [56] που εντοπίζει χωρο-χρονικά σημεία ενδιαφέροντος (Σχήμα 5.7) και υπολογίζει τους τοπικούς χωρο-χρονικούς περιγραφείς Ιστογράμματα Προσανατολισμένης Κλίσης (Histograms of Oriented Gradient-HOG) και Ιστογράμματα Οπτικής Ροής (Histograms of Optical Flow - HOF). Όπως στο [56], χρησιμοποιούμε την έκδοση εκείνη του κώδικα που δεν χρησιμοποιεί επιλογή κλίμακας, αλλά αντίθετα ένα σύνολο πολλαπλών συνδυασμών από χωρικές και χρονικές κλίμακες.

Μέθοδος Dense Trajectories

Η δεύτερη μέθοδος αναγνώρισης ανθρώπινης δραστηριότητας σε βίντεο, που εφαρμόστηκε στα βίντεο της βάσης δεδομένων THETIS (βίντεο που απεικονίζουν το βάθος και τον σκελετό σε τρεις διαστάσεις) είναι η μέθοδος Dense Trajectories ή πυκνές τροχιές που προτάθηκε από τον Wang και τους συνεργάτες του [120]. Η μέθοδος Dense



Σχήμα 5.8: *Εφαρμογή της μεθόδου Dense Trajectories σε βίντεο της βάσης THETIS που απεικονίζει το βάθος.*

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 βίντεο (Σχήμα 5.8).

Επιπλέον για την αντιμετώπιση των προβλημάτων που προέρχονται από την κίνηση της κάμερας, στο [120] 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) [17].

5.5.2 Αποτελέσματα μεθόδου STIPs

Στην ενότητα αυτή παρουσιάζονται τα αποτελέσματα που προέκυψαν από την εφαρμογή της μεθόδου STIPs στα δεδομένα των συνόλων THETIS_Depth και THETIS_Skelet3D για τον συνδυασμό των δυο περιγραφέων HOG και HOF. Στον Πίνακα 5.2 καταγράφεται ο μέσος όρος ακρίβειας (accuracy) των αποτελεσμάτων ταξινόμησης και για τα τρία σύνολα δεδομένων. Στο συγκεκριμένο πίνακα όπως και σε όλους τους ακόλουθους που αφορούν σε αποτελέσματα στη βάση THETIS, προς χάριν χώρου, η αρίθμηση των κινήσεων συμφωνεί με την αρίθμηση που παρουσιάστηκε στην ενότητα 5.3.

Στον Πίνακα 5.3 παρουσιάζεται ο πίνακας σύγχυσης (confusion matrix) από την εφαρμογή της μεθόδου STIPs στο σύνολο δεδομένων THETIS_Depth. Στον Πίνακα 5.4 δίνεται

Πίνακας 5.2: *Μέσος όρος ακρίβειας ταξινόμησης για τη μέθοδο STIPs.*

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

Πίνακας 5.3: *Πίνακας σύγχυσης σε ποσοστά % για το σύνολο THETIS_Depth.*

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

ο πίνακας σύγχυσης σε ποσοστά % για το σύνολο THETIS_Skelet3D.

Στους Πίνακες 5.5 παρουσιάζονται πιο αναλυτικά τα αποτελέσματα ταξινόμησης για κάθε κλάση κινήσεων, από την εφαρμογή της μεθόδου STIPs στο σύνολο δεδομένων 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 5.3 και 5.4. 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.

Πίνακας 5.4: *Πίνακας σύγχυσης σε ποσοστά % για το σύνολο THETIS_Skelet3D.*

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

Πίνακας 5.5: *Πίνακας σύγχυσης σε απόλυτες τιμές για το σύνολο KTH.*

Κινήσεις	Boxing	Handclapping	Handwaving	Jogging	Running	Walking
Boxing	99	1	0	0	0	0
Handclapping	2	97	0	0	0	0
Handwaving	3	1	96	0	0	0
Jogging	0	0	0	86	10	4
Running	0	0	0	18	80	2
Walking	0	0	0	1	0	99

Πίνακας 5.6: *Μέσος όρος ακρίβειας ταξινόμησης (average accuracy) με τη μέθοδο Dense Trajectories, για διαφορετικούς συνδυασμούς περιγραφέων.*

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%

Πίνακας 5.7: *Πίνακας σύγχυσης 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	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

5.5.3 Αποτελέσματα μεθόδου Dense Trajectories

Στην ενότητα αυτή παρουσιάζονται τα αποτελέσματα που προέκυψαν από την εφαρμογή της μεθόδου Dense Trajectories στα δεδομένα των συνόλων THETIS_Depth, THETIS_Skelet3D και KTH. Για κάθε σύνολο πραγματοποιήθηκαν τρία 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. Στον Πίνακα 5.6 καταγράφεται ο μέσος όρος ακρίβειας (accuracy) των αποτελεσμάτων ταξινόμησης και για τα τρία σύνολα δεδομένων, για κάθε διαφορετικό περιγραφέα.

Στους Πίνακες 5.7, 5.8, 5.9, 5.10, 5.11 και 5.12, 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.

Στους Πίνακες 5.13, 5.14, 5.15 και 5.16 παρουσιάζονται πιο αναλυτικά τα αποτελέσματα ταξινόμησης για κάθε κλάση κινήσεων για την εφαρμογή των διαφόρων περιγραφέων της μεθόδου Dense Trajectories στο σύνολο δεδομένων 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 χαμηλό-

Πίνακας 5.8: *Πίνακας σύγχυσης 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	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

Πίνακας 5.9: *Πίνακας σύγχυσης σε απόλυτες τιμές για το σύνολο THETIS_Depth, με περιγραφείς Trajectory, HOG, HOF και 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

Πίνακας 5.10: *Πίνακας σύγχυσης 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	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

Πίνακας 5.11: *Πίνακας σύγχυσης 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	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

Πίνακας 5.12: *Πίνακας σύγχυσης σε απόλυτες τιμές για το σύνολο THETIS_ Skelet3D, με περιγραφείς Trajectory, HOG, HOF και 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

Πίνακας 5.13: *Ποσοστά precision και accuracy κάθε κλάσης για το σύνολο KTH, για τους διάφορους περιγραφείς.*

	Precision			**Accuracy**		
Είδος Κίνησης	**Trajectory, HOF, MBH**	**MBH**	**Trajectory, HOG, HOF, MBH**	**Trajectory**	**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%

Πίνακας 5.14: *Πίνακας σύγχυσης σε απόλυτες τιμές για το σύνολο KTH, με περιγραφέα Trajectory.*

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

Πίνακας 5.15: *Πίνακας σύγχυσης σε απόλυτες τιμές για το σύνολο KTH, με περιγραφέα MBH.*

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

Πίνακας 5.16: *Πίνακας σύγχυσης σε απόλυτες τιμές για το σύνολο KTH, με περιγραφείς Trajectory, HOG, HOF και 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

Πίνακας 5.17: *Σύγκριση των αποτελεσμάτων των μεθόδων Dense Trajectories και STIPs, για όλα τα σύνολα δεδομένων που χρησιμοποιήθηκαν στην πειραματική διαδικασία.*

Σύνολο Δεδομένων	**Dense Trajectories**			**STIPs**
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%

τερα. Επίσης, 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. 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.

5.5.4 Συγκριτικά αποτελέσματα

Στην ενότητα αυτή, παρουσιάζονται τα αποτελέσματα της μεθόδου STIPs σε αντιπαράθεση με τα αποτελέσματα της μεθόδου Dense Trajectories. Όπως προκύπτει και από τον συγκριτικό Πίνακα 5.17, η εφαρμογή της μεθόδου STIPs οδήγησε σε καλύτερα αποτελέσματα τόσο για τα βίντεο της βάσης KTH, όσο και για τα βίντεο της παρουσιαζόμενης βάσης δεδομένων THETIS.

Ειδικότερα, παρατηρούμε πως η καλύτερη επίδοση του συνόλου THETIS_Depth σημειώνεται για τη μέθοδο STIPs (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 STIPs. Επιπροσθέτως, και στα τρία 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 τρία είδη service που είναι διαχωρίσιμα μόνο από έμπειρους παίκτες, ενώ είναι πολύ δύσκολο να γίνει διάκριση μεταξύ τους από ένα μέσο θεατή.

Στο σημείο αυτό, κρίνεται απαραίτητο να σημειωθεί πως στα βίντεο 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 STIPs και Dense Trajectories του συνόλου δεδομένων KTH με εκείνα του συνόλου THETIS_Skelet3D, καθώς το είδος της πληροφορίας που έχει καταγραφεί στα βίντεο 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.

Κεφάλαιο 6

Μετασχηματισμός 3D Cylindrical Trace για την κατηγοριοποίηση χωρο-χρονικών ακολουθιών

6.1 Εισαγωγή

Η δουλειά η οποία περιγράφεται στο τρέχον κεφάλαιο, είναι μια προέκταση της δουλειάς η οποία δημοσιεύθηκε στο [35] και παρουσιάστηκε στο Κεφάλαιο 3. Στη προηγούμενη πρότασή μας, εξετάσαμε τη δυνατότητα του μετασχηματισμού Trace για τη χρήση του στην αναγνώριση ανθρώπινης δραστηριότητας και προτείναμε δύο καινοτόμες τεχνικές εξαγωγής χαρακτηριστικών για τη συγκεκριμένη διεργασία. Η προτεινόμενη τεχνική επιτύγχανε να παράγει πολύ εύρωστα στο θόρυβο χαρακτηριστικά και αποδείχθηκε ιδιαιτέρως ικανοποιητική για την επιτυχή αναγνώριση ανθρώπινης δραστηριότητας όταν εξετάστηκε με την εφαρμογή της σε δυο δημοφιλείς βάσεις δεδομένων. Επιπρόσθετα, η δεύτερη μέθοδος η οποία προτάθηκε, παράγει εύρωστα στο θόρυβο χαρακτηριστικά τα οποία είναι αμετάβλητα σε διάφορες συνήθεις σε βίντεο παραμορφώσεις, όπως η περιστροφή, η μετατόπιση και η κλιμάκωση και εξετάστηκε με επιτυχία σε τέσσερα απαιτητικά σύνολα δεδομένων.

Στο συγκεκριμένο κεφάλαιο, παρουσιάζουμε το σχεδιασμό μιας νέας μορφής του μετασχηματισμού Trace επεκτείνοντας τις δυνατότητές του στον τρισδιάστατο χώρο και προτείνοντας μια καινοτόμο μέθοδο για την εξαγωγή χαρακτηριστικών για την αναγνώριση δραστηριοτήτων από βίντεο. Πιο συγκεκριμένα, προτείνουμε μια κυλινδρική μορφή του μετασχηματισμού Trace η οποία μπορεί να εφαρμοστεί σε τρισδιάστατα δεδομένα όπως είναι μια χωρο-χρονική ακολουθία. Με την εφαρμογή διαφορετικών συναρτησιακών για τον υπολογισμό διαφορετικών μετασχηματισμών Trace και υπολογίζοντας τα τριπλά χαρακτηριστικά (Triple features) όπως είδαμε στο κεφάλαιο 3.4, ένα σύνολο χαρακτηριστι-

κών αμετάβλητων στις προαναφερθείσες στρεβλώσεις, μπορεί να περιγράψει ολόκληρο το εξεταζόμενο βίντεο. Η προτεινόμενη μέθοδος συνδυάζεται με τα *επιλεκτικά χωρο-χρονικά σημεία ενδιαφέροντος* (Selective Spatio-Temporal Interest Points (STIPs)), τα οποία προτάθηκαν στο [13], προκειμένου να προσαρμοστεί ακόμα καλύτερα στη χρονική φύση της κίνησης και με σκοπό να τονίσει τη σημασία των διακριτών χρονικών χαρακτηριστικών στο τελικό αντιπροσωπευτικό διάνυσμα.

Η τεχνική δοκιμάστηκε επάνω σε δύο διαφορετικά σενάρια. Αυτό της αναγνώρισης ανθρώπινης δραστηριότητας και αυτό του εντοπισμού πτώσεων. Οι βάσεις που χρησιμοποιήθηκαν για την περίπτωση της αναγνώρισης δράσης, ήταν η KTH, η Weizmann και η THETIS ενώ για την περίπτωση των πτώσεων, η UR Fall και η Le2i Fall αντίστοιχα. Τα αποτελέσματα έδωσαν μια εντυπωσιακή απόδοση σε όλα τα διαφορετικά σύνολα δεδομένων αναδεικνύοντας τις δυνατότητες της προτεινόμενης μεθόδου.

Με την καλύτερη γνώση που μπορεί να έχουν οι συγγραφείς, αυτή είναι η πρώτη φορά που προτείνεται μια τρισδιάστατη μορφή του μετασχηματισμού Trace και που χρησιμοποιείται για οποιαδήποτε διεργασία στην αναγνώριση προτύπων. Είναι επίσης η πρώτη φορά που χρησιμοποιείται μια τέτοια μορφή του μετασχηματισμού για την εξαγωγή χαρακτηριστικών ικανών να χρησιμοποιηθούν για τη κατηγοριοποίηση δράσεων από βίντεο.

Το υπόλοιπο του κεφαλαίου έχει οργανωθεί ως εξής: Οι βασικές αρχές που διέπουν τον μετασχηματισμό Trace και το μετασχηματισμό 3D Radon, ο οποίος αποτελεί την πηγή έμπνευσης για τον προτεινόμενο μετασχηματισμό 3D Cylindric Trace, παρουσιάζονται στην ενότητα 6.2. Η παρουσίαση και η σημειογραφία για τον προτεινόμενο μετασχηματισμό περιγράφονται στην ενότητα 6.3. Η πειραματική διεργασία παρέχεται στην ενότητα 6.4.

6.2 3D CTT και σχετικοί μετασχηματισμοί

Ο μετασχηματισμός Trace όπως αναφέρθηκε και σε προηγούμενο κεφάλαιο, είναι μια γενίκευση του μετασχηματισμού Radon [20] ενώ την ίδια ώρα ο Radon αποτελεί υποπερίπτωσή του. Ο μετασχηματισμός Radon μιας εικόνας είναι η δισδιάστατη αναπαράστασή του σε συντεταγμένες ϕ και p με την τιμή του ολοκληρώματος της εικόνας να υπολογίζεται κατά μήκος κάθε αντίστοιχης γραμμής τοποθετημένης σε κελί (ϕ, p). Ο μετασχηματισμός Trace από την άλλη, υπολογίζει ένα συναρτησιακό επάνω στην παράμετρο t κατά μήκος μιας γραμμής, το οποίο δεν είναι απαραίτητα το ολοκλήρωμα. Ο μετασχηματισμός Trace στην ουσία δημιουργείται διατρέχοντας μια εικόνα με ευθείες γραμμές επάνω στις οποίες υπολογίζονται συγκεκριμένα συναρτησιακά της συνάρτησης της εικόνας. Από την ίδια εικόνα μπορούν να δημιουργηθούν διαφορετικοί μετασχηματισμοί με διαφορετικές ιδιότητες. Ο παραγόμενος μετασχηματισμός, είναι στην ουσία μια δισδιάστατη συνάρ-

τηση των παραμέτρων που χαρακτηρίζουν κάθε ιχνο-γραμμή. Ο αναγνώστης μπορεί να βρει τον ορισμό των αναφερθέντων παραμέτρων στο Σχήμα 3.1. Δείγματα των μετασχηματισμών Trace και Radon για διαφορετικά στιγμιότυπα παρμένα από ακολουθίες δράσεων, δίνονται στο Σχήμα 3.2. Στη συνέχεια, γίνεται σύντομη επεξήγηση της λογικής πίσω από τον μετασχηματισμό Trace και το γενικευμένο μετασχηματισμό 3D-Radon. Περισσότερες λεπτομέρειες σχετικά με τη θεωρία που αφορά το μετασχηματισμό Trace, παρέχονται στην ενότητα 3.2 της διατριβής, ενώ σχετικά με τον μετασχηματισμό 3D-Radon στη δημοσίευση [4].

6.2.1 Γενικευμένος 3D Radon μετασχηματισμός

Ο σχεδιασμός του προτεινόμενου μετασχηματισμού έχει εμπνευστεί από τον γενικευμένο 3D Radon [4]. Με άλλα λόγια ο προτεινόμενος μετασχηματισμός είναι μια επεκτεταμένη αναπαράσταση του μετασχηματισμού 3D Radon. Για να κάνουμε την κατανόησή του ευκολότερη, παραθέτουμε το πλαίσιο του 3D Radon όπως παρουσιάζεται στο [131]. Ο μετασχηματισμός 3D Radon ορίζεται με τη χρήση μονοδιάστατων προβολών ενός τρισδιάστατου αντικειμένου $f(x, y, z)$ όπου παρατηρούνται οι συγκεκριμένες προβολές, ολοκληρώνοντας το $f(x, y, z)$ επάνω σε ένα επίπεδο, του οποίου ο προσανατολισμός μπορεί να περιγραφεί από το μοναδιαίο διάνυσμα $\vec{a}$. Γεωμετρικά,ο συνεχής 3D Radon μετασχηματισμός τοποθετεί μια συνάρτηση στον $\Re^3$ χώρο, στο σύνολο των επίπεδων ολοκληρώσεών της στον $\Re^3$. Οι ορισμοί και οι βασικές ιδιότητες της συνεχόμενης μορφής του μετασχηματισμού όπως παρουσιάστηκε και αποδείχθηκε στο [4], ισχύουν και για τη διακριτή μορφή του που παρατίθεται στη συνέχεια.

Έστω ότι *M* είναι ένα 3D μοντέλο και $f(x)$ η ογκομετρική συνάρτηση του *M*, η οποία ορίζεται ως:

$$f(x) = \begin{cases} 1, & \text{όταν το x βρίσκεται στον όγκο του 3D μοντέλου} \\ 0, & \text{διαφορετικά} \end{cases} \tag{6.1}$$

Ο τρισδιάστατος διακριτός μετασχηματισμός Radon του 3D μοντέλου $f(x)$ δίνεται από:

$$T_f(\boldsymbol{\eta}, \rho) = \sum_{j=1}^{J} f(x_j)\delta(x_j^T \boldsymbol{\eta} - \rho) \tag{6.2}$$

όπου $\boldsymbol{\eta}$ είναι ένα μοναδιαίο διάνυσμα στον $\Re^3$, ρ είναι ένας πραγματικός αριθμός και $\delta(\cdot)$ είναι η συνάρτηση Dirac δέλτα. Το μοναδιαίο διάνυσμα $\boldsymbol{\eta}$ μπορεί να γραφτεί σε σφαιρικές συντεταγμένες ως: $\boldsymbol{\eta} = [\cos\phi \sin\theta, \sin\phi \sin\theta, \cos\theta]$. Έτσι, η εξίσωση (6.2) ξαναγράφεται ως:

$$T_f(\rho,\theta,\phi) = \sum_{n=1}^{N} f(x_j, y_j, t_j) \cdot \delta(x_j \cos\phi \sin\theta + y_j \sin\phi \sin\theta + t_j \cos\theta - \rho). \quad (6.3)$$

Ο εν λόγω μετασχηματισμός μπορεί εύκολα να υπολογιστεί, δεν είναι παρ' ολ' αυτά αμετάβλητος σε κλιμακώσεις, μετατοπίσεις και περιστροφές.

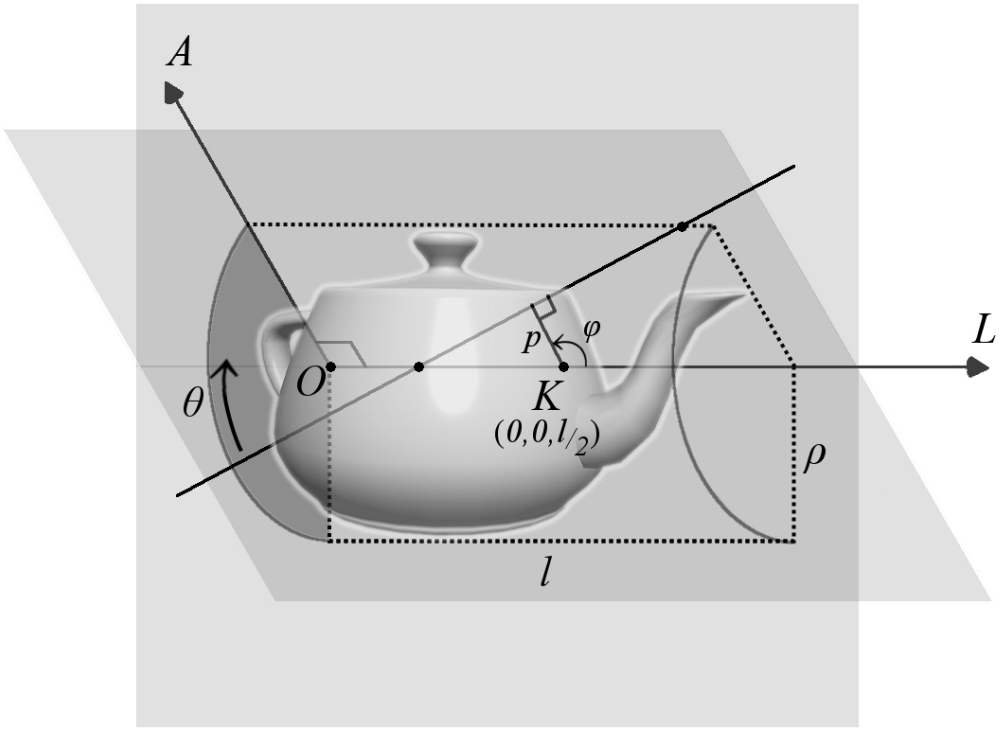
6.2.2 3D Cylindrical Trace Transform (CTT)

Ο προτεινόμενος 3D Cylindrical Trace Transform (CTT) είναι μια επέκταση του μετασχηματισμού Trace στο τρισδιάστατο χώρο. Ο κυλινδρικός μετασχηματισμός Trace $CTT_f(p,\varphi,\theta)$ μιας συνάρτησης $f(x)$ ενός τρισδιάστατου μοντέλου M, συσχετίζει σε κάθε ιχνο-γραμμή τοποθετημένη σε κελί (φ, p, θ) ενός επιπέδου που διασταυρώνεται με το κέντρο της μάζας του 3D μοντέλου, ένα συναρτησιακό T. Ένας μετασχηματισμός Trace δημιουργείται συνεχόμενα κατά τη φορά περιστροφής του διαμετρικά αντίθετου άξονα (polar axis) A, ο οποίος δημιουργεί γωνία θ με το αρχικό επίπεδο ($\theta = 0$). Ο εικονικός κύλινδρος ο οποίος σχηματίζεται από τη συνεχή περιστροφή του Trace στη διεύθυνση περιστροφής τους άξονα A, είναι ακτίνας ρ και μήκους l και έχει αρχή αξόνων $O : (0,0,0)$. Η ακτίνα ρ ορίζεται ως η απόσταση του πιο απομακρυσμένου σημείου $x(\rho_{max}, \theta_x, l_x)$ του 3D μοντέλου από το διαμήκη άξονα (longitudinal axis) L του κυλίνδρου. Ως μήκος l ορίζεται η παράλληλη με τον διαμήκη άξονα απόσταση μεταξύ των δύο πιο απομακρυσμένων αντιδιαμετρικών σημείων του τρισδιάστατου μοντέλου. Ως σημείο ορισμού τον αξόνων του κάθε μετασχηματισμού Trace στα περιστρεφόμενα πεδία, ορίζεται το σημείο K επάνω στον διαμήκη άξονα L με κυλινδρικές συντεταγμένες $(0,0,l/2)$. Το σημείο K συμπίπτει με το κέντρο του εξεταζόμενου όγκου M. Ο κάθε μετασχηματισμός Trace $\breve{g}_n(p,\varphi,\theta)$ υπολογίζεται ως προς το κέντρο K του εικονικού κυλίνδρου το οποίο σημαίνει ότι μετά τις 180 μοίρες περιστροφής ο $\breve{g}_j(p,\varphi,\theta)$ επαναλαμβάνεται. Η τελική αναπαράσταση του τρισδιάστατου μοντέλου είναι ο προτεινόμενος 3D CTT ο οποίος δίνεται από το άθροισμα των υπολογισμών των διαρκώς περιστρεφόμενων Trace μετασχηματισμών στη διεύθυνση των γωνιακών συντεταγμένων που ορίζονται από τη γωνία θ και δίνονται από:

$$CTT_f(p,\varphi,\theta) = \sum_{n=1}^{N} \breve{g}_n(p,\varphi,\theta). \quad (6.4)$$

όπου $\breve{g}_n$ είναι ο n^{th} μετασχηματισμός Trace με $N \geq 2$ και $\theta_{max} = 180^o$. Μια απεικόνιση του μετασχηματισμού 3D Cylindrical Trace δίνεται στο Σχήμα 6.1.

έστω $k(x)$ είναι η $f(x)$ περιστραμμένη από ένα περιστροφικό πίνακα , $k(x) = f(Ax)$. Τότε:



Σχήμα 6.1: *Απεικόνιση του προτεινόμενου μετασχηματισμού 3D Cylindrical Trace.*

$$CTT_k(\rho) = CTT_f(\rho). \tag{6.5}$$

μόνο εάν η $f(x)$ είναι περιστραμμένη κατά θ. Οπότε, ο CTT ενός 3D μοντέλου είναι αμετάβλητος περιστροφής, μόνο όταν το μοντέλο περιστρέφεται κατά γωνία θ. Επιπλέον, έστω ότι $k(x)$ είναι η $f(x)$ κλιμακωμένη κατά ένα συντελεστή a, $a > 0$, π.χ. $k(x) = f\frac{x}{a}$. Τότε:

$$CCT_k(\rho) = CTT_f(\frac{\rho}{a}). \tag{6.6}$$

Οπότε, ο CTT ενός κλιμακωμένου μοντέλου κλιμακώνεται επίσης από τον ίδιο συντελεστή. Για να κάνουμε τον μετασχηματισμό αμετάβλητο στις κλιμακώσεις, υπολογίζουμε τη μέγιστη απόσταση d_{max} μεταξύ του κέντρου του όγκου και του πιο απομακρυσμένου σημείου του από αυτό. Στη συνέχεια, η $f(x)$ κλιμακώνεται ούτως ώστε $d_{max} = 1$.

Ο CTT υπολογίζεται πάντα ως προς το κέντρο του εξεταζόμενου όγκου ο οποίος όπως αναφέρθηκε, συμπίπτει με το σημείο K του εικονικού κυλίνδρου. Συνεπώς, ο CTT είναι αυτομάτως και αμετάβλητος στις μετατοπίσεις. Οι ίδιες ιδιότητες ισχύουν τόσο στη συνεχή όσο και στη διακριτή μορφή του μετασχηματισμού.

6.3 Επισκόπηση του προτεινόμενου συστήματος

Στο Κεφάλαιο 3, παρουσιάστηκαν δύο διαφορετικές μέθοδοι με τη χρήση του μετασχηματισμού Trace για την εξαγωγή χαρακτηριστικών από εικονοσειρές που αναπαριστούν ανθρώπινες δράσεις. Και οι δύο μέθοδοι, (*History Trace Templates* (HTTs) και *History Triple Features* (HTFs)), έχουν αποδειχθεί ικανές να δημιουργήσουν τελικές αναπαραστάσεις μικρής διάστασης για κάθε δράση από σχετικά βίντεο. Επίσης, και οι δυο μέθοδοι απέδειξαν την ευρωστία τους τόσο στο θόρυβο όσο και στις μεταβολές της φωτεινότητας. Ωστόσο, τα HTFs παρέχουν μεγάλη προσαρμοστικότητα καθώς με τη χρήση ενός συνόλου διαφορετικών συναρτησιακών, κάποιος μπορεί να κατασκευάσει μια πιο εύρωστη τελική αναπαράσταση προσαρμοσμένη σε συγκεκριμένες ανάγκες μιας απαιτούμενης εφαρμογής. Μπορεί επίσης να κατασκευάσει χαρακτηριστικά αμετάβλητα στη μετατόπιση, την περιστροφή και την κλιμάκωση, δίνοντας έτσι λύσεις σε μερικά από τα πιο σημαντικά προβλήματα στο πεδίο της αναγνώρισης δραστηριότητας.

Το προτεινόμενο σύστημα έχει σχεδιαστεί με γνώμονα το σενάριο της αναγνώρισης και του εντοπισμού της ανθρώπινης κίνησης σε βίντεο ακολουθίες. Συνδυάζει τον προτεινόμενο 3D CTT με έναν αλγόριθμο τελευταίας τεχνολογίας, τον γνωστό και ως *επιλεκτικά χωρο-χρονικά σημεία ενδιαφέροντος* Selective Spatio-Temporal Interest Points (SSTIPs)

[13] και δημιουργεί μια ελάχιστη αναπαράσταση πολύ μικρής διάστασης με τη χρήση των HTFs. Η βασική ιδέα πίσω από το συγκεκριμένο συνδυασμό τεχνικών είναι, να ωφεληθούμε από τα πιο πολύτιμα χαρακτηριστικά που έχει να προσφέρει η κάθε μέθοδος και να τα συνδυάσουμε σε μια τελική αναπαράσταση χαμηλής διάστασης.

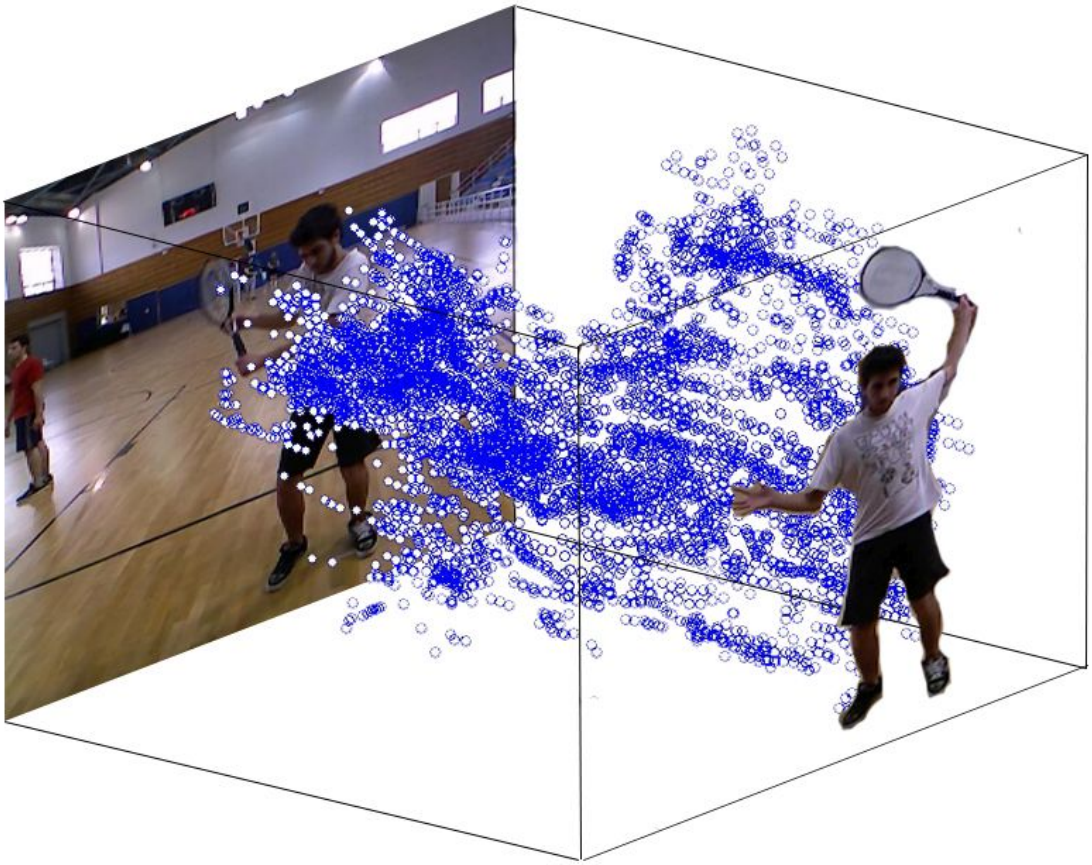
Οι μέθοδοι βασισμένες στα χωρο-χρονικά σημεία ενδιαφέροντος όπως τα γνωστά Bag of Visual Words (BOVW), έχουν κινήσει το έντονο ενδιαφέρον των ερευνητών τελευταία στο πεδίο της αναγνώρισης δράσεων. Ωστόσο, αυτού του είδους οι αναπαραστάσεις, αγνοούν πιθανή πολύτιμη πληροφορία η οποία αναφέρεται στη συνολική χωρο-χρονική κατανομή των σημείων ενδιαφέροντος [131]. Με την εισαγωγή του CTT η μέθοδος καταφέρνει να συλλάβει τη λεπτομερή γεωμετρική κατανομή και τη γεωμετρική πληροφορία των σημείων ενδιαφέροντος ενώ την ίδια ώρα παρέχει την ευελιξία της δημιουργίας πιο αποτελεσματικών και πιο κατάλληλων για μια πληθώρα διαφορετικών, συνθηκών λήψης, περιβαλλόντων και εφαρμογών. Η χρήση και ο συνδυασμός διαφορετικών και κατάλληλων συναρτησιακών για τον υπολογισμό διαφορετικών CTTs και HTFs μπορεί να παρέχει μια πολύ εύρωστη αναπαράσταση μιας κίνησης αναπαριστάμενης σε ένα βίντεο. Περισσότερες πληροφορίες για τις επιμέρους τεχνικές και το προτεινόμενο σύστημα παρέχονται στις ενότητες που ακολουθούν.

6.3.1 Επιλεκτικά χωρο-χρονικά σημεία ενδιαφέροντος (SSTIPs)

Όπως αναφέρθηκε πιο πάνω, το προτεινόμενο σύστημα ενσωματώνει μια καινοτόμο προσέγγιση στο πρόβλημα με το οποίο ασχολείται ο αλγόριθμος STIPs, την τεχνική γνωστή και ως Selective Spatio-Temporal Interest Points που προτάθηκε στο [13]. Σε αυτή τη μελέτη, οι συγγραφείς πρότειναν μια μεθοδολογία εξαγωγής χωρο-χρονικών σημείων ενδιαφέροντος, η οποία επικεντρώνεται στη συνολική κίνηση αντί της τοπικής χωροχρονικής πληροφορίας. Με αυτό τον τρόπο αποτρέπει τον άστοχο εντοπισμό σημείων ενδιαφέροντος εξαιτίας αχρείαστων φόντων και κίνησης της κάμερας. Επιπλέον, έδειξαν ότι η μέθοδός τους αποδίδει καλά παράγοντας σταθερά, επαναλαμβανόμενα STIPs, ανθεκτικά στις τοπικές ιδιότητες του ανιχνευτή (τον οποίο χρησιμοποιεί) διαμέσου της ακολουθίας της κίνησης.

Η μέθοδος των επιλεκτικών χωρο-χρονικών σημείων ενδιαφέροντος μπορεί να περιγραφεί ως μια διαδικασία η οποία:

1. ανιχνεύει χωρικά σημεία ενδιαφέροντος
2. καταστέλλει αχρείαστα σημεία φόντου
3. επιβάλλει χωρικούς και χρονικούς περιορισμούς στο αποτέλεσμα



Σχήμα 6.2: *Επιλεκτικά STIPs (SSTIPs) τα οποία έχουν εξαχθεί από ένα βίντεο της βάσης THETIS που αναπαριστά ένα κτύπημα τένις (backhand shot).*

Το πρώτο βήμα γίνεται ουσιαστικά με την εφαρμογή ενός ανιχνευτή γωνιών Harris. Η ιδέα πίσω από το δεύτερο βήμα είναι η παρατήρηση πως τα γωνιακά σημεία που ανιχνεύθηκαν στο φόντο ακολουθούν συγκεκριμένα γεωμετρικά πρότυπα, ενώ αυτά που ανιχνεύθηκαν επάνω σε ανθρώπους δεν ακολουθούν αυτή την ιδιότητα. Τελικά, επιβάλλονται χωρικοί και χρονικοί περιορισμοί βασισμένοι στην αντίληψη ότι για να θεωρηθεί ένα σημείο ενδιαφέροντος ως ακριβές και επαναλαμβανόμενο STIP, πρέπει να δείχνει μια αλλαγή θέσης μέσα στην ακολουθία της κίνησης. Παράδειγμα εξαγωγής SSTIPs από ένα δείγμα της βάσης δεδομένων THETIS δίνεται στο Σχήμα 6.2.

6.3.2 Εφαρμογή του 3D CTT σε Selective Spatio Temporal Interest Points

Οι συγγραφείς του [49] έχουν δείξει ότι όχι μόνο τα ολοκληρώματα (περίπτωση του μετασχηματισμού Radon), αλλά επίσης άλλα επιλεγμένα συναρτησιακά υπολογιζόμενα κατά μήκος ευθειών ορισμένων στο πεδίο δισδιάστατων συναρτήσεων, μπορούν ολοκληρωτικά να την αναπαραγάγουν (περίπτωση μετασχηματισμού Trace). Όπως αναφέρθηκε

νωρίτερα, ο μετασχηματισμός Trace παράγεται διατρέχοντας μια εικόνα με ευθείες γραμμές επάνω στις οποίες έχουν υπολογιστεί συγκεκριμένα συναρτησιακά της συγκεκριμένης συνάρτησης. Το αποτέλεσμα είναι μια άλλη 2D συνάρτηση και εξαρτάται από τις παραμέτρους (ϕ, p) οι οποίες χαρακτηρίζουν κάθε γραμμή. Με τη χρήση διαφορετικών συναρτησιακών μπορεί κάποιος να παραγάγει διαφορετικούς μετασχηματισμούς Trace.

Αντίστοιχα, ο μετασχηματισμός 3D CTT παράγεται διατρέχοντας με ευθείες γραμμές διαρκώς περιστρεφόμενα κατά γωνία θ επίπεδα τα οποία ανήκουν στο ίδιο ελάχιστο παράθυρο (κύλινδρο) που περιβάλει ένα τρισδιάστατο μοντέλο, με αρχή αξόνων το σημείο *O*. Όπως συμβαίνει και με τον Trace, επιλεγμένα συναρτησιακά μπορούν να υπολογιστούν κατά μήκος των ιχνο-γραμμών. Ο 3D CTT είναι το άθροισμα αυτών των ιδιαζόντων μετασχηματισμών και έχει ως αποτέλεσμα μια νέα δισδιάστατη συνάρτηση η οποία εξαρτάται από τις παραμέτρους (ϕ, p, θ) οι οποίες χαρακτηρίζουν κάθε γραμμή. Όπως και στον Trace, κάποιος μπορεί να παραγάγει διαφορετικούς 3D CTTs οι οποίοι με τη σειρά τους θα έχουν αποκτήσει διαφορετικές ιδιότητες με τον υπολογισμό διαφορετικών συναρτησιακών.

Καθώς οι ανθρώπινες κινήσεις είναι στην ουσία χωρο-χρονικοί όγκοι, σκοπός είναι η επίτευξη μιας αναπαράστασης η οποία να έχει συλλάβει όσο το δυνατόν περισσότερη από τη δυναμική και και δομική πληροφορία μια κίνησης. Σε αυτό το σημείο, όπως είδαμε και στο Κεφάλαιο 3, ο μετασχηματισμός Trace παρουσιάζει εξαιρετική καταλληλότητα. Μετασχηματίζει με γραμμές εικόνες δυο διαστάσεων, σε ένα χώρο πιθανών παραμέτρων γραμμών, όπου κάθε γραμμή στην εικόνα θα δώσει μια κορυφή τοποθετημένη στις αντίστοιχες γραμμικές παραμέτρους.

Ο 3D CTT που παρουσιάζεται σε αυτό το κεφάλαιο, είναι μια εξέλιξη του μετασχηματισμού Trace καθώς παρέχει όλα τα πλεονεκτήματα της απλής μορφής και προεκτείνει τις δυνατότητές του στον τρισδιάστατο χώρο. Αυτό τον κάνει πιο κατάλληλο για την αναπαράσταση όγκων όπως για παράδειγμα είναι οι ακολουθίες κινήσεων. Συνδυασμένος με τα Spatio-Temporal Interest Points, παρέχει το πλεονέκτημα της σύλληψης της τρισδιάστατης γεωμετρικής διασποράς των STIPs τα οποία έχουν προηγουμένως κάνει τη διάκριση των σημείων ενδιαφέροντος σε μια δράση τόσο χωρικά, όσο και χρονικά. Όταν πολλαπλοί περιστραμμένοι μετασχηματισμοί Trace έχουν υπολογιστεί με σημείο αναφοράς το κέντρο του STIPs νέφους, συγκεκριμένοι συντελεστές θα έχουν συλλάβει πολλή από την πληροφορία η οποία θα σχετίζεται με το χώρο και το χρόνο που κατέλαβε η δράση. Αυτοί οι συντελεστές θα ποικίλουν κατά την περιστροφή και θα παρέχουν σημαντικές διαφορές από μια κίνηση σε μια άλλη για τους αντίστοιχους χωρο-χρονικούς όγκους.

Με αυτό τον τρόπο, τα χαρακτηριστικά που παράγονται θα είναι συνάρτηση πολλαπλών εμφαντικών χωρο-χρονικών διακρίσεων οι οποίες περιέχονται σε πολλαπλούς μετασχηματισμούς παραγόμενους για το 3D νέφος το οποίο αναπαριστά μια ολόκληρη κίνηση. Ένα άλλο πλεονέκτημα του συστήματος αυτού, είναι ότι δεν απαιτείται εξαγωγή ανθρώ-

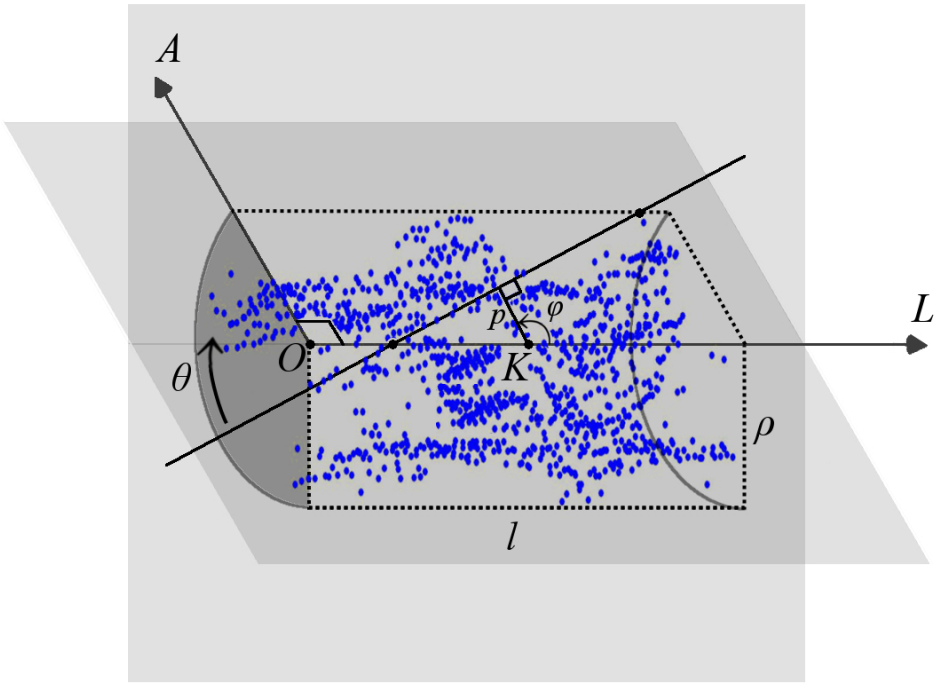
πινων σιλουετών ή χρονική στοίχιση των ακολουθιών. Τα STIPs εξάγονται βασισμένα στη χωρική και χρονική αξιολόγηση ενός αριθμού σχετικών διακρίσεων. Υπολογίζοντας τον CTT σε ένα τέτοιο νέφος σημείων, η γεωμετρική διασπορά των σχετικών σημείων είναι ενσωματωμένη σε μια τελική αναπαράσταση, ενώ την ίδια στιγμή πολλαπλοί περιστρεφόμενοι συντελεστές του συγκεκριμένου μετασχηματισμού θα παρέχουν κορυφές για τα ξεχωριστά και επίμονα σημεία που χαρακτηρίζουν κάθε κίνηση. Με αυτό τον τρόπο, χρονική αντιστοίχιση δεν απαιτείται. Παρ' ολ' αυτά, πειράματα έχουν διεξαχθεί και για τα δύο σενάρια (με την εξαγωγή σιλουετών και απευθείας εφαρμογή) για να αναδείξουμε την παραπάνω ικανότητα. Τα αποτελέσματα παρουσιάζονται στην ενότητα 6.4. Μια απεικόνιση του πως υπολογίζεται ο 3D CTT εφαρμοζόμενος στα SSTIPs δίνεται στο Σχήμα 6.3.

Σε αυτό το σημείο πρέπει επίσης να αναφέρουμε, ότι καθώς ο CTT έχει σχεδιαστεί με γνώμονα την εφαρμογή του στην εξαγωγή χαρακτηριστικών από χωρο-χρονικές ακολουθίες, διαφοροποιείται από τον μετασχηματισμό 3D Radon και κατά έναν άλλο τρόπο. Καθώς τα 3D μοντέλα στα οποία εφαρμόζεται ο CTT είναι στην πράξη χωρο-χρονικοί όγκοι, η επεξεργασία του μετασχηματισμού λαμβάνει χώρα πάντα στην διεύθυνση του χρόνου. Καθώς αυτή είναι πάντα η περίπτωση εφαρμογής του προτεινόμενου μετασχηματισμού και η διάσταση του χρόνου δεν μετατοπίζεται στον τρισδιάστατο χώρο, ο CTT έχει τη δυνατότητα να πάρει τη μορφή ενός εικονικού κυλίνδρου και όχι αυτό της σφαίρας. Με αυτό τον τρόπο ο συγκεκριμένος μετασχηματισμός καταφέρνει να είναι πιο αποτελεσματικός χρονικά ενώ ταυτόχρονα επιτυγχάνει να ενσωματώσει όλους τους πολύτιμους διακριτούς συσχετισμούς των STIPs στη τελική αναπαράσταση.

6.3.3 History Triple Features (HTFs)

Στο Κεφάλαιο 3 παρουσιάστηκε μια άλλη καινοτόμος απεικόνιση της ανθρώπινης δράσης. Η συγκεκριμένη τεχνική με το όνομα *History Triple Features* (HTFs) κάνει χρήση των χαρακτηριστικών που εξάγονται από διάφορους μετασχηματισμούς Trace αναπαριστώντας σε ένα διάνυσμα πολύ μικρής διάστασης μια συγκεκριμένη κίνηση. Έχει αποδειχθεί ότι η μέθοδος είναι ικανή να παρέχει μια ουσιώδη λύση σε παραμορφώσεις που συνήθως παρουσιάζονται σε προβλήματα κατηγοριοποίησης δράσεων και τα οποία οφείλονται σε περιορισμούς που παρατηρούνται κατά τη διάρκεια λήψης (π.χ. διαφορετικές συνθήκες φωτισμού, κλιμακώσεις/αποκλιμακώσεις και συμβάντα περιστροφών).

Με περισσότερες λεπτομέρειες, τα HTFs είναι μια απλοποιημένη αναπαράσταση ενός συνόλου πολλαπλών Trace μετασχηματισμών οι οποίοι στην ουσία καθένας από αυτούς είναι μια πολύ πλούσια αναπαράσταση μιας δισδιάστατης συνάρτησης (εικόνας). Τα HTFs αποτελούνται από ένα σύνολο υπολογισμένων λόγων (*triple features*), τα οποία αρχικά προτάθηκαν [49] για την κατηγοριοποίηση εικόνων που απεικόνιζαν πολύ συγγενείς κλάσεις όπως διαφορετικά είδη ψαριών. Η θεωρία πίσω από τα triple features περιγράφηκε



Σχήμα 6.3: *Εφαρμογή του προτεινόμενου 3D CTT σε SSTIPs τα οποία έχουν εξαχθεί από ένα βίντεο δράσης.*

στην ενότητα 3.2. Έχοντας εξαγάγει το χωρο-χρονικό νέφος σημείων ενδιαφέροντος, αρχικά μετασχηματίζουμε συνεχή περιστρεφόμενα επίπεδα στο χώρο του μετασχηματισμού Trace. Για κάθε επίπεδο το οποίο συντελεί στον σχηματισμό του εικονικού κυλίνδρου του CTT, υπολογίζεται ένα σύνολο Trace μετασχηματισμών. Ακολουθώντας τη διαδικασία που περιγράφηκε στην ενότητα 3.2, παράγεται ένα σύνολο από triple features. Ο λόγος ενός ζεύγους τέτοιων χαρακτηριστικών όπως έχει δειχθεί, μπορεί να είναι αμετάβλητος σε διάφορα είδη παραμορφώσεων και εξαρτάται από τα συναρτησιακά που έχουν χρησιμοποιηθεί για τον υπολογισμό τους. Αυτά τα συναρτησιακά μπορεί να επιλεχθούν ούτως ώστε να είναι ευαίσθητα ή αμετάβλητα σε πιθανές μεταβολές που συμβαίνουν συχνά κατά τη λήψη βίντεο δράσεων, ενώ παράλληλα να διατηρούν τη διακριτική τους ικανότητα.

Με απλό τρόπο τα triple features κατασκευάζοντας ως εξής:
α) Ο μετασχηματισμός Trace παράγεται εφαρμόζοντας ένα Trace συναρτησιακό T κατά μήκος γραμμών που διατρέχουν την εικόνα.
β) Η κυκλική (circus) συνάρτηση μιας εικόνας παράγεται εφαρμόζοντας ένα διαμετρικό συναρτησιακού P κατά μήκος των στηλών του μετασχηματισμού Trace.
γ) Το τριπλό χαρακτηριστικό παράγεται τελικά εφαρμόζοντας ένα circus συναρτησιακό Φ κατά μήκος μιας σειράς αριθμών που παράγεται από το βήμα .

Στη συνέχεια, παρουσιάζεται η εξαγωγή των HTFs όπως προτάθηκε στο [35] και περιγράφηκε στην ενότητα 3.4 προσαρμοσμένο στο προτεινόμενο σύστημα που περιγράφεται στο τρέχον κεφάλαιο.

6.3.4 Εξάγοντας χαρακτηριστικά με τη χρήση της προτεινόμενης μεθοδολογίας

Έστω ότι M είναι ένα 3D μοντέλο αποτελούμενο από ένα νέφος SSTIPs το οποίο έχει δημιουργηθεί για μια αναπαράσταση δράσης σε βίντεο. Έστω ρ και l ότι είναι η ακτίνα και το μήκος αντίστοιχα του μικρότερου κυλίνδρου που περιβάλλει το νέφος. Η διακριτή δυαδική συνάρτηση όγκου $f(u)$ του M αποτελείται από N χωρο-χρονικά σημεία u_i και ορίζεται ως:

$$f(u) = \begin{cases} 1, & \text{εαν } u_i \text{ είναι σημείο ενδιαφέροντος} \\ 0, & \text{διαφορετικά} \end{cases} \tag{6.7}$$

Έστω z είναι ένα επίπεδο με μέγεθος 2ρ x l όπου ρ είναι η ακτίνα και l είναι το μήκος του μικρότερου κυλίνδρου που μπορεί να περιβάλλει το τρισδιάστατο νέφος σημείων. Ο μετασχηματισμός Trace $\breve{g}_f(p, \varphi, \theta)$, είναι η συνάρτηση που ορίζεται στο χώρο των ευθειών γραμμών επάνω στο επίπεδο z από μια συνάρτηση κατά μήκος κάθε τέτοιας ευθείας.

p και φ είναι οι παράμετροι οι οποίες ορίζουν τη θέση της γραμμής η οποία βρίσκεται στο επίπεδο z και η οποία έχει περιστραφεί κατά θ. Έτσι, ο $\check{g}_f(p, \varphi)$ είναι το αποτέλεσμα του συναρτησιακού T επάνω στη γραμμή $p = x\cos\theta + y\sin\varphi$ η οποία βρίσκεται επάνω στο επίπεδο. Ως σημείο αναφοράς ορίζεται το κέντρο του όγκου των SSTIPs. Ο πιο πάνω μετασχηματισμός, εφαρμόζεται σε J αριθμό επιπέδων με σημείο αναφοράς (0, 0, l/2) τα οποία περιστρέφονται στη διεύθυνση περιστροφής του πολικού άξονα του κυλίνδρου κατά γωνία $0 < \theta \leq 180$. Όπου J είναι ο αριθμός των επιπέδων και συνεπώς και ο αριθμός των εφαρμοζόμενων Trace μετασχηματισμών στο τρισδιάστατο νέφος. Έτσι, ο διακριτός 3D CTT ενός 3D μοντέλου δίνεται από:

$$CTT_f(p, \varphi, \theta) = \sum_{j=1}^{J} \check{g}_j(p, \varphi, \theta). \tag{6.8}$$

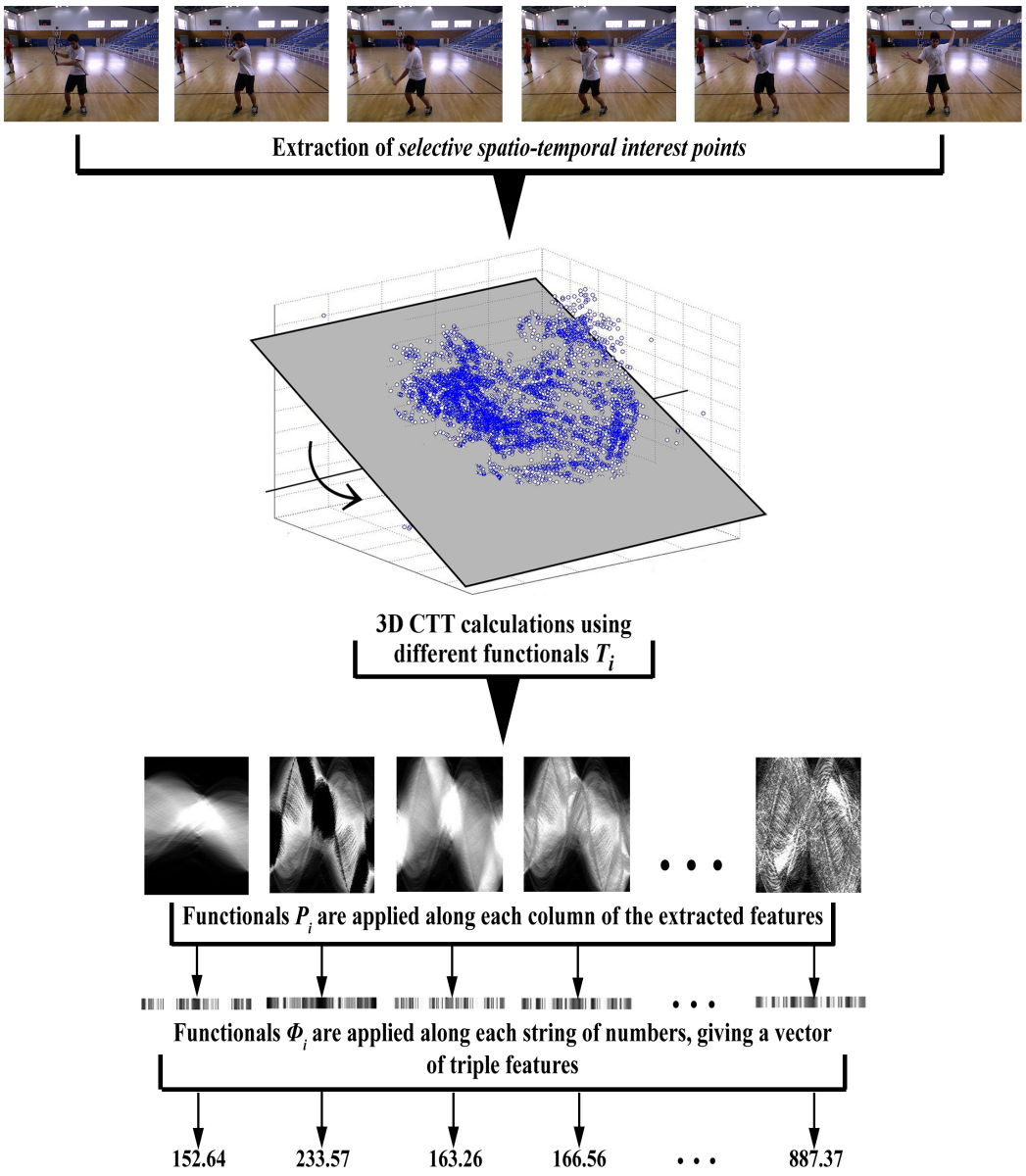
Εφαρμόζοντας διαφορετικά συναρτησιακά T στο νέφος M των SSTIPs, παράγεται ένα σύνολο $CTT_{f_i}(p, \varphi, \theta)$ μετασχηματισμών. Όπου $i = 1...I$ και I είναι ο αριθμός των μετασχηματισμών που κάποιος επιλέγει να υπολογίσει. Για κάθε $CTT_{f_i}(p, \varphi, \theta)$ υπολογίζεται ένα σύνολο από Π_{norm} κανονικοποιημένα triple features.

Διαιρώντας όλα τα Π_{norm} μεταξύ τους, παράγεται ένα νέο σύνολο αμετάβλητων χαρακτηριστικών. Έτσι, ολόκληρη η κίνηση αναπαριστάται τελικά από ένα διάνυσμα **v** το οποίο αποτελείται από το σύνολο όλων των λόγων των triple features που έχουν υπολογιστεί για τα SSTIPs της βιντεο-ακολουθίας μιας δράσης. Η διαδικασία απεικονίζεται στο Σχήμα 6.4.

$$\mathbf{v} = (\Pi_{rat_1}, \Pi_{rat_2}, ..., \Pi_{rat_{g-1}}, \Pi_{rat_g}) \tag{6.9}$$

όπου Π_{rat} είναι ο λόγος των κανονικοποιημένων triple features και g είναι ο αριθμός των υπολογιζόμενων λόγων. Τα συναρτησιακά που χρησιμοποιήθηκαν για την εξαγωγή των HTFs διανυσμάτων που χρησιμοποιήθηκαν στην πειραματική διαδικασία αξιολόγησης του αλγορίθμου, δίνονται στον πίνακα 6.3.4 και για λόγους συνέπειας συμπίπτουν με αυτά της πειραματικής διαδικασίας που περιγράφηκε στο Κεφάλαιο 3.

Αυτή η μέθοδος επιτρέπει την κατασκευή πολλών χαρακτηριστικών πολύ εύκολα. Αν υποθέσουμε ότι κάποιος κάνει χρήση 10 συναρτησιακών για κάθε στάδιο κατασκευής των triple features (π.χ. 10 συναρτησιακά T, 10 συναρτησιακά P και 10 συναρτησιακά Φ) σε έναν CTT 10 βημάτων (που να δημιουργείται από 10 περιστρεφόμενα επίπεδα), μπορεί να κατασκευάσει $10x10x10x10 = 10000$ χαρακτηριστικά για μια ακολουθία. Σε αυτό το σημείο, πρέπει να επισημάνουμε ότι οι αριθμοί που προκύπτουν μπορεί να μην έχουν φυσική σημασία, εμπεριέχουν όμως όλες τις απαραίτητες μαθηματικές ιδιότητες για τους σκοπούς της κατηγοριοποίησης.



Σχήμα 6.4: *Σύνοψη της εξαγωγής χαρακτηριστικών με την εφαρμογή του προτεινόμενου συστήματος*

Πίνακας 6.1: *Διάφορα συναρτησιακά που υπολογίστηκαν για την πειραματική διαδικασία.*

Trace Transform	Functional
1	$T(f(x)) = \int_{[0,\infty]} rf(r)dr$ where $r = x - c$ and $c = median_x \{x, f(x)\}$
2	$T(f(x)) = \int_{[0,\infty]} r^2 f(r)dr$ where $r = x - c$ and $c = median_x \{x, f(x)\}$
3	$T(f(x)) = median_{r\geq 0}\{f(r), (f(r))^{\frac{1}{2}}\}$ where $r = x - c$ and $c = median_x \{x, f(x)\}$
4	$T(f(x)) = median_{r\geq 0}\{rf(r), (f(r))^{\frac{1}{2}}\}$ where $r = x - c$ and $c = median_x \{x, f(x)\}$
5	$T(f(x)) = \int_{[0,\infty]} e^{iklogr} r^p f(r)dr, (p = 0.5, k = 4)$ where $r = x - c$ and $c = median_x \{x, (f(x))^{\frac{1}{2}}\}$
6	$T(f(x)) = \int_{[0,\infty]} e^{iklogr} r^p f(r)dr, (p = 0, k = 3)$ where $r = x - c$ and $c = median_x \{x, (f(x))^{\frac{1}{2}}\}$
7	$T(f(x)) = \int_{[0,\infty]} e^{iklogr} r^p f(r)dr, (p = 1, k = 5)$ where $r = x - c$ and $c = median_x \{x, (f(x))^{\frac{1}{2}}\}$

Καθώς η διακριτική ικανότητα των παραγόμενων χαρακτηριστικών οπωσδήποτε θα ποικίλει, η εφαρμογή μια τεχνικής για τη μείωση της διάστασης θα μπορούσε να πραγματοποιήσει μια επιλογή των πιο διακριτών χαρακτηριστικών ενώ ταυτόχρονα να κάνει και το πρόβλημα ευκολότερα διαχειρίσιμο. Στο σύστημα το οποίο προτείνεται, στα διανύσματα HTFs τα οποία παράγονται με τον παραπάνω τρόπο, εφαρμόζεται PCA με σκοπό να οριστεί ο κατάλληλος υποχώρος για την κατηγοριοποίηση. Στην πράξη, κρατάμε μόνο ένα μικρό υποσύνολο από το αρχικό διάνυσμα HTFs το οποίο περιέχει τα πιο διακριτά από τα υπολογιζόμενα χαρακτηριστικά, ικανά να περιγράψουν ικανοποιητικά ολόκληρη την κινησιο-ακολουθία.

6.4 Πειραματικά αποτελέσματα

Στη συνέχεια, δίνονται τα αποτελέσματα της πειραματικής διαδικασίας με σκοπό να αναδειχθεί η αποτελεσματικότητα της παρουσιαζόμενης τεχνικής σε δύο διαφορετικά σενάρια εφαρμογής. Αυτό της αναγνώρισης ανθρώπινης δραστηριότητας και αυτό του εντοπισμού ανθρώπινης πτώσης. Θα δοθούν τα αποτελέσματα για μια σειρά διαφορετικών γνωστών και απαιτητικών βάσεων και θα παρουσιαστεί ο τρόπος με τον οποίον συμπεριφέρεται ο αλγόριθμος κάτω από διαφορετικούς τύπους βίντεο και διαφορετικές συνθήκες λήψης των βίντεο.

Σε αυτό το σημείο και προτού περιγράψουμε το πειραματικό πρωτόκολλο το οποίο χρησιμοποιήθηκε, είναι χρήσιμο να αναφέρουμε ότι κατά τους συγγραφείς του [27], υπάρχει μια ποικιλία εφαρμοζόμενων πρωτοκόλλων τα οποία εφαρμόζονται από τους ερευνητές επάνω στις ίδιες βάσεις δεδομένων για την αναγνώριση κινήσεων από βίντεο. Αναφέ-

ρεται επίσης, ότι μέθοδοι που αξιολογούνται με τη χρήση γνωστών βάσεων όπως η KTH [100] και Weizmann [6] μπορούν να παρουσιάζουν διαφορές αποτελεσμάτων της τάξεως του 10.67% όταν εφαρμόζονται διαφορετικές προσεγγίσεις τεκμηρίωσης. Παρ' όλ' αυτά, για την ώρα δεν υπάρχει ενιαίο πρότυπο αξιολόγησης.

Στα πειράματα που ακολουθούν, όταν αυτό μπορεί να εφαρμοστεί (π.χ. βάσεις δεδομένων αναγνώρισης δράσης), για την αξιολόγηση της απόδοσης του συστήματος εφαρμόζεται το πρωτόκολλο leave-one-person-out cross validation. Είναι ένα ιδιαίτερα απαιτητικό πρωτόκολλο το οποίο αναπαριστά με τον πλησιέστερο τρόπο τις συνθήκες που επικρατούν σε πραγματικές συνθήκες εφαρμογής. Σε αντιστοιχία με της πραγματικές συνθήκες, η φυσική δυναμική συμπεριφορά ενός άγνωστου υποκειμένου συλλαμβάνεται και συγκρίνεται από ένα σύστημα αναγνώρισης δράσεων και στη συνέχεια γίνεται επεξεργασία και σύγκριση με ένα προ-εγγεγραμμένο σύνολο δεδομένων το οποίο έχει προηγουμένως χρησιμοποιηθεί για την εκπαίδευση του συστήματος. Η τελική απάντηση λαμβάνεται βάσει της σχετικότητας της εξεταζόμενης κίνησης με ένα από τα δεδομένα τα οποία αποτελούν το σύνολο εκπαίδευσης και βάσει του κανόνα που εφαρμόζει το σύστημα. Για την ακρίβεια, το πιο πάνω πρωτόκολλο, κάνει χρήση των δειγμάτων ενός προσώπου για εξέταση και τα υπόλοιπα δείγματα της βάσης χρησιμοποιούνται για εκπαίδευση. Η διαδικασία επαναλαμβάνεται N φορές, όπου N είναι ο αριθμός των υποκειμένων μέσα στη βάση. Η απόδοση αναφέρεται ως η μέση ακρίβεια $I = \sum_{n=1}^{N} H_n$ επαναλήψεων, όπου H_n είναι ο αριθμός των δειγμάτων του n^{th} δείγματος μέσα στη βάση.

Παρά το ότι ο εντοπισμός πτώσεων είναι αρκετά συναφής με την αναγνώριση ανθρώπινης δραστηριότητας, τα αποτελέσματα υπολογίζονται βασισμένα στη λογική του *ναι* ή *όχι* καθώς απαιτείται μια διαφορετική προσέγγιση. Έτσι, ένα πιθανό εφαρμοσμένο σύστημα παρακολουθεί διαρκώς ένα υποκείμενο και καταγράφει τη φυσική δυναμική του. Η συμπεριφορά η οποία έχει καταγραφεί, αναλύεται κατά τακτά χρονικά διαστήματα και συγκρίνεται με ένα προεγγεγραμμένο σύνολο από πιθανές πτώσεις οι οποίες έχουν χρησιμοποιηθεί για την εκπαίδευση του συστήματος. Η τελική απόφαση λαμβάνεται βάσει της σχετικότητας της εξεταζόμενης συμπεριφοράς με ένα από τα διαφορετικά δείγματα πτώσεων και αναφέρεται μια κατάσταση τύπου *πτώση* ή *μη πτώση*. Το πρωτόκολλο που χρησιμοποιήθηκε κατά την πειραματική διεργασία, χρησιμοποιεί ένα δείγμα για τον έλεγχο και τα υπόλοιπα δείγματα του συνόλου χρησιμοποιούνται για εκπαίδευση. Η απόφαση η οποία λαμβάνεται, είναι δυαδικού τύπου (0 ή 1) και επαναλαμβάνεται H φορές όπου H είναι ο αριθμός των υποκειμένων μέσα στη βάση. Η απόδοση αναφέρεται ως ο λόγος των επιτυχών κατηγοριοποιήσεων προς τον H αριθμό των ελέγχων.

Σχήμα 6.5: *Δείγματα δράσεων από τη βάση δεδομένων Weizmann για τις δράσεις wave1, wave2, walk, pjump, side, run, skip, jack, jump και bend.*

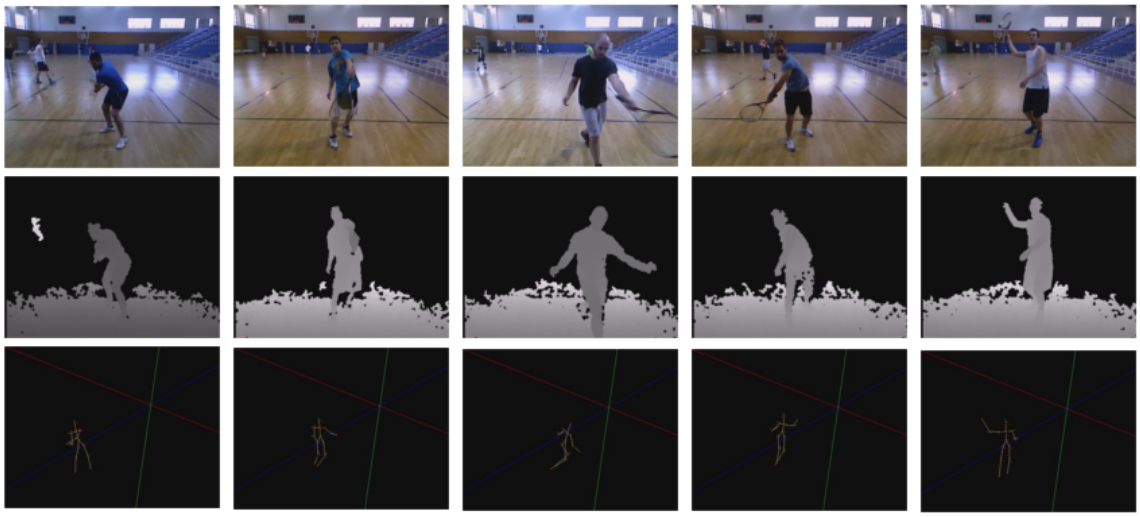
Σχήμα 6.6: *Δείγματα δράσεων από τη βάση δεδομένων KTH για τις δράσεις walking, jogging, running, boxing, hand, waving και hand clapping.*

6.4.1 Πειράματα αναγνώρισης κίνησης

Για τα πειράματα χρησιμοποιήθηκαν τρεις διαφορετικές βάσεις δεδομένων. Η KTH, η Weizmann και η THETIS [37]. Τα Σχήματα 6.5, 6.6 και 6.7 απεικονίζουν διαφορετικά δείγματα από τις τρεις προαναφερθείσες βάσεις δεδομένων. Η βάση δεδομένων KTH περιέχει 6 τύπους κινήσεων (walking, jogging, running, boxing, hand waving και hand clapping) εκτελεσμένες αρκετές φορές από 25 άτομα σε 4 διαφορετικά σενάρια, κάτω από διάφορες συνθήκες φωτισμού: εξωτερικοί χώροι, εξωτερικοί χώροι με μεταβαλλόμενη κλιμάκωση (zoom-in, zoom-out), εξωτερικοί χώροι με διαφορετικά ρούχα και εσωτερικοί χώροι. Η βάση αποτελείται από 600 ακολουθίες. Όλες οι ακολουθίες έχουν κινηματογραφηθεί σε ομοιόμορφο φόντο, με στατική κάμερα και ταχύτητα λήψης 25 καρέ το δευτερόλεπτο.

Η βάση βίντεο δεδομένων Weizmann αποτελείται από 90 βίντεο χαμηλής ανάλυσης (180 x 144, ταχύτητα κλείστρου 50 καρέ το δευτερόλεπτο) που παρουσιάζουν 9 διαφορετικούς ανθρώπους. Κάθε άτομο έχει εκτελέσει 10 φυσικές δραστηριότητες όπως τρέξιμο βάδισμα και άλλα. Στη βάση οι κινήσεις που εμπεριέχονται αναφέρονται με τις ακόλουθες ονομασίες: run, walk, skip, jumping-jack (ή shortly jack), jump-forward-on-two-legs (ή jump), jump-in-place-on-two-legs (ή pjump), gallopsideways (ή side), wave-two-hands (ή wave2), waveone- hand (ή wave1), ή bend.

Το σύνολο δεδομένων THETIS αποτελείται από τα 12 βασικά κτυπήματα του τένις εκτελεσμένα από 31 αρχάριους και 24 έμπειρους παίκτες. Όλες οι κινήσεις έχουν κινημα-



Σχήμα 6.7: *Δείγματα διαφορετικών τύπων δεδομένων από τη βάση THETIS.*

τογραφηθεί με τη χρήση του αισθητήρα Kinect ο οποίος ήταν τοποθετημένος μπροστά στους αθλητές. Κάθε κτύπημα έχει εκτελεστεί επαναλαμβανόμενα έχοντας σαν αποτέλεσμα 8734 (κομμένα σε περιόδους) βίντεο της μορφής AVI. Η συνολική διάρκεια των βίντεο είναι 7 ώρες και 15 λεπτά. Τα κτυπήματα που έχουν εκτελεστεί δοσμένα με τη διεθνή τους ονομασία είναι τα ακόλουθα: Backhand with two hands, Backhand, Backhand slice, Backhand volley Forehand flat, Forehand open, stands, Forehand slice, Forehand volley, Service flat, Service kick, Service slice και Smash. Δείγματα της βάσης THETIS παρέχονται στο Σχήμα 6.7.

Στα πειράματα τα οποία διεξήχθησαν οι βίντεο-ακολουθίες για την KTH και τη Weizmann έχουν μετατραπεί στη χωρική ανάλυση των 240*180 εικονοστοιχείων και έχουν μήκος τεσσάρων δευτερολέπτων κατά μέσο όρο. Τα δείγματα εκπαίδευσης δεν έχουν στοιχηθεί χρονικά ή χωρικά και αφαίρεση φόντου δεν έχει πραγματοποιηθεί με οποιαδήποτε τεχνική. Για τη γενικευμένη αξιολόγηση της μεθόδου με την εφαρμογή πολλαπλών κατηγοριοποιητών εφαρμόστηκε η προσέγγιση leave-one-person-out cross-validation.

Τα βίντεο της βάσης THETIS για λόγους ταχύτητας, έχουν υποκλιμακωθεί στην ανάλυση των 320*240 εικονοστοιχείων από την αρχική των 640*480. Στα πειράματα τα οποία διεξήχθησαν, χρησιμοποιήθηκαν τρεις τύποι δεδομένων από τη συνολική βάση. Οι: RGB, depth και 3D skeletons. Και σε αυτή την περίπτωση, δεν έχει προηγηθεί οποιαδήποτε επεξεργασία για την αφαίρεση φόντου (στην περίπτωση του τύπου 3D skeletons δεν υπάρχει φόντο) ούτε έχει πραγματοποιηθεί στοίχιση χωρική ή χρονική. Στην ίδια λογική και με τις βάσεις KTH και Weizmann τα SSTIPs έχουν υπολογιστεί στην κλίμακα του γκρι (RGB τύπος ο οποίος έχει μετατραπεί στην κλίμακα του γκρι) και στον τύπο βάθους (ο οποίος είναι στην κλίμακα του γκρι). Το πρωτόκολλο leave-one-person-out cross-validation έχει εφαρμοστεί και σε αυτή την περίπτωση.

Μετά τη μετατροπή των βίντεο στην επιθυμητή ανάλυση, το αποτέλεσμα της συνολικής διαδικασίας που περιγράφηκε στο 6.3.3 χρησιμοποιείται ως είσοδος για την εκπαίδευση μιας σειράς (τόσα όσα και οι κλάσεις σε κάθε περίπτωση) από SVMs. Για να πειραματιστούμε ως προς τις αλλαγές που επιφέρει η αλλαγή του βήματος περιστροφής του επιπέδου στο μετασχηματισμό CTT, το σύστημα εξετάστηκε με την εφαρμογή διαφορετικών γωνιών, 9^o και 6^o μοιρών αντίστοιχα. Μπορεί κανείς διαισθητικά να αντιληφθεί πως όσο μικρότερο είναι το βήμα περιστροφής, τόσο πιο πολύ πλησιάζουμε τη συνεχή μορφή του μετασχηματισμού. Αυτό μπορεί ενδεχομένως να προσφέρει πιο εύρωστα χαρακτηριστικά, είναι κάτι όμως που συμβαίνει σε βάρος της χρονικής αποτελεσματικότητας. Τελικά, με την εφαρμογή PCA στα παραγόμενα διανύσματα, κρατάμε ένα μικρό υποσύνολο των πιο διακριτών χαρακτηριστικών μειώνοντας δραστικά τη διάσταση.

Στη συνέχεια, τα διανύσματα που έχουν παραχθεί χρησιμοποιούνται για την εκπαίδευση ενός αριθμού SVMs όπως προστάζει το πρωτόκολλο που ακολουθεί. Για την εκπαίδευση των SVMs χρησιμοποιήθηκε πυρήνας Gaussian Radial Basis Function. Σε αυτό το σημείο πρέπει να σημειώσουμε ότι η κατηγοριοποίηση της ανθρώπινης κίνησης είναι πρόβλημα πολλαπλών κλάσεων. Αντιμετωπίζουμε το ζήτημα αυτό κατασκευάζοντας το πρόβλημα ως μια γενίκευση δυαδικής κατηγοριοποίησης. Πιο συγκεκριμένα, για κάθε σύνολο δεδομένων, εκπαιδεύσαμε 6, 10 και 12 διαφορετικά SVMs (ένα για κάθε κλάση για την KTH, την Weizmann και την THETIS αντίστοιχα) χρησιμοποιώντας το πρωτόκολλο one-against-all. Η τελική απόφαση λαμβάνεται με την ανάθεση κάθε δείγματος σε μια κλάση $\mathcal{C}_a$, βάσει της απόστασης d του εξεταζόμενου διανύσματος από τα διανύσματα υποστήριξης. Όπου $\mathcal{C}_a$ είναι το σύνολο των δειγμάτων εκπαίδευσης που απαρτίζουν μια κλάση (π.χ. τρέξιμο). Ωστόσο, καθώς θέλουμε να αξιολογήσουμε τον αλγόριθμο κατά ένα πιο γενικευμένο τρόπο, υπολογίζουμε τις επιτυχημένες δυαδικές κατηγοριοποιήσεις για κάθε δείγμα σε κάθε διαφορετικό SVM επάνω στο οποίο εξετάζεται.

6.4.2 Πειράματα στον εντοπισμό πτώσης

Σε αυτή την ενότητα, θα παρουσιάσουμε τα πειραματικά αποτελέσματα και τη διαδικασία η οποία ακολουθήθηκε για την αξιολόγηση της προτεινόμενης τεχνικής εφαρμοσμένη στο σενάριο του εντοπισμού ανθρωπίνων πτώσεων. Γενικά, υπάρχει περιορισμένος αριθμός διαθέσιμων βάσεων δεδομένων που να αφορούν αποκλειστικά στην ανίχνευση και αναγνώριση πτώσεων. Παρ' ολ' αυτά προκειμένου να έχουμε ένα σημείο αναφοράς, έχουμε αξιολογήσει την τεχνική σε δυο πρόσφατα και ευρέως διαθέσιμα σύνολα: Το UR Fall [52],[55] και το Le2i Fall [14].

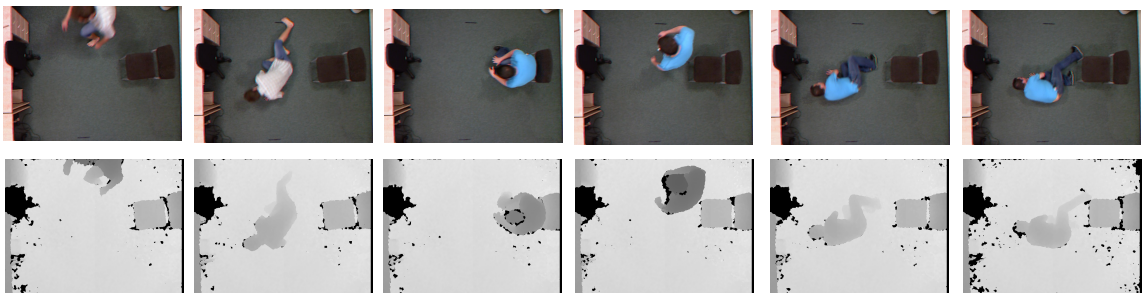
Η βάση δεδομένων UR Fall περιέχει 60 ακολουθίες βιντεοσκοπημένες με τη χρήση 2 καμερών Kinect και τις σχετικές μετρικές επιτάχυνσης. Τα δεδομένα αισθητήρα συλλέχθηκαν με τη βοήθεια των συσκευών PS Move (60Hz) και x-IMU (256Hz). Η βάση περιέχει

ακολουθίες από δεδομένα βάθους και RGB δεδομένα από δυο κάμερες τοποθετημένες (η μια παράλληλα στο δάπεδο και η άλλη στην οροφή αντίστοιχα) και μετρικά δεδομένα επιτάχυνσης. Κάθε βιντεο-ακολουθία είναι αποθηκευμένη σε διαφορετικούς φακέλους με τη μορφή png αρχείων. Από τη συγκεκριμένη βάση δεδομένων έχουμε χρησιμοποιήσει τα δεδομένα βάθους που προέρχονται από την κάμερα που είναι τοποθετημένη στην οροφή και ακολουθήσαμε το πειραματικό πρωτόκολλο που δίνεται από τους συγγραφείς του [52]. Δείγματα από τη βάση URFall δίνονται στο Σχήμα 6.8.

Τα πειράματα επάνω στη βάση UR fall χωρίστηκαν σε δύο φάσεις. Η πρώτη είχε ως στόχο να αξιολογήσει το σενάριο κατά το οποίο η κάμερα είναι τοποθετημένη στην οροφή του δωματίου, κατ' αντιστοιχία με τη μεθοδολογία που παρουσιάστηκε στο [52]. Πιο συγκεκριμένα, χρησιμοποιήθηκε ένα σύνολο από 60 τμηματοποιημένες ακολουθίες κίνησης από το υποσύνολο των βίντεο πληροφορίας βάθους. Οι συγκεκριμένες ακολουθίες κίνησης περιέχουν τόσο ακούσιες πτώσεις, όπως για παράδειγμα γλίστρημα και πέσιμο από καρέκλα, όσο και καθημερινές δραστηριότητες όπως περπάτημα στο δωμάτιο. Για να κάνουμε το πρόβλημα ακόμα πιο δύσκολο, εισαγάγαμε στη βάση δεδομένων τεχνητά δημιουργημένες δραστηριότητες τύπου πτώσης. Ακολουθίες κίνησης προσώπων τα οποία σχεδόν πέφτουν από την καρέκλα αλλά επανέρχονται στην αρχική του θέση, δημιουργήθηκαν χειροκίνητα και προστέθηκαν στο σύνολο δεδομένων. Για την εξαγωγή της δυαδικής σιλουέτας και της σιλουέτας πληροφορίας βάθους, έγινε αφαίρεση φόντου ενώ χρησιμοποιήθηκαν τεχνικές κατοφλίωσης και μείωσης θορύβου. Τα χωρο-χρονικά σημεία ενδιαφέροντος υπολογίστηκαν στις συγκεκριμένες σιλουέτες.

Στη δεύτερη φάση, πειραματιστήκαμε με τα έγχρωμα βίντεο RGB τα οποία αφορούν σε εμπρόσθιες λήψεις και κατ' αντιστοιχία με τα πειράματα που διενεργήθηκαν στο [55]. Στο συγκεκριμένο στάδιο, έγινε μια προσπάθεια να αξιολογήσουμε πλήρως τις δυνατότητες του προτεινόμενου συστήματος σε περιβάλλοντα όπου η αφαίρεση φόντου δεν αποτελεί μια ασήμαντη διαδικασία μη επιρρεπή στα σφάλματα. Για αυτό το λόγο, δεν έγινε κανένας διαχωρισμός σιλουέτας και τα πειράματα βασίστηκαν αποκλειστικά στη χωροχρονική πληροφορία από την κίνηση των υποκειμένων μέσα στο βίντεο. Αυτό το τμήμα της βάσης χρησιμοποιήθηκε χωρίς κατάτμηση, περιλαμβάνοντας για παράδειγμα σε κάθε βίντεο ένα ολοκληρωμένο σύνολο κινήσεων όπως κάποιον ο οποίος εισέρχεται σε ένα δωμάτιο, περπατάει και στη συνέχεια πραγματοποιεί μια εκούσια ή ακούσια πτώση. Δραστηριότητες η οποίες μοιάζουν με πτώσεις όπως κάθομαι οκλαδόν σε ένα καναπέ, ξαπλώνω σε ένα κρεβάτι ή σκύβω να δέσω κορδόνια, έχουν προστεθεί στη βάση.

Η βάση Le2i Fall έχει ληφθεί σε ρεαλιστικά περιβάλλοντα με τη χρήση μιας απλής RGB κάμερας. Η ταχύτητα λήψης είναι 25 καρέ/δευτερόλεπτο και η ανάλυση είναι 320x240 εικονοστοιχεία (pixels). Τα δεδομένα βίντεο που έχουν ληφθεί απεικονίζουν τις κύριες δυσκολίες που μπορούν να παρουσιαστούν στο χώρο ενός ηλικιωμένου ή σε ένα συνη-



Σχήμα 6.8: *Δείγματα καρέ από τη βάση δεδομένων UR Fall για δύο διαφορετικές πτώσεις. Η επάνω σειρά παρέχει δείγματα RGB ενώ η κάτω δίνει τις αντίστοιχες εικόνες βάθους.*

Σχήμα 6.9: *Δείγματα καρέ από τη βάση Le2i. Η επάνω σειρά αναπαριστά δείγματα από καθημερινές δραστηριότητες στο περιβάλλον "CoffeeRoom" ενώ η κάτω σειρά παρέχει δείγματα από μια πτώση που έχει λάβει χώρα στο περιβάλλον "Home".*

θισμένο περιβάλλον εργασίας (γραφείο). Τα βίντεο παρουσιάζουν διαφοροποιήσεις στις φωτεινότητα και τυπικές δυσκολίες όπως επικαλύψεις ή ανεπιθύμητο και ιδιαίτερης υφής φόντο. Οι ηθοποιοί εκτέλεσαν ποικίλες δραστηριότητες καθημερινής φύσης καθώς και πτώσεις. Το σύνολο δεδομένων περιέχει 130 υποσημειωμένα βίντεο, με επιπλέον πληροφορία που αναπαριστά τη στάθμη αληθείας (groundtruth) του σημείου πτώσης στην ακολουθία εικόνων. Η βάση δεδομένων παρουσιάζει διαφορετικές τοποθεσίες για έλεγχο και για εκπαίδευση ενώ οι συγγραφείς [14] έχουν ορίσει διάφορα πρωτόκολλα για τον έλεγχο της μεθόδου τους. Δουλεύοντας με τη συγκεκριμένη βάση, ακολουθήσαμε το πρωτόκολλο P1 όπως δίνεται στην προαναφερθείσα δημοσίευση, όπου τα σύνολα για έλεγχο και εκπαίδευση έχουν δημιουργηθεί με βίντεο από τα υποσύνολα "Home" και "Coffee room". Δείγματα από τη βάση Le2i παρέχονται στο Σχήμα 6.9.

Και στα δύο πρώτα πειραματικά σενάρια τα οποία εφαρμόστηκαν στις βάσεις UR και Le2i με την εφαρμογή του προτεινόμενου συστήματος, προηγήθηκε εξαγωγή σιλουέτας. Στην περίπτωση του συνόλου UR, αυτό πραγματοποιήθηκε με τον υπολογισμό των δια-

φορών μεταξύ των εικονοστοιχείων βάθους σε κάθε τρέχον καρέ και τα αντίστοιχα εικονοστοιχεία σε ένα προ-υπολογισμένο καρέ αναφοράς. Το καρέ αναφοράς δημιουργήθηκε υπολογίζοντας τη μέση τιμή κάθε διάμεσου (median) εικονοστοιχείου και η χρήση του είχε σαν αποτέλεσμα την αφαίρεση μιας σημαντικής ποσότητας θορύβου η οποία δημιουργείται στο φόντο από τον αισθητήρα βάθους.

Κάποιος μπορεί να συσχετίσει την ανθρώπινη παρουσία σε ένα συγκεκριμένο καρέ με τις περιπτώσεις όπου η διαφορά μεταξύ των τιμών των εικονοστοιχείων του καρέ και του καρέ αναφοράς, ξεπερνάει ένα συγκεκριμένο κατώφλι. Στην περίπτωση της βάσης UR, προκειμένου να προσδώσουμε ευρωστία, έγινε χρήση συνολικά τεσσάρων κατωφλίων. Τα πρώτα δυο χρησιμοποιήθηκαν για να φιλτράρουν θορυβώδη και άκυρα εικονοστοιχεία. Προκειμένου η τιμή ενός εικονοστοιχείου να θεωρηθεί έγκυρη, (π.χ. να είναι πιθανώς τμήμα μιας ανθρώπινης σιλουέτας) απαιτείτο η τιμή αυτή να βρίσκεται μεταξύ 1100 και 3620mm. Πρέπει να σημειωθεί ότι αυτό το εύρος τιμών αντιπροσωπεύει την απόσταση μεταξύ του αισθητήρα που είναι τοποθετημένος στην οροφή, για παράδειγμα το Kinect και του οποίου οι τιμές βάθους μετριούνται σε μιλιμέτρ, στο εύρος τιμών 800-4000. Στη συνέχεια, για να αναδειχθεί μια ανθρώπινη κίνηση, η τιμή ενός εικονοστοιχείου στο τρέχον καρέ και του αντίστοιχου εικονοστοιχείου αναφοράς, θα πρέπει να είναι μεταξύ 50 και 2200mm. Αυτές οι τιμές βρέθηκαν να είναι οι καταλληλότερες προσφέροντας τη μέγιστη ανεκτικότητα στον τυχαίο θόρυβο.

Δεδομένου ότι η βάση Le2i αποτελείται από βίντεο τύπου RGB χαμηλής ανάλυσης, έγινε χρήση της βάσης με ένα διαφορετικό τρόπο. Επιπλέον, οι συνθήκες φωτισμού στις περισσότερες των περιπτώσεων (ειδικά στο υποσύνολο "Home") κατέστησαν τη χρήση διαφορών μεταξύ τρέχοντος καρέ και καρέ αναφοράς, αναξιόπιστη. Για τη εξαγωγή της ανθρώπινης σιλουέτας και την αφαίρεση του φόντου, χρησιμοποιήθηκε η μέθοδος που προτάθηκε από τον Zivkovic στο [136] και στο [137]. Σε αυτή την τεχνική, πραγματοποιείται αφαίρεση μεταξύ του τρέχοντος καρέ και ενός μοντέλου φόντου. Αυτό το μοντέλο επικαιροποιείται διαρκώς σε επίπεδο εικονοστοιχείου, με τη χρήση μια μεθόδου βασιμένης σε μίξη Γκαουσιανής (gaussian mixture), ούτως ώστε να περιέχει ότι θεωρείται ως το στατικό μέρος της σκηνής και προσαρμόζεται στις αλλαγές της σκηνής μέσα στη βιντεοακολουθία.

Στα δικά μας πειράματα, οι ακολουθίες και των δύο συνόλων δεδομένων, έχουν υποκλιμακωθεί στην ανάλυση 320*240 εικονοστοιχείων με χρονική διάρκεια 26 και 12 καρέ κατά μέσο όρο για την UR και την Le2i αντίστοιχα. Για το σενάριο της προσαρμοσμένης στην οροφή κάμερα, στις βάσεις UR και Le2i, τα δείγματα εκπαίδευσης και ελέγχου δημιουργήθηκαν με χειροκίνητη τμηματοποίησή τους σε κινησιο-ακολουθίες οι οποίες να περιέχουν μόνο το τμήμα των πτώσεων ή των παραλίγο πτώσεων. Η μέθοδος που χρησιμοποιήθηκε για την εξαγωγή των διανυσμάτων χαρακτηριστικών είναι η ίδια η οποία

ακολουθήθηκε και στα πειράματα της αναγνώρισης δραστηριότητας και περιγράφηκε νωρίτερα.

Σε αυτό το σημείο, θα πρέπει να επισημάνουμε ότι δεν υπάρχει κάποιο ενιαίο πρότυπο για την αξιολόγηση των αλγορίθμων εντοπισμού πτώσεων. Στα πειράματα τα οποία διεξήχθησαν σε αυτή τη μελέτη, χρησιμοποιήθηκε ένα απλό πρωτόκολλο leave-one-sample-out. Σε αντίθεση με τη διαδικασία που ακολουθήθηκε στα πειράματα της αναγνώρισης δραστηριότητας, η έλλειψη πολλαπλών δειγμάτων ανά συμμετέχοντα στις βάσεις δεδομένων, αποτρέπει την υιοθέτηση του πρωτοκόλλου leave-one-person-out. Σε κάθε επανάληψη, χρησιμοποιείται ένα διαφορετικό δείγμα δραστηριότητας ανεξάρτητα από το αν εμπίπτει σε πτώση ή όχι το οποίο στη συνέχεια ελέγχεται σε σχέση με τη βάση. Η βάση έχει πρωτίστως εκπαιδεύσει το σύστημα και περιέχει όλα τα εναπομείναντα δείγματα του αρχικού συνόλου. Τα αποτελέσματα της πειραματικής διαδικασίας παρατίθενται στην ακόλουθη ενότητα (6.4.3).

6.4.3 Αποτελέσματα

Συγκριτικά αποτελέσματα για το σενάριο της αναγνώρισης δραστηριότητας, παρέχονται στον Πίνακα 6.2. Επιπρόσθετα, οι Πίνακες 6.3, 6.4, 6.5, 6.6 και 6.7 παρέχουν τους πίνακες σύγχυσης (confusion matrices) οι οποίοι δημιουργήθηκαν με την εφαρμογή της προτεινόμενης μεθόδου στις εξεταζόμενες βάσεις δεδομένων. Οι σειρές των πινάκων αντιστοιχούν στην ακρίβεια που επετεύχθη (σωστές απαντήσεις/όλες τις απαντήσεις) από όλα τα εκπαιδευμένα SVMs. Οι στήλες από την άλλη, δείχνουν την απόδοση που παρέχεται από τα επί μέρους SVMs για κάθε κλάση ξεχωριστά. Τα συνολικά αποτελέσματα για την περίπτωση του εντοπισμού πτώσης παρέχονται στον Πίνακα 6.8.

Όπως φαίνεται στον Πίνακα 6.2, η μέθοδος εξαγωγής χαρακτηριστικών η οποία βασίζεται στον 3D CTT και τα SSTIPs επιτυγχάνει εντυπωσιακή ακρίβεια σε όλα τα εξεταζόμενα σύνολα δεδομένων και παρέχει ισότιμα ή συνήθως ανώτερα αποτελέσματα σε σύγκριση με άλλες προτεινόμενες μεθόδους που έχουν αξιολογηθεί στα ίδια σύνολα. Πίο συγκεκριμένα, στη βάση KTH η παρουσιαζόμενη τεχνική επιτυγχάνει μια εντυπωσιακή ακρίβεια της τάξεως του 99.8% η οποία πιστοποιείται και από τον αντίστοιχο πίνακα σύγχυσης (Πίνακας 6.3). Σε μια αξιοσημείωτη σύγκριση, η μέθοδος που προτάθηκε από τους Yuan et al. στο [131] και η οποία βασίζεται σε χαρακτηριστικά που εξάγονται με τη χρήση μιας μορφής του 3D Radon και ένα συνδιασμό STIPs και BoVW, επιτυγχάνει 95.49% ακρίβεια στην ίδια βάση δεδομένων. Τα αποτελέσματα στην Weizmann, όπου επετεύχθη ακρίβεια 96.34%, καταδεικνύουν μια μικρή σύγχυση σε επιμέρους κατηγοριοποιητές και ειδικότερα σε αυτούς που έχουν εκπαιδευτεί στην κλάση ”jump” και στην συγγενή της ”pjump”.

Όπως μπορεί να παρατηρηθεί υπάρχει μια αξιοσημείωτη διαφορά στην απόδοση της μεθόδου εφαρμοζόμενη επάνω στους τρεις διαφορετικούς τύπους δεδομένων του συνό-

Πίνακας 6.2: *Ποσοστά κατηγοριοποίησης (%) που επετεύχθησαν από διαφορετικές δημοσιευμένες μεθόδους στις βάσης δεδομένων KTH, Weizmann και THETIS.*

Method	Dataset				
	KTH	Weizmann	THETIS-Skelet3D	THETIS-Depth	THETIS-RGB
3D CTT	**99.98**	**96.34**	**86.06**	**98.03**	**100**
Selective STIPs + BoV [13]	96.35	99.5	-	-	-
Dense Trajectories: MBH [120] (rep. in [37])	92.32	-	46.84	51.59	-
Dense Trajectories: Combination [120] (rep. in [37])	90.65	-	50.78	54.32	-
Dense Trajectories: Trajectory [120] (rep. in [37])	86.98	-	53.08	57.5	-
HTFs [35]	93.14	95.42	-	-	-
Yuan et al. [131]	95.49	-	-	-	-
Vainstein et al. [118]	-	-	-	-	86.44
Wong and Cipolla [124]	86.5	-	-	-	-
Sun et al. [110]	94	97.8	-	-	-
Liu and Shah [61]	94.16	-	-	-	-
Dollar et al. [21]	81.2	-	-	-	-
Schuldt et al. [100] (reported in [91])	50.33	-	-	-	-
Rapantzikos et al. [91]	88.3	-	-	-	-
Oikonomopoulos et al. [81] (reported in [124])	74.79	-	-	-	-
Ke et al. [50]	80.9	-	-	-	-
Schuldt et al. [100]	71.7	-	-	-	-
Niebles et al. [78]	81.5	72.8	-	-	-
Jiang et al. [46]	84.4	-	-	-	-
Laptev et al. [58] (reported in [37]	92.99	-	54.4	60.23	-
Klasser et al. [53]	-	84.3	-	-	-
Jhuang et al. [44]	-	96.3	-	-	-
Thurau [112]	-	86.66	-	-	-
Thurau et al. [113]	-	94.4	-	-	-

Πίνακας 6.3: *Πίνακας σύγχυσης παραγόμενος από την προτεινόμενη μέθοδο στη βάση δεδομένων KTH.*

	boxing	handclapping	handwaving	jogging	running	walking
boxing	**1**	0	0	0	0	0
handclapping	0.0009	**0.9991**	0	0	0	0
handwaving	0	0	**1**	0	0	0
jogging	0	0	0	**1**	0	0
runing	0	0	0,01	0	**1**	0
walking	0	0	0	0	0	**1**

Πίνακας 6.4: *Πίνακας σύγχυσης παραγόμενος από την προτεινόμενη μέθοδο στη βάση δεδομένων Weizmann.*

	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	**1**	0	0.1111	0	0	0	0	0	0	0
jack	0	**1**	0.0455	0	0	0	0	0	0	0
jump	0.0588	0.0294	**0.9706**	0.0294	0.1176	0.0294	0.0588	0.0294	0.0588	0.0294
pjump	0.0303	0.0303	0.0303	**0.9697**	0.0303	0.0303	0.0303	0.0303	0.0303	0.0303
run	0.1053	0.1579	0.2105	0.1053	**0.9474**	0.0526	0.1579	0.0526	0.1053	0.0526
side	0	0	0	0	0	**1**	0.0357	0	0	0
skip	0.0526	0.0526	0.1053	0.0526	0.0789	0.0526	**0.9474**	0.0526	0.0526	0.0526
walk	0.0455	0.0455	0.1364	0	0.0455	0	0.0455	**1**	0	0
wave1	0.0476	0.0476	0.0476	0	0.0952	0.0476	0.0952	0.0476	**0.9524**	0.0476
wave2	0	0	0	0	0	0	0	0	0	**1**

λου THETIS. Στα συγκριτικά αποτελέσματα του Πίνακα 6.2 μπορεί εύκολα να παρατηρηθεί ότι η τεχνική που προτάθηκε ξεπερνά όλες τις μεθόδους που έχουνε εφαρμοστεί σε

αυτή τη βάση όπως τα γνωστά HOG-HOF τα οποία βασίζονται στα STIPs του Laptev [58] και τα dense trajectories των Wang et al. [120] (στα σύνολα depth και Skelet3D) και την τεχνική dynamic phases των Vainstein et al. [118] (στο σύνολο RGB).

Παρ' όλο που η απόδοση του αλγορίθμου στα δεδομένα Skelet3D δεν μπορεί να θεωρηθεί μη ικανοποιητική, μπορεί να παρατηρηθεί στον αντίστοιχο Πίνακα 6.6, μια σύγχυση στους κατηγοριοποιητές που αφορούν στις κατηγορίες "forehand open", "forehand flat" και "forehand volley". Ως ένα σημείο, η σύγχυση αυτή είναι αναμενόμενη βάσει της φύσεως των εξεταζόμενων δεδομένων. Οι σκελετοί είναι ένα υπεραπλουστευμένο μοντέλο του ανθρώπινου σώματος με ελάχιστη επιφάνεια η οποία δυσχεραίνει την ακριβή εξαγωγή των STIPs. Ωστόσο, συγκρινόμενη με άλλη μέθοδο που κάνει χρήση των STIPs [58], η προτεινόμενη μέθοδος δείχνει να παρέχει σημαντική αναβάθμιση. Από την άλλη, στο σύνολο δεδομένων τύπου depth, τα αποτελέσματα είναι εντυπωσιακά. Το γεγονός αυτό πιστοποιείται από τον Πίνακα 6.5. Τελικά, ακρίβεια της τάξεως του 100% αναφέρεται στην εφαρμογή του αλγορίθμου στα δεδομένα τύπου RGB. Ειδικά τα τελευταία αποτελέσματα, πιστοποιούν την καταλληλότητα της μεθόδου για δεδομένα τύπου RGB καθώς και για δεδομένα της κλίμακας του γκρι.

Πίνακας 6.5: *Πίνακας σύγχυσης παραγόμενος από την προτεινόμενη μέθοδο στη βάση δεδομένων THETIS στα δεδομένα τύπου Depth.*

	backhand	backhand2h	bslice	bvolley	foreflat	foreopen	fslice	fvolley	serflat	serkick	serslice	smash
backhand	**0.7758**	0	0	0	0	0	0	0	0	0	0	0
backhand2h	0.2182	**1**	0	0	0	0	0	0	0	0	0	0
bslice	0.2182	0	**1**	0	0	0	0	0	0	0	0	0
bvolley	0.1939	0	0	**1**	0	0	0	0	0	0	0	0
foreflat	0.2485	0	0	0	**1**	0	0	0	0	0	0	0
foreopen	0.2727	0	0	0	0	**1**	0	0	0	0	0	0
fslice	0.2545	0	0	0	0	0	**1**	0	0	0	0	0
fvolley	0.2121	0	0	0	0	0	0	**1**	0	0	0	0
serflat	0.2121	0	0	0	0	0	0.0061	0	**1**	0	0	0.0061
serkick	0.2303	0	0	0	0	0	0	0	0	**1**	0	0
serslice	0.2242	0	0	0	0	0	0	0	0	0	**1**	0
smash	0.3091	0	0	0	0	0	0	0	0	0	0	**1**

Πίνακας 6.6: *Πίνακας σύγχυσης παραγόμενος από την προτεινόμενη μέθοδο στη βάση δεδομένων THETIS στα δεδομένα τύπου Skelet3D.*

	backhand	backhand2h	bslice	bvolley	foreflat	foreopen	fslice	fvolley	serflat	serkick	serslice	smash
backhand	**1**	0	0	0	0.6404	0.2584	0	0.7303	0	0	0.0112	0.1011
backhand2h	0	**1**	0	0	0.6262	0.2991	0	0.729	0	0	0	0.0654
bslice	0	0.01	**1**	0	0.63	0.26	0	0.71	0	0	0	0.05
bvolley	0	0	0	**1**	0.6778	0.3556	0	0.6778	0	0	0.0111	0.0222
foreflat	0	0	0	0	**0.4091**	0.2909	0	0.7545	0	0	0	0.0455
foreopen	0	0	0	0	0.5227	**0.75**	0	0.6932	0	0	0	0.0455
fslice	0	0	0	0	0.6495	0.2784	**1**	0.8247	0	0	0	0.0515
fvolley	0	0.0253	0	0.0127	0.5949	0.2405	0	**0.3165**	0	0	0	0.1266
serflat	0	0	0	0	0.5698	0.314	0	0.686	**1**	0	0	0.0465
serkick	0	0.0183	0	0	0.5506	0.2661	0	0.7706	0	**1**	0	0.0642
serslice	0	0.01	0	0	0.53	0.34	0	0.8	0	0	**0.98**	0.06
smash	0	0	0	0	0.3333	0.2667	0	0.8444	0	0	0	**0.9889**

Σχετικά με το σενάριο του εντοπισμού πτώσης, όπως φαίνεται και από τον Πίνακα 6.8, η προτεινόμενη μέθοδος επιτυγχάνει αποτελέσματα που είναι συγκρίσιμα με τις τεχνι-

Πίνακας 6.7: *Πίνακας σύγχυσης παραγόμενος από την προτεινόμενη μέθοδο στη βάση δεδομένων THETIS στα δεδομένα τύπου RGB.*

	backhand	backhand2h	bslice	bvolley	foreflat	foreopen	fslice	fvolley	serflat	serkick	serslice	smash
backhand	**1**	0	0	0	0	0	0	0	0	0	0	0
backhand2h	0	**1**	0	0	0	0	0	0	0	0	0	0
bslice	0	0	**1**	0	0	0	0	0	0	0	0	0
bvolley	0	0	0	**1**	0	0	0	0	0	0	0	0
foreflat	0	0	0	0	**1**	0	0	0	0	0	0	0
foreopen	0	0	0	0	0	**1**	0	0	0	0	0	0
fslice	0	0	0	0	0	0	**1**	0	0	0	0	0
fvolley	0	0	0	0	0	0	0	**1**	0	0	0	0
serflat	0	0	0	0	0	0	0	0	**1**	0	0	0
serkick	0	0	0	0	0	0	0	0	0	**1**	0	0
serslice	0	0	0	0	0	0	0	0	0	0	**1**	0
smash	0	0	0	0	0	0	0	0	0	0	0	**1**

κές που ανήκουν στην τελευταία λέξη της τεχνολογίας και οι οποίες έχουνε εφαρμοστεί στα σύνολα δεδομένων UR και Le2i. Αυτές οι μέθοδοι και ειδικότερα αυτές που παρουσιάστηκαν στα [52], [14], δείχνουν να περιορίζονται αρκετά στο πεδίο δράσης και εξαρτώνται αυστηρά από τη μορφή της ανθρώπινης σιλουέτας και του πλαισίου οριοθέτησης (bounding box), έχοντας αυστηρούς συσχετισμούς με την τοποθέτηση του μέσου λήψης. Αυτό έρχεται σε αντίθεση με την τεχνική που βασίζεται στον 3D CTT και η οποία παρέχει την εξαγωγή γενικευμένων χαρακτηριστικών που επιτρέπουν τη κατηγοριοποίηση ανεξάρτητα από την τοποθέτηση του μέσου λήψης.

Αυτό που είναι επίσης αξιοσημείωτο, είναι η ικανότητα του αλγορίθμου που βασίζεται στον 3D CTT, να επιβεβαιώσει μια ακούσια πτώση στα βίντεο τύπου RGB τα οποία περιέχουν επίσης και άλλες δραστηριότητες. Η ακρίβεια φτάνει το 95.71% και επιτυγχάνεται βασισμένη αποκλειστικά στα δεδομένα εικόνας. Αντίστοιχα, η μέθοδος που προτείνεται στο [55] επιτυγχάνει, εφαρμοζόμενη στα ίδια δεδομένα, ακρίβεια 90% η οποία αυξάνει στο 98.33% μόνο όταν εφαρμοστούν συνδυαστικά και δεδομένα επιταχυνσιόμετρου. Ένα άλλο γεγονός που αξίζει προσοχής, είναι η ικανότητα του αλγορίθμου να κατηγοριοποιήσει ένα περιστατικό ως πτώση, σε σιλουέτες οι οποίες έχουν εξαχθεί αυτοματοποιημένα και οι οποίες έχουν ληφθεί από κάμερες αυθαίρετα τοποθετημένες. Σε αυτού του είδους την εφαρμογή, η προτεινόμενη τεχνική επιτυγχάνει ακρίβεια 96.34%, ενώ η μέθοδος που προτάθηκε στο [14] επιτυγχάνει 95.06%, πιθανότατα ελαχιστοποιώντας το συνολικό σφάλμα κατά 2.5% με την χειροκίνητη επισημείωση των πλαισίων οριοθέτησης σε κάθε σκηνή.

Πίνακας 6.8: *Ακρίβεια κατηγοριοποίησης (%) που επετεύχθη από το προτεινόμενο σύστημα και από άλλες δημοσιευμένες μεθόδους, για το σενάριο του εντοπισμού πτώσης.*

	UR		Le2i
	Ceiling mounted	RGB frontal	
3D CTT on binary shil.	**100**	-	**96,34**
3D CTT on depth shil.	**95.92**	-	-
3D CTT on RGB sequences	-	**95.71**	-
Kepski and Kwolek [52]	100	-	-
Kepski and Kwolek [55]	-	90	-
Charfi et al. [14]	-	-	95.06

Κεφάλαιο 7

Σύνοψη-Συμπεράσματα

Η αναγνώριση της ανθρώπινης δραστηριότητας αποτελεί ένα από τα πιο καυτά θέματα στο πεδίο της μηχανικής μάθησης και της αναγνώρισης προτύπων καθώς αφορά ολοένα και περισσότερες εφαρμογές από την απλή αλληλεπίδραση ανθρώπου μηχανής, έως την ρομποτική. Εκτός από τα συνήθη προβλήματα που παρουσιάζονται σε τεχνικές που αφορούν στα συγκεκριμένα πεδία, η αναγνώριση της ανθρώπινης δραστηριότητας εισάγει νέα προβλήματα που πηγάζουν από τον ανθρώπινο παράγοντα καθώς και την ιδιαιτερότητα του κάθε ατόμου που συμμετέχει σε μια τέτοιου είδους διαδικασία. Ένα από τα σημαντικότερα θέματα που απασχολούν τους ερευνητές σε σχέση με το πρόβλημα της αναγνώρισης της ανθρώπινης δραστηριότητας, είναι το πως θα ενσωματωθεί όσο το δυνατόν μεγαλύτερο μέρος της χωρικής και χρονικής πληροφορίας που περιέχεται σε μια κίνηση, σε ένα διάνυσμα εξαχθέντων χαρακτηριστικών που στη συνέχεια θα χρησιμοποιηθεί για την κατηγοριοποίησή της.

Στην παρούσα διατριβή, παρουσιάστηκε η συνεισφορά μας στο πεδίο της όρασης και της τεχνητής νοημοσύνης και πιο συγκεκριμένα σε εφαρμογές που αφορούν στην αναγνώριση της ανθρώπινης δραστηριότητας και την εξαγωγή χαρακτηριστικών για το σκοπό αυτό. Αρχικά, προτείναμε τη χρήση του μετασχηματισμού Radon (Κεφάλαιο2) για τη δημιουργία εύρωστων στο θόρυβο χαρακτηριστικών εκμεταλλευόμενοι τις φυσικές του ιδιότητες. Στη συνέχεια, προτείναμε μια μέθοδο η οποία καταφέρνει να ενσωματώσει τα χαρακτηριστικά αυτά και να αναπαραστήσει ολόκληρη την κίνηση σε ένα τελικό δισδιάστατο template το οποίο εμπεριέχει πολλή από την προαναφερθείσα πληροφορία.

Πιο συγκεκριμένα, μια περίοδος μιας κίνησης που εκτελείται από κάποιο άτομο, εισέρχεται στο σύστημα με τη μορφή μιας εικονοσειράς (βίντεο). Από κάθε καρέ του βίντεο, εξάγεται η σιλουέτα του ατόμου σε δυαδική μορφή χωρίς την εφαρμογή κάποιου εξεζητημένου αλγορίθμου. Στη συνέχεια, από κάθε δυαδική σιλουέτα εξάγεται ένας μετασχηματισμός Radon που με τη σειρά του ενσωματώνεται σε ένα τελικό template. Αυτή η τελική αναπαράσταση περιγράφει το σύνολο της κίνησης ενώ καταφέρνει να εμπεριέχει

πολλή από την πολύτιμη χωρο-χρονική πληροφορία της κίνησης.

Πειράματα στη βάση δεδομένων KTH έδειξαν ότι ο αλγόριθμος μπορεί να ανταπεξέλθει σε πολύ απαιτητικές συνθήκες όπως αυτές της χρησιμοποιούμενης βάσης. Αξίζει να σημειωθεί, πως προκειμένου να εξετάσουμε την ευρωστία του αλγορίθμου στο θόρυβο, η εξαγωγή των σιλουετών γίνεται με τον απλούστερο δυνατό τρόπο έχοντας σαν αποτέλεσμα αρκετά θορυβώδεις, πολλές φορές ανολοκλήρωτες ή και παραμορφωμένες σιλουέτες. Εξετάζοντας τον αλγόριθμο με τη χρήση SVM πυρήνα RBF προέκυψε ένα ποσοστό κατηγοριοποίησης της τάξεως του 87.7% αναδεικνύοντας τις δυνατότητες της προτεινόμενης μεθόδου. Αν και η μέθοδος λόγω του ότι ο μετασχηματισμός εφαρμόζεται στο κέντρο της σιλουέτας είναι ανεπηρέαστη σε μετατοπίσεις, δεν είναι εξίσου ανθεκτική σε άλλου είδους παραμορφώσεις όπως κλιμακώσεις και περιστροφές.

Στη συνέχεια της εργασίας μας (Κεφάλαιο 3), βασιζόμενοι στην ιδέα της δημιουργίας εύρωστων χαρακτηριστικών με τη χρήση του μετασχηματισμού Radon, επεκτείναμε την αρχική ιδέα προτείνοντας τη χρήση ενός νεότερου και πιο ευμετάβλητου μετασχηματισμού, του ονομαζόμενου μετασχηματισμού Trace. Για την ακρίβεια, ο μετασχηματισμός Trace αποτελεί την εξέλιξη του Radon, ενώ ταυτόχρονα ο Radon αποτελεί υποπερίπτωση του. Εκμεταλλευόμενοι την ευελιξία που παρέχει ο Trace, εφαρμόζοντας μια πληθώρα διαφορετικών συναρτησιακών, δημιουργήσαμε μια σειρά μετασχηματισμών με διαφορετικές ιδιότητες που με τη σειρά τους προσδίδουν διαφορετική συμπεριφορά στο τελικό template που δημιουργείται με τον τρόπο που περιγράφεται παραπάνω.

Εξετάζοντας πειραματικά την απόδοση του αλγορίθμου σε διάφορες απαιτητικές συνθήκες βιντεοσκόπησης, αναδείξαμε την ικανότητα να δημιουργούμε διαφορετικά χαρακτηριστικά με διαφορετική συμπεριφορά για αντίστοιχα εξειδικευμένες απαιτήσεις. Ακολουθώντας την προαναφερθείσα πειραματική διαδικασία για τις βάσεις δεδομένων KTH και Weizmann και τις ίδιες θορυβώδεις σιλουέτες, επιτύχαμε ποσοστό ορθής κατηγοριοποίησης ίσο με 90,22% και 93,4% (για τις βάσεις KTH και Weizmann αντίστοιχα) δοκιμάζοντας μια σειρά διαφορετικών συναρτησιακών, χωρίς αυτό να σημαίνει σε καμία περίπτωση ότι εξαντλήσαμε τις πιθανές επιλογές.

Στο πρώτο μέρος του Κεφαλαίου 3, παρουσιάσαμε πως ο μετασχηματισμός Trace μπορεί να χρησιμοποιηθεί για την αναπαράσταση μιας ολόκληρης κίνησης. Ωστόσο, η συγκεκριμένη αναπαράσταση μπορεί να χρησιμοποιηθεί κατά κύριο λόγο στην περίπτωση που οι κινήσεις έχουν βιντεοσκοπηθεί κάτω από τις ίδιες συνθήκες (γωνία λήψης, κλίμακα, περιστροφή κάμερας κτλ.). Καθώς αυτό δεν είναι κάτι σύνηθες στις περισσότερες από τις εφαρμογές που αφορούν στην αναγνώριση της ανθρώπινης δραστηριότητας, προτείνουμε στη συνέχεια μια πιο εξελιγμένη τεχνική η οποία ξεπερνά πολλούς από τους παραπάνω περιορισμούς.

Ερευνώντας τις δυνατότητες του μετασχηματισμού Trace στην αναγνώριση της αν-

θρώπινης δραστηριότητας, δημιουργήσαμε και προτείναμε έναν ακόμη τρόπο για την δημιουργία χαρακτηριστικών, αυτή τη φορά όχι μόνο εύρωστων στο θόρυβο, αλλά και αμετάβλητων σε διάφορες παραμορφώσεις που παρατηρούνται συχνά στη λήψη ενός βίντεο, όπως η περιστροφή, η μετατόπιση και η κλιμάκωση των αντικειμένων (Δεύτερο μέρος του Κεφαλαίου 3). Η τεχνική που προτείναμε εκμεταλλεύεται τις ιδιότητες που μπορεί να προσδώσει η εφαρμογή διαφόρων συναρτησιακών στο μετασχηματισμό Trace και εισαγάγει έναν νέο, απλό και αποτελεσματικό τρόπο για τη δημιουργία χαρακτηριστικών που μπορεί να δώσει λύση σε πολλά από τα γνωστά προβλήματα της αναγνώρισης της ανθρώπινης δραστηριότητας.

Οι συγγραφείς του [49] έχουν αποδείξει ότι εξάγοντας τα ονομαζόμενα "triple features", μπορούν να παραχθούν πολύ εύρωστα χαρακτηριστικά για την κατηγοριοποίηση διαφορετικών αλλά πολύ συγγενών κλάσεων (π.χ. αναγνώριση διαφορετικών ειδών ψαριών). Κάνοντας χρήση της μεθόδου για την εξαγωγή των triple features, εξελίξαμε την τεχνική και προτείναμε μια νέα μέθοδο για την αναπαράσταση κινήσεων από βιντεο-ακολουθίες. Έχοντας εξαγάγει τις σιλουέτες, μετασχηματίζουμε το χώρο αυτό (που περιέχει τις σιλουέτες), στο χώρο του μετασχηματισμού Trace. Ακολουθώντας την διαδικασία που περιγράφεται στο 3.2.1 για την εξαγωγή των triple features, δημιουργείται ένα σύνολο αντίστοιχων χαρακτηριστικών. Ο λόγος ενός ζεύγους τέτοιων χαρακτηριστικών όπως έχει δειχθεί, μπορεί να είναι αμετάβλητος σε διάφορα είδη παραμορφώσεων, κατά αντιστοιχία με τα συναρτησιακά που έχουν χρησιμοποιηθεί. Αυτά τα συναρτησιακά μπορούν να έχουν επιλεχθεί ούτως ώστε να είναι ευαίσθητα ή σχετικά ανεπηρέαστα από τις πιθανές μεταβολές που συμβαίνουν στα βίντεο κινήσεων, ενώ την ίδια στιγμή διατηρούν την διακριτικότητα τους. Με την εφαρμογή της μεθόδου που προτείναμε, ολόκληρη η κίνηση εκφράζεται από ένα νέο διάνυσμα, το οποίο ονομάστηκε History Triple Features (HTFs) και το οποίο αποτελείται από ένα σύνολο ευαίσθητων ή σχετικά ανεπηρέαστων χαρακτηριστικών.

Αξίζει να σημειωθεί πως με την εφαρμογή μιας απλουστευμένης τεχνικής για τη μείωση των διαστάσεων, όπως η γραμμική διακριτική ανάλυση (LDA), η αναπαράσταση της κίνησης γίνεται τελικά με ένα διάνυσμα πάρα πολύ μικρής διάστασης. Το διάνυσμα αυτό αν και δεν παρουσιάζει στοιχεία που είναι κατανοητά στην ανθρώπινη αντίληψη, αποδεικνύεται πειραματικά πως εμπεριέχει όλα τα απαραίτητα μαθηματικά στοιχεία για τον επιτυχή διαχωρισμό των αντίστοιχων κλάσεων. Ακολουθώντας την ίδια πειραματική διαδικασία με τις προαναφερθείσες μεθόδους, επιτύχαμε ορθή κατηγοριοποίηση της τάξεως του 94,14% και 95,42% (για τις βάσεις KTH και Weizmann αντίστοιχα), δείχνοντας ότι η τεχνική έχει μεγάλες δυνατότητες ενώ προσφέρει λύση σε πολλά από τα συνήθη προβλήματα στο συγκεκριμένο ερευνητικό πεδίο. Σε αυτό το σημείο θα πρέπει να τονίσουμε πως η εξαγωγή των ανθρώπινων σιλουετών έχει γίνει και σε αυτή την περίπτωση με τον απλούστερο δυνατό τρόπο, έχοντας σαν αποτέλεσμα η προτεινόμενη τεχνική να έχει εφαρμοστεί σε πολύ

θορυβώδη δεδομένα.

Στην προσπάθεια διερεύνησης των πραγματικών δυνατοτήτων της προαναφερθείσας τεχνικής, εξετάσαμε την εφαρμογή της και για το σενάριο του εντοπισμού ανθρώπινης πτώσης (Κεφάλαιο 4). Ο εντοπισμός πτώσης, είναι μια αρκετά συγγενής εφαρμογή ως προς την αναγνώριση των ανθρωπίνων κινήσεων. Ωστόσο, αντιμετωπίζεται στη βιβλιογραφία ως διαφορετική κατηγορία και επιδέχεται πολλές φορές διαφορετικής προσέγγισης. Παραμετροποιόντας το πρόβλημα της κατηγοριοποίησης για τον εντοπισμό πτώσεων, εφαρμόσαμε τη μέθοδο των HTFs σε δύο νέες και απαιτητικές βάσεις (UR και Le2i). Το σύνολο δεδομένων UR περιέχει πτώσεις οι οποίες έχουν ληφθεί με τη βοήθεια αισθητήρα υπερύθρων και περιέχει πληροφορία βάθους, ενώ το σύνολο δεδομένων Le2i περιέχει καθημερινές δραστηριότητες και ηθελημένες ή αθέλητες πτώσεις, οι οποίες έχουν ληφθεί σε ποικίλα περιβάλλοντα εσωτερικού χώρου και κάτω από διαφορετικές συνθήκες.

Η πειραματική διαδικασία απέδειξε ότι με τη χρήση των HTFs, μπορέσαμε να επιτύχουμε 100% ακρίβεια στην αναγνώριση πτώσεων σε σειρές κινήσεων προερχόμενες και από τα δύο προαναφερθέντα σύνολα δεδομένων. Παράλληλα, τα πειραματικά αποτελέσματα ενισχύουν την πεποίθησή μας, ότι η προτεινόμενη μεθοδολογία μπορεί εύκολα να γενικευτεί στην επίλυση πιο περίπλοκων ζητημάτων αναγνώρισης πτώσης, χωρίς να εξαρτάται από τη φύση των υπό εξέταση δεδομένων.

Προς αυτή την κατεύθυνση, τα μελλοντικά μας πλάνα για τη μέθοδο αυτή περιλαμβάνουν εκτεταμένους πειραματισμούς στο ζήτημα του διαχωρισμού ακούσιων-εκούσιων πτώσεων, καθώς και μια προσπάθεια για βαθύτερη κατηγοριοποίηση άλλων ενεργειών που μπορούν να οδηγήσουν σε ή να ακολουθήσουν μια πτώση, με στόχο την καλύτερη μοντελοποίηση της ανθρώπινης συμπεριφοράς σε ένα νοικοκυριό ηλικιωμένων ατόμων. Σκέψη για περαιτέρω προσπάθειες υπάρχουν επίσης, προς την κατεύθυνση της ένταξης της προτεινόμενης μεθοδολογίας σε ένα ολοκληρωμένο σύστημα βοηθητικού περιβάλλοντος πραγματικού χρόνου.

Στην ενασχόλησή μας στο πεδίο της αναγνώρισης της ανθρώπινης δραστηριότητας, βρεθήκαμε αντιμέτωποι με ποικίλα προβλήματα και νέες προκλήσεις. Στην προσπάθεια διερεύνησης των ικανοτήτων των προτεινόμενων μεθόδων αλλά και της ανάδειξης των εν λόγω προκλήσεων, δημιουργήσαμε και παρουσιάσαμε μια νέα βάση δεδομένων (Κεφάλαιο5). Η βάση THETIS, περιλαμβάνει 12 βασικές κινήσεις του αθλήματος της αντισφαίρισης (tennis) βιντεοσκοπημένες με τη χρήση της κάμερας τρισδιάστατης λήψης Kinect. Η καταγραφή είναι διάρκειας μερικών ωρών και περιέχει συγκεκριμένες και αρκετά συγγενείς κινήσεις αντισφαίρισης εκτελεσμένες από 55 διαφορετικά άτομα. Η βάση στην τελική της μορφή αποτελείται από 8374 βίντεο καταγεγραμμένης κίνησης σε 5 διαφορετικούς τύπους δεδομένων: RGB, πληροφορία βάθους, σιλουέτες, σκελετούς και σκελετούς σε τρεις διαστάσεις. Η δημιουργία της βάσης THETIS, έγινε με την ελπίδα να αποτελέ-

σει ένα χρήσιμο 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 γεγονότων κ.α.

Η βάση δεδομένων THETIS συνδυάζει έναν αριθμό πλεονεκτημάτων που μπορούν να αξιοποιηθούν από μελλοντικές εφαρμογές στα πλαίσια της αναγνώρισης της ανθρώπινης δραστηριότητας. Πρώτα απ΄ όλα παρουσιάζει πρωτοτυπία 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, όπως αναφέρθηκε, περιλαμβάνει βίντεο όχι μόνο εικόνας RGB, αλλά και βίντεο 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 βίντεο σκελετού 3D, σκελετού 2D και περιγράμματος, παρέχουν σε γενικές γραμμές το πλεονέκτημα της ανεξαρτησίας από το φόντο.

Ένα ακόμη πλεονέκτημα του συνόλου δεδομένων THETIS, είναι τα βίντεο 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 βίντεο, 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 βίντεο και δίνει τη δυνατότητα αξιολόγησης σε μεθόδους που ασχολούνται με αυτού του είδους την πληροφορία.

Στο σημείο αυτό, θα πρέπει να θυμίσουμε πως στη δημιουργία της βάσης συμμετείχαν τόσο έμπειροι όσο και αρχάριοι στο άθλημα της αντισφαίρισης. Συγκεκριμένα, συμμετείχαν 31 αρχάριοι και 24 έμπειροι αντισφαιριστές. Το γεγονός ότι στη διαδικασία εκπαίδευσης του SVM χρησιμοποιήθηκαν και τα δείγματα των αρχαρίων στην αντισφαίριση, απέτρεψε την κατασκευή ενός 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. Για το λόγο αυτό, παρατηρούμε πως τα αποτελέσματα ακρί-

βειας που αφορούν στις συγκεκριμένες κλάσεις, είναι έως και 20% χαμηλότερα από τις υπόλοιπες κινήσεις.

Μια επιπλέον δυνατότητα που παρέχει το συγκεκριμένο σύνολο δεδομένων, είναι η εναλλακτική προσέγγιση και αναγνώριση τύπου "expert-non expert", με βάση τα βίντεο του συνόλου, που όπως αναφέρθηκε και πιο πάνω περιέχουν δείγματα τόσο έμπειρων, όσο και αρχάριων παικτών. 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 βίντεο 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.

Όπως έχει ήδη αναφερθεί, οι ανθρώπινες δραστηριότητες που αναπαριστώνται σε ένα βίντεο, είναι στην ουσία χωρο-χρονικοί όγκοι. Μπορούν δηλαδή να εξεταστούν σαν μοντέλα τριών διαστάσεων. Προς την κατεύθυνση αυτή και προκειμένου να συλλάβουμε τη δομή της κίνησης όσο το δυνατόν πιο κοντά στην πραγματική της διάσταση, σχεδιάσαμε και προτείναμε μια τρισδιάστατη μορφή του μετασχηματισμού Trace για πρώτη φορά στη βιβλιογραφία (3D Cylindrical Trace Transform (CTT)) (Κεφάλαιο 6). Ο προτεινόμενος CTT είναι στην ουσία μια επέκταση του μετασχηματισμού Trace, η οποία μεταφέρει όλα τα πλεονεκτήματα του μετασχηματισμού, τα οποία έχουν αναδειχθεί στα προηγούμενα κεφάλαια, στον τρισδιάστατο χώρο ενώ παράλληλα επιτυγχάνει ταχύτητα εκτέλεσης σε σχέση με άλλους τρισδιάστατους μετασχηματισμούς.

Στη συνέχεια του ίδιου κεφαλαίου, προτείνεται ένα σύστημα για την εξαγωγή χαρακτηριστικών το οποίο συνδυάζει τον προτεινόμενο CTT και τα χωρο-χρονικά σημεία ενδιαφέροντος (STIPs). Η μέθοδος αναπαριστά το σύνολο της κινησιο-ακολουθίας με ένα διάνυσμα πολύ μικρού μεγέθους αποτελούμενο όπως και στις παραπάνω προτεινόμενες μεθόδους, από αμετάβλητα χαρακτηριστικά (HTFs). Η μέθοδος καταφέρνει να συλλάβει αποτελεσματικά πολύ μεγάλο μέρος τόσο των χωρικών, όσο και των χρονικών διακριτών

στοιχείων μιας κίνησης και να αποδώσει με επιτυχία τους τρισδιάστατους γεωμετρικούς συσχετισμούς που διαφοροποιούν μια δράση από μια άλλη, σε μια πολύ μικρής διάστασης διαχειρίσιμη μορφή. Τα πειραματικά αποτελέσματα σε βάσεις δεδομένων για την αναγνώριση δραστηριότητας αλλά και σε βάσεις για τον εντοπισμό πτώσεων και ενώ έχει εφαρμοστεί το πιο απαιτητικό πρωτόκολλο (leave-one-out cross validation) για την αξιολόγησή του, έδειξαν ότι η προτεινόμενη μέθοδος ανταποκρίνεται πολύ καλά ξεπερνώντας σε πολλές περιπτώσεις τις μεθόδους που βρίσκονται σήμερα στην αιχμή της τεχνολογίας.

Για την ακρίβεια, η μέθοδος δοκιμάστηκε σε όλα τα παραπάνω σύνολα δεδομένων καθώς και στο σύνολο THETIS το οποίο συνιστά μια ιδιαίτερα απαιτητική βάση δεδομένων. Είναι χαρακτηριστικό ότι στα πειράματα τα οποία αφορούν το σενάριο του εντοπισμού πτώσεων, η προτεινόμενη τεχνική επιτυγχάνει ακρίβεια κατηγοριοποίησης της τάξεως του 100% όπως και σε υποσύνολα της βάσης δεδομένων THETIS για την περίπτωση αναγνώρισης συγγενών αθλητικών δράσεων. Σε όλα τα υπόλοιπα πειράματα που αφορούν στην αναγνώριση ανθρώπινης δραστηριότητας, η μέθοδος επιτυγχάνει ποσοστά πολύ κοντά στο 100% ξεπερνώντας στις περισσότερες περιπτώσεις τον ανταγωνισμό. Αξίζει να σημειωθεί ότι ακόμη και στο ”ιδιότροπο” (από πλευράς τύπου δεδομένων) υποσύνολο ”Skelet3D” της βάσης THETIS, όπου ο πολύ γνωστός αλγόριθμος τελευταίας γενιάς STIPs επιτυγχάνει ποσοστό 54.4%, ως συνδυαστική τεχνική με τον 3D CTT, καταφέρνει κατηγοριοποίηση της τάξεως του 86.06%.

Σε αυτό το σημείο θα θέλαμε να επισημάνουμε ότι οι δυνατότητες του αλγορίθμου δεν έχουν εξεταστεί σε όλο τους το εύρος. Περαιτέρω μελέτη για την εύρεση των κατάλληλων συναρτησιακών βάσει των οποίων υπολογίζεται ο CTT καθώς και παραμετροποίηση του βήματος περιστροφής του, μπορεί να αυξήσουν τα ποσοστά επιτυχίας του ειδικά εάν αυτό γίνει έχοντας ως γνώμονα την εφαρμογή του σε συγκεκριμένο τύπο δεδομένων. Τέλος, πιστεύουμε ότι ο συγκεκριμένος μετασχηματισμός θα μπορούσε να εφαρμοστεί με επιτυχία και σε άλλα πεδία της όρασης και της αναγνώρισης προτύπων, όπως για παράδειγμα στην περίπτωση της ανάκτησης τρισδιάστατων μοντέλων (3D model retrieval).

Τέλος, πρόθεσή μας είναι να εξετάσουμε την παραπάνω τεχνική για την ευρύτερη εφαρμογή του προβλήματος, που είναι τα βίντεο ”χωρίς περιορισμούς” (unconstrained videos) ή αλλιώς η αναγνώριση δραστηριότητας ”in the wild” όπως συνηθίζεται να αναφέρεται στη βιβλιογραφία. Οι συνθήκες που αντιμετωπίζονται σε αυτή την εκδοχή του προβλήματος είναι μη προβλέψιμες καθώς για παράδειγμα, η γωνία λήψης μιας δραστηριότητας μπορεί να είναι τελείως διαφορετική από αυτή με την οποία μπορεί να έχει εκπαιδευτεί το σύστημα, οι κινήσεις δεν θεωρούνται απαραίτητα περιοδικές και γενικότερα οι συνθήκες βιντεοσκόπησης δεν έγκεινται σε κανενός είδους κανόνα.

Βιβλιογραφία

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

[2] Saad Ali, Arslan Basharat, and Mubarak Shah. Chaotic invariants for human action recognition. *ICCV*, 206, 2007.

[3] James F. Allen. Maintaining knowledge about temporal intervals. *Commun. ACM*, 26(11):832–843, November 1983.

[4] Amir Averbuch and Yoel Shkolnisky. 3d fourier based discrete radon transform. *Applied and Computational Harmonic Analysis*, 15(1):33 – 69, 2003.

[5] Sermetcan Baysal, Mehmet Can Kurt, and Pinar Duygulu. Recognizing human actions using key poses. *Pattern Recognition, International Conference on*, pages 1727–1730, 2010.

[6] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In *The Tenth IEEE International Conference on Computer Vision (ICCV'05)*, pages 1395–1402, 2005.

[7] V. Bloom, D. Makris, and V. Argyriou. G3d: A gaming action dataset and real time action recognition evaluation framework. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 7–12, 2012.

[8] Aaron F. Bobick and James W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:257–267, 2001.

[9] Aaron F. Bobick, Stephen S. Intille, James W. Davis, Freedom Baird, Claudio S. Pinhanez, Lee W. Campbell, Yuri A. Ivanov, Arjan Schütte, and Andrew D. Wilson. The kidsroom. *Commun. ACM*, 43(3):60–61, 2000.

[10] Aaron F. Bobick and Andrew D. Wilson. A state-based approach to the representation and recognition of gesture. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 19:1325–1337, 1997.

[11] Nikolaos V. Boulgouris and Zhiwei X. Chi. Gait recognition using radon transform and linear discriminant analysis. *IEEE Transactions on Image Processing*, 16(3):731–740, 2007.

[12] L. W. Campbell and A. F. Bobick. Recognition of human body motion using phase space constraints. In *Proceedings of the Fifth International Conference on Computer Vision*, ICCV '95, pages 624–. IEEE Computer Society, 1995.

[13] B. Chakraborty, M.B. Holte, T.B. Moeslund, and J. Gonzalez. Selective spatio-temporal interest points. *Computer Vision and Image Understanding*, 116(3):396–410, 2012.

[14] Imen Charfi, Johel Miteran, Julien Dubois, Mohamed Atri, and Rached Tourki. Definition and performance evaluation of a robust svm based fall detection solution. In *SITIS'12*, pages 218–224, 2012.

[15] Olivier Chomat and James L. Crowley. Probabilistic recognition of activity using local appearance. In *CVPR*, pages 2104–2109. IEEE Computer Society, 1999.

[16] Peng Dai, Huijun Di, Ligeng Dong, Linmi Tao, and Guangyou Xu. Group interaction analysis in dynamic context. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 39(1):34–42, 2009.

[17] Navneet Dalal, Bill Triggs, and Cordelia Schmid. Human detection using oriented histograms of flow and appearance. In *Proceedings of the 9th European Conference on Computer Vision - Volume Part II*, ECCV'06, pages 428–441. Springer-Verlag, 2006.

[18] Dima Damen and David Hogg. Recognizing linked events: Searching the space of feasible explanations. In *CVPR*, pages 927–934. IEEE, 2009.

[19] T. Darrell and A. Pentland. Space-time gestures. In *CVPR '93: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1993*, pages 335–340, 1993.

[20] Stanley R. Deans. *The Radon Transform and Some of Its Applications*. Krieger Publishing Company, 1983.

[21] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Proceedings of the 14th International Conference on*

Computer Communications and Networks, pages 65–72, Washington, DC, USA, 2005. IEEE Computer Society.

[22] Charalampos N Doukas and Ilias Maglogiannis. Emergency fall incidents detection in assisted living environments utilizing motion, sound, and visual perceptual components. *Information Technology in Biomedicine, IEEE Transactions on*, 15(2):277–289, 2011.

[23] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. Wiley, 2001.

[24] Alexei A. Efros, Alexander C. Berg, Er C. Berg, Greg Mori, and Jitendra Malik. Recognizing action at a distance. In *In ICCV*, pages 726–733, 2003.

[25] Weiguo Feng, Rui Liu, and Ming Zhu. Fall detection for elderly person care in a vision-based home surveillance environment using a monocular camera. *Signal, Image and Video Processing*, 8(6):1129–1138, 2014.

[26] H. Foroughi, H.S. Yazdi, H. Pourreza, and M. Javidi. An eigenspace-based approach for human fall detection using integrated time motion image and multi-class support vector machine. In *Intelligent Computer Communication and Processing, 2008. ICCP 2008. 4th International Conference on*, pages 83–90, Aug 2008.

[27] Zan Gao, Ming-Yu Chen, Alexander G. Hauptmann, and Anni Cai. Comparing evaluation protocols on the kth dataset. In *Proceedings of the First international conference on Human behavior understanding*, HBU'10, pages 88–100, Berlin, Heidelberg, 2010. Springer-Verlag.

[28] D. M. Gavrila and L. S. Davis. Towards 3-d model-based tracking and recognition of human movement: a multi-view approach. In *In International Workshop on Automatic Face- and Gesture-Recognition. IEEE Computer Society*, pages 272–277, 1995.

[29] Andrew Gilbert, John Illingworth, and Richard Bowden. Scale invariant action recognition using compound features mined from dense spatio-temporal corners. In *Proceedings of the 10th European Conference on Computer Vision: Part I*, ECCV '08, pages 222–233, Berlin, Heidelberg, 2008. Springer-Verlag.

[30] R. Goldenberg, R. Kimmel, E. Rivlin, and M. Rudzsky. Behavior classification by eigendecomposition of periodic motions. In *Pattern Recognition*, pages 38:1033–1043, 2005.

[31] Shaogang Gong and Tao Xiang. Recognition of group activities using dynamic probabilistic networks. In *ICCV*, pages 742–749. IEEE Computer Society, 2003.

[32] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. *Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253, December 2007.

[33] G. Goudelis, K. Karpouzis, and S. Kollias. Robust human action recognition using history trace templates. In *12th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), Delft, The Netherlands, 13-15 April.*, 2011.

[34] G. Goudelis, G. Tsatiris, K. Karpouzis, and S. Kollias. 3d cylindrical trace transform based feature extraction for effective spatiotemporal sequence classification. *Image and Vision Computing. Special Issue on Handcrafted vs. Learned Representations for Human Action Recognition. Submitted*, 2015.

[35] Georgios Goudelis, Konstantinos Karpouzis, and Stefanos Kollias. Exploring trace transform for robust human action recognition. *Pattern Recognition*, 46(12):3238 – 3248, 2013.

[36] Georgios Goudelis, Anastasios Tefas, and Ioannis Pitas. Automated facial pose extraction from video sequences based on mutual information. *IEEE Trans. Circuits Syst. Video Techn.*, 18(3):418–424, 2008.

[37] S. Gourgari, G. Goudelis, K. Karpouzis, and S. Kollias. Thetis: Three dimensional tennis shots - a human action dataset. In *IEEE Workshop on Behaviour Analysis in Games and modern Sensing devices*. CVPR 2013, Portland, Oregon., 2013.

[38] M. Grundmann, F. Meier, and I. Essa. 3d shape context and distance transform for action recognition. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1 –4, dec. 2008.

[39] W. Hu, D. Xie, T. Tan, and S. Maybank. Learning activity patterns using fuzzy self-organizing neural network. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, 34(3):1618–1626, jun 2004.

[40] Raul Igual, Carlos Medrano, and Inmaculada Plaza. Challenges, issues and trends in fall detection systems. *BioMedical Engineering OnLine*, 12(1):66, 2013.

[41] N. Ikizler and P. Duygulu. Human action recognition using distribution of oriented rectangular patches. In *(ICCV)*, pages 271–284. Human Motion, 2007.

[42] Stephen S. Intille and Aaron F. Bobick. A framework for recognizing multi-agent action from visual evidence. In *In AAAI-99*, pages 518–525. AAAI Press, 1999.

[43] Yuri A. Ivanov and Aaron F. Bobick. Recognition of visual activities and interactions by stochastic parsing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):852–872, August 2000.

[44] Hueihan Jhuang, Thomas Serre, Lior Wolf, and Tomaso Poggio. A biologically inspired system for action recognition. In *ICCV*, pages 1–8. IEEE, 2007.

[45] Kui Jia and Dit-Yan Yeung. Human action recognition using local spatio-temporal discriminant embedding. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, 2008.

[46] Hao Jiang, Mark S. Drew, and Ze nian Li. Successive convex matching for action detection. In *In Proc. CVPR*, pages 1646–1653. Press, 2006.

[47] Seong-Wook Joo and Rama Chellappa. Attribute grammar-based event recognition and anomaly detection. In *Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop*, CVPRW '06, pages 107–, Washington, DC, USA, 2006. IEEE Computer Society.

[48] Imran N. Junejo, Emilie Dexter, Ivan Laptev, and Patrick Perez. View-independent action recognition from temporal self-similarities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:172–185, 2011.

[49] Alexander Kadyrov and Maria Petrou. The trace transform and its applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23:811–828, August 2001.

[50] Yan Ke, Rahul Sukthankar, and Martial Hebert. Spatio-temporal shape and flow correlation for action recognition. In *In 7th Int. Workshop on Visual Surveillance*, 2007.

[51] V. Kellokumpu, G.Y. Zhao, and M. Pietikainen. Human activity recognition using a dynamic texture based method. In *Proc. The British Machine Vision Conference (BMVC), Leeds, UK*, page 10, 2008.

[52] B. Kepski. M., Kwolek. Fall detection using ceiling-mounted 3d depth camera. In *Proc. 9th Int. Conf. on Computer Vision Theory and Applications (VISAPP)*, pages vol. 2, 640–647, 2014.

[53] Alexander Kläser, Marcin Marszałek, and Cordelia Schmid. A spatio-temporal descriptor based on 3d-gradients. In *British Machine Vision Conference*, pages 995–1004, sep 2008.

[54] Irene Kotsia and Ioannis Patras. Relative margin support tensor machines for gait and action recognition. In *CIVR*, pages 446–453, 2010.

[55] Bogdan Kwolek and Michal Kepski. Human fall detection on embedded platform using depth maps and wireless accelerometer. *Computer Methods and Programs in Biomedicine*, 117(3):489–501, 2014.

[56] I. Laptev and T. Lindeberg. Space-time interest points. In *IEEE Int. Conf. on Computer Vision (ICCV'03)*, Nice, France, October 2003.

[57] Ivan Laptev. On space-time interest points. *Int. J. Comput. Vision*, 64:107–123, September 2005.

[58] Ivan Laptev, Marcin Marszałek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *Conference on Computer Vision & Pattern Recognition (CVPR)*. IEEE, 2008.

[59] T. Lee and Mihailidis A. An intelligent emergency response system: preliminary development and testing of automated fall detection. *Journal of telemedicine and telecare*, 11(4):194?198, 2005.

[60] Jingen Liu, Jiebo Luo, and Mubarak Shah. Recognizing realistic actions from videos "in the wild", 2009.

[61] Jingen Liu and Mubarak Shah. Learning human action via information maximization, 2008.

[62] David G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2*, ICCV '99, pages 1150–. IEEE Computer Society, 1999.

[63] Wei-Lwun Lu and James J. Little. Simultaneous tracking and action recognition using the pca-hog descriptor. In *(CRV)*, page 6. IEEE Computer Society, 2006.

[64] Roberto Lublinerman, Necmiye Ozay, Dimitrios Zarpalas, and Octavia I. Camps. Activity recognition from silhouettes using linear systems and model (in)validation techniques. In *ICPR (1)*, pages 347–350. IEEE Computer Society, 2006.

[65] Xin Ma, Haibo Wang, Bingxia Xue, Mingang Zhou, Bing Ji, and Yibin Li. Depth-based human fall detection via shape features and improved extreme learning machine. *IEEE J. Biomedical and Health Informatics*, 18(6):1915–1922, 2014.

[66] Marcin Marszałek, Ivan Laptev, and Cordelia Schmid. Actions in context. In *IEEE Conference on Computer Vision & Pattern Recognition*, 2009.

[67] Georgios Mastorakis and Dimitrios Makris. Fall detection system using Kinect's infrared sensor. *Journal of Real-Time Image Processing*, 9(4):635–646, 2014.

[68] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocitym histories of tracked keypoints. In *IEEE 12th Conference on Computer Vision*, pages 104–111, 2009.

[69] K. Mikolajczyk and H. Uemura. Action recognition with motion appearance vocabulary forest. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

[70] David Minnen, Irfan A. Essa, and Thad Starner. Expectation grammars: Leveraging high-level expectations for activity recognition. In *CVPR (2)*, pages 626–632. IEEE Computer Society, 2003.

[71] Darnell Moore and Irfan Essa. Recognizing multitasked activities from video using stochastic context-free grammar. In *In Proc. AAAI National Conf. on AI*, pages 770–776. AAI, 2002.

[72] L. P. Morency, A. Quattoni, and T. Darrell. Latent-dynamic discriminative models for continuous gesture recognition. *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, June 2007.

[73] M. Mubashir, L. Shao, and L. Seed. A survey on fall detection: Principles and approaches. *Neurocomputing*, 100:144–152, 2012.

[74] Pradeep Natarajan and Ramakant Nevatia. Coupled hidden semi markov models for activity recognition. In *Proceedings of the IEEE Workshop on Motion and Video Computing*, WMVC '07, pages 10–. IEEE Computer Society, 2007.

[75] Ram Nevatia, Tao Zhao, and Somboon Hongeng. Hierarchical language-based representation of events in video streams. In *Proceedings of the IEEE Workshop on Event Mining. IEEE*, 2003.

[76] Anh-Tuan Nghiem, Edouard Auvinet, and Jean Meunier. Head detection using Kinect camera and its application to fall detection. In *11th International Conference on Information Science, Signal Processing and their Applications, ISSPA, Montreal, QC, Canada, July 2-5*, pages 164–169, 2012.

[77] Nam T. Nguyen, Dinh Q. Phung, Svetha Venkatesh, and Hung Bui. Learning and detecting activities from movement trajectories using the hierarchical hidden markov models. In *Proceedings of the 2005 IEEE Computer Society Conference on*

Computer Vision and Pattern Recognition (CVPR'05) - Volume 2 - Volume 02, pages 955–960. IEEE Computer Society, 2005.

[78] Juan Carlos Niebles, Hongcheng Wang, and Li Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision (IJCV)*, 79:299–318, September 2008.

[79] J.C. Nieblesand, C.-W. Chen, and L. Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *IEEE 11th European Conference on Computer Vision*, pages 392–405, 2010.

[80] N. Noury, P. Rumeau, A.K. Bourke, G. Olaighin, and J.E. Lundy. A proposal for the classification and evaluation of fall detectors. *IRBM journal of Alliance for engineering in Biology an Medicine*, 26(6):340–349, 2008.

[81] Antonios Oikonomopoulos, Ioannis Patras, and Maja Pantic. Spatiotemporal salient points for visual recognition of human actions, 2006.

[82] Nuria Oliver, Eric Horvitz, and Ashutosh Garg. Layered representations for human activity recognition. In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, ICMI '02, pages 3–. IEEE Computer Society, 2002.

[83] Nuria M. Oliver, Barbara Rosario, and Alex P. Pentland. A bayesian computer vision system for modeling human interactions. *IEEE Transaction on Pattern Analysis and Machine Intellignece*, 22(8):831–843, 2000.

[84] World Health Organization. on falls prevention in older age who library cataloguing-in-publication data, 2007.

[85] Sangho Park. A hierarchical bayesian network for event recognition of human actions and interactions. In *Association For Computing Machinery Multimedia Systems Journal*, pages 164–179, 2004.

[86] Alonso Patron, Marcin Marszalek, Andrew Zisserman, and Ian Reid. High five: Recognising human interactions in tv shows. In *Proceedings of the British Machine Vision Conference*, pages 50.1–50.11. BMVA Press, 2010.

[87] J. T. Perry, S. Kellog, S. M. Vaidya, Jong-Hoon Youn, H. Ali, and H. Sharif. Survey and evaluation of real-time fall detection approaches. In *High-Capacity Optical Networks and Enabling Technologies (HONET), 2009 6th International Symposium on*, pages 158–164. IEEE, December 2009.

[88] Claudio Pinhanez and Aaron Bobick. Human action detection using pnf propagation of temporal constraints. In *In Proc. of the Conference on Computer Vision and Pattern Recognition*, pages 898–904, 1997.

[89] Rainer Planinc and Martin Kampel. Introducing the use of depth data for fall detection. *Personal Ubiquitous Comput.*, 17(6):1063–1072, August 2013.

[90] Cen Rao and Mubarak Shah. View-invariance in action recognition. In *CVPR (2)*, pages 316–322. IEEE Computer Society, 2001.

[91] K. Rapantzikos, Y. Avrithis, and S. Kollias. Dense saliency-based spatiotemporal feature points for action recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[92] Mikel D. Rodriguez, Javed Ahmed, and Mubarak Shah. Action mach: a spatio-temporal maximum average correlation height filter for action recognition. In *In Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2008.

[93] Caroline Rougier, Jean Meunier, Alain St-Arnaud, and Jacqueline Rousseau. Fall detection from human shape and motion history using video surveillance. In *AINA Workshops (2)*, pages 875–880. IEEE Computer Society, 2007.

[94] Caroline Rougier, Jean Meunier, Alain St-Arnaud, and Jacqueline Rousseau. Robust video surveillance for fall detection based on human shape deformation. *IEEE Trans. Circuits Syst. Video Techn.*, 21(5):611–622, 2011.

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

[96] M. S. Ryoo and J. K. Aggarwal. Recognition of composite human activities through context-free grammar based representation. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, pages 1709–1718, Washington, DC, USA, 2006. IEEE Computer Society.

[97] Michael S. Ryoo and Jake K. Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *ICCV*, pages 1593–1600. IEEE, 2009.

[98] Silvio Savarese, Andrey Delpozo, Juan Carlos Niebles, and Li Fei-fei. Spatial-temporal correlatons for unsupervised action classification, 2008.

[99] K. Schindler and L.J. Van Gool. Action snippets: How many frames does human action recognition require? In *CVPR08*, pages 1–8, 2008.

[100] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: A local svm approach. In *In Proc. ICPR*, pages 32–36, 2004.

[101] Eli Shechtman and Michal Irani. Space-time behavior based correlation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 405–412, June 2005.

[102] Eli Shechtman and Michal Irani. Space-time behavior based correlation -or- how to tell if two underlying motion fields are similar without computing them? In *In IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, volume 29, pages 2045–2056, November 2007.

[103] Yaser Sheikh, Mumtaz Sheikh, and Mubarak Shah. Exploring the space of a human action. In *ICCV*, pages 144–149. IEEE Computer Society, 2005.

[104] Yifan Shi, Yan Huang, David Minnen, Aaron F. Bobick, and Irfan A. Essa. Propagation networks for recognition of partially ordered sequential action. In *CVPR (2)*, pages 862–869, 2004.

[105] Huang Shirlena, Thang Leng Leng, and Toyota Mika. Transnational mobilities for care: Rethinking the dynamics of care in asia. *Global Networks*, 12(2):129–134, 2012.

[106] Cristian Sminchisescu, Atul Kanaujia, and Dimitris Metaxas. Conditional models for contextual human motion recognition. *Comput. Vis. Image Underst.*, 104:210–220, November 2006.

[107] S. Srisuk, M. Petrou, W. Kurutach, and A. Kadyrov. A face authentication system using the Trace transform. *Pattern Anal. Appl.*, 8(1):50–61, September 2005.

[108] Thad Starner and Alex Pentland. Real-time american sign language recognition from video using hidden markov models, 1996.

[109] Chris Stauffer and W. Eric L. Grimson. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:747–757, 2000.

[110] Xinghua Sun, Mingyu Chen, and A. Hauptmann. Action recognition via local descriptors and holistic features. *Computer Vision and Pattern Recognition Workshop*, 0:58–65, 2009.

[111] Tanveer Syeda-Mahmood, A. Vasilescu, and S. Sethi. Recognizing action events from multiple viewpoints. *Detection and Recognition of Events in Video, IEEE Workshop on*, 0:64, 2001.

[112] Christian Thurau. Behavior histograms for action recognition and human detection. In *Proceedings of the 2nd conference on Human motion: understanding, modeling, capture and animation*, pages 299–312, Berlin, Heidelberg, 2007. Springer-Verlag.

[113] Christian Thurau and Vaclav Hlavac. Pose primitive based human action recognition in videos or still images. In *Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, page 8. IEEE Computer Society, Madison, USA, Omnipress, 2008.

[114] GHY Ting and J Woo. Elder care: is legislation of family responsibility the solution. *Asian J Gerontol Geriatr*, 4:72–75, 2009.

[115] B. Uğur Töreyin, Yiğithan Dedeoğlu, and A. Enis Çetin. Hmm based falling person detection using both audio and video. In *Proceedings of the 2005 International Conference on Computer Vision in Human-Computer Interaction*, ICCV'05, pages 211–220, Berlin, Heidelberg, 2005. Springer-Verlag.

[116] Son D. Tran and Larry S. Davis. Event modeling and recognition using markov logic networks. In *Proceedings of the 10th European Conference on Computer Vision: Part II*, ECCV '08, pages 610–623. Springer-Verlag, 2008.

[117] C. Mandal V. Vishwakarma and S. Sural. Automatic detection of human fall in video. In *in Proc. Int. Conf. Pattern Recogn. Mach. Intell.*, page 616?623, 2007.

[118] Jonathan Vainstein, Josef Manera, Pablo Negri, Claudio Delrieux, and Ana Maguitman. Modeling video activity with dynamic phrases and its application to action recognition in tennis videos. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, volume 8827 of *Lecture Notes in Computer Science*, pages 909–916. Springer International Publishing, 2014.

[119] Ashok Veeraraghavan and Amit K. Roy-chowdhury. The function space of an activity. In *in Proc. Comput. Vis. Pattern Recognit*, pages 959–968, 2006.

[120] Heng Wang, A. Klaser, C. Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3169–3176, June 2011.

[121] Liang Wang and David Suter. Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model. In *CVPR*, 2007.

[122] Daniel Weinland and Edmond Boyer. Action recognition using exemplar-based embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage*, pages 1–7, 2008.

[123] Daniel Weinland, Remi Ronfard, and Edmond Boyer. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, 104(2-3):249–257, November/December 2006.

[124] Shu-Fai Wong and Roberto Cipolla. Extracting spatiotemporal interest points using global information. *Computer Vision, IEEE International Conference on*, 0:1–8, 2007.

[125] Yaser Yacoob and Michael J. Black. Parameterized modeling and recognition of activities. In *Proceedings of the Sixth International Conference on Computer Vision*, ICCV '98, pages 120–. IEEE Computer Society, 1998.

[126] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden markov model. In *Conference on Computer Vision and Pattern Recognition., 1992 IEEE Computer Society*, pages 379–385, 1992.

[127] C.J. Yang, Y.L. Guo, H.S. Sawhney, and R.T. Kumar. Learning actions using robust string kernels. In *HUMO07*, pages 313–327. Springer, 2007.

[128] L. Yeffet and L. Wolf. Local trinary patterns for human action recognition. In *(ICCV)*. IEEE, 2009.

[129] Alper Yilmaz and Mubarak Shah. Actions sketch: A novel action representation. In *CVPR (1)'05*, pages 984–989, 2005.

[130] G. Yo, J. Yuan, and Z. Liu. Unsupervised random forest indexing for fast action search. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 865–872, 2010.

[131] Chunfeng Yuan, Xi Li, Weiming Hu, Haibin Ling, and S. Maybank. 3d r transform on spatio-temporal interest points for action recognition. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 724–730, June 2013.

[132] Lihi Zelnik-manor and Michal Irani. Event-based analysis of video. In *In Proc. CVPR*, pages 123–130, 2001.

[133] Dong Zhang, Daniel Gatica-perez, Samy Bengio, Iain Mccowan, and Guillaume Lathoud. Modeling individual and group actions in meetings: a two-layer hmm framework. In *In Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Workshop on Event Mining in Video (CVPREVENT), Washington DC*, 2004.

[134] L. Zhang, B. Wu, and R. Nevatia. Detection and tracking of multiple humans with extensive pose articulation. In *(ICCV)*, pages 1–8. Computer Vision, IEEE International Conference, 2007.

[135] Zhong Zhang, Weihua Liu, Vangelis Metsis, and Vassilis Athitsos. A viewpoint-independent statistical method for fall detection, 2012.

[136] Zoran Zivkovic. Improved adaptive gaussian mixture model for background subtraction. In *Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 2 - Volume 02*, ICPR '04, pages 28–31, 2004.

[137] Zoran Zivkovic and Ferdinand van der Heijden. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recogn. Lett.*, 27(7):773–780, May 2006.

Κατάλογος Δημοσιεύσεων

Διεθνή Περιοδικά

G. Goudelis, G. Tsatiris, K. Karpouzis, S. Kollias, 3D Cylindrical Trace transform based feature extraction for effective spatiotemporal sequence classification, *Image and Vision Computing. Special Issue on Handcrafted vs. Learned Representations for Human Action Recognition. Elsevier. Submitted*

G. Goudelis, Konstantinos Karpouzis, Stefanos Kollias, Exploring Trace transform for robust human action recognition, *Pattern Recognition, Vol. 46, Issue 12, December 2013, p. 3238-3248, Elsevier*

G. Goudelis, A. Tefas and I. Pitas, Emerging Biometric Modalities: A Survey, *Journal on Multimodal User Interfaces": Volume 2, Issue 3 (2009), Page 217, Springer*

G. Goudelis, A. Tefas and I. Pitas, Facial Pose Extraction based on Mutual Information, *IEEE Transactions on Circuits and Systems for Video Technology". March 2008 volume: 3, pages: 418-424*

G. Goudelis, S. Zafeiriou A. Tefas and I. Pitas, Class-Specific Kernel Discriminant Analysis for Face Verification, In *IEEE Transactions on Information Forensics and Security" Sept. 2007 Volume: 2, Issue: 3, Part 2, page(s): 570-587*

Διεθνή Συνέδρια

G. Goudelis, G. Tsatiris, K. Karpouzis, S. Kollias, Fall detection Using History Triple Features, *International Conference on Pervasive Technologies Related to Assistive Environments (PETRA 15), July 1-3 2015. Corfu, Greece*

Sofia Gourgari, Georgios Goudelis, Konstantinos Karpouzis, Stefanos Kollias, THETIS: Three Dimensional Tennis Shots-A Human Action Dataset, In *CVPR workshop in Behavior Analysis in Games and modern Sensing. Portland, Oregon June 2013*

G. Goudelis, K. Karpouzis, S. Kollias, Robust Human Action Recognition Using History Trace Templates, In *International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 11), April 13-15, 2011. Delft, the Netherlands (Best Paper Award)*

G. Goudelis, Tefas and I. Pitas, Using Mutual Information to Indicate Facial Poses in Video Sequences, In *ACM International Conference on Image and Video Retrieval (CIVR 09), July 8-10, 2009*

G. Goudelis, S. Zafeiriou A. Tefas and I. Pitas, Motivating Class-Specific Nonlinear Projections for Single and Multiple View Face Verification, In *IEEE International Conference on Image Processing (ICIP 08) October 12–15, 2008. San Diego, California, U.S.A*

G. Goudelis, S. Zafeiriou, A. Tefas and I. Pitas, A Novel Kernel Discriminant Analysis for Face Verification, In *International Conference on Image Processing 2007 (ICIP07)". 16-19 September 07, pages: 493-496, San Antonio, Texas*

G. Goudelis, A. Tefas and I. Pitas, On emerging biometric technologies, In *Cost 275 workshop. Hertfordshire 27-28 October 2005 UK*

Κεφάλαια σε βιβλία

G. Goudelis, A. Tefas, I. Pitas, Intelligent Multimedia Analysis for Emerging Biometrics. Intelligent Multimedia Analysis for Security Applications, In *2010: 97-125. Studies in Computational Intelligence Vol. 282 Springer 2010, ISBN 978-3-642-11754-1*

□

Βιογραφικό σημείωμα

Ο Γεώργιος Γουδέλης αποφοίτησε από το τμήμα Ηλεκτρονικών Μηχανικών του πανεπιστημίου του Κεντ στο Κάντερμπερι (Ηνωμένο Βασίλειο) το 2003 και το 2005 έλαβε το μεταπτυχιακό του από το ίδιο τμήμα. Το ίδιο έτος, ξεκίνησε να δουλεύει σαν ερευνητής στο εργαστήριο "τεχνητής νοημοσύνης και ανάλυσης εικόνας" του τμήματος πληροφορικής του Αριστοτελείου πανεπιστημίου Θεσ/κης ενώ το 2008 μετέφερε τις ερευνητικές του δραστηριότητες στο εργαστήριο "ψηφιακής επεξεργασίας εικόνας βίντεο και πολυμέσων" της σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του ΕΜΠ. Το 2009 έγινε δεκτός από τη Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του ΕΜΠ για εκπόνηση διδακτορικής διατριβής υπό την επίβλεψη του καθηγητή κ. Σ. Κόλλια και του ερευνητή Δρ. Κωνσταντίνου Καρπούζη. Στη διατριβή του ασχολήθηκε με θέματα αυτοματοποιημένης αναγνώρισης ανθρώπινης δραστηριότητας ενώ έχει ασχοληθεί και με βιομετρικά συστήματα ασφαλείας και πιο συγκεκριμένα, με την αυτοματοποιημένη αναγνώριση προσώπου. Τον Αύγουστο του 2015 ολοκλήρωσε με επιτυχία την διδακτορική του διατριβή.