ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΠΜΣ: Προηγμένα Συστήματα Πληροφορικής

Διπλωματική Εργασία

«Εντοπισμός Ανθρώπων σε Video:

Βασικές Έννοιες, Μέθοδοι και Προοπτικές»

Γαλάνης Δημήτρης, Α.Μ: ΜΠΣΠ/09014

Επιβλέπων Καθηγητής**:** Γ. Τσιχριντζής

Πειραιάς 2013

ΠΕΡΙΕΧΟΜΕΝΑ

Εισαγωγή

Η κατανόηση της ανθρώπινης συμπεριφοράς αφορά ένα ευρύ φάσμα μελλοντικών εφαρμογών που κυμαίνονται από ευφυή συστήματα επιτήρησης μέχρι προηγμένα συστήματα διεπαφής με τον χρήστη. Συστήματαεπιτήρης καταγραφής ήχου και εικόνας είναι ήδη διαθέσιμα σε τράπεζες, ξενοδοχεία, καταστήματα, εθνικές οδούς και εμπορικά κέντρα, όπου τα καταγεγραμένα τηλεοπτικά δεδομένα ελέγχονται από φρουρούς ασφάλειας και αποθηκεύονται στα αρχεία προς δικανική αξιολόγηση. Σε ένα χαρακτηριστικό σύστημα, μια φρουρά ασφάλειας προσέχει 16 τηλεοπτικά κανάλια συγχρόνως και μπορεί να χάσει πολλά σημαντικά γεγονότα. Θα ήταν επιθυμητή η ύπαρξη ενός ευφυούς συστήματος με ικανότητα ανάλυσης του βίντεο σε πραγματικό χρόνο και προειδοποίησης του προσωπικού ασφάλειας ή ενός ημιαυτόματου συστήματος που να δίνει έμφαση στις πιθανές περιοχές ενδιαφέροντος για την περιοχή εξέτασης των μηχανών επιτήρησης για πρόληψη εγκλήματος και έλεγχο πλήθους. Η επιτήρηση ηλικιωμένων και ασθενών ανθρώπων είναι ένας άλλος σημαντικός τομέας εφαρμογών που απαιτούν αυτόματη ανάλυση *video*. Η ανίχνευση πραγματικού χρόνου ενός πεσμένου ατόμου είναι εξαιρετικά σημαντική για την παροχή άμεσης βοήθειας έκτακτης ανάγκης και μπορεί να αυξήσει την ποιότητα ζωής των ηλικιωμένων και ατόμων με αναπηρία. Ο ήχος ενός πεσμένου σώματος μπορεί να είναι απαραίτητος για να επιτύχει ένα ισχυρό αυτόματο σύστημα. Άλλες εφαρμογές περιλαμβάνουν την επιμέρους παρακολούθηση των μερών του σώματος και της κίνησης των ενώσεων, σε ένα βίντεο για ιατρική διάγνωση, υποστήριξη θεραπείας, έλεγχο πρόσβασης, και βελτίωση αθλητικών, καλλιτεχνικών επιδόσεων.

Η συνεχής ανάπτυξη των ψηφιακών βιβλιοθηκών θα απαιτήσει προηγμένα συστήματα διεπαφής με τον χρήστη, τα οποία θα μπορούν να ερμηνεύσουν στοιχεία πολυμέσων κατά τρόπο αυτόματο. Προηγμένα συστήματα διεπαφής με τον χρήστη που θα χρησιμοποιούν ομιλία και κατανόηση ανθρώπινων κινήσεων, μπορούν να παρέχουν τον έλεγχο στην υψηλού επιπέδου αλληλεπίδραση με αυτοματοποιημένα συστήματα και βάσεις δεδομένων πολυμέσων. Η πρόσφατη έρευνα για τα συστήματα ανάκτησης εικόνας είναι βασισμένη στην εισαγωγή κειμένων, όπου οι εικόνες σχολιάζονται από το κείμενο και η ανάκτηση εκτελείται στο κείμενο. Εντούτοις, ο χειρωνακτικός σχολιασμός απαιτεί εντατική εργασία και γίνεται μη πρακτικός όταν η συλλογή είναι μεγάλη. Ένα άλλο πρόβλημα είναι η υποκειμενική φύση των **λέξεων-κλειδιά** (*keywords*). Η ίδια εικόνα ή το ίδιο βίντεο μπορεί να σχολιαστεί διαφορετικά από διαφορετικούς σχολιαστές. Επομένως, είναι επιθυμητό να υπάρξει ένα σύστημα διαχείρισης βάσεων δεδομένων βίντεο (video database management system), συμπεριλαμβανομένου ενός **δικτυοκεντρικού** (*WEB-based*) γραφικού συστήματος διεπαφής **ερωτήσεων** (*queries*), η οποία μπορεί να χειριστεί την εκτέλεση ερωτήσεων για χωροχρονικές και σημασιολογικές ιδιότητες των τηλεοπτικών δεδομένων. Μια χωροχρονική ερώτηση μπορεί να περιέχει οποιοδήποτε συνδυασμό όρων: κατευθυντικών, τοπολογικών, εμφάνισης αντικειμένου, προβολής τροχιάς, τροχιάς αντικειμένου με βάση την ομοιότητα, κ.ά. Κατ' αυτό τον τρόπο, βίντεο που περιέχουν ανθρώπινη δραστηριότητα και εικόνες όπου απεικονίζονται άνθρωποι, μπορούν να ανακτηθούν αποδοτικά από μια βάση δεδομένων. Ένα τέτοιο σύστημα μπορεί επίσεις να αλληλεπιδράσει με απομακρυσμένους χρήστες μέσω Διαδικτύου, μέσω μιας γραφικής διεπαφής ερωτήσεων. Είναι επίσης επιθυμητό να υπάρξουν συστήματα διεπαφής φυσικής γλώσσας για **μηχανές διαδικτυακής αναζήτησης** (*web search engines*). Ο χρήστης θα είναι σε θέση να πραγματοποιήσει ερωτήσεις χρησιμοποιώντας φυσική γλώσσα ή λεκτική εισαγωγή στο σύστημα. Ένα τέτοιο σύστημα μπορεί να προσελκύσει το χρήστη σε πολυμορφικό διάλογο με στόχο να περιορίσει ή να διευρύνει περαιτέρω την ερώτηση και να επιτύχει καλύτερη αντιστοιχία με τις απαιτήσεις αναζήτησης του χρήστη, και θα είναι

απαραίτητο για το συνοψισμό των αποτελεσμάτων ερώτησης και την οπτικοποίηση των αποτελεσμάτων. Η επεξεργασία φυσικής γλώσσας και η κατανόηση ομιλίας έχουν χρησιμοποιηθεί ήδη πρόσφατα σε διεπαφές ανθρώπου-μηχανής. Η μηχανική όραση καλείται να συμπληρώσει τη λεκτική αναγνώριση και την κατανόηση φυσικής γλώσσας για τη φυσικότερη και ευφυέστερη επικοινωνία μεταξύ ανθρώπων και μηχανών.

Περισσότερο λεπτομερή δείγματα μπορούν να ληφθούν από χειρονομίες, στάσεις του σώματος, εκφράσεις του προσώπου, κ.λπ. Ως εκ τούτου, τα μελλοντικά συστήματα πρέπει να είναι σε θέση να αισθάνονται αυτόνομα το περιβάλλον π.χ., την ανίχνευση ανθρώπινης παρουσίας και ερμηνεία της ανθρώπινης συμπεριφοράς. Προκειμένου να εξαχθούν οι σημασιολογικές πληροφορίες οι άνθρωποι στο βίντεο πρέπει πρώτα να ανιχνευθούν. Οι άνθρωποι μπορούν να είναι στάσιμοι ή κινούμενοι. Η ανίχνευση των στάσιμων ατόμων σε βίντεο παραμένει ισοδύναμη με ανίχνευση ατόμων στις εικόνες. Αυτό περιλαμβάνει την ανίχνευση των μερών του ανθρώπινου σώματος συμπεριλαμβανομένου του προσώπου, του κορμού, των χεριών κ.λπ. σε μια δεδομένη εικόνα. Εάν υπάρχει κινούμενο πρόσωπο(-α) στο βίντεο τότε το πρόβλημα της ανίχνευσης είναι ευκολότερο υπό την έννοια ότι κάποιος μπορεί να εξαγάγει τα όρια του κινούμενου αντικειμένου και να εκτελέσει την ανίχνευση των μερών ανθρώπων ή σωμάτων μέσα στο όριο αντικειμένου.

Επιπλέον, τα όρια του αντικειμένου παρέχουν πρόσθετες πληροφορίες για την κατηγοριοποίηση (ταξινόμηση σε κλάσεις) και ανίχνευσή του. Προκειμένου να επιτευχθεί μια ακόμα πιο αξιόπιστη απόφαση, μπορεί επίσης να εκμεταλλευθεί ο σχετικός ήχος ή/και το κείμενο (εάν είναι διαθέσιμα). Εάν υπάρχει ανθρώπινη ομιλία στο ακουστικό μέρος του βίντεο μπορεί έπειτα να χρησιμοποιηθεί ως πρόσθετη ένδειξη για να επιτευχθεί μια τελική απόφαση. Επομένως, προτού εξαχθεί χρήσιμη γνώση από τα δεδομένα πολυμέσων, πρέπει πρώτα να λυθούν σε ένα σύστημα σημασιολογικών πληροφοριών, τα ακόλουθα προβλήματα:

- Ανίχνευση στάσιμων ανθρώπινων σωμάτων και μερών σε εικόνα και βίντεο
- Ανίχνευση και **κατηγοριοποίηση** (*classification*) κινούμενων αντικειμένων σε βίντεο.
- Ανίχνευση κινούμενων αντικειμένων και βηματισμού σε ανάλυση βίντεο
- Ανίχνευση ομιλίας στο ηχητικό μέρος, και χρήση των σχετικών δεδομένων για σχολιασμός του βίντεο, και
- Ανάλυση του διαθέσιμου κειμένου για την κατανόηση ανθρώπινης συμπεριφοράς στο βίντεο

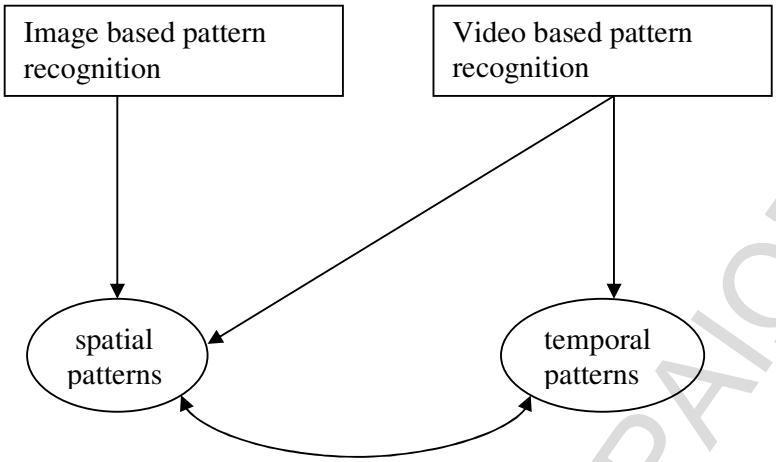
Αν και υπάρχουν πρακτικά συστήματα εντοπισμού και αναγνώρισης προσώπου σε ελεγχόμενα περιβάλλοντα, τα παραπάνω προβλήματα, δυστυχώς, δεν λύνονται επαρκώς σε περιπτώσεις όπου υπάρχουν σύνθετες φυσικές σκηνές. Ο εντοπισμός στάσιμης ανθρώπινης παρουσίας σε εικόνες ή βίντεο στοχεύει στην τομή των περιοχών όπου αντιστοιχούν άνθρωποι από το υπόλοιπο μιας εικόνας.

1. Αναγνώριση προτύπων σε video

1.1 Εισαγωγή

Οι εικόνες συνιστούν δεδομένα που ανήκουν σε πολυδιάστατους χώρους, της τάξης των εκατοντάδων χιλιάδων διαστάσεων. Η εξαγωγή συμπερασμάτων από συσχετισμό δεδομένων τέτοιων διαστάσεων οργανώνεται δύσκολα. Επομένως αρκετά από τα προβλήματα όπως η αναγνώριση προσώπων ή αντικειμένων, η κατανόηση σκηνών (δηλαδή πράξεων ή γεγονότων που συμβαίνουν σε κάποιο περιβάλλον) κ.ά, παραδοσιακά προσεγγίζονται με χρήση μεθόδων αναγνώρισης προτύπων. Τέτοιες μέθοδοι, σε συνδυασμό με άλλες τεχνικές μείωσης των διαστάσεων, υπήρξαν ιδιαίτερα δημοφιλείς και επιτυχείς στην αντιμετώπιση προβλημάτων επεξεργασίας εικόνας. Η εξάπλωση των φθηνών, υψηλής ποιότητας συσκευών *video camera* έχει προκαλέσει το ενδιαφέρον επέκτασης των μεθοδολογιών ανάλυσης εικόνας, σε *video*. Δηλαδή από την ανάλυση σταθερών εικόνων, στην ανάλυση των **ακολουθιών εικόνων** που καταγράφει μια τηλεοπτική **εικονοληπτική μηχανή** (*video camera*). Η προστιθέμενη διάσταση του χρόνου σε αυτά τα video δημιουργεί νέα προβλήματα όπως: αναγνώριση ανθρώπων με βάση το πρόσωπο και το βηματισμό, ανίχνευση γεγονότων και κατανόηση δραστηριοτήτων. Επίσης, τα **πρότυπα** (*patterns*) των **τηλεοπτικών ακολουθιών** (*video streams*) αναφέρονται τόσο σε πρότυπα που ενυπάρχουν στη δομή του εικονιζόμενου χώρου τα , τα λεγόμενα **χωρικά** ή **σταθερά πρότυπα** που περιβάλλουν τα σημεία ενδιαφέροντος, όσο και σε **μεταβαλλόμενα** ή **χρονικά πρότυπα** που προκύπτουν λόγω κίνησης της κάμερας ή των αντικειμένων. Στο εισαγωγικό αυτό κεφάλαιο, θα ασχοληθούμε με τις εφαρμογές της αναγνώρισης προτύπων σε *video* σε θέματα όπως: αναγνώριση προσώπου, αναγνώριση ανθρώπινου βηματισμού, κατηγοριοποίηση συμπεριφοράς και αναγνώριση ανθώπινης δραστηριότητας.

Η αναγνώριση προτύπων ασχολείται με την κατηγοριοποίηση δεδομένων σε διαθέσιμες κλάσεις. Προκειμένου να γίνει αυτό, πρέπει πρώτα να αποφασίζουμε σχετικά με το σύνολο των χαρακτηριστικών γνωρισμάτων (διάνυσμα χαρακτηριστικών) που θα αντιπροσωπεύουν τα στοιχεία, με τρόπο που θα απλοποιεί την ταξινόμηση. Μόλις αυτό καθοριστεί, τότε περιγράφουμε κάθε κλάση ή **κατηγορία μέσω δεσμευμένων πυκνοτήτων** (*class conditional densities*). Επόμενος στόχος είναι η κατηγοριοποίηση των δεδομένων στις διαθέσιμες κλάσεις εφαρμόζοντας τη Μπεϋζιανή θεωρία αποφάσεων. Αυτός ο στόχος αναγνώρισης, περιγραφής και αναγνώρισης οπτικών προτύπων έχει οδηγήσει σε πρόοδο την αυτοματοποίηση αρκετών διαδικασιών όπως η αναγνώριση οπτικών χαρακτήρων, ανάλυση σκηνών, ταυτοποίηση δαχτυλικών αποτυπομάτων, αναγνώριση προσώπου κτλ. Τα τελευταία χρόνια, η εξάπλωση των φθηνών, υψηλής ποιότητας συσκευών video, έχει προκαλέσει το ενδιαφέρον επέκτασης των μεθόδων αναγνώρισης προτύπων σε video. Στις ακολουθίες εικόνων πλέον, υπάρχουν διακρίνονται δύο κατηγορίες προτύπων. Η πρώτη είναι τα **χωρικά πρότυπα** (*spatial patterns*), τα οποία αντιστοιχούν στα ήδη γνωστά προβλήματα αναγνώρισης προτύπων σε εικόνες, και συναντάμε σε εφαρμογές όπως αναγνώριση προσώπων ή δακτυλικών αποτυπωμάτων.Εκτός από τα χωρικά πρότυπα, τα οποία μπορούν να χαρακτηριστούν και «σταθερά», σε εικόνες video παρουσιάζεται ένα πλήθος από **χρονικά πρότυπα** (*temporal patterns*). Σε αρκετές εφαρμογές όπως, αναγνώριση δραστηριοτήτων, ανίχνευση ή/και κατηγοριοποίηση συμβάντων, ανίχνευση ανωμαλιών, αναγνώριση δραστηριοτήτων ατόμων κτλ, υπάρχουν προσωρινές ακολουθίες, στις οποίες παρουσιάζονται διάφορα από τα χωρικά πρότυπα. Είναι ιδιαίτερα σημαντική λοιπόν η ανίχνευση των προσωρινών προτύπων των video.



Σχήμα 1.1: *χωρικά και χρονικά πρότυπα και αναγνώριση προτύπων.*

1.2 Αναπαράσταση χαρακτηριστικών γνωρισμάτων

Στα περισσότερα προβλήματα **αναγνώρισης προτύπων** (*Pattern Recognition*), ιδιαίτερα σημαντική είναι η διαδικασία εξαγωγής των χαρακτηριστικών. Είναι πολύ στενά συνδεδεμένη με την **αναπαράσταση προτύπων** (*pattern representation*). Είναι επίσης δύσκολο να επιτευχθεί γενίκευση των προτύπων χωρίς την εφαρμογή μιας σωστής αναπαράστασης. Η επιλογή αναπαράστασης όχι μόνο επηρεάζει τη μέθοδο προσέγγισης του προβλήματος σε μεγάλο βαθμό, αλλά και περιορίζει την απόδοση του συστήματος ανάλογα με την καταλληλότητα της επιλογής. Παραδείγματος χάριν, αν υποτεθεί ένα επίπεδο μοντέλο δύο διαστάσεων, δε μπορεί να γίνει αξιόπιστη ανάκτηση των οπτικών γωνιών ενός προσώπου σε κλίση, ή και που αλλάζει κλίσεις.

Ανάλογα με τη φύση του προβλήματος, η αναπαράσταση χαρακτηριστικών μπορεί να γίνει με διαφορετικούς τρόπους. Αν και στην περίπτωση των σταθερών εικόνων αρκεί η μοντελοποίηση του χώρου, ωστόσο στην περίπτωση των video, είναι απαραίτητη η σωστή αναπαράσταση της μεταβαλλόμενης χρονικής πληροφορίας. Κάποιες φορές, απαιτείται αναλυτική αναπαράσταση, όπως σε γεωμετρικά πρότυπα, ωστόσο υπάρχουν προβλήματα PR όπου αυτό δεν είναι απαραίτητο.

1.2.1 Μοντέλο Συσχετισμού Εμφανίσεων

Η αναγνώριση αντικειμένων σε video απαιτεί τη μοντελοποίηση της κίνησης των αντικειμένων και των μεταβολών της εμφάνισης αυτών. Αυτό καθιστά τον **εντοπισμό αντικειμένων** (*image tracking*), ένα κρίσιμο βήμα πριν την αναγνώριση. Στους συμβατικούς αλγόριθμους, το **μοντέλο εμφάνισης** (*appearance model*) είναι είτε σταθερό, είτε ταχύτατα μεταβαλλόμενο, ενώ το **μοντέλο κίνησης** (*motion model*) είναι ένα τυχαίο μοντέλο βηματισμού σταθερής διαφοράς. Ένα σταθερό πρότυπο εμφάνισης δεν είναι ικανό για να χειριστεί τις μεταβολές εμφάνισης σε βίντεο, ενώ ένα γρήγορα μεταβαλλόμενο πρότυπο είναι ευαίσθητο στις μετατοπίσεις. Όλοι αυτοί οι παράγοντες μπορούν να οδηγήσουν σε αστάθεια τον οπτικό εντοπισμό οδηγώντας σε ανεπαρκή αποτελέσματα αναγνώρισης. Για το λόγο αυτό χρησιμοποιούνται προσαρμοστικά μοντέλα εμφάνισης προκειμένου ο εντοπισμός να σταθεροποιείται, καθώς και να παρακολουθούνται στενά οι μεταβολές στην εμφάνιση εξαιτίας της κίνησης αντικειμένων. Έτσι, η εμφάνιση μοντελοποιείται ως μίξη τριών διαφορετικών προτύπων:

1. Η εμφάνιση των αντικειμένων σε ένα **πλαίσιο-κανόνα** (*canonical frame*), συνήθως το πρώτο πλαίσιο της ακολουθίας.
2. Η διακύμανση της σταθερής εμφάνισης μέσα στην παρατήρηση.
3. Τα ταχέως μεταβαλλόμενα συστατικά που καθιστούν δύο πλαίσια διαφορετικά.

Οι πιθανότητες κάθε ενός από τα τρία αυτά «μίγματα», ανανεώνονται σε κάθε **πλαίσιο** (*frame*) της παρατήρησης. Επιπλέον, εφαρμόζεται ένα πρότυπο προσαρμοστικής ταχύτητας, όπου η προσαρμοστική ταχύτητα προβλέπεται χρησιμοποιώντας μια first-order γραμμική προσέγγιση με βάση τις αλλαγές τις εμφάνισης μεταξύ της εισερχόμενης παρατήρησης και της προηγούμενης διαμόρφωσης. Ο στόχος εδώ είναι να προσδιοριστεί μια περιοχή ενδιαφέροντος για κάθε πλαίσιο του βίντεο και όχι η τρισδιάστατη θέση του αντικειμένου. Επιπλέον, τα προσαρμοστικά μοντέλα εμφάνισης μπορούν εύκολα να απορροφήσουν τις μεταβολές εμφάνισης που οφείλονται σε αλλαγές πόζας ή φωτισμού. Επομένως, χρησιμοποιούνται επίπεδα-δισδιάστατα πρότυπα και επιτρέπονται μονάχα μετασχηματισμοί που να σχετίζονται μεταξύ τους. Το *σχήμα 1.2*, αναπαριστά ένα παράδειγμα εντοπισμού προσώπου σε βίντεο. Στην πρώτη εικόνα του σχήματος το αντικείμενο-πρόσωπο εμφανίζεται στο πλαίσιο-κανόνας. Σε κάθε ένα από τα επόμενα πλαίσια γίνεται συσχετισμός με το πρώτο που ορίστηκε ως «κανόνας» (δηλαδή μέτρο σύγκρισης).

Σχήμα 1.2: *παράδειγμα εντοπισμού προσώπου σε βίντεο*

1.2.2 Τρισδιάστατες Γραφικές Παραστάσεις Χαρακτηριστικών

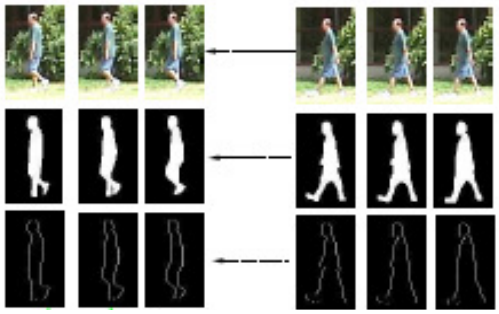
Το πρότυπο συσχετισμού αρκεί για την εντόπιση της θέσης του αντικειμένου στην εικόνα, αλλά δεν έχει την ικανότητα να επαληθεύσει την τρισδιάστατη διαμόρφωση του αντικειμένου για κάθε στιγμή ξεχωριστά. Παραδείγματος χάριν, εάν ο στόχος είναι να χρησιμοποιηθούν οι τρισδιάστατες πληροφορίες για την αναγνώριση προσώπου στο βίντεο, η αναπαράσταση συσχετισμών δεν θα είναι επαρκής. Κάποιοι αλγόριθμοι χρησιμοποιούν ένα κυλινδρικό πρότυπο με ελλειπτική διατομή που εκτελεί εντοπισμό και αναγνώριση προσώπου σε τρισδιάστες παραστάσεις. Η κυρτή επιφάνεια του κυλίνδρου διαιρείται σε ορθογώνια πλέγματα και το διάνυσμα που περιέχει τις μέσες τιμές έντασης για κάθε ένα από τα πλέγματα αυτά χρησιμοποιείται ως χαρακτηριστικό γνώρισμα. Όπως προηγουμένως, το πρότυπο εμφάνισης είναι ένα μίγμα από το σταθερό στοιχείο (που παράγεται από το πρώτο πλαίσιο) και το δυναμικό (εμφάνιση προηγούμενου πλαισίου). Το σχήμα 1.3 παρουσιάζει μερικά πλαίσια ενός βίντεο με τον κύλινδρο να επικαλύπτει την εικόνα μιας υπολογισμένης πόζας.

Σχήμα 1.3: *εντοπισμός προσώπου με χρήση κυλινδρικού προτύπου*

Μια άλλη δυνατότητα είναι να χρησιμοποιηθεί ένα πιο ρεαλιστικό πρότυπο προσώπου (π.χ., τρισδιάστατο πρότυπο ενός μέσου προσώπου) αντί ενός κυλίνδρου. Τέτοιες λεπτομερείς τρισδιάστατες αναπαραστάσεις καθιστούν τη διαδικασία δύσκολότερη. Υπάρχουν πειράματα όπου διαταραχές στις παραμέτρους των προτύπων έχουν επιπτώσεις στην απόδοση μεθόδων εντοπισμού που χρησιμοποιούν ένα σύνθετο τρισδιάστατο πρότυπο, ενώ το απλό κυλινδρικό πρότυπο εμφανίζεται πιο σταθερό σε τέτοιες διαταραχές. Αυτό δίνει έμφαση στη σημασία της γενικευμένης αναπαράστασεις των χαρακτηριστικών. Είναι, σε τελευταία ανάλυση, προτιμότερη η χρήση ενός λιγότερου ρεαλιστικού και ταυτόχρονα περισσότερο γενικευμένου προτύπου σε αυτού του είδους τις εφαρμογές αναγνώρισης προσώπου σε video.

1.2.3 Αναγνώριση με βάση το Βηματισμό

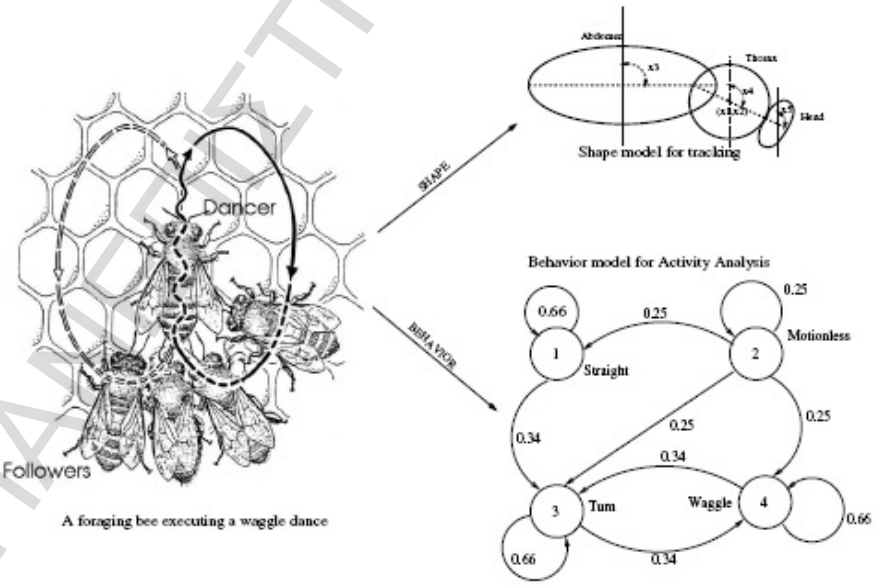
Ο βηματισμός είναι μια ιδαίτερα δομημένη δραστηριότητα με ορισμένα κύρια γεγονότα όπως το χτύπημα της φτέρνας, τα οποία εκτελούνται διαδοχικά σε ένα επαναλαμβανόμενο πρότυπο. Πρόσφατες έρευνες υποστηρίζουν ότι ο βηματισμός ενός ατόμου μπορεί να είναι ευδιάκριτος και χαρακτηριστικός και επομένως μπορεί να χρησιμοποιηθεί ως βιομετρικό για τον προσδιορισμό ατόμων. Χαρακτηριστικές αναπαραστάσεις για τον προσδιορισμό ατόμων με βάση το βηματισμό περιλαμβάνουν τη χρήση της ολόκληρης δυαδικής σκιαγραφίας του σώματος, και άλλες αναπαραστάσεις, όπως διάνυσμα πλάτους ή μορφή του εξωτερικού περιγράμματος. Μια επίσης βιώσιμη αναπαράσταση για την ανάλυση βηματισμού, αποτελούν οι τρισδιάστατες περιγραφές μερών του ανθρώπινου σώματος.
Η ανάλυση ανθρώπινου βηματισμού/δραστηριότητας σε video προσελκύει το ενδιαφέρον πολλών ερευνητών του τομέα της μηχανικής όρασης. Εφαρμόζοντας μεθόδους αναγνώρισης προτύπων, γίνεται ανάλυση δύο εκ των κυριότερων θέσεων-στάσεων του ανθρώπινου σώματος κατά το βηματισμό. Έτσι, εφαρμόζοντας και ιδέες από την κινηματική, μοντελοποιείται το περίγραμμα ενός ατόμου που βαδίζει ως μια ακολουθία σχημάτων που αλλάζουν μορφή. Επίσης, προτείνονται μετρικές σύγκρισης δύο τέτοιων ακολουθιών μέσω του **αλγορίθμου DTW** (*Dynamic Time Warping*). Οι ακολουθίες σχημάτων μοντελοποιούνται είτε μέσω **αυτοανάδρομων προτύπων** (*autoregressive*) είτε με πρότυπα **κινούμενου μέσου όρου** (*moving average*). Το σχήμα δείχνει μια γραφική απεικόνιση αναγνώρισης βηματισμού με σύγκριση των ακολουθιών μορφής.



Σχήμα 1.4: *παράδειγμα εντοπισμού ανθρώπινου βηματισμού*

1.2.4 Μοντέλα Συμπεριφοράς

Η στατιστική διαμόρφωση της κίνησης των αντικειμένων μας επιτρέπει να συλλάβουμε τα χρονικά πρότυπα ενός βίντεο. Η αναλυτική μοντελοποίηση τέτοιων συμπεριφορών είναι χρήσιμη για την επίτευξη ακριβέστερου και αποδοτικότερου εντοπισμού. Θεωρητικά κάθε αντικείμενο θα μπορούσε να επιδείξει τις πολλαπλές συμπεριφορές. **Μοντέλα Markov** χρησιμοποιούνται (κυρίως σε περιπτώσεις κινητικότητας χαμηλού επιπέδου), για την αναπαράσταση κάθε μιας συμπεριφοράς του αντικειμένου. Αυτό δημιουργεί ένα αναμεμιγμένο πλαίσιο μοντελοποίησης της κίνησης του αντικειμένου. Για παράδειγμα, ένα χαρακτηριστικό μοντέλο Markov για ένα είδος κινητικής συμπεριφοράς μιας μέλισσας βλέπουμε στο σχήμα 1.4.



Σχήμα 1.5: *ανάλυση κινητική συμπεριφορά μέλισσας με χρήση μοντέλου Markov*

1.3 Particle Filtering για την αναγνώριση αντικειμένων σε video.

Έχουμε εξετάσει μέχρι τώρα ζητήματα που αφορούν την αναπαράσταση των προτύπων σε video, και ασχοληθήκαμε με το πώς μπορούν να αναπαρασταθούν τα χωρικά και χρονικά πρότυπα με τρόπο που να απλοποιεί τον προσδιορισμό αυτών. Αλλά, μιας και επιλέχτηκε ένα ορισμένο σύνολο αναπαραστάσεων για χωρικά πρότυπα και πρότυπα κινήσεων, χρειαζόμαστε τους κατάλληλους αλγόριθμους για τον υπολογισμό αυτών των παραμέτρων. Μια μέθοδος για να εφαρμοστεί αυτό είναι η αντιμετώπιση του προβλήματος εκτίμησης των παραμέτρων, ως πρόβλημα ελαχιστοποίησης των ενεργειών, κι έτσι να γίνει χρήση των δημοφιλών μεθόδων ελαχιστοποίησης που βασίζονται σε υπολογισμούς διακύμανσης. Τέτοιες μέθοδοι είναι: *αλγόριθμοι καθόδου αρνητικής βαθμίδας (gradient descent)*, η *προσομοιωμένη-ντετερμινιστική ανόπτηση* (*simulated/deterministic annealing*), η μέθοδος *Προσδοκίας-Μεγιστοποίησης (Expectation-Maximization*), κ.ά. Οι περισσότερες από αυτές είναι τοπικές και ως εκ τούτου δεν εγγυώνται ότι θα συγκλίνουν στο σφαιρικό βέλτιστο. Η προσομοιωμένη ανόπτηση μπορεί να συγκλίνει στο σφαιρικό βέλτιστο εάν ακολουθείται από τον κατάλληλο σχεδιασμό αλλά αυτό καθιστά τον αλγόριθμο εξαιρετικά αργό και υπολογιστικά κοστοβόρο. Όταν η περιγραφή παρατήρησης της κατάστασης του συστήματος είναι γραμμική και γκαουσσιανή, τότε ο υπολογισμός των παραμέτρων μπορεί να εκτελεσθεί χρησιμοποιώντας το φίλτρο Kalman. Αλλά ο σχεδιασμός του φίλτρου Kalman γίνεται περίπλοκος για τα πραγματικά, μη γραμμικά, προβλήματα και είναι ακατάλληλος για τον υπολογισμό των μεταγενέστερων πυκνοτήτων που δεν ακολουθούν την Γκαουσσιανή κατανομή. Με τη μέθοδο *φίλτρου μορίων* (*particle filter*), εκτιμώνται οι αυθαίρετες μεταγενέστερες πυκνότητες, αναπαριστώντας αυτές ως ένα σύνολο σταθμισμένων μορίων.

1.3.1 Το πρόβλημα

Σε ένα σύστημα ϑ παραμέτρων, οι παράμετροι αυτές ακολουθούν μια συγκεκριμένη χρονική δυναμική που δίνεται από την Ft (θ, D, N). (Εδώ η δυναμική του συστήματος αλλάζει με το χρόνο).

Δυναμική Συστήματος: $\boldsymbol{\vartheta t = Ft(\vartheta t-1, Dt, Nt)}$

όπου *N* ο θόρυβος του συστήματος. Η βοηθητική μεταβλητή *D* δείχνει το σύνολο προτύπων κινήσεων ή συμπεριφορών που παρουσιάζει το αντικείμενο. Αυτή η βοηθητική μεταβλητή αποκτά σημασία σε προβλήματα όπως η αναγνώριση δραστηριότητας ή ανάλυση συμπεριφοράς, ενώ σε τυπικά προβλήματα ανίχνευσης παραλείπεται. Κάθε πλαίσιο του βίντεο περιέχει τις εντάσεις των pixels που ενεργούν ως επιμέρους Z παρατηρήσεις, μιας ϑ κατάστασης του συστήματος.

Εξίσωση Παρατηρήσεων: $\boldsymbol{Zt = G(\vartheta t, I, Wt)}$

όπου, *W* ο θόρυβος των παρατηρήσεων. Η βοηθητική μεταβλητή *I* δηλώνει τις διάφορες κατηγορίες (κλάσεις) αντικειμένων που διαμορφώνονται, δηλ., αντιπροσωπεύει την ταυτότητα του αντικειμένου. Το ενδιαφέρον πρόβλημα εδώ είναι να εκτιμώνται οι παράμετροι του συστήματος κατά τη διάρκεια του χρόνου, όποτε οι παρατηρήσεις γίνονται διαθέσιμες. Ποσοτικά, ενδιαφερόμαστε για τον υπολογισμό

της εκ των υστέρων πιθανότητας της κατάστασης του συστήματος (η οποία εκφράζεται από τις παραμέτρους) δεδομένων των παρατηρήσεων, δηλαδή: ***P*(*ϑt*/*Z*1:*t*).**

1.3.2 Φίλτρο Μορίων

Το «φιλτράρισμα μορίων» είναι μια τεχνική εξαγωγής συμπερασμάτων για τον υπολογισμό μιας άγνωστης δυναμικής κατάστασης ϑ ενός συστήματος από μια συλλογή παρατηρήσεων που εμφανίζουν θόρυβο. Το φίλτρο προσεγγίζει την επιθυμητή εκ των υστέρων συνάρτηση πυκνότητας πιθανότητας (σ.π.π.): ***p*(*ϑt*|*Z*1:*t*)**, από ένα σύνολο σταθμισμένων μορίων: $\left\{\theta_t^{(j)}, w_t^{(j)}\right\}_{j=1}^{M}$, όπου το Μ δηλώνει το πλήθος τον μορίων. Το $\hat{\theta}_t$ μπορεί να ανακτηθεί από τη σ.π.π. ως εκτίμηση μέγιστης πιθανοφάνειας (ML), ή ελαχιστοποίησης του μέσου τετραγωνικού σφάλματος (MMSE) ή με οποιοαδήποτε κατάλληλη τεχνική εκτίμησης, με βάση τη σ.π.π.

1.3.3 Εντοπισμός και Αναγνώριση Ατόμου

Έστω μια συλλογή *P* αντικειμένων, και ότι ένα από αυτά τα αντικείμενα περιέχεται στο video. Στόχος είναι ο εντοπισμός των παραμέτρων θέσης ϑ του αντικειμένου και ταυτόχρονα η αναγνώριση της ταυτότητάς του. Για κάθε αντικείμενο *i* , η εξίσωση της παρατήρησης δίνεται από την $Z_t = G(\theta_t, i, W_t)$. Έστω ότι γνωρίζουμε ότι ψάχνουμε το *p-οστό* αντικείμενο, κατόπιν, θα μπορούσε να βρεθεί μεσω ενός particle filter υπολογίζοντας την εκ των υστέρων πιθανότητα της $P(\theta_t / Z_{1:t}, p)$, ως ένα σύνολο *M* σταθμισμένων μορίων: $\left\{\theta_t^{(j)}, w_t^{(j)}\right\}_{j=1}^{M}$. Εάν όμως η ταυτότητα του αντικειμένου που αναζητείται δεν είναι γνωστή (δηλαδή δεν είναι γνωρίζουμε εκ των προτέρων πιο αντικείμενο της συλλογής θέλουμε να βρούμε), τότε πρέπει να εκτιμηθεί καί αυτή επίσης. Υποθέτουμε ότι η ταυτότητα του αντικειμένου δεν αλλάζει σε όλη τη χρονική διάρκεια του video, δηλαδή ότι: $I_t = p$, όπου $p = \{1,2,...,P\}$. Τότε, αφού η ταυτότητα παραμένει σταθερή στο χρόνο, έχουμε:

$$P(X_t, I_t = i / X_{t-1}, I_{t-1} = j) = P(X_t / X_{t-1}) P(I_t = i / I_{t-1} = j)$$

$$= \begin{cases} 0 & \text{όταν } i \neq j \\ P(X_t / X_{t-1}) & \text{όταν } i = j \text{ όπου } i = \{1,2,...,P\} \end{cases}$$

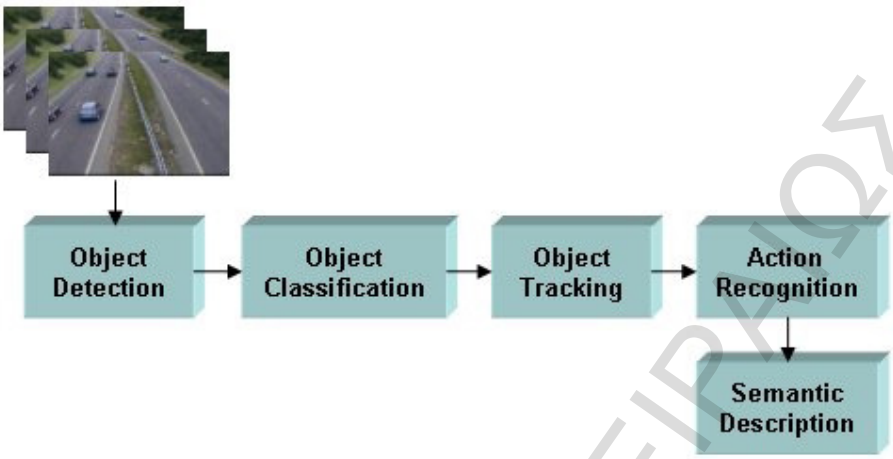
Σύμφωνα με την προηγούμενη παράγραφο, η εκ των υστέρων σ.π.π. $P(X_t, I = p / Z_{1:t})$, υπολογίζεται μέσω ενός συνόλου M_p σταθμισμένων μορίων: $\left\{\theta_{t,p}^{(j)}, w_{t,p}^{(j)}\right\}^{j=1:M_p}$. Διατηρείται ένα τέτοιο σύνολο *M p* μορίων για κάθε αντικείμενο *p* = 1 , 2, ..*P* . Τώρα το σύνολο των σταθμισμένων μορίων $\left\{\theta_{t,p}^{(j)}, w_{t,p}^{(j)}\right\}_{j=1:M_p}^{p=1:P}$ με βάρη τέτοια ώστε: $\sum_{p=1:P} \sum_{j=1:M_i} w_{t,p}^{(j)} = 1$, αντιπροσωπεύει την κατανομή της $P(\theta_t / Z_{1:t})$.

Οι εκτιμήσεις των μεθόδων ελαχιστοποίησης του μέσου τετραγωνικού σφάλματος, και μέγιστης πιθανοφάνειας, για τις παραμέτρους εντοπισμού $\hat{\theta}_t$, μπορούν να υπολογισθούν ελαχιστοποιώντας την κατανομή $P(\theta_t / Z_{1:t})$, ως προς τη μεταβλητή ταυτότητας. Ομοίως, μέσω της ελαχιστοποίησης της εκ των υστέρων κατανομής ως προς τις παραμέτρους $\hat{\theta}_t$, μπορεί να εκτιμηθεί η μεταβλητή ταυτότητας.

1.3.4 Αναγνώριση Συμπεριφοράς

Η ταυτόχρονη εκτέλεση εντοπισμού και ανάλυσης συμπεριφοράς/δραστηριότητας μπορεί επίσης να γίνει χρησιμοποιώντας μια βοηθητική μεταβλητή *D* κατά τρόπο παρόμοιο με την ταυτόχρονη εκτέλεση εντοπισμού και επαλήθευσης. Ουσιαστικά, το σύνολο των σταθμισμένων μορίων $\{\theta_t^{(j)}, w_t^{(j)}, D_t^{(j)}\}$ αντιπροσωπεύει την κατανομή της εκ των υστέρων πιθανότητας $P(\theta_t, D_t / Z_{1:t})$. Έπειτα, υπολογίζοντας το σχετικό ελάχιστο κατανομής μέσω της εκ των υστέρων κατανομής, μπορούν να εξαχθούν συμπεράσματα, τόσο για το διάνυσμα παραμέτρων $\hat{\theta}_t$, όσο και για τη συμπεριφορά του αντικειμένου D_t.

2. Εντοπισμός Ανθρώπων σε Video



Σχήμα 2.1 *Διάγραμμα ενός χαρακτηριστικού ευφυούς συστήματος ανάλυσης video*

Ένα διάγραμμα ενός χαρακτηριστικού ευφυούς συστήματος ανάλυσης video παρουσιάζεται στο *σχήμα 2.1*. Λαμβάνοντας υπόψη ένα ψηφιακό τηλεοπτικό σήμα, το πρώτο βήμα είναι να ανιχνευθούν οι περιοχές ενδιαφέροντος, οι οποίες είναι συνήθως κινούμενα αντικείμενα σε πολλά πρακτικά προβλήματα. Η ανίχνευση των κινούμενων περιοχών υπό μορφή *σταγόνων εικόνας (image blobs)* παρέχει μια στενότερη περιοχή ενδιαφέροντος για τις πιό πρόσφατες διαδικασίες όπως η ανίχνευση και η ανάλυση δραστηριότητας, επειδή μόνο τα pixels της κινούμενης περιοχής χρειάζεται να αναλυθούν. Αυτό βέβαια δεν το καθιστά τετριμμένο πρόβλημα, λόγω των αλλαγών φωτισμού, και των σκιών στις περισσότερες φυσικές σκηνές.

Οι μέθοδοι *εκτίμησης φόντου ή υποβάθρου (background estimation)* για την ανίχνευση ανθρώπων, χρησιμοποιούνται ευρέως και στην ανίχνευση κινούμενου αντικειμένου σε video. Σε αυτές τις μεθόδους το φόντο υπολογίζεται προσαρμοστικά από τα περασμένα πλαίσια εικόνων και τα *pixels (εικονοστοιχεία)* των κινούμενων αντικειμένων υπολογίζονται με την αφαίρεση της τρέχουσας εικόνας-πλαίσιο του video από το κατ' εκτίμηση φόντο. Το *υπόβαθρο (background)* του video ορίζεται ως η ένωση όλων των στάσιμων αντικειμένων και το *«πρώτο πλάνο» (foreground)* αποτελείται από τα κινούμενα αντικείμενα. Μια απλή προσέγγιση για τον υπολογισμό της εικόνας του φόντου είναι να υπολογίσουμε κατά μέσο όρο όλα τα προηγούμενα πλαίσια του video, επειδή η επίδραση της κίνησης των αντικειμένων θα είναι αμελητέα στην πορεία του χρόνου με την πρυπόθεση ότι η κάμερα είναι στάσιμη. Σε εργασίες των R.T. Collins, A. Lipton, T. Kanade (βλ. βιβλιογραφία [1-3]), το τρέχον φόντο του video υπολογίζεται επαναληπτικά από τα προηγούμενα πλαίσια εικόνας χρησιμοποιώντας επαναλαμβανόμενης πρώτης σειράς (recursive first order) φίλτρα *Infinite-duration Impulse Response (IIR)* που ενεργούν παράλληλα σε κάθε pixel του video. Μια στατιστική μέθοδος εκτίμησης υποβάθρου περιγράφεται αναλυτικά στο άρθρο του Stauffer [23] στο οποίο, κάθε εικονοκύτταρο μοντελοποιείται ως μίγμα Γκαουσιανών συναρτήσεων και το πρότυπο ενημερώνεται με επαναληπτική μέθοδο. Αυτή η μέθοδος οδηγεί σε έναν αξιόπιστο, ανιχνευτή υπαίθριου χώρου, πραγματικού χρόνου, ο οποίος μπορεί να αντιμετωπίσει τις αλλαγές φωτισμού. Αναπτύχθηκαν επίσης στατιστικές μέθοδοι διάταξης (Order statistical methods) για εκτίμηση υποβάθρου. Ο αλγόριθμος των Yang και Levine [24] χρησιμοποιούν τη μέση αξία ενός τρέχοντος pixel μέσα σε μια σειρά προηγούμενων εικόνων. Στο άρθρο [4], ένα

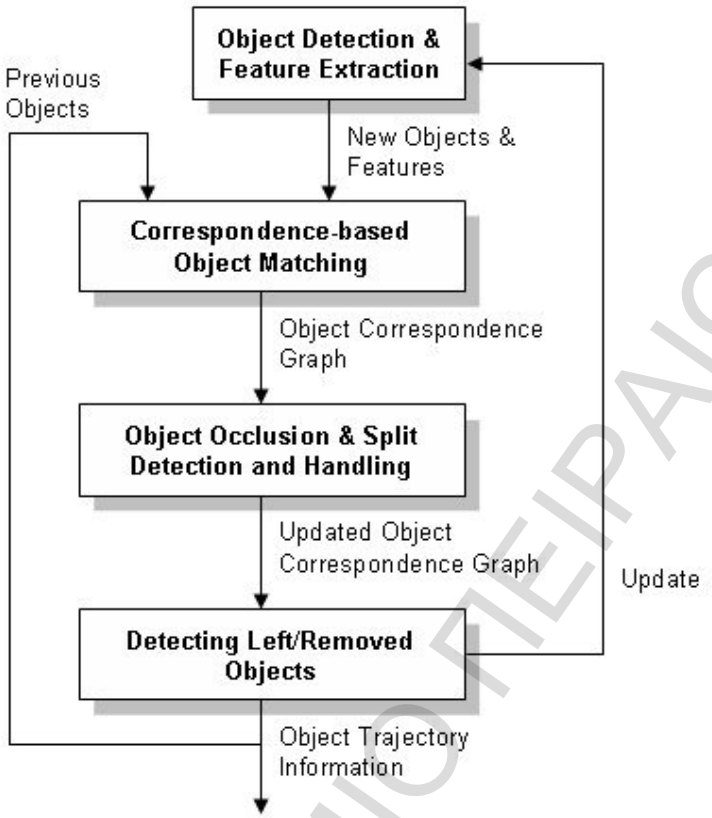
στατιστικό πρότυπο κατασκευάζεται με την αντιπροσώπευση κάθε ενός pixel από τις ελάχιστες και μέγιστες τιμές έντασής του, και η μέγιστη διαφορά έντασης των διαδοχικών πλαισίων που παρατηρούνται μεταξύ αυτών κατά τη διάρκεια μιας περιόδου εκπαίδευσης. Οι παράμετροι του μοντέλου ενημερώνονται περιοδικά. Οι πολυάριθμες μέθοδοι που περιγράφονται στη βιβλιογραφία για το πρόβλημα εκτίμησης υποβάθρου, διαφέρουν κυρίως ως προς στον τύπο του υποβάθρου και τη διαδικασία που χρησιμοποιείται για την ενημέρωση του πρότυπου του υποβάθρου. Σε όλες αυτές τις μεθόδους, τα pixels των αντικειμένων του foreground υπολογίζονται με αφαίρεση-αποκοπή της τρέχουσας εικόνας από την κατ' εκτίμηση εικόνα background. Οι κινούμενες «σταγόνες» κατασκευάζονται από τα pixels με την εκτέλεση μιας ανάλυσης των συνδεδεμένων τμημάτων. Εάν η συσκευή καταγραφής video είναι μη στατική, τότε πρέπει να χρησιμοποιηθεί επίσης ένας αλγόριθμος *αντιστάθμισης κινήσης (camera motion compensation).* Άλλη κατηγορία μεθόδων εκτίμησης κίνησης αντικειμένου περιλαμβάνει μεθόδους βασισμένες στην *οπτική εκτίμηση ροής (optical flow estimation).* Τέτοιες μέθοδοι χρησιμοποιούν *διανύσματα οπτικής ροής (optical flow vectors)* της κίνησης των αντικειμένων κατά τη διάρκεια του χρόνου ώστε να ανιχνευθούν αλλαγές περιοχών σε μια ακολουθία εικόνων.

2.1 Ανίχνευση Κινούμενων Αντικειμένων

Πρέπει να ανιχνεύονται οι κινούμενες σταγόνες του video για να επιτευχθούν αξιόπιστα αποτελέσματα αναγνώρισης. Η ανίχνευση αντικειμένου είναι απαραίτητη για την ανάλυση και αναγνώριση ανθρώπινης συμπεριφοράς σε video. Βασικά, ανίχνευση (ή, θα μπορούσαμε να πούμε, ιχνηλασία) των κινούμενων σταγόνων (blobs) από μια ακολουθία εικόνων περιλαμβάνει το συσχετισμό ανάμεσα σε ανιχνευμένα κινούμενα αντικείμενα μέσα σε διαδοχικά πλαίσια, χρησιμοποιώντας τα διάφορα χαρακτηριστικά γνωρίσματα των σταγόνων. Μερικές μέθοδοι ανίχνευσης μπορούν επίσης να προβλέψουν τη θέση της κινούμενης σταγόνας στο επόμενο πλαίσιο εικόνας. Τυπικοί αλγόριθμοι ανίχνευσης είναι βασισμένοι σε στατιστικές μεθόδους συσχετισμού, *φίλτρα Kalman* (*Kalman filtering*) [28], αλγόριθμος συμπύκνωσης (condensation algorithm) [29], ανίχνευση μέσης μετατόπισης (mean-shift tracking) [30], δυναμικό Μπεϋζιανό δίκτυο (dynamic Bayesian network) [31], *φιλτράρισμα στοιχείων (particle filtering)* [28]. Με τη μέθοδο *particle filtering* ασχοληθήκαμε στο κεφάλαιο 1. Οι μέθοδοι συσχετισμού δουλεύουν καλά εάν δεν υπάρχει καμία αλλαγή κλίμακας στα αντικείμενα που ανιχνεύονται. Οι μέθοδοι φιλτραρίσματος Kalman αναπτύσσονται για προβλήματα εντοπισμού ραντάρ και λειτουργούν πολύ καλά στην ανίχνευση σημειακών στόχων.

Ο αλγόριθμος συμπύκνωσης αναπτύχθηκε πρόσφατα και είναι βασισμένος στην διάδοση δεσμευμένης πυκνότητας σε ένα σύνολο από διαδοχικά πλαίσια εικόνων. Σε αυτόν τον αλγόριθμο μια *«εκ των υστέρων» (posterior)* κατανομή
υπολογισμένη σε ένα προηγούμενο πλαίσιο επεκτείνεται στα επερχόμενα πλαίσια κατά τρόπο επαναληπτικό. Ένα δυναμικό μοντέλο μπορεί να συνδυαστεί με οπτικές παρατηρήσεις για να επιτύχει αξιόπιστη ανίχνευση ενός αντικειμένου. Εντούτοις, χρειάζεται ένας μεγάλος αριθμός δειγμάτων για να υπολογίστούν οι κατανομές ώστε να εξαχθούν αξιόπιστες εκτιμήσεις μέγιστης πιθανότητας μιας δοσμένης κατάστασης. Η ανίχνευση *μέσης μετατόπισης* είναι μια άλλη πρόσφατα αναπτυγμένη μέθοδος για κινούμενα αντικείμενα σε video. Αποτελεί σημαντικό εργαλείο για οποιοδήποτε σύστημα ανίχνευσης. Χρησιμοποιεί ένα *ομαλοποιημένο και λειασμένο ιστόγραμμα χρώματος (normalized and smoothed color histogram)* των κινούμενων αντικειμένων. Η λείανση του ιστογράμματος χρώματος είναι ουσιαστικά ισοδύναμη με την αποκαλούμενη διαδικασία *εκτίμησης πυκνότητας πυρήνα (kernel density estimation).* Ο τυπικός αλγόριθμος ανίχνευσης *μέσης μετατόπισης* καθορίζει τη θέση του κινούμενου αντικειμένου μέσα το

επόμενο πλαίσιο εικόνας μέσω μιας επαναληπτικής διαδικασίας. Μόλις κατασκευαστεί το ιστόγραμμα που αντιπροσωπεύει ένα κινούμενο αντικείμενο χειρωνακτικά ή αυτόματα από ένα αρχικό πλαίσιο εικόνας (n frame) στο οποίο το αντικείμενο εμφανίζεται για πρώτη φορά, η θέση αυτού του αντικειμένου στο επόμενο πλαίσιο (*n+1* frame) υπολογίζεται με τον καθορισμό της ομοιότητας του ιστογράμματος *Hn* του αντικειμένου, με διάφορα ιστόγραμμα που εξάγονται από το επόμενο πλαίσιο (*n+1*). Επίσης, το αρχικό ιστόγραμμα επαναλαμβάνεται στην ίδια θέση στο επόμενο πλαίσιο. Ωστόσο, καλό θα ήταν επίσης να υπολογιστεί το αρχικό ιστόγραμμα H1 μέσα στην κοντινότερη περιοχή *ρ σταγόνων κίνησης* που λαμβάνεται ύστερα από την αφαίρεση-αποκοπή του υποβάθρου της εικόνας και να συγκριθεί μέσω του *μέτρου Bhattacharya*. Εάν η απόσταση είναι μεγαλύτερη από το προκαθορισμένο κατώτατο όριο, αυτό σημαίνει ότι η σταγόνα κίνησης που λαμβάνεται στο επόμενο πλαίσιο δεν συσχετίζεται με το αντικείμενο *O* και το ιστόγραμμα δημιουργείται, στο *n+1* πλαίσιο εικόνας, στη θέση του αντικειμένου *O* στο *n-οστό* πλαίσιο. Η απόσταση μεταξύ των δύο ιστογραμμάτων μπορεί να βρεθεί χρησιμοποιώντας και άλλα μέτρα απόστασης επίσης. Οι συγκρίσεις ιστογραμμάτων εκτελούνται κατά τρόπο επαναληπτικό έως ότου καθιερωθεί μια ικανοποιητική αντιστοιχία. Μια διαδικασία, που ανέπτυξε ο Comaniciu, καθορίζει ένα *διάνυσμα συμψηφισμού* (offset vector) που ονομάζεται *διάνυσμα μέσης μετατόπισης (mean-shift vector)*. Φαίνεται ότι οι επαναλήψεις συγκλίνουν σε ένα τοπικό μέγιστο, το οποίο ορίζεται ως το κέντρο της μάζας του αντικειμένου *O* στο *n+1* πλαίσιο. Με τη συνέχιση αυτής της διαδικασίας καθορίζεται η τροχιά του αντικειμένου *O* στην ακολουθία των εικόνων που απαρτίζουν το βίντεο. Εάν υπάρχουν διάφορα αντικείμενα στο βίντεο η διαδικασία ανίχνευσης πραγματοποιείται ξεχωριστά για κάθε αντικείμενο, ώστε να υπολογιστούν οι τροχιές τους. Οι σχετικές αποστάσεις μεταξύ αντικείμενων γνωστού background και κινούμενων αντικειμένων υπολογίζονται συνεχώς από τις τροχιές αυτών. Εδώ πρέπει να επισημανθεί ότι ένας μοναδικός αλγόριθμος, που να είναι αποδοτικός σε όλες τις πιθανές περιπτώσεις που μπορούν να αντιμετωπιστούν, δεν υπάρχει. Πρέπει να είναι ενσωματωμένος στον αλγόριθμο ένας μηχανισμός ανατροφοδότησης για διόρθωση λαθών όπως φαίνεται στο σχήμα 2. Παραδείγματος χάριν, στην ανωτέρω συζήτηση συνδυάζονται οι μέθοδοι ανίχνευσης κίνησης: αφαίρεσης υποβάθρου και μέσης μετατόπισης. Η επαναληπτική διαδικασία μέσης μετατόπισης ξεκινά όχι μόνο από την αρχική θέση της σταγόνας του προηγούμενου πλαισίου αλλά και από τις σταγόνες που καθορίζονται με τη μέθοδο ανίχνευσης κινήσεων.

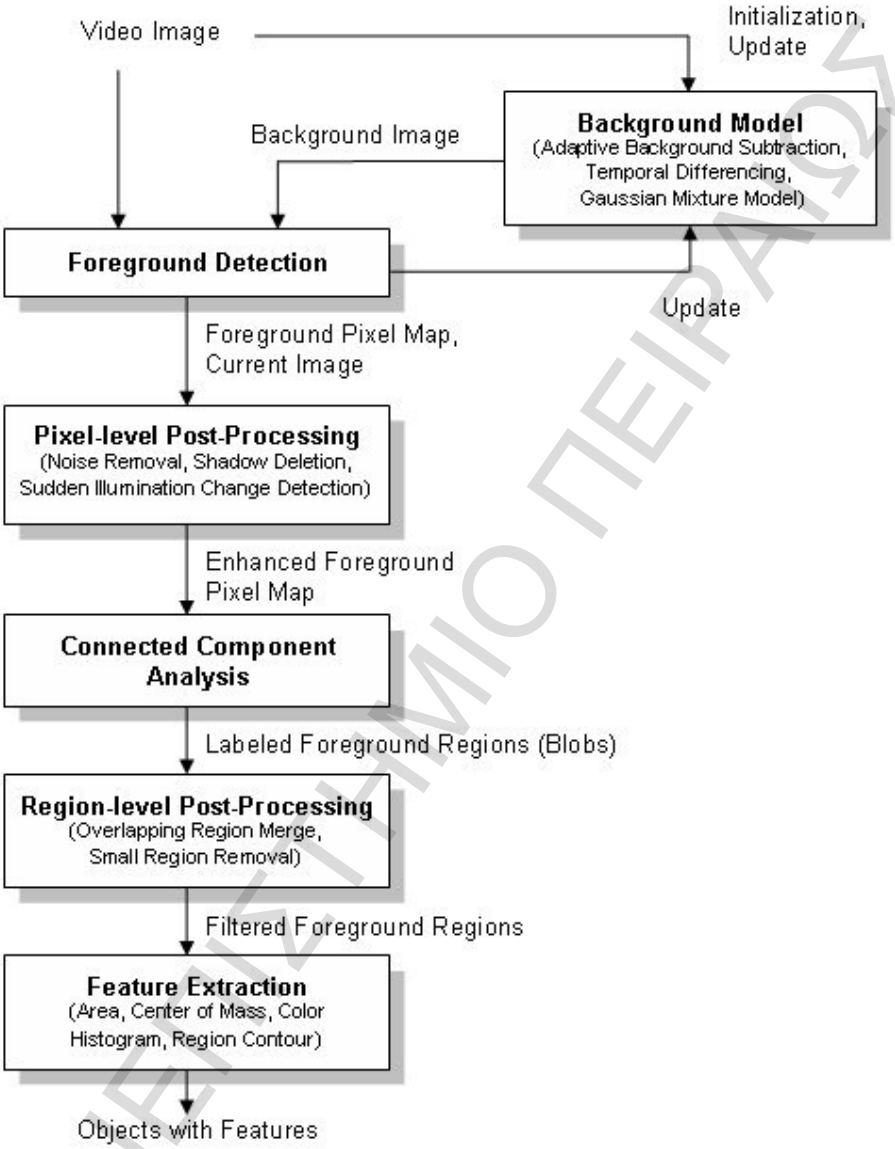


Σχήμα 2.2 *Διάγραμμα ενός αλγορίθμου εντοπισμού αντικειμένου*

2.2 Ταξινόμηση Αντικειμένων

Το επόμενο βήμα είναι να καθοριστεί *τί είναι* (σε τί αντιστοιχούν) οι υπολογισμένες κινούμενες σταγόνες. Συνήθως, το πρόβλημα τίθεται σαν πρόβλημα ταξινόμησης. Παραδείγματος χάριν, το βίντεο που καταγράφεται από μια μηχανή επιτήρησης εσωτερικού χώρου μπορεί να περιέχει μεμονωμένα άτομα, ομάδες ανθρώπων, ακόμα και κατοικίδια ζώα και ποντίκια σε μερικές περιπτώσεις. Ενώ το βίντεο μιας μηχανής επιτήρησης υπαίθριου χώρου μπορεί να περιλαμβάνει πεζούς, οχήματα, πουλιά, σύννεφα, και άλλα ζώα. Για να εξαχθούν σημασιολογικές πληροφορίες είναι απαραίτητο να διακριθούν οι άνθρωποι από τα άλλα κινούμενα αντικείμενα και να ανιχνευθεί η κίνησή τους. Τα κινούμενα αντικείμενα ταξινομούνται σύμφωνα με *τη μορφή, το χρώμα* και *την κίνηση.* Διάγραμμα ενός χαρακτηριστικού αλγορίθμου ανίχνευσης αντικειμένου παρουσιάζεται στο σχήμα 2.2. Η μορφή των κινούμενων περιοχών μπορεί να χαρακτηριστεί με διάφορους τρόπους, π.χ ως δισδιάστατο (2-D) περίγραμμα, σκιαγραφία,σκελετός, και *πληροφορίες χωρίων σταγόνων* (blob box information). Οι σταγόνες ενός κινούμενου αντικειμένου είναι ταξινομημένες σε τέσσερις κατηγορίες που αντιπροσωπεύουν απομονωμένα ανθρώπινα όντα, οχήματα, ομάδες ανθρώπων και σωρούς, χρησιμοποιόντας ως ταξινομητή ένα *νευρωνικό δίκτυο (neural network)* τριών στρωμάτων. Η ταξινόμηση πραγματοποιείται χρησιμοποιώντας ένα σύνολο χαρακτηριστικών γνωρισμάτων συμπεριλαμβανομένου ενός μίγματος image-based και scene-based παραμέτρων αντικειμένου όπως η κίνηση των διασκορπισμένων σταγόνων, περιοχή σταγόνων, λόγος

διάστασης του οριοθετημένου κιβωτίου, και οι πληροφορίες που παρέχονται μέσω λειτουργίας μεγέθυνσης (zoom) της μηχανής.



Σχήμα 2.3 *Διαδικασία εξαγωγής χαρακτηριστικών*

Τα κινούμενα αντικείμενα μοντελοποιούνται χρησιμοποιώντας καμπύλες, τις αποκαλούμενες *κορδέλλες* (ribbons), δισδιάστατα ελλειψοειδή κ.λπ. Οι παράμετροι υπολογίζονται από τις σταγόνες, και διάφορες μέθοδοι ταξινόμησης χρησιμοποιούνται για την αναγνώριση της κινούμενης σταγόνας. Η ύπαρξη χρωμάτων ανθρώπινου προσώπου και χεριών στο ιστόγραμμα χρώματος της κινούμενης σταγόνας προσδίδει αξιοπιστία τη διαδικασία απόφασης. Η ανίχνευση χρώματος ανθρώπινου δέρματος μπορεί να πραγματοποιηθεί σε διάφορα διαστήματα χρωμάτων συμπεριλαμβανομένου του κόκκινου, πράσινου, και μπλε (RGB) διαστήματος, του διαστήματος φωτεινότητας και διαφοράς

χρώματος ή chrominance (YUV), και διαστήματα χρώματος κορεσμού (saturation) (HSV). Πειραματικά διαπιστώνεται ότι το απλό διάστημα χρώματος YUV είναι το ίδιο αποτελεσματικό με οποιαδήποτε άλλη μέθοδο αναπαράστασης χρώματος για το ανθρώπινο δέρμα.

Η χρήση των χρονικών πληροφοριών για την ταξινόμηση αντικειμένου έχει μελετηθεί από πολλούς ερευνητές επίσης. Στην πραγματικότητα, το σχέδιο ταξινόμησης που περιγράφεται από κάποιους από αυτούς επαναλαμβάνεται μέσα σε διάφορα πλαίσια για να φθάσει σε μια τελική απόφαση με τη βοήθεια ενός αλγορίθμου ανίχνευσης. Οι περιπλοκότερες μέθοδοι εκμεταλλεύονται την περιοδική συμπεριφορά που εκτίθεται από την εύκαμπτη αρθρωμένη ανθρώπινη κίνηση.Κάποιος μπορεί επίσης να εκμεταλλευθεί τις προγενέστερες πληροφορίες για τη σκηνή που ελέγχεται από μια μηχανή επιτήρησης. Για παράδειγμα, η κίνηση των πεζών είναι πολύ πιό αργή από των οχημάτων σε μια περίπτωση ελέγχου κυκλοφορίας. Οι έννοιες συγχώνευσης δεδομένων (data fusion) μπορούν να χρησιμοποιηθούν για να συνδυάσουν τις πληροφορίες μορφής, χρώματος και κινήσεων για την ταξινόμηση κίνησης αντικειμένου. Υπάρχουν καλά αποτελέσματα αναγνώρισης στα ελεγχόμενα περιβάλλοντα εντούτοις οι φυσικές περιοχές είναι ακόμα ένα άλυτο πρόβλημα αναγνώρισης. Το πρόβλημα γίνεται ακόμη πιο δύσκολο εάν υπάρχει οπτικός περιορισμός π.χ., ένα άτομο που κάθεται πίσω από ένα τραπέζι. Η ανίχνευση κίνησης αντικειμένων μπορεί να παρέχει κάποια μερική λύση στο πρόβλημα αυτό.

Είναι πιθανώς αδύνατο να λυθεί ολοκληρωτικά το πρόβλημα αλλά μια μέθοδος που αποτελείται από τη συνδυασμένη χρήση διάφορων αλγορίθμων μπορεί να δώσει μια λύση.

2.3 Ανίχνευση και Ταξινόμηση Προτύπων

Η μοντελοποίηση προτύπων ανθρώπινου σώματος και οι σχετικές μέθοδοι ανίχνευσης ανθρώπινων σωμάτων περιλαμβάνουν:

- Μεθόδους αναπαράστασης καμπυλών
- Μοντελοποίηση δισδιάστατου περιγράμματος (2-D contour)
- Τρισδιάστατα ογκομετρικά πρότυπα (3-D volumetric models)

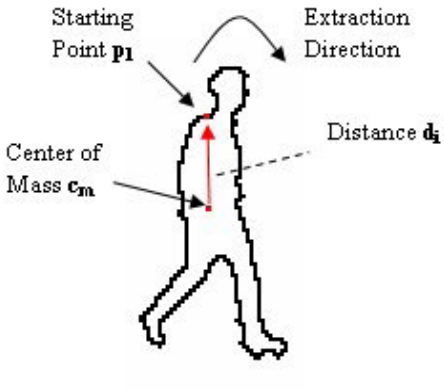
Οι οποίες αντιπροσωπεύουν τη γεωμετρική δομή του ανθρώπινου σώματος με διάφορους μαθηματικούς τρόπους. Η αναπαράσταση καμπυλών (stick-figure) ενός ανθρώπινου σώματος είναι στην πραγματικότητα ένας συνδυασμός τμημάτων γραμμών που συνδέονται με ενώσεις. Μια τέτοια αναπαράσταση είναι κατάλληλη για να μοντελοποιήσει τις μετακινήσεις του κεφαλιού, κορμού και άκρων. Οι διάφορες ανθρώπινες κινήσεις που αντιπροσωπεύονται από μια ακολουθία μετακινήσεων καμπυλών διαμορφώνονται μέσω των μοντέλων ταξινόμησης Hidden Markov (HMMs). Αυτή η προσέγγιση μπορεί να χρησιμοποιηθεί όχι μόνο για αναγνώριση της ύπαρξης ανθρώπων σε βίντεο αλλά και για να ταξινομήσει τις ανθρώπινες κινήσεις. Ένα μοντέλο Hidden Markov που αντιπροσωπεύει κάθε χαρακτηριστική ανθρώπινη κίνηση και στάση μπορεί να είναι σχεδιασμένο και εκπαιδευμένο a priori χρησιμοποιώντας τυπικές ακολουθίες βίντεο. Κατά τη διάρκεια της φάσης αναγνώρισης η κίνηση του ανθρώπινου σώματος αναγνωρίζεται σύμφωνα με το HMM που παράγει τη μεγαλύτερη πιθανότητα.

Η αδυναμία αυτής της προσέγγισης είναι η υπερβολικά απλουστευμένη αντιπροσώπευση καμπυλών του ανθρώπινου σώματος. Τα λάθη ταξινόμησης οφείλονται κυρίως στο απλουστευμένο πρότυπο. Τα περιπλοκότερα πρότυπα περιλαμβάνουν όχι μόνο απλές καμπύλες αλλά και δισδιάστατες «κορδέλες» (2-D ribbons) για να επιτύχει καλύτερα πρότυπα σχηματισμού της κίνησης των ανθρώπινων άκρων. Προσεγγίσεις που χρησιμοποιούν δισδιάστατα περιγράμματα περιλαμβάνουν τα λεγόμενα *περιγράμματα*

φιδιού (snakes) ή τα *ενεργά πρότυπα περιγράμματος* (active contour) και διάφορες λειτουργίες 1-D που αντιπροσωπεύουν τα 2-D περιγράμματα των κινούμενων σταγόνων. Τεχνικές ανίχνευσης βασισμένες στο ενεργό περίγραμμα και τη *διαμόρφωση ορίων αντικειμένου* έχουν μελετηθεί ιδιαίτερα στα πλαίσια της κωδικοποίησης video. Αναπτύχθηκαν επίσης προσαρμοστικά πρότυπα περιγράμματος, τα οποία ενημερώνονται δυναμικά για κάθε πλαίσιο -video frame. Λαμβάνοντας υπόψη μια ακολουθία εικόνας που περιέχει ένα κινούμενο αντικείμενο, αρκεί η κατασκευή ενός *φιδιού* γύρω από το περίγραμμα του κινούμενου αντικειμένου, το οποίο μπορεί να εξαχθεί με τη χρησιμοποίηση ενός αλγορίθμου αφαίρεσης υποβάθρου, ή συνόλων επιπέδων (level sets). Για κάθε νέα εικόνα, το *φίδι* τοποθετείται σε αυτήν, στην ίδια θέση που κατείχε στην προηγούμενη εικόνα. Το *φίδι* έπειτα εξάγει το περίγραμμα του αντικειμένου στη νέα εικόνα, εάν η ταχύτητα του αντικειμένου είναι αργή έτσι ώστε να υπάρχει ένα λογικό ποσό επικάλυψης στα pixels του αντικειμένου μεταξύ των δύο διαδοχικών πλαισίων εικόνας. Ένας αλγόριθμος πρόβλεψης, όπως το φίλτρο Kalman μπορεί να χρησιμοποιηθεί για να υπολογίσει την επόμενη θέση του *φιδιού* σε μια τηλεοπτική ακολουθία. Αυτό καθιστά τον *ανιχνευτή-ιχνηλάτη* πιό αξιόπιστο και ικανό να εντοπίσει αντικείμενα που κινούνται με υψηλή ταχύτητα.

Χρησιμοποιούνται επίσης νευρωνικά δίκτυα για την ταξινόμηση των ανιχνευμένων κινούμενων αντικειμένων ως άνθρωπο ή όχι. Ένα νευρωνικό δίκτυο μπορεί να εκπαιδευθεί χρησιμοποιώντας τα αποκαλούμενα snaxels (snake pixels) του ενεργού-δυναμικού περιγράμματος. Λόγω του σχετικά μικρού μεγέθους του διανύσματος χαρακτηριστικών που αποτελείται από τα snaxels, το υπολογιστικό φορτίο ταξινόμησης είναι χαμηλό. Εντούτοις, η εκτίμηση ορίων μέσω ενεργού περιγράμματος είναι μια επαναληπτική διαδικασία, και ως εκ τούτου είναι δαπανηρή υπολογιστικά. Η κύρια αρνητική πτυχή του ενεργού περιγράμματος, ή γενικά των αλγορίθμων ταξινόμησης μέσω των ορίων του αντικειμένου, είναι το πρόβλημα της παρεμβολής (occlusion). Εάν η ύπαρξη ενός ανθρώπου παρεμβάλλεται τότε αυτές οι μέθοδοι οδηγούν σε ανακριβή αποτελέσματα.

Νευρωνικά δίκτυα, HMMs και άλλοι μηχανισμοί ταξινόμησης μπορούν να εκπαιδευθούν επίσης χρησιμοποιώντας τις συναρτήσεις 1-D εξαγόμενες από 2-D περιγράμματα. Παραδείγματος χάριν, μια συνάρτηση μπορεί να κατασκευαστεί με τον υπολογισμό της απόστασης: των ορίων του αντικειμένου από το κέντρο μάζας του αντικειμένου, σε διάφορες γωνίες, όπως φαίνεται στο σχήμα 4. Μια άλλη μονοδιάστατη συνάρτηση μπορεί να κατασκευαστεί με την προβολή του κινούμενου αντικειμένου επάνω στους οριζόντιους και κάθετους άξονες και τη σύνδεση των προβολών αυτών.



Σχήμα 2.4-α: *Μια συνάρτηση μπορεί να κατασκευαστεί με τον υπολογισμό της απόστασης των ορίων του αντικειμένου από το κέντρο μάζας του αντικειμένου σε διάφορες γωνίες*

Σχήμα 2.4-β: *Τα όρια ενός κινούμενου ατόμου και παραγόμενη συνάρτηση*

Τρισδιάστατα *ογκομετρικά πρότυπα (3-D volumetric models*) συμπεριλαμβανομένων των ελλειπτικών κυλίνδρων, κώνων, σφαιρών και άλλα περιπλοκότερα πρότυπα χρησιμοποιούνται για να μοντελοποιήσουν το ανθρώπινο σώμα. Η έρευνα σε αυτόν τον τομέα είναι βασισμένη σε μέθοδους μοντελοποίησης ανθρώπινου σώματος που αναπτύσσονται στον τομέα των ηλεκτρονικών γραφικών (computer graphics). Εάν είναι επίσης διαθέσιμο ένα τρισδιάστατο πρότυπο του εκάστοτε σκηνικού, τότε τα τρισδιάστατα ογκομετρικά πρότυπα παρέχουν μια ακριβή περιγραφή κινούμενων αντικειμένων και ανθρώπων. Προκειμένου να υπολογιστούν οι τρισδιάστατες παράμετροι θα ήταν επιθυμητό να υπάρχουν και άλλες μηχανές που να επιβλέπουν το ίδιο σκηνικό. Προφανώς, τα τρισδιάστατα πρότυπα απαιτούν την εκτίμηση, και ενημέρωση κατά τη διάρκεια της διαδικασίας ανίχνευσης, περισσότερων παραμέτρων κι αυτό οδηγεί σε υπολογιστικά κοστοβόρες μεθόδους, συγκριτικά με άλλες μοντελοποιήσεις ανθρώπινων προτύπων, κατά τη διάρκεια της διαδικασίας ταξινόμησης. Γενικά, ο αριθμός παραμέτρων που αντιπροσωπεύει τις συνδέσεις του ανθρώπινου σώματος ελαχιστοποιείται για να μειώσει την υπολογιστική πολυπλοκότητα. Στην περίπτωση των παραμορφώσιμων συνδέσεων, οι παράμετροι παραμόρφωσης μειώνονται με τον περιορισμό των παραμορφώσεων σε γραμμικού τύπου.

Υπάρχουν πολλές έρευνες για την εύρεση των κατάλληλων προτύπων για σκοπούς αναγνώρισης αντικειμένων, ειδικά για την αναγνώριση και ανίχνευση ανθρώπων σε εικόνες και βίντεο. Ο Biederman [80] προτείνει παραμετροποιημένα πρότυπα μερών, τα οποία αντιλαμβάνονται τη δομή των αντικειμένων μέσα από μερικές παραμέτρους. Προτείνει ότι η αποκατάσταση των ρυθμίσεων μερικών σημαντικών πρωταρχικών συστατικών οδηγεί σε ταχύτερη αναγνώριση του αντικειμένου. Αυτά τα πρότυπα είναι υπολογιστικά αποδοτικά λόγω του ότι η χρησιμοποίηση ενός μικρού αριθμού παραμέτρων απλοποιεί το πρόβλημα της αναζήτησης σε μια μικρή βάση δεδομένων των αποθηκευμένων προς ταίριασμα προτύπων. Οι Terzopoulos, Witkin, και Kass [81] πρότειναν έναν παραμορφώσιμο κύλινδρο για την ανάκτηση τρισδιάστατης μορφής και κίνησης. Η εργασία τους εκμεταλλεύεται την υπολογιστική φυσική στη διαδικασία αναγνώρισης. Εντούτοις, ένα πρότυπο με μικρό αριθμό παραμέτρων δεν αποδείχθηκε αρκετό για να παρέχει μια αφαιρετική αναπαράσταση μορφής αντικειμένου.

Οι Gupta και Bacjy [82] χρησιμοποιούν ένα *υπερτετραγωνικό ελλειψοειδές (superquadric ellipsoid)* με παραμετροποιημένες παραμορφώσεις για την περιγραφή και κατάτμηση μερών ενός αντικειμένου. Ο Pentland και άλλοι [83, 84] πρότειναν μια μέθοδο βασισμένη στη φυσική για τη συναρμολόγηση παραμορφώσιμων μερών προτύπων μέσω superquadric ellipsoids. Χρησιμοποίησαν τροπική ανάλυση για να μειώσουν υπολογιστικό κόστος της διαδικασίας συναρμολόγησης. Ελαχιστοποίησαν τα τμήματα υψηλής συχνότητας που δεν έχουν καμία σημαντική επίδραση. Οι Μεταξάς και Τερζόπουλος συνδύασαν τις καλύτερες πτυχές των μεθόδων Τερζόπουλου, Witkin, Kass και Pentland για την αναγνώριση προτύπων που αποτελούνται από παραμορφώσιμα μέρη. Συνδύασαν επίσης δυναμικές παραμόρφωσης άκαμπτων σωμάτων. Δεδομένου ότι χρησιμοποιούν σφαιρικές παραμορφώσεις που καθορίζονται από μη γραμμικές εξισώσεις παραμέτρων, οι μέθοδοί τους είναι γενικότερες.

2.4 Εντοπισμός Ανθρώπινου Προσώπου και Σώματος

Ο εντοπισμός ανθρώπινου προσώπου και σώματος μπορεί να πραγματοποιηθεί όχι μόνο σε βίντεο αλλά και σε εικόνες. Σε αυτήν την περίπτωση δεν υπάρχουν πληροφορίες κινήσεων που να περιορίζουν το διάστημα αναζήτησης. Αυτό το πρόβλημα είναι σημαντικός σε εφαρμογές όπως η αναγνώριση προσώπου για εφαρμογές ασφάλειας, εκτίμηση συναισθημάτων, διαχείριση βάσεων δεδομένων εικόνων προσώπου, και σε μερικές ακόμα εφαρμογές ασφάλειας. Το πρόβλημα του εντοπισμού ανθρώπινου προσώπου και σώματος επιβαρύνεται επίσης από αλλαγές φωτισμού, ακατάστατο φωτισμό υποβάθρου, μερική παρεμβολή, παραλλαγές κλίμακας και θέσεων. Ένα πλάνο εντοπισμού ανθρώπων πρέπει να αντιμετωπίζει τις παραλλαγές κλίμακας, προσανατολισμού, θέσης και φωτισμού. Στα υπάρχοντα συστήματα, μερικές από αυτές τις παραλλαγές υποτίθεται ότι καθορίζονται με τρόπο ώστε να μειώσουν την πολυπλοκότητα του προβλήματος. Παρ'όλα αυτά το πρόβλημα παραμένει ουσιαστικά άλυτο. Αν και επιβαρύνεται από αλλαγές φωτισμού, η ανίχνευση χρώματος δέρματος είναι ένα σημαντικό σημείο έναρξης ή τουλάχιστον ένα βασικό στοιχείο σε πολλές μεθόδους εντοπισμού ανθρώπων σε εικόνες. Διάφορα διαστήματα χρωμάτων έχουν χρησιμοποιηθεί για τη μοντελοποίηση χρώματος δέρματος. Αλλά οι πρόσφατες μελέτες δείχνουν το ευρέως χρησιμοποιημένο διάστημα Y,U και V (φωτεινότητας, χρωμιότητας ή διαφορά χρώματος) δεν υστερεί συγκριτικά με άλλες περίπλοκες μεθόδους αναπαράστασης χρώματος.

Ο αποτελεσματικότερος τρόπος να αντιμετωπιστούν οι παραλλαγές στην κλίμακα είναι να χρησιμοποιηθεί ένα *multiresolution wavelet* ή μια *πυραμιδική* (pyramidal) στρατηγική ανάλυσης εικόνας. Με την αναπαράσταση της εικόνας εισαγωγής μέσω διαφόρων αναλύσεων, είναι δυνατά να έχουμε διαθέσιμο ένα ανθρώπινο πρόσωπο ή σώμα

σε διαφορετικά μεγέθη έτσι ώστε ένας αντιστοιχος αλγόριθμος- φίλτρο εντοπισμού αντικειμένων μπορεί να το εντοπίσει, μέσω εφαρμογής-τρεξίματός του σε όλες τις υπο-εικόνες πολλαπλής ανάλυσης (multiresolution subimages) [87-91]. Διαφορτικά, τα χαρακτηριστικά του αντικειμένου μπορούν να εκπαιδευθούν σε μια σταθερή κλίμακα με την προσδοκία ότι μία από τις υπο-εικόνες που δημιουργούνται από την αποσύνθεση πολλαπλής ανάλυσης έχει ένα παρόμοιου μεγέθους αντικείμενο.

Άλλη σχετικά πρόσφατη προσέγγιση είναι να χρησιμοποιηθούν τα φίλτρα κυμάτων Haar στις διάφορες κλίμακες ως *αντιστοιχημένα φίλτρα (matched filters)* σε μια εικόνα και να εξάγουν μερικές παραμέτρους χαρακτηριστικών για να τροφοδοτηθεί μια μηχανή ταξινόμησης προτύπων. Αυτή η προσέγγιση είναι επίσης αξιόπιστη σε παραλλαγές προσανατολισμού προσώπου. Η εφαρμογή διάφορων φίλτρων ή προτύπων στις διάφορες αναλύσεις σε μια εικόνα είναι μια συνδυασμένη εκδοχή της λήψη των υπο-εικόνων πολλαπλής ανάλυσης από μια δεδομένη εικόνα και της επεξεργασίας αυτών με κάποιο σταθερό φίλτρο. Αυτές οι δύο προσεγγίσεις είναι ισοδύναμες από την άποψη της εφαρμοσμένης μηχανικής συστημάτων. Στην πραγματικότητα μια παρόμοια προσέγγιση μπορεί επίσης να χρησιμοποιηθεί για να αντιμετωπίσει τις παραλλαγές στον προσανατολισμό προσώπου. Μπορούν να χρησιμοποιηθούν φίλτρα αντιστοιχημένα σε πρότυπα, ανεξάρτητα κλίμακας και προσανατολισμού [92], [93]. Μια άλλη προσέγγιση είναι η ανίχνευση των χαρακτηριστικών γνωρισμάτων του προσώπου και η εφαρμογή ορισμένων συγγενικών γεωμετρικών μετασχηματισμών για να μετατραπούν σε ένα σταθερό διάστημα [94], [94]. Το πρόβλημα με αυτήν την προσέγγιση είναι ότι το στάδιο εξαγωγής χαρακτηριστικών γνωρισμάτων μπορεί να μην παράγει αξιόπιστα αποτελέσματα ή να εντοπίσει υπερβολικά πολλές υποψήφιες περιοχές.

Πολλαπλές σταθερές εικόνες μπορούν να χρησιμοποιηθούν επίσης για να λύσουν τις παραλλαγές θέσης και κλίμακας. Αλλά αυτό απαιτεί τη χρήση των πολλαπλών φωτογραφικών μηχανών (ή video). Είναι μάλλον αδύνατο να βρεθεί ένας αλγόριθμος που να δουλεύει τέλεια για όλες τις πιθανές περιπτώσεις. Ωστόσο, τα αποτελέσματα διάφορων αλγορίθμων μπορούν να συνδυαστούν με στόχο την επίτευξη μιας αξιόπιστης μεθόδου εντοπισμού ανθρώπινου προσώπου και σώματος ή μερών σώματος, σε εικόνες ή video.

3. Συνδυασμός Μεθόδων Εντοπισμού Ανθρώπων σε Video

3.1 Εισαγωγή

Ο εντοπισμός ανθρώπων καθώς και η ανίχνευση της τροχιάς αυτών σε video ακολουθίες εικόνων, και μάλιστα σε μή ελεγχόμενα περιβάλλοντα, είναι αδύνατο να επιτευχθεί με την εφαρμογή ενός μοναδικού αλγορίθμου. Γνωρίζοντας αυτό καλά, οι ερευνητές στράφηκαν σε συνδυασμένες μεθόδους εντοπισμού, που να μπορούν ανταποκριθούν στις απαιτήσεις των φυσικών καταγραφών video, μή ελεγχόμενων και ταυτόχρονα απρόβλεπτων καταστάσεων και περιβάλλοντος. Στο σημείο αυτό παρουσιάζεται μια προσέγγιση για αυτόματο εντοπισμό και ανίχνευση πορείας πολλαπλών, και ενδεχομένως μερικώς παρεμβαλλόμενων άνθρωπων (είτε από άλλους ανθρώπους είτε από αντικείμενα) σε περπάτημα ή στάση, από μια και μοναδική συσκευή, η οποία μπορεί να είναι στάσιμη ή και κινούμενη. Ένα ανθρώπινο σώμα αναπαριστάται ως σύνθεση μερών σώματος. Οι ανιχνευτές μερών εκπαιδεύονται με την ώθηση διάφορων αδύναμων ταξινομητών που χρησιμοποιούν τα χαρακτηριστικά ενός *edgelet* (περίγραμμα ακμών). Τα αποτελέσματα των ανιχνευτών μερών συνδυάζονται για να διαμορφώσουν ένα κοινό πιθανοτικό μοντέλο που περιλαμβάνει μια ανάλυση των πιθανών occlusions (παρεμβολών). Οι συνδυασμένες μέθοδοι εντοπισμού και εντοπισμού μερών παρέχουν παρατηρήσεις που χρησιμοποιούνται για την ανίχνευση πορείας. Η πορεία ενός αντικειμένου ανιχνεύεται μέσω συσχετισμένων δεδομένων και μεθόδους *μέσης μετατόπισης (meanshift*). Η προσέγγιση αυτή, μπορεί να ανιχνεύει την τροχιά της κίνησης ενός ατόμου, ακόμα και σε περιπτώσεις όπου παρεμβάλλεται το σκηνικό ή αντικείμενα, είτε σε στατικό είτε σε δυναμικό υπόβαθρο. Επίσης, γίνεται σύγκριση με παλαιότερες μεθόδους.

3.2 Βασικές Έννοιες της Μεθόδου

Η μέθοδος στηρίζεται σε αναπαράσταση των μερών του σώματος. Τα πλεονεκτήματά της είναι:

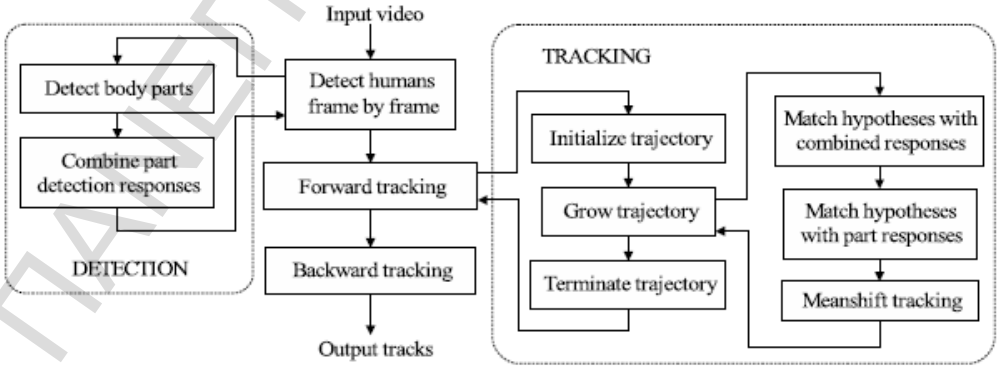
- Μπορεί να αντιμετωπίσει περιπτώσεις μερικής παρεμβολής, π.χ., όταν δεν φαίνονται τα πόδια, το άτομο μπορεί να εντοπιστεί από το άνω μέρος του σώματος.
- Η τελική απόφαση εντοπισμού στηρίζεται σε πολλαπλές ενδείξεις, οπότε ελαχιστοποιούνται οι εσφαλμένοι εντοπισμοί
- Παρουσιάζει ανθεκτικότητα σε αλλαγές οπτικών γωνιών και στη μεταβλητότητα θέσης και στάσης των σχηματιζόμενων αντικειμένων.

Στο *σχήμα 3.1* βλέπουμε το γενικό διάγραμμα της μεθόδου. Ο εντοπισμός γίνεται *πλαίσιο προς πλαίσιο* (frame by frame detection) και εκτελείται σε δύο στάδια: εντοπισμός μερών, και στη συνέχεια σχεδιασμός τους. Ενώ η παρακολούθηση της πορείας από τα εξής τρία: έναρξη, ανάπτυξη και τερματισμός τροχιάς. Στο πρώτο στάδιο εντοπισμού, χρησιμοποιείται ένα εκτεταμένο σύνολο χαρακτηριστικών περιγράμματος σώματος που ονομάζονται χαρακτηριστικά *edgelet*. Τα χαρακτηριστικά αυτά είναι εύχρηστα στον εντοπισμό ανθρώπων, καθώς μένουν σχετικά αμετάβλητα σε διαφορές ρουχισμού, αντίθετα με χαρακτηριστικά έγχρωμης ή ασπρόμαυρης εικόνας που χρησιμοποιούνται για τον εντοπισμό προσώπων. Οι εντοπιστές μερών δομημένοι, κατά δέντρα, εκπαιδεύονται

μέσω ενός αλγόριθμου που πρότεινε ο Huang (2004, 2005), η οποία αποτελεί εξέλιξη της προσέγγισης των Viola και Jones (2001).

Κατά το δεύτερο στάδιο εντοπισμού συνδυάζονται τα αποτελέσματα εντοπισμού των διαφόρων μερών. Καθορίζεται μια συνάρτηση πιθανοφάνειας κοινής εικόνας για πολλαπλούς ανθρώπους, ενδεχομένως που να παρεμβάλλονται μεταξύ τους. Το πρόβλημα εντοπισμού πολλαπλών ατόμων διατυπώνεται ως πρόβλημα εκτίμησης MAP και ψάχνουμε διάστημα λύσης για να βρεθεί η καλύτερη ερμηνεία της παρατηρηθείσας εικόνας. Η απόδοση του συνδυασμένου εντοπιστή είναι καλύτερη από αυτήν οποιουδήποτε μεμονωμένου εντοπιστή μερών μέσα σε όρους μείωσης του ποσοστού λανθασμένων εντοπισμών. Εντούτοις ο συνδυασμένος εντοπιστής είναι λειτουργικός μόνο για περιπτώσεις παραμβολής μεταξύ αντικειμένων, ενώ οι ανιχνευτές μερών μπορούν να λειτουργήσουν και παρουσία παρεμβολών σκηνής. Οι παλιότερες προσεγγίσεις, όπως των Mohan, 2001; Mikolajczyk, 2004; Shashua, 2004, θεωρούσαν τα άτομα ανεξάρτητα μεταξύ τους και δεν προσομοίωναν μεταξύ τους παρεμβολές.

Η μέθοδος ανίχνευσης πορείας βασίζεται στην παρακολούθηση μερών του ανθρώπινου σώματος. Τα αποτελέσματα των εντοπιστών μερών και του συνδυασμένου εντοπιστή, λαμβάνονται ως δεδομένα εισαγωγής για τον ανιχνευτή πορείας. Οι πορείες των ανθρώπουν ανιχνεύονται μέσω συσχετισμού δεδομένων, όπως ταίριασμα των υποθέσεων αντικειμένου με τις αποκρίσεις εντοπισμού, όποτε αντίστοιχες αποκρίσεις είναι διαθέσιμες. Ταιριάζουμε τις υποθέσεις με τα αποτελέσματα του συνδυασμένου εντοπιστήπρώτα, δεδομένου ότι είναι πιό αξιόπιστος από αυτούς των μεμονωμένων μερών. Εάν για μια υπόθεση δεν βρίσκεται καμία συνδυασμένη απόκριση με παρόμοια εμφάνιση και κοντά στην προβλεφθείσα θέση, κατόπιν προσπαθούμε να τη συνδέσουμε με τις αποκρίσεις εντοπιστής μερών. Εάν αυτό αποτύχει πάλι, ένας «ιχνηλάτης» meanshift (Comaniciu, 2001) χρησιμοποιείται για να ακολουθήσει την πορεία του αντικειμένου. Τις περισσότερες φορές τα αντικείμενα παρακολουθούνται επιτυχώς από την ένωση στοιχείων. Ο «ιχνηλάτης» meanshift παίρνει θέση μόνο περιστασιακά και για μικρές χρονικές περιόδους. Επίσης, δεδομένου ότι τα τεκμήρια για ανίχνευσης πορείας είναι ισχυρά, δε χρησιμοποιούνται στατιστικές τεχνικές δειγματοληψίας όπως σε μερικές από τις παλιότερες εργασίες (Isard, MacCormick, 2001; Zhaoand Nevatia, 2004; Smith, 2005). Η έναρξη τροχιάς γίνεται όταν τα στοιχεία από τις νέες παρατηρήσεις δεν μπορούν εξηγηθούν από τις τρέχουσες υποθέσεις, όπως επίσης σε πολλές προηγούμενες μεθόδους (Davis, 2000; Isard, Mac-Cormick, 2001; Zhao, Nevatia, 2004a; Smith, 2005; Peter, 2005). Ομοίως, μια τροχιά ολοκληρώνεται όταν χάνεται από τους ανιχνευτές για μια ορισμένη περίοδο.



Σχήμα 3.1: *Το γενικό διάγραμμα του συστήματος εντοπισμού και ανίχνευσης τροχιάς ανθρώπων*

Σχήμα 2. *Παραδείγματα αποτελεσμάτων ανίχνευσης του συστήματος των Wu, Nevatia*

3.3 Εντοπισμός Μερών του Ανθρώπινου Σώματος

Ο εντοπισμός ανθρώπων γίνεται με συνδυασμό των αποτελεσμάτων ενός συνόλου από εντοπιστές μερών σωμάτων που εκπαιδεύονται από τα χαρακτηριστικά τοπικής μορφής.

3.3.1 Χαρακτηριστικά Περιγράμματος Ακμών (edgelet)

Με βάση την παρατήρηση ότι οι σκιαγραφίες είναι από τα περισσότερο εμφανή πρότυπα των ανθρώπων, δημιουργήθηκε μια νέα κατηγορία χαρακτηριστικών τοπικής μορφής που αποκαλούμε χαρακτηριστικά περιγράμματος ακμών, ή *edgelet features*. Ένα edgelet είναι ένα μικρό τμήμα μιας γραμμής ή καμπύλης. Έστω ότι οι θέσεις και τα κανονικά διανύσματα των σημείων ενός edgelet, *E*, δίνονται από τα $\{u_i\}_{i=1}^{k}$ και $\{n_i^E\}_{i=1}^{k}$, όπου *k* το μήκος του edgelet, βλ. σχήμα 3. Δοσμένης μιας εικόνας εισαγωγής *I* , έχουμε $M^I(p)$ και $n^I(p)$, ένταση ακμών και κανονικό διάνυσμα αντίστοιχα**,** στη θέση *p* του *I*. Η σχέση ανάμεσα στο edgelet *E* και την εικόνα *I* στη θέση *w* υπολογίζεται από τον τύπο:

$$f(E;I,w)=\frac{1}{k}\sum_{i=1}^{k}M^I(u_i+w)\left|\left\langle n^I(u_i+w),n_i^E\right\rangle\right| \qquad (1)$$

Στον παραπάνω τύπο, το u_i βρίσκεται μέσα στο πλαίσιο συντεταγμένων του υποπαραθύρου, και *w* είναι ο αντισταθμιστικός παράγων του υποπαραθύρου μέσα στο πλαίσιο-frame του video. Η συνάρτηση συσχέτισης του edgelet, συλλαμβάνει τόσο την ένταση όσο και πληροφορίεες για το σχήμα των ακμών (μπορεί αυτό να θεωρηθεί παραλλαγή της κλασσικής μεθόδου αντιστοίχησης Chamfer, Barrow, 1977).

Η ένταση ακμών $M^I(p)$ και το κανονικό διάνυσμα $n^I(p)$ υπολογίζονται μέσω συνελίξεων πυρήνα Sobel 3x3, που εφαρμόζονται σε ασπρόμαυρες (gray scale) εικόνες (βλ. σχήμα 3). Δεν χρησιμοποιούνται έγχρωμες πληροφορίες για τον εντοπισμό. Δεδομένου ότι χρησιμοποιούμε τα χαρακτηριστικά edgelet μόνο ως αδύναμα χαρακτηριστικά (weak features) σε έναν ενισχυτικό (boosting) αλγόριθμο, τα απλοποιούμε για καλύτερη υπολογιστική αποδοτικότητα. Πρώτα, ποσοτικοποιούμε τον προσανατολισμό (orientation quantization) του κανονικού διανύσματος σε έξι διακριτές τιμές, βλ. σχήμα 3. Το εύρος [0◦, 180◦) διαιρείται σε έξι ομοιόμορφα κατανεμημένα «δοχεία», τα οποία αντιστοιχούν σε

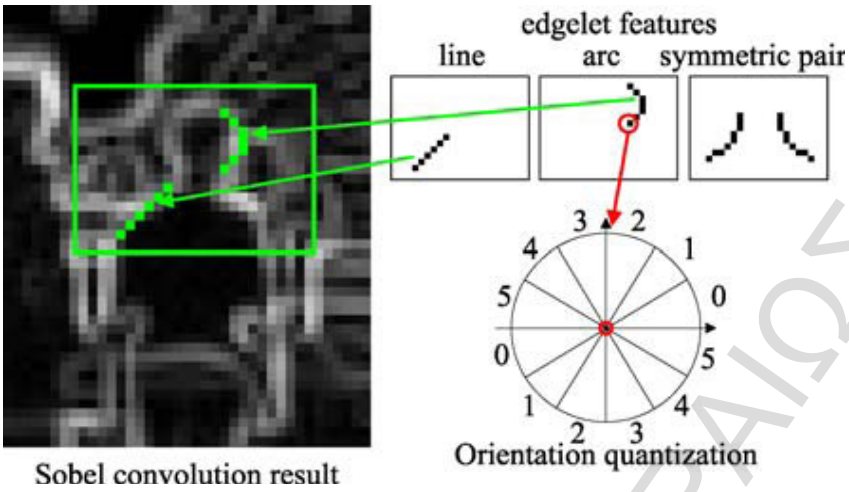
ακέραιους αριθμούς από *0* έως *5* αντίστοιχα. Μια γωνία ϑ μέσα στο εύρος [180◦, 360◦) έχει ποσοτικοποιημένη αξία ίδια με 360◦ − ϑ. Δεύτερον, το σημειακό γινόμενο δύο κανονικών διανυσμάτων υπολογίζεται από την ακόλουθη συνάρτηση:

$$l[x]=\begin{cases}1 & x=0\\ 4/5 & x=\pm1,\pm5\\ 1/2 & x=\pm2,\pm4\\ 0 & x=\pm3\end{cases} \qquad (2)$$

όπου η είσοδος *x* είναι η διαφορά μεταξύ δύο ποσοτικοποιημένων προσανατολισμών. Έχουμε επίσης, $\{V_i^E\}_{i=1}^k$ και $V^I(p)$ τους ποσοτικοποημένους προσανατολισμούς ακμών του edgelet και την εικόνα *I* της εισαγωγής αντίστοιχα. Η συνάρτηση απλοποιημένης συσχέτισης είναι:

$$\tilde{f}(E;I,w)=\frac{1}{k}\sum_{i=1}^{k}M^I(u_i+w)\cdot l\left[V^I(u_i+w)-V_i^E\right] \qquad (3)$$

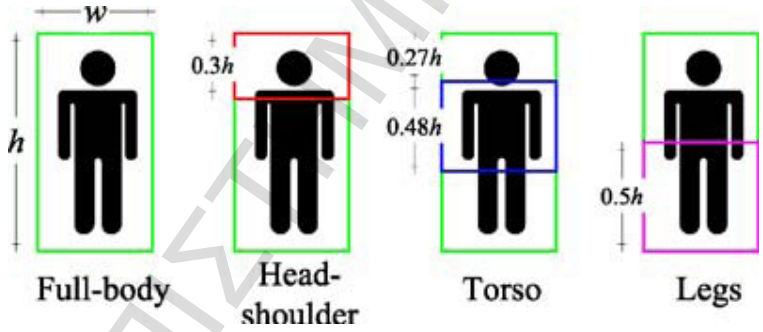
Όπως φαίνεται, ο υπολογισμός των χαρακτηριστικών του edgelet περιλαμβάνει μονάχα λειτουργίες short integer. Κάθε edgelet μπορεί να έχει μήκος από 4 μέχρι 12 pixels. Τα χαρακτηριστικά edgelet που χρησιμοποιούνται, αποτελούνται από (ή αναλύονται σε) μεμονομένα edgelet, τα οπία μπορεί να είναι γραμμές, 1/8, 1/4, 1/2 του κύκλου, ή συμμετρικά ζευγάρια αυτών. Ένα συμμετρικό ζευγάρι είναι η ένωση ενός μεμονομένου edgelet και του συμμετρικού του. Το σχήμα 3 επεξηγεί τον καθορισμό των χαρακτηριστικών γνωρισμάτων edgelet. Για ένα μέγεθος δειγμάτων 24 x 58, ο συνολικός αριθμός των πιθανών χαρακτηριστικών edgelet είναι 857.604.



Σχήμα 3.3. *Χαρακτηριστικά περιγράμματος ακμών (edgelet)*

3.3.2 Αδύναμοι Ταξινομητές και Edgelet

Τα μέρη ανθρώπινων σωμάτων που χρησιμοποιούνται σε αυτήν την εργασία είναι κεφάλι-ώμοι, κορμός, και πόδια. Εκτός από τους εντοπιστές των τριών αυτών μερών, γίνεται επίσης εκμάθηση ενός εντοπιστή πλήρους σώματος. Το σχήμα 4 παρουσιάζει τις χωρικές σχέσεις των μερών του σώματος.



Σχήμα 3.4: *Χωρικές σχέσεις των μερών του σώματος.*

Χρησιμοποιούμε μια ενισχυμένη έκδοση (Huang et Al, 2004) της αρχικής μεθόδου των Viola και Jones (2001) για να γίνει εκμάθηση στους εντοπιστές μερών. Έστω ότι οι τιμές των χαρακτηριστικών που υπολογίζονται από τον τύπο (3) είναι κανονικοποιημένες σε [0 , 1]. Διαιρούμε το διάστημα σε *ν* υποδιαστήματα:

$$bin_j = \left[\frac{j-1}{n}, \frac{j}{n} \right) \qquad (4)$$

Εδώ, το *n* = 16. Η ισομερής τμηματοποίηση του χώρου των χαρακτηριστικών αντιστοιχεί στην τμηματοποίηση του χώρου εικόνας. Για τον εντοπισμό αντικειμένου, ένα δείγμα αντιπροσωπεύεται ως *{x, y}* όπου το *I* είναι το κανονικοποιημένο χωρίο εικόνας και το *y* είναι η ετικέτα της κλάσης της οποίας η τιμή μπορεί να είναι + 1 (αντικείμενο) ή -1 (μη-αντικείμενο).

Σύμφωνα με τη real-valued έκδοση του αλγορίθμου AdaBoost (Schapire και Singer, 1999), ο αδύναμος ταξινομητής (weak classifier) $h^{(w)}$ είναι βασισμένος σε ένα χαρακτηριστικό edgelet *ε* και ορίζεται ως:

Αν: $\tilde{f}(E;x,O)\in bin_j$ **τότε:** $h^{(w)}(x)=\frac{1}{2}\ln\left(\frac{\overline{W}_{+1}^{j}+\varepsilon}{\overline{W}_{-1}^{j}+\varepsilon}\right)$ (5)

όπου *O* είναι η προέλευση του **x**, και *ε* είναι ένας παράγοντας λείανσης (Schapire και Singer, 1999), και:

$$\overline{W}_c^{\,j}=P\left(\tilde{f}(E;x,O)\in bin_j, y=c\right),$$
$$c=\pm 1, j=1\ldots n \quad (6)$$

Δοσμένης της χαρακτηριστικής συνάρτησης

$$B_n^j(u)=\begin{cases}1, & u\in\left[\frac{j-1}{n},\frac{j}{n}\right), j=1\ldots n\\ 0, & \alpha\lambda\lambda\iota\acute{\omega}\varsigma\end{cases} \quad (7)$$

ο αδύναμος ταξινομητής που βασίζεται στο χαρακτηριστικό edgelet *E*, διαμορφώνεται ως εξής:

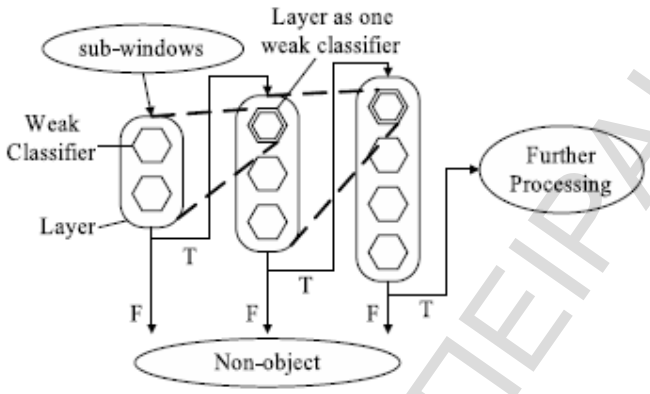
$$h^{(w)}(x)=\frac{1}{2}\sum_{j=1}^{n}\ln\left(\frac{\overline{W}_{+1}^{j}+\varepsilon}{\overline{W}_{-1}^{j}+\varepsilon}\right)B_n^j\left(\tilde{f}(E;xO)\right) \quad (8)$$

Για κάθε χαρακτηριστικό edgelet, δημιουργείται ένας αδύναμος ταξινομητής. Κατόπιν ο αλγόριθμος AdaBoost (Schapire και Singer, 1999) χρησιμοποιείται για την εκμάθηση ισχυρών ταξινομητών, που ονομάζονται *στρώματα* (layers), από την πηγή των αδύναμων ταξινομητών. Ο ισχυρός ταξινομητής $h^{(s)}$ είναι ένας γραμμικός συνδυασμός μιας σειράς επιλεγμένων αδύναμων ταξινομητών:

$$h^{(s)}(x)=\sum_{i=1}^{T}h_i^{(w)}(x)-b \quad (9)$$

όπου *T* είναι το πλήθος των αδύναμων ταξινομητών του $h^{(s)}$ και *b* ένα *κατώφλι* (threshold). Η διαδικασία εκμάθησης ενός στρώματος αναφέρεται ως *στάδιο ενίσχυσης* (boosting stage). Στο τέλος κάθε σταδίου, το κατώτατο όριο *b* τίθεται έτσι ώστε το $h^{(s)}$ να υπάρχει υψηλό ποσοστό εντοπισμού (πειραματικά περίπου 99. 8%) και να απορίπτονται όσο το δυνατόν περισσότερα αρνητικά δείγματα. Τα αποδεκτά *θετικά δείγματα* (positive samples) χρησιμοποιούνται ως το *θετικό σύνολο* (positive set) που τίθεται για την εκμάθηση του επόμενου σταδίου, ενώ οι *εσφαλμένοι εντοπισμοί* (ή false alarms) που λαμβάνονται με τον εντοπισμό αρνητικών εικόνων με τον τρέχοντα εντοπιστή χρησιμοποιείται ως *αρνητικό σύνολο* (negative set) για το επόμενο στάδιο. Τελικά, εντοπιστές σε *δομή φωλιάς* (nested structure), κατασκευάζονται από αυτά τα στρώματα. Η εκπαίδευση τερματίζεται όταν η συχνότητα εσφαλμένων εντοπισμών στο σύνολο των δεδομένων εκπαίδευσης φτάσει στο 10^{-6}. Η *δομή φωλιάς* (Huang, 2004) διαφέρει από τη *δομή καταρράκτη* (cascade structure,

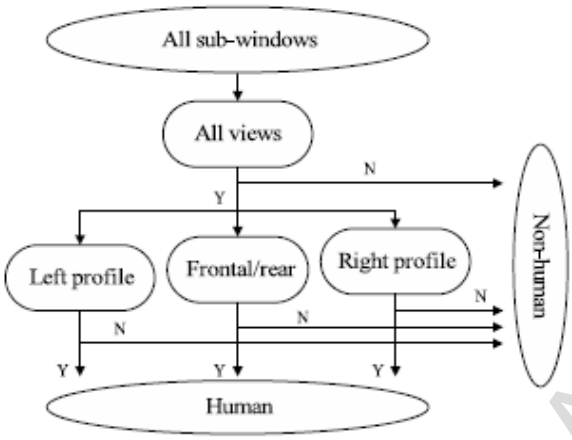
Viola and Jones, 2001). Σε μια δομή φωλιάς, κάθε στρώμα χρησιμοποιείται ως ο πρώτος αδύνατος ταξινομητής του διαδοχικού του έτσι ώστε οι πληροφορίες
της ταξινόμηση να κληρονομούνται αποτελεσματικά. Στο σχήμα 5 φαίνεται μια τέτοια δομή φωλιάς. Το κύριο πλεονέκτημα αυτής της δομής είναι ότι ο αριθμός των χαρακτηριστικών που απαιτείται για την επίτευξη ενός επιπέδου απόδοσης μειώνεται πολύ, συγκριτικά με τον αριθμό που απαιτεί ένας εντοπιστής σε δομή καταρράκτη.



Σχήμα 5. *Δομή φωλιάς (nested structure).*

3.3.2 Εντοπιστές Μερών Πολλαπλών Απόψεων (Multi-View Part Detectors)

Προκειμένου να καλύπτονται όλες οι πιθανές γωνίες περιστροφής, τα δείγματα ανθρώπων χωρίζονται σε τρεις κατηγορίες: αριστερό προφίλ, εμπρόσθιο/οπίσθιο τμήμα, δεξί προφίλ, ανάλογα με τις οπτικές γωνίες. Για κάθε μέρος, εκπαιδεύεται ένας εντοπιστής σε *δομή δέντρου*. Στο σχήμα 6 απεικονίζεται η δομή ενός εντοπιστή πολλαπλών απόψεων. Ο *κόμβος-ρίζα* (root node) του δέντρου μαθαίνεται από το διάνυσμα ενός ενισχυτικού αλγορίθμου που προτείνεται από τον Huang (2005). Το κύριο πλεονέκτημα αυτού του αλγορίθμου είναι ότι τα χαρακτηριστικά που επιλέγονται, μοιράζονται μεταξύ των διαφορετικών κατηγοριών άποψης του ίδιου τύπου αντικειμένου. Αυτό είναι αποδοτικότερο από την εκμάθηση των εντοπιστών για κάθε μεμονωμένη άποψη χωριστά. Κάνουμε έναν ανιχνευτή να καλύψει ένα εύρος της γωνίας κλίσης μιας κάμερας, περίπου [0° , 45°] που είναι κοινό για τα περισσότερα συστήματα επιτήρησης , συμπεριλαμβανομένων των δειγμάτων που συλλαμβάνονται από διαφορετικές γωνίες κλίσης στο συγκεκριμένο *training set*. Εάν θέλουμε να καλύψουμε έναν μεγαλύτερο εύρος γωνίας κλίσης, θα ήταν απαραίτητη η κατηγοριοποίηση απόψεων μέσα σε μία γωνία κλίσης. Κατά τη διάρκεια του εντοπισμού, ένα δείγμα εικόνας στέλνεται αρχικά στη ρίζα του δέντρου, από την οποία εξάγεται ένα διάνυσμα τριών καναλιών που αντιστοιχούν στις τρεις κατηγορίες άποψης. Εάν όλα τα κανάλια είναι αρνητικά τότε το δείγμα ταξινομείται άμεσα ως μη ανθρώπινο. Διαφορετικά, το δείγμα στέλνεται στους κόμβους-φύλλα που αντιστοιχούν στα θετικά κανάλια, για περαιτέρω επεξεργασία. Εάν οποιοσδήποτε από τους τρεις κόμβους-φύλλα δίνει θετικό αποτέλεσμα, τότε το δείγμα ταξινομημείται ως ανθρώπινο, δηλ. *άνθρωπος*. Διαφορετικά απορρίπτεται.

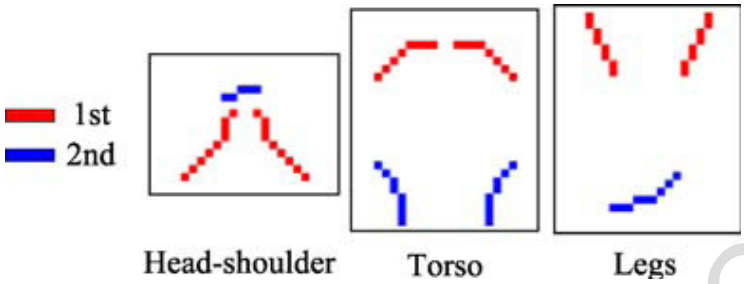


Σχήμα 6.Ένας εντοπιστής μερών πολλαπλών απόψεων σε δομή δέντρου.

Θα μπορούσαν να προκύψουν περισσότερα από ένα, θετικά κανάλια για ένα δείγμα εισαγωγής. Προκειμένου να εντοπιστούν τα μέρη του σώματος σε διαφορετικές κλίμακες, η εικόνα εισαγωγής λαμβάνεται ως δείγμα εκ νέου προκειμένου να δημιουργήσει μια πυραμίδα κλίμακας (scale pyramid) με παράγοντα κλίμακας (scale factor) 1,2, έπειτα η εικόνα σε κάθε κλίμακα «σκανάρεται» από τον εντοπιστή κατά βήματα των 2 pixels. Τα εξαγόμενα των εντοπιστών μερών ονομάζονται *part responses* (αποκρίσεις μερών). Το σχήμα δείχνει τα αποτελέσματα ενός εντοπισμού μερών.

Σχήμα 3.7: *Αποκρίσεις εντοπισμού μερών (με κίτρινο για ολόκληρο το σώμα, με κόκκινο για το τμήμα κεφάλι-ώμοι, με μωβ γαι τον κορμό, με μπλε για τα πόδια).*

Συλλέγεται ένα μεγάλο σύνολο ανθρώπινων δειγμάτων, από τα οποία γίνεται εκμάθηση εντοπιστών δομημένων κατά φωλιές για εμπρόσθια/οπίσθια άποψη ανθρώπων, και εντοπιστών δομημένοι κατά δέντρα για πολλαπλές απόψεις. Το *σχήμα 3.8* παρουσιάζει τα πρώτα δύο χαρακτηριστικά, των οποίων έγινε εκμάθηση, των μερών: κεφάλι-ώμοι, κορμός, και πόδια της εμπρόσθιας/οπίσθιας άποψης. Είναι αρκετά σημαντικά.



Σχήμα 3.8: *Τα δύο πρώτα χαρακτηριστικά τύπου edgelet τα οποία «έμαθαν» οι εντοπιστές.*

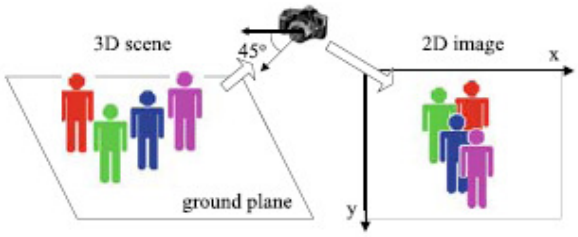
Ο *πίνακας 3.1* απαριθμεί τις περιπλοκές, δηλ., τον αριθμό χαρακτηριστικών που χρησιμοποιούνται, από τους εντοπιστές μερών και πλήρους-σώματος, εμπρόσθιας/οπίσθιας άποψης και πολλαπλών απόψεων. Ο εντοπιστής κεφαλιού-ώμων χρειάζεται περισσότερα χαρακτηριστικά από τους άλλους, ενώ ο εντοπιστής πλήρους-σώματος χρειάζεται πολύ λιγότερα χαρακτηριστικά από οποιονδήποτε εντοπιστή μεμονωμένων μερών.

	FB	*HS*	*T*	*L*
Nested detector	227	1,157	767	753
Tree detector	1,059	3,047	2,546	2,256

Πίνακας 3.1: *Το πλήθος των χαρακτηριστικών που χρησιμοποιείται από τους εντοπιστές. Εντοπιστής δομής φωλιάς (nested detector) για εμπρόσθια/οπίσθια άποψη και εντοπιστής δομής δέντρου (tree detector) για πολλαπλές απόψεις. Όπου FB, HS, T, L αντίστοιχα για πλήρες σώμα, κεφάλι-ώμοι, κορμός και πόδια.*

3.4 Μπεϋζιανός Συνδυασμός των Εντοπιστών Μερών

Για το συνδυασμό των αποτελεσμάτων των εντοπιστών μερών, υπολογίζουμε την πιθανοφάνεια παρουσίας πολλαπλών ανθρώπων στις υποτιθέμενες περιοχές. Εάν παρουσιάζεται παρεμβολή αντικειμένων , τότε η υπόθεση της υπό όρους ανεξαρτησίας μεταξύ των μεμονωμένων ανθρώπινων εμφανίσεων μιας δεδομένης κατάστασης (Mikolajczyk, 2004), δεν ισχύει, και μια περισσότερο σύνθετη διατύπωση κρίνεται απαραίτητη. Αρχίζουμε με τη διατύπωση των μεταβλητών κατάστασης (state) και παρατήρησης (observation). Για να μοντελοποιηθούν οι παρεμβολές μεταξύ αντικειμένων, εκτός από την υπόθεση ότι οι άνθρωποι είναι σε ένα πλάνο, υποθέτουμε επίσης αυτό η κάμερα κοιτάζει προς τα κάτω σε αυτό το πλάνο, βλέπε σχήμα 9. Η υπόθεση αυτή ισχύει για τα περισσότερα συστήματα επιτήρησης. Αυτή η διαμόρφωση φέρνει δύο παρατηρήσεις: (1) εάν ένας άνθρωπος μέσα στην εικόνα είναι ορατός τότε τουλάχιστον το κεφάλι του/της είναι ορατό, και
(2) όσο μακρύτερα είναι ο άνθρωπος από την κάμερα, τόσο μικρότερη είναι η συντεταγμένη *Y* της θέσης της εικόνας των ποδιών του/της.



Σχήμα 9.*Υπόθεση ενός συστήματος επιτήρησης.*

Με τη δεύτερη παρατήρηση, μπορούμε να βρούμε το σχετικό βάθος των ανθρώπων μεσω της σύγκρισης των συντεταγμένων *Y* του καθενός, και να δημιουργηθεί έτσι ένας *χάρτης κατοχής* (occupancy map), ο οποίος καθορίζει ποιο pixel προέρχεται από ποιον άνθρωπο, βλέπει το σχήμα 10 (β). Η γενική μορφή εικόνας ενός ατόμου διαμορφώνεται ως έλλειψη που είναι πιο «στενή» από το ορθογώνιο που λαμβάνεται από τους ανιχνευτές μερών. Από τον χάρτη κατοχής, η αναλογία της ορατής περιοχής προς το συνολικό τομέα του μέρους του σώματος υπολογίζεται ως *κλάσμα διαφάνειας u*. Εάν το *u* είναι μεγαλύτερο από ένα κατώτατο όριο, θ_u (εδώ τίθεται 0,7), τότε το μέρος σώματος ταξινομείται ως ορατό, διαφορετικά ως παρεμβαλλόμενο. Μια υπόθεση μερών (part hypothesis) αναπαρίσταται ως $sp = \{l, p, s, u\}$, όπου το *l* είναι μια ετικέτα που δείχνει τον τύπο μερών (label), *p* είναι η θέση εικόνας (position), *s* είναι το μέγεθος (size), και *u* είναι το κλάσμα διαφάνειας. Η υπόθεση ενός ανθρώπου (human hypothesis) σε ένα frame αποτελείται από τέσσερα μέρη, $H^f = \{sp_i | l_i = FB, HS, T, L\}$, όπου *FB* για το πλήρες σώμα (full-body), *HS* για κεφάλι-ώμους (head-shoulder), *T* για κορμό (torso) και *L* για πόδια (legs). Το σύνολο των υποθέσεων ανθρώπου σε ένα frame είναι $S = \{H_i^{(f)}\}_{i=1}^m$, όπου *m* είναι ένας άγνωστος αριθμός ανθρώπων. Αναπαριστούμε το σύνολο των *υποθέσεων των ορατών μερών* (visible part hypotheses) ως

$$\tilde{S} = \{sp_i \in S | u_i > \theta_u\} \quad (10)$$

το $\tilde{S}$ είναι ένα υποσύνολο του *S*, που προκύπτει από την αφαίρεση των *υποθέσεων παρεμβαλλόμενων μερών* (occluded part hypotheses). Θεωρούμε ότι οι πιθανοφάνειες των υποθέσεων ορατών μερών του συνόλου $\tilde{S}$, είναι ανεξάρτητες. Έστω

$$RP = \{rp_i\}_{i=1}^n \quad (11)$$

το σύνολο των *αποκρίσεων εντοπισμού μερών* (part detection responses), όπου *n* το συνολικό πλήθος των αποκρίσεων και rp_i μια μοναδική απόκριση, που γίνεται στον ίδιο χώρο με μια υπόθεση sp_i. Με το *RP* ως παρατήρηση και το $\tilde{S}$ ως κατάσταση, ορίζουμε την ακόλουθη πιθανοφάνεια προκειμένου να «ερμηνεύσει» την έκβαση των εντοπιστών μερών για μια εικόνα *I*:

$$P(I|S) = P\left(RP\middle|\tilde{S}\right) = \prod_{p\in PT} P\left(RP^{(p)}\middle|\tilde{S}^{(p)}\right) \qquad (12)$$

όπου $PT = \{FB, HS, T, L\}$, $RP^{(p)} = \{rp_i \in RP | l_i = p\}$, και $\tilde{S}^{(p)} = \{sp_i \in \tilde{S} | l_i = p\}$.

Για το «ταίριασμα» υποθέσεων και αποκρίσεων (hypotheses and responses matching), θα μπορούσε να χρησιμοποιηθεί ένας αλγόριθμος "Hungarian" (Kuhn, 1955), ωστόσο είναι αρκετά πολύπλοκος. Εφόσον η ασάφεια απόκρισης-υπόθεσης είναι περιορισμένη, προτιμάται η εφαρμογή ενός αλγορίθμου «απληστείας». Πρώτα υπολογίζεται η *μήτρα απόστασης* **B** όλων των πιθανών ζευγαριών *απόκρισης-μέρους* (response-part). Δηλαδή, **B***(i, j)* είναι η Ευκλίδεια απόσταση μεταξύ της *i-οστής* απόκρισης και της *j-οστής* «υπόθεσης» μέρους (ενν. σώματος). Έπειτα, σε κάθε βήμα ευρίσκεται το ζεύγος (i^*, j^*) με τη μικρότερη Ευκλίδεια απόσταση ενώ διαγράφεται από τη μήτρα **B** η *i-οστή* γραμμή και η *j-οστή* στήλη. Αυτή η επιλογή γίνεται επαναληπτικά έως ότου δεν υπάρχει άλλο έγκυρο ζευγάρι διαθέσιμο. Για κάθε «ταίριασμα», οι αποκρίσεις του συνόλου *RP* και οι υποθέσεις του συνόλου $\tilde{S}$, ταξινομούνται σε τρεις κατηγορίες: **έγκυροι εντοπισμοί**, δηλαδή αποκρίσεις που «έχουν ταιριάξει» με μια υπόθεση (συμβολίζονται με SD, από το successful detection), **άκυροι** ή **εσφαλμένοι εντοπισμοί**, δηλαδή αποκρίσεις που δεν «ταίριαξαν» με υπόθεση (ή FA, από το false alarm), και **αρνητικές υποθέσεις**, δηλαδή υποθέσεις που δεν «ταίριαξαν» με εντοπισμό (ή FN, από το false negative) οι οποίες είναι εντοπισμοί απέτυχαν να γίνουν. Οι κατηγορίες αυτές σημειώνονται αντίστοιχα με T_{SD}, T_{FA} και T_{FN}. Η πιθανοφάνεια ενός τύπου μερών υπολογίζεται από

$$P\left(RP^{(p)}\middle|\tilde{S}^{(p)}\right) \propto \prod_{rp_i\in T_{SD}^{(p)}} P_{SD}^{(p)} P\left(rp_i\middle|\overline{sp}_i\right) \cdot \prod_{rp_i\in T_{FA}^{(p)}} P_{FA}^{(p)} \cdot \prod_{rp_i\in T_{FN}^{(p)}} P_{FN}^{(p)} \qquad (13)$$

όπου $\overline{sp}_i$ είναι η αντίστοιχη υπόθεση μιας απόκρισης rp_i, P_{SD} είναι η «ανταμοιβή» ενός έγκυρου εντοπισμού, P_{FA} και P_{FN} είναι οι «ποινές» ενός άκυρου εντοπισμού και μιας αρνητικής υπόθεσης αντίστοιχα, ενώ $P\left(rp_i\middle|\overline{sp}_i\right) = P\left(p_{rp}\middle|p_{\overline{sp}}\right)P\left(s_{rp}\middle|s_{\overline{sp}}\right)$ είναι η δεσμευμένη πιθανότητα μιας *απόκρισης εντοπισμού*, δεδομένης της αντίστοιχης *υπόθεσης μέρους*. Οι $P\left(p_{rp}\middle|p_{\overline{sp}}\right)$ και $P\left(s_{rp}\middle|s_{\overline{sp}}\right)$ ακολουθούν Γκαουσσιανή κατανομή. Έχουμε, N_{FA}, N_{SD}, και N_G το πλήθος των άκυρων εντοπισμών, το πλήθος των έγκυρων εντοπισμών, και το πλήθος των εξ'ορισμού υπαρκτών αντικειμένων (ground-truth objects) αντίστοιχα, οπότε οι P_{FA}, P_{SD} υπολογίζονται από

$$P_{FA} = \frac{1}{\alpha} e^{-\beta} \frac{N_{FA}}{N_{FA} + N_{SD}}, P_{SD} = \frac{1}{\alpha} e^{\beta} \frac{N_{SD}}{N_{FA} + N_{SD}} \qquad (14)$$

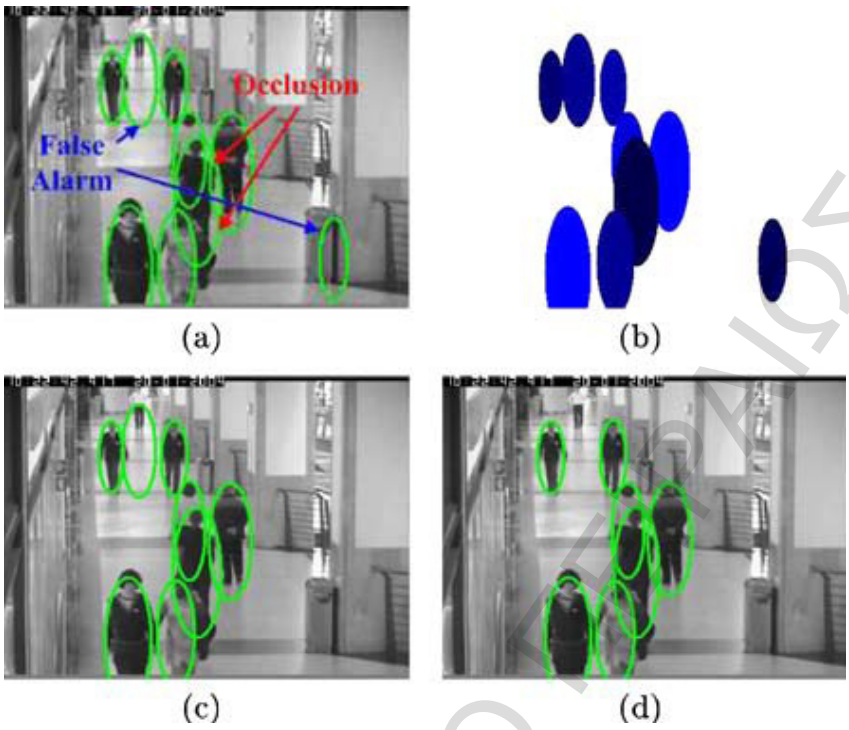
όπου α είναι ένας παράγοντας κανονικοποίησης έτσι ώστε $P_{FA} + P_{SD} = 1$ και β είναι ένας παράγοντας που ορίζεται για τον έλεγχο της «σχετικής σημασίας» του ποσοστού εντοπισμών προς άκυρων εντοπισμών (εδώ ορίζεται στο 0,5). Η P_{FN} υπολογίζεται από τον τύπο

$$P_{FN} = \frac{N_G - N_{SD}}{N_G} \qquad (15)$$

τα N_{FA}, N_{SD}, N_G, $P\left(p_{rp}\middle|p_{s\overline{p}}\right)$ και $P\left(s_{rp}\middle|s_{s\overline{p}}\right)$ *μαθαίνονται* από ένα *σύνολο επαλήθευσης* (verification set). Για διαφορετικούς εντοπιστές, μπορεί να προκύψουν διαφορετικές τιμές των P_{SD}, P_{FA}, P_{FN} και $P\left(rp\middle|s\overline{p}\right)$. Τελικά χρειάζεται μια μέθοδος που να προτείνει τις υποθέσεις που θα διαμορφώσουν μια πιθανή κατάσταση (state) *S*, και να αναζητήσει τον *χώρο λύσεων* (solution space) προκειμένου να γίνει μεγιστοποίηση της *εκ των υστέρων πιθανότητας* $P\left(S\middle|I\right)$ (maximum a posteriori, MAP). Σύμφωνα με τον κανόνα του Bayes έχουμε

$$P\left(S\middle|I\right) \propto P\left(I\middle|S\right)P(S) = P\left(RP\middle|\tilde{S}\right)P(S) \qquad (16)$$

Θεωρώντας μια ομοιόμορφη κατανομή της εκ των προτέρων *P(S)*, η παραπάνω εκτίμηση MAP ισοδυναμεί με τη μεγιστοπίηση της από κοινού πιθανοφάνειας $P\left(RP\middle|\tilde{S}\right)$. Στη μέθοδο αυτή, το αρχικό σύνολο των *υποθέσεων S* προτείνεται από τις αποκρίσεις των εντοπιστών κεφαλιού-ώμων και πλήρους σώματος. Κάθε απόκριση των δύο αυτών εντοπιστών παράγει μια *υπόθεση ανθρώπου*. Έπειτα οι υποθέσεις επαληθεύονται από το παραπάνω μοντέλο πιθανοφάνειας κατά σειρά βάθους. Τα βήματα αυτής της διαδικασίας φαίνονται στο *σχήμα 3.10*. Το *σχήμα 3.9* δίνει ένα παράδειγμα αποτελεσμάτων του *συνδυασμένου αλγορίθμου*. Στην αρχική κατάσταση, υπάρχουν δύο άκυροι εντοπισμοί που ωστόσο δεν υποστηρίζονται από αρκετά στοιχεία και αργότερα απορρίπτονται. Τα πόδια του ανθρώπου στη μέση της εικόνας δεν εντοπίζονται από τον εντοπιστή ποδιών καθώς κρύβονται από άλλον άνθρωπο. Πρόκειται δηλαδή για παρεμβολή μεταξύ αντικειμένων, οπότε δεν προστίθεται «ποινή» στον εντοπιστή. Σε αυτόν το συνδυασμένο αλγόριθμο, οι ανιχνευτές κορμού και ποδιών δεν χρησιμοποιούνται για να προτείνουν υποθέσεις ανθρώπων. Κι αυτό γιατί, οι εντοπιστές που χρησιμοποιούνται κατά την έναρξη, πρέπει να «σκανάρουν» ολόκληρη την εικόνα, ενώ οι εντοπιστές επαλήθευσης μόνο την περιοχή των προτεινόμενων υποθέσεων. Έτσι, χρησιμοποιώντας και τους τέσσερις εντοπιστές, το σύστημα γίνεται τουλάχιστον δύο φορές πιο αργό. Επίσης, έχει βρεθεί ότι η ένωση των εντοπιστών πλήρους σώματος και κεφαλιού-ώμων αποφέρει πολύ υψηλά ποσοστά εντοπισμού, και ότι, τις περισσότερες φορές, οι παρεμβολές αφορούν το κάτω μέρος του σώματος. Ο παραπάνω αλγόριθμος Μπεϋζιανού συνδυασμού ονομάζεται *συνδυασμένος εντοπιστής* (combined detector), και τα αποτελέσματά του *συνδυασμένες αποκρίσεις* (combined responses).



Σχήμα 3.9: *Αναζήτηση της βέλτιστης ερμηνείας της εικόνας: (α) αρχική κατάσταση, (β) «χάρτης κατοχής» της αρχικής κατάστασης (γ) μια ενδιάμεση κατάσταση και (δ) τελική κατάσταση*

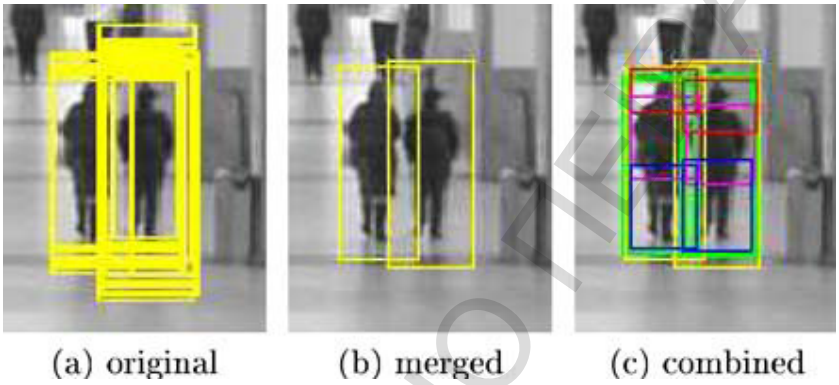
---

1. «Σάρωσε» την εικόνα με τους εντοπιστές πλήρους σώματος και μερών σώματος
2. Πρότεινε το διάνυσμα αρχικής κατάστασης *S* από τις αποκρίσεις των εντοπιστών πλήρους σώματος και κεφαλιού-ώμων
3. Ταξινομήστε τους ανθρώπους σύμφωνα με τις συντεταγμένες *Y* με κατιούσα σειρά
4. Για *i=1* έως *m* κάνε
   (α) «Ταίριασμα» των αποκρίσεων των εντοπιστών με τα αντίστοιχα ορατά μέρη
   (β) Υπολογισμό της πιθανοφάνειας της εικόνας $P\left(RP\middle|S\right)$ και $P\left(RP\middle|S-H_i^{(f)}\right)$
   (γ) Αν $P\left(RP\middle|S-H_i^{(f)}\right) > P\left(RP\middle|S\right)$, τότε $S \leftarrow S-H_i^{(f)}$
5. Εξαγωγή του *S* ως αποτέλεσμα

Σχήμα 3.10: *Αλγόριθμος για το συνδυασμό των αποκρίσεων των επιμέρους ανιχνευτών.*

Τα δεδομένα που εξάγονται από το σύστημα ανήκουν σε τρία επίπεδα. Το πρώτο επίπεδο είναι ένα σύνολο των *αρχικών αποκρίσεων* (original responses) των εντοπιστών. Σε αυτό το σύνολο, ένα αντικείμενο μπορεί να έχει πολλαπλές αντίστοιχες αποκρίσεις, βλ. σχήμα 12 (α). Το δεύτερο επίπεδο είναι αυτό των *συγχωνευμένων αποκρίσεων* (merged responses), οι οποίες προκύπτουν από την εφαρμογή ενός αλγορίθμου *ομαδοποίησης*

(clustering) στις αρχικές αποκρίσεις. Ο αλγόριθμος ομαδοποίησης επιλέγει τυχαία μια αρχική απόκριση ως «σπόρο» (seed) και συγχωνεύει τις απαντήσεις που έχουν μεγάλη επικάλυψη με αυτόν. Αυτή η διαδικασία εκτελείται επαναληπτικά μέχρι όλες οι αρχικές απαντήσεις να έχουν υποβληθεί σε επεξεργασία. Στο σύνολο των συγχωνευμένων αποκρίσεων, κάθε αντικείμενο έχει το πολύ μία αντίστοιχη απόκριση , βλ σχήμα 12 (β). Το τρίτο επίπεδο είναι αυτό των συνδυασμένων αποκρίσεων. Μια συνδυασμένη απόκριση έχει διάφορες «ταιριασμένες» αποκρίσεις μερών, το *σχήμα 3.12* (γ) δείχνει ένα παράδειγμα. Η αποκρίσεις εντοπισμού μπορεί χωρικά να μην είναι ιδιαίτερα ακριβής, επειδή τα δείγματα εκπαίδευσης του αλγορίθμου περιλαμβάνουν μερικές περιοχές υποβάθρου προκειμένου να καλυφθούν κάποιες διακυμάνσεις θέσης και μεγέθους.



Σχήμα 3.12: *Αποκρίσεις εντοπισμού. (α) και (β) από τον εντοπιστή πλήρους σώματος (γ) από το συνδυασμένο εντοπιστή (με πράσινο ο συνδυασμένος, με κίτρινο για πλήρες σώμα, με κόκκινο για κεφάλι-ώμους, με μωβ για κορμό, και μπλε για πόδια).*

3.5 Παρακολούθηση με βάση τον Εντοπισμό Μερών

Ο αλγόριθμος *παρακολούθησης πορείας ανθρώπου* (human tracking algorithm) , εκλαμβάνει τις αποκρίσεις του εντοπιστή μερών και του συνδυασμένου εντοπιστή ως παρατηρήσεις *υποθέσεων ανθρώπων*.

3.5.1 Συνάφεια Αποκρίσεων Εντοπισμού

Τόσο οι αρχικές όσο και οι συγχωνευμένες αποκρίσεις εντοπισμού είναι αποκρίσεις μερών. Για την παρακολούθηση πορείας (tracking), προστίθενται δύο ακόμα στοιχεία στην αναπαράσταση των αποκρίσεων μερών, $rp = \{l, p, s, \upsilon, f, c\}$, όπου το καινούριο στοιχείο *f* η τιμή *εμπιστοσύνης ανίχνευσης* (detection confidence), ενώ το *c* είναι ένα πρότυπο εμφάνισης. Τα πρώτα πέντε στοιχεία, *l, p, s, υ,* και *f* λαμβάνονται άμεσα από τη διαδικασία εντοπισμού. Το πρότυπο εμφάνισης *c* είναι ένα *ιστόγραμμα χρώματος* (color histogram), του οποίου ο υπολογισμός και η ανανέωση περιγράφονται λεπτομερώς στην παράγραφο 4.3. Η αναπαράσταση μιας συνδυασμένης απόκρισης είναι η ένωση των αναπαραστάσεων των μερών της, δηλ. $rc = \{rp_i | l_i = FB, HS, T, L\}$.

Ο εντοπισμός ανθρώπων γίνεται πλαίσιο προς πλαίσιο (frame by frame). Προκειμένου να διευκρινιστεί το αν δύο αποκρίσεις rp_1 και rp_2, εντοπισμού ίδιου τύπου μέρους σώματος από δύο διαφορετικά frames ανήκουν στο ίδιο αντικείμενο (π.χ στο μέρος κεφάλι-ώμοι του ίδιου ανθρώπου), ορίζεται ένα *μέτρο συνάφειας* (affinity measure)

$$A(rp_1, rp_2) = A_{pos}(p_1, p_2) A_{size}(s_1, s_2) A_{appr}(c_1, c_2) \qquad (17)$$

όπου A_{pos}, A_{size} και A_{appr} είναι οι συνάφειες θέσης, μεγέθους και εμφάνισης αντίστοιχα. Οι ορισμοί αυτών είναι

$$\begin{aligned} A_{pos}(p_1, p_2) &= \gamma_{pos} \exp\left[\frac{(x_1 - x_2)^2}{\sigma_x^2}\right] \exp\left[\frac{(y_1 - y_2)^2}{\sigma_y^2}\right] \\ A_{size}(s_1, s_2) &= \gamma_{size} \exp\left[-\frac{(s_1 - s_2)^2}{\sigma_s^2}\right] \\ A_{appr}(c_1, c_2) &= B(c_1, c_2) \end{aligned} \qquad (18)$$

όπου $B(c_1, c_2)$ είναι η απόσταση *Bhattachayya* ανάμεσα στα δύο ιστογράμματα και τα γ_{size}, γ_{pos} είναι παράγοντες κανονικοποίησης. Η συνάφεια μεταξύ των δύο συνδυασμένων αποκρίσεων, rc_1 και rc_2, είναι μέσος όρος της συνάφειας των κοινών τους ορατών μερών

$$A(rc_1, rc_2) = \frac{\sum_{l_i \in PT} A(Pt_i(rc_1), Pt_i(rc_2)) I(\upsilon_{i1}, \upsilon_{i2} > \theta_\upsilon)}{\sum_{l_i \in PT} I(\upsilon_{i1}, \upsilon_{i2} > \theta_\upsilon)} \qquad (19)$$

όπου το $Pt_i(rc)$ επιστρέφει την απόκριση του μέρους *i* της συνδυασμένης απόκρισης *rc*, υ_{ij} είναι ο βαθμός διαφάνειας του $Pt_i(rc_j)$, *j=1,2*, και *I* μία συνάρτηση δεικτών. Οι παραπάνω συναρτήσεις συνάφειας κωδικοποιούν τις πληροφορίες θέσης, μεγέθους και εμφάνισης.

Δοσμένης της συνάφειας, το «ταίριασμα» των *αποκρίσεων εντοπισμού* με τις αντίστοιχες *υποθέσεις ανθρώπων* γίνεται όμοια με το «ταίριασμα» των υποθέσεων με τις αντίστοιχες *αποκρίσεις μερών* (παρ. 3). Υποθέτουμε ότι τη χρονική στιγμή *t* ενός δοσμένου video, έχουμε *n* υποθέσεις ανθρώπων (δηλ. υπάρξεις ανθρώπων) $H_1^{(\upsilon)}, \ldots, H_n^{(\upsilon)}$, των οποίων οι προβλέψεις κατα τον χρόνο *t + 1* είναι $\hat{rc}_{t+1,1}, \ldots, \hat{rc}_{t+1,n}$ και σε χρόνο επίσης *t + 1* έχουμε *m* αποκρίσεις $rc_{t+1,1}, \ldots, rc_{t+1,m}$. Πρώτον, υπολογίζεται η *μήτρα συνάφειας (ή συγγενικότητας) A*, διαστάσεων *m x n*, για όλα τα ζευγάρια $(\hat{rc}_{t+1,i}, rc_{t+1,j})$, δηλαδή έχουμε $A(i, j) = A(\hat{rc}_{t+1,i}, rc_{t+1,j})$. Έπειτα, σε κάθε βήμα, το ζευγάρι με τη μεγαλύτερη συνάφεια, που συμβολίζεται με *(i* , j*)*, λαμβάνεται ως «ταίριασμα» και η *i*-οστή* γραμμή με την *j*-οστή* στήλη της μήτρας *A* διαγράφονται. Η διαδικασία αυτή επαναλαμβάνεται έως ότου να μην υπάρχουν άλλα ζευγάρια.

3.5.2 Έναρξη Τροχιάς (Trajectory Initialization)

Η βασική ιδέα της στρατηγικής έναρξης είναι να ξεκινά μια παρακολούθηση τροχιάς, όταν συλλέγονται αρκετά στοιχεία από τις αποκρίσεις εντοπισμού. Η ακρίβεια, *pr*, ενός εντοπιστή, ορίζεται ως η αναλογία μεταξύ του αριθμού έγκυρων εντοπισμών και του αριθμού όλων των αποκρίσεων εντοπισμού. Αν η ακρίβεια *pr* παραμένει σταθερή ανάμεσα στα frames, και εντοπισμός σε ένα frame είναι ανεξάρτητος από το γειτονικά του, τότε κατά τη διάρκεια *T* διαδοχικών χρονικών βημάτων, η πιθανότητα ότι ο εντοπιστής θα εξάγει *T* διαδοχικούς άκυρους εντοπισμούς είναι $P_{FA} = (1 - pr)^T$. Βέβαια, τα συμπεράσματα αυτά δεν είναι ακριβή για ρεαλιστικά video, όπου υπάρχει σημαντική αλληλοεξάρτηση μεταξύ των frames. Εάν ο εντοπιστής εξάγει έναν άκυρο εντοπισμό σε μια περιοχή ενός frame του video, τότε υπάρχει μεγάλη πιθανότητα και στο επόμενο frame να εξαχθεί ένας άκυρος εντοπισμός στην ίδια περίπου περιοχή. Το πρόβλημα αυτό ονομάζεται *πρόβλημα επίμονου άκυρου εντοπισμού* (persistent false alarm problem). Οπότε, η πιθανότητα P_{FA} εκφράζεται μέσω μιας εκθετικής φθίνουσας συνάρτησης του *T*, της μορφής $e^{-\lambda_{init}\sqrt{T}}$.

Υποθέτουμε ότι έχουν βρεθεί *T (>1)* διαδογικές αποκρίσεις, $(rc_1, ..., rc_T)$ που αντιστοιχούν σε μια *υπόθεση ανθρώπου* $H^{(\upsilon)}$ μέσω συσχετισμού δεδομένων. Τότε, η αξιοπιστία έναρξης μιας τροχιάς για την υπόθεση $H^{(\upsilon)}$ ορίζεται ως

$$InitConf\left(H^{(\upsilon)}; rc_{1...T}\right) = \underbrace{\frac{1}{T-1}\sum_{t=1}^{T-1} A\left(\hat{rc}_{t+1}, rc_{t+1}\right)}_{(1)} \cdot \underbrace{\left(1 - e^{-\lambda_{init}\sqrt{T}}\right)}_{(2)} \qquad (20)$$

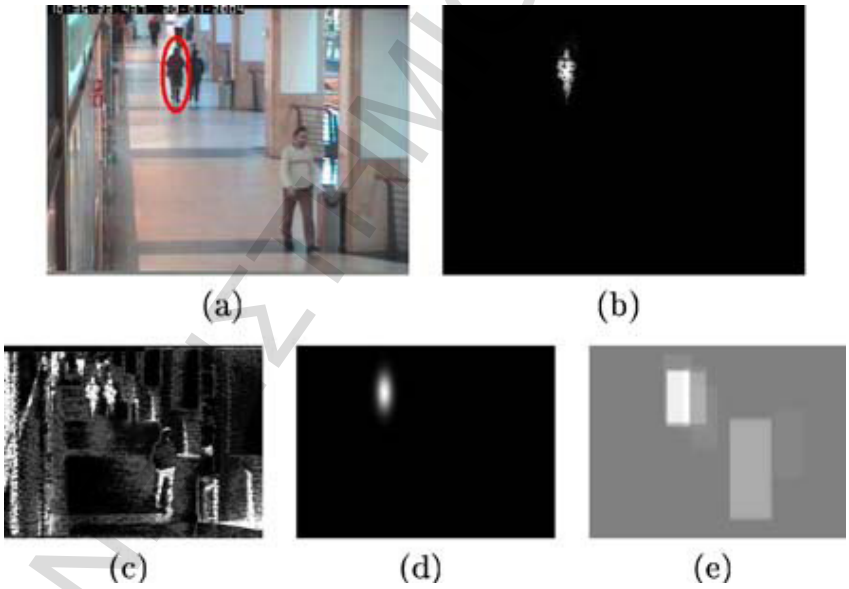
Ο πρώτος όρος (1) αριστερά της εξίσωσης (20) δηλώνει τη *μέση συνάφεια* των αποκρίσεων *T*, και ο δεύτερος όρος εξαρτάται από την ακρίβεια του εντοπιστή. Όσο πιο ακριβής είναι ο εντοπιστής, τόσο αυξάνεται η παράμετρος λ_{init}. Η στρατηγική έναρξης ανίχνευσης τροχιάς έχει ως εξής: εάν το $InitConf\left(H^{(\upsilon)}\right)$ είναι μεγαλύτερο από ένα κατώφλι θ_{init}, τότε γίνεται έναρξη ανίχνευσης τροχιάς του $H^{(\upsilon)}$, και το $H^{(\upsilon)}$ λαμβάνεται ως *αξιόπιστη τροχιά*, διαφορετικά θεωρείται μια *πιθανή τροχιά*. Στην εφαρμογή των Wu και Nevatia, 2006, είναι: $\lambda_{init} = 1,2$ και $\theta_{init} = 0,83$. Μια υπόθεση ανθρώπινης τροχιάς $H^{(\upsilon)}$, αναπαρίσταται ως $\left\{\{rc_t\}_{t=1...T}, D, \{C_i\}_{i=FB,HS,TS,L}\right\}$, όπου $\{rc_i\}$ είναι ένα σύνολο αποκρίσεων, $\{C_i\}$ το μοντέλο εμφάνισης των μερών (ενν. του σώματος), και *D* ένα *δυναμικό μοντέλο*. Στην πράξη, το $\{C_i\}$ είναι ο μέσος όρος των μοντέλων εμφάνισης όλων των αποκρίσεων εντοπισμού, και το *D* διαμορφώνεται μέσω ενός φίλτρου Kalman για κίνηση σταθερής ταχύτητας.

3.5.3 Ανάπτυξη Τροχιάς (Trajectory Growth)

Μετά την έναρξη της παρακολούθησης μιας τροχιάς, ένα αντικείμενο παρακολουθείται με δύο στρατηγικές: *συσχετισμό δεδομένων* (data association) και *μέσης μετατόπισης* (meanshift). Για κάθε νέο frame, και για όλες τις υπάρχουσες υποθέσεις, ψάχνουμε αρχικά τις αντίστοιχές τους αποκρίσεις εντοπισμού σε αυτό το frame. Εάν υπάρχει μια νέα απόκριση εντοπισμού που να αντιστοιχεί σε μια υπόθεση $H^{(\upsilon)}$, τότε η $H^{(\upsilon)}$ αναπτύσσεται βάση συσχετισμού δεδομένων, διαφορετικά εφαρμόζεται ένας ιχνηλάτης-αλγόριθμος meanshift. Ο συσχετισμός δεδομένων εκτελείται σε δύο βήματα. Πρώτα, γίνεται το «ταίριασμα» των υποθέσεων με τις συνδυασμένες αποκρίσεις, με τον τρόπο που αναφέρεται στην παράγραφο 4.1. Δεύτερον, όλες οι υποθέσεις που δεν βρήκαν «ταίρι» στο πρώτο βήμα, συσχετίζονται με τις αποκρίσεις μερών που απέμειναν και δεν

ανήκουν σε κάποια συνδυασμένη απόκριση. Το ταίριασμα των αποκρίσεων μερών με υποθέσεις, είναι μια απλοποιημένη εκδοχή της μεθόδου ταιριάσματος των συνδυασμένων αποκρίσεων με αντίστοιχες υποθέσεις. Πρέπει να εντοπιστεί τουλάχιστον ένα μέρος του αντικειμένου, προκειμένου να γίνει παρακολούθηση της τροχιάς αυτού, μέσω συσχετισμού δεδομένων. Επειδή η αντιμετώπιση των παρεμβολών, που γίνεται πριν της συσχέτισης, στις αποκρίσεις εντοπισμού στο τρέχον frame είναι πιο αξιόπιστη από τις αρχικά προβλεφθείσες και αναξιόπιστες υποθέσεις, δεν συσχετίζονται άμεσα οι αποκρίσεις μερών με τις αντίστοιχες τροχιές.

Όποτε η συσχέτιση στοιχείων αποτυγχάνει (π.χ. οι εντοπιστές δεν μπορούν βρουν το αντικείμενο ή η συνάφεια είναι χαμηλή), ένας αλγόριθμος meanshift (Comaniciu, 2001) εφαρμόζεται για να ακολουθήσει καθένα από τα μέρη χωριστά. Τα αποτελέσματα συνδυάζονται για να διαμορφώσουν την τελική εκτίμηση. Η βασική ιδέα του meanshift είναι να ακολουθηθεί μια κατανομή πιθανότητας. Αν και ο χαρακτηριστικός τρόπος να χρησιμοποιηθεί μια μέθοδος meanshift είναι μέσω μιας κατανομής χρώματος, δεν υπάρχει κανένας περιορισμός στην κατανομή που μπορεί να χρησιμοποιηθεί. Στη μέθοδο αυτή συνδυάζονται το μοντέλο εμφάνισης *C*, το δυναμικό μοντέλο *D*, και η αξιοπιστία εντοπισμού *f*, κι έτσι δημιουργείται ένας *χάρτης πιθανοφάνειας* με τον οποίο τροφοδοτείται ο meanshift. Ένας *δυναμικός χάρτης πιθανότητας*, $P_{dyn}(\upsilon)$, όπου το *υ* αντιπροσωπεύει τις συντεταγμένες της εικόνας, υπολογίζεται μέσω του δυναμικού μοντέλου *D*, βλ. *σχήμα 3.13*(d).



Σχήμα 3.13: *Χάρτης πιθανοτήτων (a) αρχικό frame, (b) τελικός χάρτης πιθανότητας (c), (d) και (e) χάρτες πιθανοτήτων μοντέλου εμφάνισης, δυναμικού μοντέλου και εντοπισμού αντίστοιχα (το αντικείμενο ενδιαφέροντος είναι σημειωμένο με κόκκινη έλλειψη) .*

Έστω $\{rp_j\}$ οι αρχικές αποκρίσεις ενός εντοπιστή μερών σε ένα frame *j*, ο *χάρτης πιθανότητας εντοπισμού* $P_{\det}(u)$ ορίζεται ως

$$P_{\det}(u) = \sum_{j:u \in \mathrm{Re}\, g(rp_j)} f_j + ms \qquad (21)$$

όπου $\mathrm{Re}\, g(rp_j)$ είναι η περιοχή της εικόνας, ένα ορθογώνιο, που αντιστοιχεί στο rp_j, f_j είναι μια τιμή αξιοπιστίας εντοπισμού για τα rp_j, και *ms* είναι μια σταθερά που αντιστοιχεί στο ποσοστό αποτυχίας (δηλ. το κλάσμα του αριθμού των χαμένων αντικειμένων προς το συνολικό αριθμό αντικειμένων). Η σταθερά *ms* υπολογίζεται μετά τη *μάθηση* των εντοπιστών. Αν κάποιο pixel ανήκει σε πολλαπλές θετικές αποκρίσεις εντοπισμού, τότε ο βαθμός εντοπισμού αυτού του pixel, ορίζεται ως το άθροισμα των βαθμών αξιοπιστίας όλων αυτών των αποκρίσεων. Διαφορετικά ο βαθμός εντοπισμού ορίζεται ως η μέση συχνότητα αποτυχίας, που είναι θετικός αριθμός. Σημειώνεται ότι, οι αρχικές αποκρίσεις χρησιμοποιούνται προς αποφυγή της επιρροής των λαθών του αλγορίθμου ομαδοποίησης (βλ. σχήμα 13(e)).

Έστω $P_{appr}(u)$ ο χάρτης πιθανότητας εμφάνισης. Δεδομένου ότι το *C* είναι ιστόγραμμα χρώματος, η $P_{appr}(u)$ είναι η bit τιμή του *C* (βλ. 13(c)). Προκειμένου να υπολογιστεί το *C*, πρέπει το αντικείμενο να κατατμηθεί ώστε να γνωρίζουμε ποια pixels ανήκουν σε αυτό. Το ορθογώνιο απόκρισης εντοπισμού δεν είναι αρκετά ακριβές για το σκοπό αυτό. Οι Zhao και Davis (2005), πρότειναν μια επαναληπτική μέθοδο κατάτμησης του άνω μέρους σώματος για επαλήθευση των εντοπισθέντων υποθέσεων ανθρώπων. Οι Wu και Nevatia εδώ προτείνουν μια μέθοδο μέσω PCA. Κατά το στάδιο εκπαίδευσης, συλλέγονται παραδείγματα και οι περιοχές αντικειμένων ονομάζονται χειρωνακτικά, βλ. σχήμα 14(a). Έπειτα γίνεται μάθηση του μοντέλου PCA με τα δεδομένα αυτά, βλ. σήμα 14(b). Θεωρούμε ότι έχουμε ένα αρχικό μοντέλο εμφάνισης, C_o. Δεδομένου ενός νέου δείγματος (βλ. σχήμα 14(c)), υπολογίζεται πρώτα ο χάρτης πιθανότητας χρώματος από το C_o (βλ. 14(d)), έπειτα χρησιμοποιείται το μοντέλο PCA ως μία global σταθερά σχήματος ανακατασκευάζοντας το χάρτη πιθανότητας (σχ. 14(e)) . Ο *thresholded* χάρτης (σχ. 14(f)) λαμβάνεται ως το τελικό αντικείμενο κατάτμησης, που χρησιμοποιείται για την ανανέωση του C_o. Το μέσο διάνυσμα το πρώτο του σχήματος 14(b), χρησιμοποιείται για τον πρώτο υπολογισμό του C_o. Γίνεται μάθηση ενός μοντέλου PCA, για κάθε μέρος σώματος. Εδώ σημειώνεται ότι, αν και αυτή η μέθοδος κατάτμησης απέχει αρκετά από την τελειότητα, είναι ωστόσο πολύ γρήγορη και επαρκής για να ανανεώσει το μοντέλο εμφάνισης.

Συνδυάζοντας τις $P_{appr}(u)$, $P_{dyn}(u)$ και $P_{\det}(u)$ ορίζουμε την πιθανοφάνεια εικόνας για ένα μέρος σώματος σε ένα pixel *u* ως εξής

$$L(u) = P_{appr}(u) P_{dyn}(u) P_{\det}(u) \qquad (22)$$

Το σχήμα 13 δείχνει ένα παράδειγμα υπολογισμού χάρτη πιθανότητας. Η αντιμετώπιση του προβλήματος των παρεμβολών, γίνεται πριν ενεργοποιηθεί ο αλγόριθμος meanshift. Γίνεται παρακολούθηση τροχιάς μονάχα των ορατών μερών που εντοπίστηκαν κατά τον τελευταίο επιτυχή συσχετισμό δεδομένων. Τελικά ανανεώνονται τα πρότυπα-μοντέλα μονάχα των αντικειμένων που εντοπίστηκαν και δεν παρεμβάλλονται από άλλα. Ωστόσο δεν εφαρμόζεται πάντα ο meanshift καθώς στηρίζεται στο χρώμα, ενώ οι εντοπιστές με βάση το σχήμα είναι πολύ πιο αξιόπιστοι.

3.5.4 Τερματισμός Τροχιάς (Trajectory Termination)

Η στρατηγική τερματισμού τροχιάς, είναι απρόμια με αυτήν της έναρξης. Σε περίπτωση που για ένα αντικείμενο $H^{(\upsilon)}$ δεν έχουν βρεθεί αποκρίσεις ανίχνευσης για για διαδοχικά *T* χρονικά βήματα, η αξιοπιστία τερματισμού ενός $H^{(\upsilon)}$ υπολογίζεται ως εξής

$$InitConf\left(H^{(\upsilon)}; rc_{1...T}\right) = \left(1 - \frac{1}{T-1}\sum_{t=1}^{T-1} A\left(r\hat{c}_{t+1}, rc_{t+1}\right)\right) \cdot \left(1 - e^{-\lambda_{init}\sqrt{T}}\right) \quad (23)$$

Σημειώνεται ότι οι συνδυασμένες αποκρίσεις rc_t αποκτώνται μέσω του meanshift και όχι μέσω του συνδυασμένου εντοπιστή. Αν το $EndConf\left(H^{(\upsilon)}\right)$ είναι μεγαλύτερο από ένα κατώφλι, θ_{end}, τότε η υπόθεση $H^{(\upsilon)}$ τερματίζεται, οπότε και ονομάζεται *νεκρή τροχιά* (dead trajectory), διαφορετικά *ζωντανή τροχιά* (alive trajectory). Στα πειράματα των Wu και Nevatia ισχύει: $\lambda_{end} = 0{,}5$, $\theta_{end} = 0{,}8$.

3.5.5 Ο Συνδυασμένος Ανιχνευτής Τροχιάς (Combined Tracker)

Εδώ τοποθετούνται μαζί και οι τρεις παραπάνων ενότητες, έναρξη τροχιάς, παρακολούθηση, και τερματισμός. Τα βήματα του προς τα εμπρός ή χωρίς οπισθοδρόμηση, αλγορίθμου παρακολούθησης, είναι:

<u>Forward Human Tracking</u>

Έστω *S*, το σύνολο των υποθέσεων, αρχικά θέτουμε *S* = *Φ*.
Για κάθε χρονικό βήμα *t*, (σημειώνεται με S_t το σύνολο των *ζωντανών τροχιών* του *S* κατά το χρόνο *t*), έχουμε

1. Στατικός εντοπισμός:
   (α) Εντόπισε μέρη. Έστω ότι το αποτέλεσμα είναι RP_t
   (β) Συνδύασε τις αποκρίσεις εντοπισμού μερών, συμπεριλαμβάνοντας τη θεώρηση της παρεμβολής αντικειμένων. Έστω ότι το αποτέλεσμα είναι RC_t
   (γ) Αφαίρεσε τα μέρη του RC_t από το RP_t
2. Συσχετισμός δεδομένων:
   (α) Συσχέτισε τις υποθέσεις του S_t με τις συνδυασμένες αποκρίσεις του RC_t. Έστω ότι το σύνολο των ταιριασμένων υποθέσεων είναι S_{t1}.
   (β) Συσχέτισε τις υποθέσεις του $S_t - S_{t1}$ με τις αποκρίσεις μερών του RP_t. Έστω ότι το σύνολο των ταιριασμένων υποθέσεων είναι S_{t2}.
   (γ) Κατασκεύασε μια καινούρια υπόθεση $H^{(\upsilon)}$ για κάθε αταίριαστη απόκριση του RC_t, και πρόσθεσε το $H^{(\upsilon)}$ στα *S* και S_t.
3. Meanshift παρακολούθηση: Για κάθε αξιόπιστη τροχιά του $S_t - S_{t1} - S_{t2}$, ανέπτυξέ την μέσω meanshift.

4. Ανανέωση του μοντέλου:

(α) Για κάθε υπόθεση του $S_{t1} + S_{t2}$, ανανέωσε το μοντέλο εμφάνισης και το δυναμικό μοντέλο.

(β) Για κάθε πιθανή τροχιά του S_{t1}, ανανέωσε την αξιοπιστία έναρξής της.

(γ) Για κάθε τροχιά του $S_{t1} + S_{t2}$, θέσε ξανά στην αξιοπιστία τερματισμού την τιμή 0.

(δ) Για κάθε τροχιά του $S_t - S_{t1} - S_{t2}$, ανανέωσε την αξιοπιστία τερματισμού της.

Εξαγωγή όλων των αξιόπιστων τροχιών του *S* ως τελικά αποτελέσματα.
(End of Forward Human Tracking)

Η έναρξη τροχιάς παρουσιάζει κάποια καθυστέρηση, οπότε για αντιστάθμισμα εφαρμόζεται μια διαδικασία οπισθοδρόμισης (ανίχνευσης προς τα πίσω) που είναι η ακριβώς αντίστροφη της ανίχνευσης προς τα εμπρός. Μετά την έναρξη μιας τροχιάς, η ανάπτυξή της μπορεί να γίνει τόσο προς τα εμπρός όσο και προς τα πίσω. Στην περίπτωση όπου καμία παρατήρηση εικόνας δεν είναι διαθέσιμη, και το ίδιο το δυναμικό μοντέλο δεν είναι αρκετά ισχυρό να ακολουθήσει το αντικείμενο, κρατάμε την *υπόθεση* στην τελευταία θέση έως ότου είτε ολοκληρώνεται η υπόθεση είτε κάποιο μέρος από αυτήν βρίσκεται πάλι. Όταν υπάρχει *ολική παρεμβολή*, αλλά για σύντομο χρόνο, τότε το άτομο (που «κρύφτηκε») μπορεί να ληφθεί μέσω συσχετισμού δεδομένων. Εάν η πλήρης παρεμβολή επιμένει, τότε η παρακολούθηση μπορεί να τερματιστεί πρόωρα (τέτοιες περιπτώσεις διακοπής τροχιάς θα μπορούσαν να συνδυαστούν και να αντιμετωπιστούν σε υψηλότερο επίπεδο ανάλυσης).

Μια απλοποιημένη εκδοχή του συνδυασμένου ανιχνευτή τροχιάς είναι η παρακολούθηση ενός μόνο μέρους σώματος κάθε φορά. Ωστόσο ο συνδυασμένος ανιχνευτής τροχιάς είναι πιο αξιόπιστος διότι:

1. Χρησιμοποιεί συνδυασμένες αποκρίσεις εντοπισμού, μεγάλης ακρίβειας, για την έναρξη τροχιάς. Αυτό έχει ως αποτέλεσμα τα πολύ χαμηλά επίπεδα άκυρων εντοπισμών κατά το στάδιο έναρξης.
2. Ο συνδυασμένος ανιχνευτής πορείας προσπαθεί να βρει τις αντίστοιχες αποκρίσεις μερών μιας υπόθεσης αντικειμένου. Η πιθανότητα ότι τουλάχιστον ένας εντοπιστής μέρους σώματος θα «ταιριάξει» είναι σχετικά υψηλή.
3. Προσπαθεί να ακολουθήσει τα αντικείμενα με ανίχνευση των μερών τους, είτε μέσω συσχετισμού δεδομένων, είτε μέσω meanshift. Αυτό κάνει τον ανιχνευτή τροχιάς να αντιμετωπίζει παρεμβολές αντικειμένων αλλά και σκηνής.
4. Λαμβάνει τη μέση τιμή των αποτελεσμάτων ανίχνευσης πορείας των μερών, ως την τελική θέση του ανθρώπου. Ως εκ τούτου, ακόμα κι αν η παρακολούθηση ενός μέρους αποτύχει, η θέση του άνθρωπος μπορεί ακόμα να βρεθεί με ακρίβεια.

4. Video Multimedia Databases, Μελλοντικές Κατευθύνσεις

Η πρόσφατη έρευνα για τα συστήματα ανάκτησης εικόνας είναι βασισμένη στην εισαγωγή κειμένων, όπου οι εικόνες σχολιάζονται από το κείμενο και η ανάκτηση εκτελείται σε αυτό. Εντούτοις, ο χειρωνακτικός σχολιασμός είναι χρονοβόρος, απαιτεί εντατική εργασία και γίνεται μη πρακτικός όταν η συλλογή είναι μεγάλη. Ένα άλλο πρόβλημα είναι η υποκειμενική φύση της λέξης κλειδιού, η ίδια εικόνα/βίντεο μπορεί να σχολιαστεί διαφορετικά από διαφορετικούς σχολιαστές. Επομένως, είναι επιθυμητό να υπάρξει ένα σύστημα διαχείρισης βάσεων τηλεοπτικών δεδομένων, συμπεριλαμβανομένου ενός γραφικού διαδικτυακού συστήματος διεπαφής ερωτήσεων που μπορεί να χειριστεί την ανάμειξη ερωτήσεων χωροχρονικών και σημασιολογικών ιδιοτήτων των πολυμεσικών δεδομένων όπως είναι τα video αρχεία.

4.1 Εξαγωγή Σημασιολογικών Πληροφοριών από Αρχεία Πολυμέσων

Στα συστήματα ανάκτησης με βάση το περιεχόμενο ο στόχος είναι ο έλεγχος και η ανάκτηση σύμφωνα με τα οπτικά χαρακτηριστικά των δεδομένων, όπως το χρώμα, η σύσταση ή η μορφή. Αρκετά συστήματα αναπτύσσονται για την αναζήτηση εικόνας και αρχείων video που χρησιμοποιούν το οπτικό περιεχόμενό τους. Εκτός από μερικά, τα οποία προσπαθούν να προσδιορίσουν πρόσωπα τα αυτοκίνητα, ανθρώπους, και πεζούς, η αντιστοίχηση εικόνων στα περισσότερα συστήματα βάσεων δεδομένων δεν κατευθύνονται συνήθως προς τη σημασιολογία αντικειμένου
αλλά στηρίζονται στα λεγόμενα *χαμηλού επιπέδου χαρακτηριστικά γνωρίσματα*, όπως το χρώμα ή/και τη σύσταση. Οι μελέτες προτιμήσεων χρηστών δείχνουν ότι οι χρήστες δείχνουν αρκετό ενδιαφέρον για τη σημασιολογική ανάκτηση δεδομένων εικόνας. Λόγω των περιορισμών των συστημάτων ανάκτησης είτε μόνο σε κείμενο είτε μόνο σε περιεχόμενο, η χρήση πολυμορφικών δεδομένων, εάν είναι διαθέσιμα, είναι μια στρατηγική για αξιόπιστα αποτελέσματα ανάκτησης. Αποδεικνύεται ότι απόδοση συστημάτων ανάλυσης και κατανόησης πολυμέσων μπορούν να ενισχυθεί σημαντικά με το συνδυασμό διαφορετικών τύπων δεδομένων όπως εικόνα, video, ήχος και κείμενο.

Ο στόχος είναι να αναπτυχθούν εύχρηστα συστήματα πολυμέσων που υιοθετούν μεθόδους αναζήτησης με βάση τις σημασιολογικές έννοιες που είναι φυσικές στο χρήστη, αλλά και με βάση τα κείμενα περιγραφής. Ωστόσο, η εξαγωγή σημασιολογίας από εικόνες είναι ένα πολύ δύσκολο και μακροχρόνιο πρόβλημα. Η εκμάθηση της σημασιολογίας που συνδέεται με τα χαρακτηριστικά γνωρίσματα μιας εικόνας απαιτεί προσεκτική ονομασία δεδομένων, κάτι που είναι πολύ δύσκολο να απαιτηθεί σε μεγάλες ποσότητες αυτών. Αντ' αυτού, στις πρόσφατες μελέτες αποδεικνύεται ότι, τέτοιες σχέσεις μπορούν να μαθευτούν από πολυμορφικά σύνολα δεδομένων που παρέχουν δεδομένα γενικού προσδιορισμού. Πιθανολογικά πρότυπα προτείνονται για την εύρεση των στατιστικών σχέσων των οπτικών και των κειμενικών χαρακτηριστικών γνωρισμάτων.Αποδεικνύεται ότι μαθαίνοντας τους συσχετισμούς μεταξύ των διαφορετικών μορφών δεδομένων παρέχεται καλύτερος σχολιασμός και απόδοση ανάκτησης και συλλαμβάνεται αποτελεσματικότερα η κρυμμένη σημασιολογία. Υπάρχουν παραδείγματα που περιλαμβάνουν αναγνώριση αντικειμένων και προσώπων σε μεγάλες
βάσεις δεδομένων.

4.2 Εξαγωγή Σημασιολογικών Πληροφοριών από Video

Μόλις ολοκληρωθούν οι λεγόμενοι στόχοι χαμηλού επιπέδου επεξεργασίας εικόνας (low-level image processing), όπως είναι ο εντοπισμός, ανίχνευση κ.τ.λ κινούμενων αντικειμένων, τότε το επόμενο βήμα είναι η εξαγωγή σημασιολογικής πληροφορίας από video. Ένας από τους βασικούς στόχους είναι η κατανόηση της ανθρώπινης συμπεριφοράς σε video. Η προσέγγιση σε αυτό το πρόβλημα είναι η ανάλυση του video για την εξαγωγή μερικών διανυσμάτων χαρακτηριστικών και η ταξινόμηση αυτών των χρονικά μεταβαλλόμενων δεδομένων χαρακτηριστικών. Κατά τη διάρκεια της φάσης αναγνώρισης, το εκτιμημένο διανυσματικό σύνολο χαρακτηριστικών συγκρίνεται με μια ομάδα ορισμένων χαρακτηριστικών γνωρισμάτων αναφοράς που αντιπροσωπεύουν χαρακτηριστικές ανθρώπινες ενέργειες. Το πρόβλημα είναι τυπικά ισοδύναμο με την επιλογή μιας φράσης ή μιας πρότασης από ένα πεπερασμένο σύνολο προτάσεων ή φράσεων. Επομένως, η φάση εκπαίδευσης αποτελείται από τη λήψη ακολουθιών διανυσμάτων δράσης αναφοράς που αντιστοιχούν σε κάθε πιθανή "πρόταση" από τα video της εκπαίδευσης. Η διαδικασία εξαγωγής χαρακτηριστικών είναι πολύ σημαντική για την επίτευξη αξιόπιστων αποτελεσμάτων. Είναι αδύνατο να εκτελεστεί εκπαίδευση και ταξινόμηση χρησιμοποιώντας τις σήμερα διαθέσιμες μηχανές ταξινόμησης. Το μέγεθος του διανύσματος χαρακτηριστικών πρέπει να είναι όσο το δυνατόν μικρότερο ώστε να έχει υπολογιστική αποδοτικότητα. Ταυτόχρονα θα πρέπει να αντιπροσωπεύει με ακρίβεια κάθε μία από τις δράσεις. Παραδείγματος χάριν, το κέντρο μάζας του ανιχνευόμενου αντικειμένου μέσα στα πλαίσια εικόνων ενός video μπορεί να χρησιμοποιηθεί ως διάνυσμα χαρακτηριστικών σε μια εφαρμογή ασφάλειας στην οποία ανιχνεύονται κινούμενα πρόσωπα που εισέρχονται ή εξέρχονται από ένα κτήριο. Σε αυτό το απλό πρόβλημα, το "λεξιλόγιο" αποτελείται
από ένα πρόσωπο που εξέρχεται από το κτήριο και ένα πρόσωπο που εισέρχεται σε ένα κτήριο και το κέντρο μάζας, του οποίου οι πληροφορίες αποτελούνται από τις οριζόντιες και κάθετες συντεταγμένες σε μια εικόνα του video και μπορούν να δώσουν ένα καλό διάνυσμα χαρακτηριστικών για αυτό το πρόβλημα. Αφ' ετέρου, ο εντοπισμός ενός πεσμένου ατόμου σε ένα δωμάτιο μπορεί να απαιτήσει μερικές άλλες πρόσθετες παραμέτρους στο σύνολο χαρακτηριστικών. Παραδείγματος χάριν, τα αποκαλούμενα snaxels, τα pixels όπου παρατηρείται το ενεργό περίγραμμα ενός ανθρώπινου σώματος μπορούν να προστεθούν στο διάνυσμα χαρακτηριστικών γνωρισμάτων για τη διάκριση ενός πεσμένου ατόμου από μια συνάντηση ατόμων σε έναν καναπέ. Η πυκνότητα των ορίων του περιγράμματος, η ταχύτητα του κέντρου μάζας κ.λπ... μπορούν να χρησιμοποιηθούν επίσης για να διακρίνουμε την κανονική πράξη του καθίσματος σ' έναν καναπέ και την ανώμαλη δράση ενός πεσμένου ατόμου. Από τη στιγμή που ένα video αποτελείται από μια ακολουθία εικόνων έτσι λαμβάνεται επίσης μια ακολουθία διανυσμάτων χαρακτηριστικών για να χαρακτηρίσουν την κίνηση ενός ατόμου.

Γενικές μέθοδοι ταξινόμησης προτύπων που χρησιμοποιούνται στην κατανόηση ανθρώπινης συμπεριφοράς σε video περιλαμβάνουν:

- Δυναμικό προγραμματισμό (Dynamic Programming)
- Μοντέλα Hidden Markov (Hidden Markov Models, HMM) και Γκαουσσιανών Μιγμάτων (Gaussian Mixture Models, GMM)
- Νευρωνικά Δίκτυα (Neural Networks, NN) και Διανυσματικές Μηχανές Υποστήριξης (Support Vector Machines SVM)
- Ανάλυση Κύριων Τμημάτων (Principal Component Analysis, PCA) και Γραμμικής Διακρίνουσας (Linear Discriminant Analysis, LDA).

Οι ανωτέρω μέθοδοι ταξινόμησης χρησιμοποιούνταν αρχικά κυρίως σε συστήματα αναγνώρισης ομιλίας και ομιλητών.
Ο δυναμικός προγραμματισμός (DP) χρησιμοποιήθηκε στην αντιστοίχηση προτύπων ανθρώπινων κινήσεων . Μπορεί να χρησιμοποιηθεί επιτυχώς για το ταίριασμα μεταξύ του διανύσματος δοκιμής και του αντίστοιχου διανύσματος αναφοράς για όσο κρατούν οι χρονικοί περιορισμοί. Ωστόσο ο DP δεν αξιοποιείται πλήρως στα προβλήματα μηχανικής όρασης, συμπεριλαμβανομένης της ταξινόμησης αντικειμένου με βάση το όριο του περιγράμματος. Το υπολογιστικό κόστος μπορεί να μειωθεί με τη χρησιμοποίηση μερικών ευρετικών μεθόδων βελτιστοποίησης στα προβλήματα DP. Τα HMM είναι πιθανολογικές μηχανές πεπερασμένης φύσης. Στα πλαίσια της ανάλυσης ανθρώπινων κινήσεων, ένα τέτοιο μοντέλο markov ορίζεται για κάθε ένα πιθανό σενάριο και οι παράμετροί του εκπαιδεύονται με διανύσματα χαρακτηριστικών αυτής της τυπικής ανθρώπινης δράσης. Η διαδικασία εκπαίδευσης είναι συνήθως ένας επαναληπτικός αλγόριθμος Baum-Welch . Κατά τη διάρκεια της φάσης ταξινόμησης ή αναγνώρισης, το διανυσματικό σύνολο χαρακτηριστικών δοκιμής εφαρμόζεται σε όλα τα μοντέλα markov και υπολογίζονται οι εξαγόμενες πιθανότητες. Καθορίζεται το μοντέλο markov που παράγει την υψηλότερη πιθανότητα, και επιλέγεται το αντίστοιχο σενάριο ανθρώπινης δράσης ως αποτέλεσμα. Τα HMM και DP δίνουν παρόμοια αποτεκλέσματα σε εφαρμογές αναγνώρισης ομιλίας. Δεδομένου, επίσης, ότι η βασισμένη σε HMM προσέγγιση είναι υπολογιστικά αποδοτικότερη από το δυναμικό προγραμματισμό χρησιμοποιείται στα περισσότερα πρακτικά συστήματα αναγνώρισης ομιλίας. HMM χρησιμοποιούνται όχι μόνο στην ταξινόμηση ανθρώπινης δράσης αλλά και σε άλλους αλγορίθμους ανάλυσης εικόνων και video, για αυτόματη ευρετηρίαση και κατάτμηση (indexing and segmentation), αλλά και επεξεργασία φυσικής γλώσσας.

Νευρωνικά Δίκτυα (Neural Networks, NN), Νευρωνικά Δίκτυα με Χρονοκαθυστέρηση (Time-Delay Neural Networks, TDNN), Κυψελοειδή Νευρωνικά Δίκτυα (Cellular Neural Networks, CNN), και άλλοι σχετικοί μηχανισμοί όπως τα SVM, χρησιμοποιούνται επίσης στην αναγνώριση και ταξινόμηση προτύπων. Εφαρμογές αυτών περιλαμβάνουν ταξινόμηση ανθρώπινων προσώπων σύμφωνα με τις συναισθηματικές εκφράσεις και πρότυπα ακανόνιστων ανθρώπινων κινήσεων.

Στην ταξινόμηση προτύπων χρησιμοποιείται επίσης η Ανάλυση Κύριων Τμημάτων και η Ανάλυση Γραμμικής Διακρίνουσας, ιδιαίτερα για την ανάλυση ανθρώπινου προσώπου.

4.3 Μελλοντικές Ερευνητικές Κατευθύνσεις

Η ανάλυση και αναγνώριση ανθρώπινης κινήσης και δράσης βρίσκεται σε αρχικά στάδια έρευνας παρά το μεγάλο το ποσό εργασίας που έχει εκτελεσθεί σε πολλά διακεκριμένα ερευνητικά κέντρα. Γενικά τα ακόλουθα προβλήματα αξίζουν περαιτέρω προσοχή:

- Πρόβλημα κατάτμησης video
- Παρεμβολή (occlusion) σε video
- Μοντελοποίηση σκηνικού, και εντοπισμός και ανίχνευση τρισδιάστατων αντικειμένων
- Εξαγωγή χαρακτηριστικών μέσω πολλαπλών συσκευών καταγραφής εικόνων
- Αναγνώριση ανθρώπινης δράσης, και αυτόματες ή ημιαυτόματες τεχνικές εκπαίδευσης
- Αυτόματος και ημιαυτόματος σχολιασμός αρχείων εικόνας και video, και μέθοδοι σημασιολογικής αναζήτησης.

- Εντοπισμός ανθρώπινου προσώπου και σώματος σε εικόνα
- Βάσεις δεδομένων αξιολόγησης απόδοσης και δοκιμής

Η τομή του βίντεο σε σημασιολογικού ενδιαφέροντος τομείς, σε πραγματικό χρόνο είναι ένα ακόμα ενεργός μη τετριμμένο ερευνητικό πρόβλημα λόγω των αλλαγών στα επίπεδα φωτισμού, και της ύπαρξης σκιών στις φυσικές σκηνές. Αυτό το βήμα είναι πολύ σημαντικό για να αρχίσει η περαιτέρω διαδικασία ανίχνευσης και εντοπισμού κινήσεων.

Οι παρεμβολές σε φυσικές σκηνές προκαλούν προβλήματα σε πολλούς αλγορίθμους εντοπισμού ανθρώπων. Επομένως απαιτούνται αξιόπιστοι αλγόριθμοι ικανοί να χειριστού παρεμβολλές. Αυτό είναι ειδικά ένα κύριο πρόβλημα στα συστήματα αλληλεπίδρασης ανθρώπου-υπολογιστή αναγνώρισης χειρονομιών. Εάν το χέρι του χρήστη εμποδίζεται μερικώς κατά τη διάρκεια της διαδικασίας αλληλεπίδρασης έπειτα η αναγνώριση χειρονομίας χεριών στα περισσότερα συστήματα αποτυγχάνει. Αλγόριθμοι πρόβλεψης ανθρώπινης στάσης και θέσης, και παρεμβολής ανθρώπινων μερών, είναι απαραίτητοι για τη δημιουργία αξιόπιστων συστημάτων.

Η αυτόματη ή ημιαυτόματη μοντελοποίηση και κατανόηση σκηνών με τη χρήση ποολαπλών συσκευών, και ο καθορισμός της θέσης κινούμενων αντικειμένων και ανθρώπων σε ένα τρισδιάστατο μοντέλο της σκηνής αποτελούν άλυτα προβλήματα. Η εύρεση της ακριβούς τρισδιάστατης θέσης μέσα σε ένα τρισδιάστατο σκηνικό είναι πολύ σημαντική για πολλές εφαρμογές ασφάλειας, και αλληλεπίδρασης ανθρώπου-υπολογιστή.

Οι πρόσφατες προσπάθειες για την ταξινόμηση ανθρώπινων δράσεων στηρίζονται σε τεχνικές παρόμοιες με αυτές που χρησιμοποιούνται σε συστήματα ταξινόμησης για την αναγνώριση ομιλίας. Τέτοια συστήματα μπορούν να κατηγοριοποιήσουν την ανθρώπινη συμπεριφορά εφαρμοσμένα σε καθαρά ελεγχόμενα περιβάλλοντα και τμηματοποιημένα video, και να αντιστοιχήσουν μια δράση σε μία μόνο από τις διαθέσιμες κλάσεις. Αυτόματες ή ημιαυτόματες τεχνικές εκπαίδευσης επίσης χρειάζονται για την ανάλυση ανθρώπινης συμπεριφοράς. Επιπροσθέτως, αλγόριθμοι εξόρυξης δεδομένων για την εύρεση προτύπων αξιόλογης ανθρώπινης συμπεριφοράς και δράσης σε βάσεις δεδομένων πολυμέσων, και συστημάτων ανάκτησης πολυμέσων, βρίσκονται ακόμα σε αρχικά στάδια ανάπτυξης.

Οι μέχρι στιγμής αλγόριθμοι εντοπισμού προσώπου δεν αποδίδουν τέλεια σε όλες τις πιθανές περιπτώσεις, στις οποίες τα πρόσωπα μπορεί να ποικίλουν σε μέγεθος, κατεύθυνση, φωτισμό και στάση. Επίσης, οι περιπτώσεις αυτές αντιμετωπίζονται καλύτερα από τις προσσεγγίσεις που στηρίζονται σε χαρακτηριστικά γνωρίσματα.

Τέλος, εκτός από κάποιες βάσεις δεδομένων ανθρώπινων προσώπων που οργανώθηκαν τα τελευταία χρόνια, δεν υπάρχουν τυπικές βάσεις δεδομένων πολυμέσων για αξιόλογηση όπως στην αναγνώριση ομιλίας και ομιλητή. Η ύπαρξη αυτών είναι όμως απαραίτητη για την αξιολόγηση της απόδοσης ενός αλγορίθμου.

Πρόσθετες Πληροφορίες:

Η ανάλυση των δεδομένων πολυμέσων για την ανίχνευση και τον εντοπισμό ανθρώπων, και την ερμηνεία της ανθρώπινη συμπεριφορά είναι ένας σημαντικός ερευνητικός τομέας στην μηχανική όραση. Τα τελευταία χρόνια, ο εντοπισμός ανθρώπων και η ανάλυση κίνησης, έχουν παρουσιαστεί στα περισσότερα πρωτοπόρα διεθνή επιστημονικά περιοδικά όπως: IJCV (International Journal of Computer Vision), CVIU (Computer Vision and Image Understanding), IEEE PAMI (IEEE Transactions on Pattern Recognition and Machine Intelligence) and IVC (Image and Vision Computing), IEEE Transactions on Image Processing, IEEE Transactions on Multimedia, καθώς επίσης και σε διεθνούς κύρους συνέδρια και εργαστήρια όπως: European Signal Processing Conference (EURASIP), ICCV (International Conference on Computer Vision), CVPR (IEEE International Conference on Computer Vision and Pattern Recognition), ECCV (European Conference on Computer Vision), WACV (Workshop on Applications of Computer Vision) and IWVS (IEEE International Workshop on Visual Surveillance), International Conference of Pattern Recognition (ICPR), IEEE International Conference on Multimedia and Expo, (ICME). IEEE International Conference on Acoustics, Speech, and Signal Processing
(ICASSP), IEEE International Conference on Image Processing (ICIP).

Βιβλιογραφία

1- R.T. Collins, A. Lipton, T. Kanade et al., “A system for video surveillance and monitoring,” Technical Report, CMU-RI-TR-00-12, , Carnegie Mellon University, 2000.

2- R.T. Collins, A.J. Lipton, T. Kanade, Introduction to the special section on video surveillance, IEEE Trans. on Pattern Analysis and Machine Intelligence, 22 (8) (2000) 745-746.

3- R.T. Collins, A.J. Lipton, T. Kanade, "A System for Video Surveillance and Monitoring," in *Proc. American Nuclear Society (ANS) Eighth International Topical Meeting on Robotics and Remote Systems,* Pittsburgh, PA, April 25-29, 1999.

4- I. Haritaoglu, D. Harwood, L.S. Davis, W 4: real-time surveillance of people and their activities, IEEE Trans. on Pattern Analysis and Machine Intelligence, 22 (8) (2000) 809-830.

5-Liang Wang, Weiming Hu, Tieniu Tan, Recent Developments in Human Motion Analysis, *Pattern Recognition*, Vol. 36, No. 3, pp.585-601, 2003.

6. D.M. Gavrila, The visual analysis of human movement: a survey, Computer Vision and Image Understanding, 73 (1) (1999) 82-98.

7- S. Maybank, T. Tan, Introduction to special section on visual surveillance, International Journal of Computer Vision, 37 (2) (2000) 173-173.

8- B.A. Boghossian, S.A. Velastin, Image processing system for pedestrian monitoring using neural classification of normal motion patterns, Measurement and Control, 32 (9) (1999) 261-264.

9- J.J. Little, J.E. Boyd, Recognizing people by their gait: the shape of motion, Videre: Journal of Computer Vision Research, The MIT Press, 1 (2), 1998.

10- J.D. Shutler, M.S. Nixon, C.J. Harris, Statistical gait recognition via velocity moments. Proc. of IEE Colloquium on Visual Biometrics. 2000, pp. 10/1-10/5.

11- P.S. Huang, C.J. Harris, M.S. Nixon, Human gait recognition in canonical space using temporal templates. Proc. of IEE Vis. Image Signal Process. 146 (2) (1999) 93-100.

12-H.M. Lakany, G.M. Haycs, M. Hazlewood, S.J. Hillman, Human walking: tracking and analysis. Proc. of IEE Colloquium on Motion Analysis and Tracking. 1999, pp. 5/1-5/14.

13- M. Köhle, D. Merkl, J. Kastner, Clinical gait analysis by neural networks: issues and experiences. Proc. of IEEE Symp. on Computer-Based Medical Systems. 1997, pp. 138-143.

14- D. Meyer, J. Denzler and H. Niemann, Model based extraction of articulated objects in image sequences for gait analysis. Proc. of IEEE Intl. Conf. on Image Processing. 1997, pp. 78-81.

15- W. Freeman et al., Computer vision for computer games. Proc. of Intl. Conf. on Automatic Face and Gesture Recognition. 1996, pp. 100-105.

16- S. Chang and A. Hsu. Image information systems: Where do we go from here? IEEE Trans. on Knowledge and Data Enginnering, 4(5):431-442, October 1992.

17- M.E. Dönderler, E. _aykol, Ö. Ulusoy, U. Güdükbay, BilVideo: A Video Database Management System, IEEE Multimedia, Vol. 10, No. 1, pp. 66-70, January/March 2003.

18- Yi Li, Songde Ma, Hanqing Lu, Human posture recognition using multi-scale morphological method and Kalman motion estimation. Proc. of IEEE Intl. Conf. on Pattern Recognition. 1998, pp. 175-177.

19- J. Segen, S. Kumar, Shadow gestures: 3D hand pose estimation using a single camera. Proc. of IEEE CS Conf. on Computer Vision and Pattern Recognition. 1999, pp. 479-485.

20- M-H. Yang, N. Ahuja, Recognizing hand gesture using motion trajectories. Proc. of IEEE CS Conference on Computer Vision and Pattern Recognition. 1999, pp. 468-472.

21- Y. Cui, J.J. Weng, Hand segmentation using learning-based prediction and verification for hand sign recognition. Proc. of IEEE CS Conf. on Computer Vision and Pattern Recognition. 1997, pp. 88-93.

22- M. Turk, Visual interaction with lifelike characters. Proc. of IEEE Intl. Conf. on Automatic Face and Gesture Recognition, Killington, 1996, pp. 368-373.
23- C. Stauffer, W. Grimson, Adaptive background mixture models for real-time tracking. Proc. of IEEE CS Conf. on Computer Vision and Pattern Recognition, Vol. 2, 1999, pp. 246-252.
24-Y.H. Yang, M.D. Levine, The background primal sketch: an approach for tracking moving objects, Machine Vision and applications, 5 (1992) 17-34.
25- A. Verri, S. Uras, E. DeMicheli, Motion Segmentation from optical flow. Proc. of the 5th Alvey Vision Conference. 1989, pp. 209-214.
26- J. Barron, D. Fleet, S. Beauchemin, Performance of optical flow techniques, International Journal of Computer Vision, 12 (1) (1994) 42-77.
27- H.A. Rowley, J.M. Rehg, Analyzing articulated motion using expectation-maximization. Proc. of Intl. Conf. on Pattern recognition. 1997, pp. 935-941.
28- Branko Ristic, Sanjeev Arulampalam, Neil Gordon, Beyond the Kalman Filter: Particle Filters for Tracking Applications, Artech House Radar Library, 2004
29- M. Isard, A. Blake, Condensation—conditional density propagation for visual tracking, International Journal of Computer Vision, 29 (1) (1998) 5-28.
30- Comaniciu, D., Ramesh, V. and Meer, P., "Real-Time Tracking of Non-Rigid Objects using Mean Shift," *IEEE Computer Vision and Pattern Recognition*, Vol II, 2000, pp.142-149.
31- V. Pavlovi_, J.M. Rehg, T-J. Cham, K.P. Murphy, A dynamic Bayesian network approach to figure tracking using learned dynamic models. Proc. of Intl. Conf. on Computer Vision. 1999, pp. 94-101.
32- J. Yang, A. Waibel, A real-time face tracker. Proc. of IEEE CS Workshop on Applications of Computer Vision. Sarasota, FL, 1996, pp. 142-147
33- C.R. Wren, A. Azarbayejani, T. Darrell, A. P. Pentland, Pfinder: real-time tracking of the human body, IEEE Trans. on Pattern Analysis and Machine Intelligence, 19 (7) (1997) 780-785. 996, pp. 142-147.
34- M-H. Yang, N. Ahuja, Recognizing hand gesture using motion trajectories. Proc. of IEEE CS Conference on Computer Vision and Pattern Recognition. 1999, pp. 468-472.
35- J. Rehg, T. Kanade, Visual tracking of high DOF articulated structures: an application to human hand tracking. Proc. of European Conference on Computer Vision. 1994, pp. 35-46.
36- D. Meyer et al., Gait classification with HMMs for Trajectories of body parts extracted by mixture densities. British Machine Vision Conference. 1998, pp. 459-468.
37- P. Fieguth, D. Terzopoulos, Color-based tracking of heads and other mobile objects at video frame rate. Proc. of IEEE CS Conf. on Computer vision and Pattern Recognition. 1997, pp. 21-27.
38- D-S. Jang, H-I. Choi, Active models for tracking moving objects, Pattern Recognition, 33 (7) (2000) 1135-1146.
39- S. Wachter, H-H. Nagel, Tracking persons in monocular image sequences, Computer Vision and Image Understanding, 74 (3) (1999) 174-192.
40- J.M. Rehg, T. Kanade, Model-based tracking of self-occluding articulated objects. Proc. of 5th Intl. Conf. on Computer Vision. Cambridge, 1995, pp. 612-617.
41. I.A. Kakadiaris, D. Metaxas, Model-based estimation of 3-D human motion with occlusion based on active multiviewpoint
selection. Proc. of IEEE Conf. on Computer Vision and Pattern Recognition. San Francisco, 1996, pp. 81-87.
42- N. Goddard, Incremental model-based discrimination of articulated movement from motion features. Proc. of IEEE Workshop on Motion of Non-Rigid and Articulated Objects. Austin, 1994, pp. 89-94.
43- J. Badenas, J. Sanchiz, F. Pla, Motion-based segmentation and region tracking in image sequences, Pattern Recognition, 34 (2001) 661-670.

44- A. Baumberg, D. Hogg, An efficient method for contour tracking using active shape models. Proc. of IEEE Workshop on Motion of Non-Rigid and Articulated Objects. Austin, 1994, pp. 194-199.
45- M. Isard, A. Blake, Contour tracking by stochastic propagation of conditional density. Proc. of European Conference on Computer Vision. 1996, pp. 343-356.
46- N. Paragios, R. Deriche, Geodesic active contours and level sets for the detection and tracking of moving objects, IEEE Trans. on Pattern Analysis and Machine Intelligence, 22 (3) (2000) 266-280.
47- R. Cutler, L.S. Davis, Robust real-time periodic motion detection, analysis, and applications, IEEE Trans. on Pattern Analysis and Machine Intelligence, 22 (8) (2000) 781-796.
48- M. Oren, C. Papageorgiou, P. Sinha, and E. Osuna. Pedestrian detection using wavelet templates, Proc. of IEEE CS Conf. Computer vision and Pattern Recognition. 1997, pp. 193-199.
49- C. Stauffer, Automatic hierachical classification using time-base co-occurrences. Proc. of IEEE CS Conf. on Computer Vision and Pattern Recognition. 1999, pp. 333-339.
50- S.A. Niyogi, E.H. Adelson, Analyzing and recognizing walking figures in XYT. Proc. of IEEE Conf. on Computer Vision and Pattern Recognition. 1994, pp. 469-474.
51- H. Fujiyoshi, A.J. Lipton, Real-time human motion analysis by image skeletonization. Proc. of IEEE Workshop on Applications of Computer Vision. 1998, pp. 15-21.
52- I-C. Chang, C-L. Huang, Ribbon-based motion analysis of human body movements. Proc. of Intl. Conf. on Pattern Recognition. Vienna, 1996, pp. 436-440.
53- A. Geurtz. Model-based Shape Estimation. PhD thesis, Department of Electrical Engineering, Polytechnic Institute of Lausanne, 1993.
54-A.J. Lipton, Local application of optic flow to analyse rigid versus non-rigid motion. In the website
http://www.eecs.lehigh.edu/FRAME/Lipton/iccvframe.html.
55- A. Selinger, L. Wixson, Classifying moving objects as rigid or non-rigid without correspondences. Proc. of DAPRA Image Understanding Workshop, Vol. 1, 1998, pp. 341-358
56-169. A. Shio and J. Sklansky. Segmentation of people in motion. In IEEE Workshop on Visual Motion, pages 325- 332, 1991.
57- A. Iketani et al., Detecting persons on changing background. Proc. of International Conference on Pattern Recognition, Volume 1: 74-76, 1998.
58- C. Cedras, M. Shah, Motion-based recognition: a survey, Image and Vision Computing, 13 (2) (1995) 129-155.
59- A.M. Elgammal, L.S. Davis, Probabilistic framework for segmenting people under occlusion. Proc. of International Conference on Computer Vision, 2001.
60- I.A. Karaulova, P.M. Hall, A.D. Marshall, A hierarchical model of dynamics for tracking people with a single video camera. British Machine Vision Conference. 2000, pp. 352-361.
61- Y. Guo, G. Xu, S. Tsuji, Tracking human body motion based on a stick figure model, Visual communication and Image Representation, 1994, 5: 1-9.
62- C.R. Wren, B.P. Clarkson, A. Pentland, Understanding purposeful human motion. Proc. of Intl. Conf. on Automatic Face and Gesture Recognition. France, March 2000.
63- Y. Luo, F.J. Perales, J. Villanueva, An automatic rotoscopy system for human motion based on a biomechanic graphical model, Comput. Graphics 16 (4) (1992) 355-362.
64- C. Yaniz, J. Rocha, F. Perales, 3D region graph for reconstruction of human motion. Proc. of Workshop on Perception of Human Motion at ECCV, 1998.
65- C. Vogler, D. Metaxas, ASL recognition based on a coupling between HMMs and 3D motion analysis. Proc. of International Conference on Computer Vision. 1998, pp. 363-369.
66- M.K. Leung, Y.H. Yang, First sight: a human body outline labeling system, IEEE Trans. on Pattern Analysis and Machine Intelligence, 17 (4) (1995) 359-377.

67- I-C. Chang, C-L. Huang, Ribbon-based motion analysis of human body movements. Proc. of Intl. Conf. on Pattern Recognition. Vienna, 1996, pp. 436-440.
68- S. Ju, M. Black, Y. Yaccob, Cardboard people: a parameterized model of articulated image motion. Proc. of IEEE Intl. Conf. on Automatic Face and gesture Recognition. 1996, pp. 38-44.
69- C. Hu et al., Extraction of parametric human model for posture recognition using generic algorithm. Proc. of the Fourth Intl. Conf. on Automatic Face and Gesture Recognition. France, March 2000.
70- M. Kass, A. Witkin, D. Terzopoulos, Snakes: Active contour models, International Journal of Computer Vision, Vol. 2, No. 3, pp. 321-331, 1988
71- D. Meyer, J. Denzler, H. Niemann, Model based extraction of articulated objects in image sequences. Proc. of the Fourth IEEE International Conference on Image Processing, 1997.
72- Y. Zhong, A.K. Jain, M. P. Dubuisson-Jolly, Object tracking using deformable templates, IEEE Trans. on Pattern Analysis and Machine Intelligence, 22 (5) (2001) 544-549.
73- S. Wachter, H-H. Nagel, Tracking persons in monocular image sequences, Computer Vision and Image Understanding, 74 (3) (1999) 174-192.
74- J.M. Rehg, T. Kanade, Model-based tracking of self-occluding articulated objects. Proc. of 5th Intl. Conf. on Computer Vision. Cambridge, 1995, pp. 612-617.
75- I.A. Kakadiaris, D. Metaxas, Model-based estimation of 3-D human motion with occlusion based on active multiviewpoint selection. Proc. of IEEE Conf. on Computer Vision and Pattern Recognition. San Francisco, 1996, pp. 81-87.
76- C. Bregler, J. Malik, Tracking people with twists and exponential maps. Proc. of IEEE CS Conf. on Computer Vision and Pattern Recognition, 1998.
77- O. Munkelt et al., A model driven 3D image interpretation system applied to person detection in video images. Proc. of International Conference on Pattern Recognition, 1998.
78- Q. Delamarre, O. Faugeras, 3D articulated models and multi-view tracking with silhouettes. Proc. of International Conference on Computer Vision. Greece, September 1999.
79- D. Metaxas and D. Terzopoulos. Shape and nonrigid motion estimation through physics-based synthesis. IEEE Transactions on Pattern Analysis and Machine Intelligence , 15(6):580-591, 1993.
80- I. Biedermann. Human image understanding. *Computer Vision, Graphics, and Image Processing*, 32:29-73, 1985
81- D. Terzopoulos, Andrew Witkin, and Michael Kass. Constraints on deformable models: Recovering 3d shape and nonrigid motion. Artificial Intelligence, 36,1, 1988.
82- Anil K. Jain**,**Chitra Dorai, 3D object recognition: Representation and matching, Statistics and Computing, 10 (2): 167-182, April 2000
83- A. Pentland. Automatic extraction of deformable models. International Journal of Computer Vision, 4:107-126, 1990.
84- A. Pentland, B. Horowitz, *Recovery of non-rigid motion and structure*, IEEE Transactions on Pattern Analysis and Machine Intelligence 13 (1991) 730 -- 742.
85-Rein-Lien Hsu, Mohamed Abdel-Mottaleb, Anil K. Jain, Face Detection in Color Images, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 707–711, May 2003.
86- E. Saber and A.M. Tekalp, “Frontal-view face detection and facial feature extraction using color, shape and symmetry based cost functions,” *Pattern Recognition Letters*, vol. 19, no. 8, pp. 669–680, 1998.
87-Q. Chen, H. Wu, and M. Yachida. Face detection by fuzzy pattern matching. In *Proc. 5th Int. Conf. On Comp. Vision*, pages 591-596, MIT, Cambridge, MA., 1995.
88- Y. Dai, Y. Nakano, and H. Miyao. Extraction of facial images from a complex background using SGLD matrices. In *Proc. Int. Conf. on Pattern Recognition*, volume A, pages 137-141, Jerusalem, 1994. IEEE Comp. Soc. Press.

89- A. Lanitis, C. J. Taylor, and T.F. Cootes. An automatic face identification system using flexible appearnace models. *Image and Vision Computing*, 13(5):393-401, 1995.
90-K. K. Sung and T. Poggio. Example-based learning for view-based human face detection. Technical Report A.I. Memo 1521, CBLC Paper 112, MIT, Dec. 1994.
91-G. Yang and T. S. Huang. Human face detection in a complex background. *Pattern Recognition*, 27(1):53-63, 1994.
92- P. Perona. Steerable-scalable kernels for edge detection and junction analysis. In G. Sandini, editor, *Proc. 2nd European Conf. on Comp. Vision*, pages 3-18, Italy, 1992. Springer-Verlag.
93- W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE Trans. On Patt. Analy. And Machine Intell.*, 13(9):891-906, 1991.
94- T. Leung, M.Burl, and P. Perona. Finding faces in cluttered scenes using labelled randim graph matching. In *Proc. 5th Int. Conf. on Comp. Vision*, pages 637-644, MIT, Cambridge, MA., 1995.
95- K. C. Yow and R.Cipolla. finding initial estimates of human face location. In *Proc. 2nd Asian Conf. on Comp. Vision*, volume 3, pages 514-518, Singapore, 1995.
96- Y. Rui, T. S. Huang, and S. Chang. Image retrieval: Current techniques, promising directions, and open issues. Journal of Visual Communication and Image Representation, 10(4):39, 62, 1999.
97- H. Schneiderman and T.Kanade. A statistical approach to 3D object recognition applied to faces and cars. In IEEE Conference on Computer Vision and Pattern Recognition, page 100, 2000.
98- M.M. Fleck, D.A. Forsyth, and C.Bregler. Finding naked people. In 4th European Conference on Computer vision, volume 2, pages 591-602, 1996.
99- S. Ornager, View a picture. Theoretical image analysis and empirical studies on idexing and retrieval, Swedish Library Research, vol.2, pp.31-41, 1996.
100- M. Markkula, E. Sormunen, End-user searching challenges indexing pictures in the digital newspaper photo archive, Information retrieval, vol.1, pp.259-285, 2000.
101- C.G.M. Snoek and M. Worring, Multimodal Video Indexing: A Review of the State-of-the-art, Multimedia Tools and Applications, 2005 (in press).
102- I. A. Erdem, M. E. Erdem, Volkan Atalay, A. E. Cetin, Vision-based continuous Graffiti-like text entry system, Optical Engineering, 43(03) p. 553-558, March 2004.
103- A. Hauptmann et.al., Informedia at TRECVID 2003: Analyzing and Searching Broadcast News Video Proceedings of (VIDEO) TREC 2003 (Twelfth Text Retrieval Conference), Gaithersburg, MD, November 17-21, 2003
104- C.-Y. Lin et.al., IBM Research TRECVID-2003 Video Retrieval System, Proceedings of (VIDEO) TREC 2003 (Twelfth Text Retrieval Conference), Gaithersburg, MD, November 17-21, 2003
105- O. Maron and A. L. Ratan Multiple-Instance Learning for Natural Scene Classification, International Conference on Machine Learning (ICML-98), 1998.
106- J. Jeon, V. Lavrenko and R. Manmatha, Automatic Image Annotation and Retrieval using cross-media relevance models, ACM SIGIR, 2003.
107- J. Li, J. Z. Wang, Automatic linguistic indexing of pictures by a statistical modeling approach, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 25, no. 10, 14 pp., 2003.
108- Y. Mori, H. Takahashi, R. Oka, Image-to-word transformation based on dividing and vector quantizing imageswith words, MISRM'99 First International Workshop on Multimedia Intelligent Storage and Retrieval Management October 30th, 1999 Orlando, Florida, in conjonction with ACM Multimedia Conference 1999

109- P. Duygulu, K. Barnard, N.d. Freitas, and D. A. Forsyth, Object recognition as machine translation: learning a lexicon for a fixed image vocabulary, In Seventh European Conference on Computer Vision (ECCV), volume 4, pages
97-112, Copenhagen, Denmark, May 27 - June 2 2002.
110- K. Barnard, P. Duygulu, N. de Freitas, D. A. Forsyth, D. Blei, and M. Jordan, Matching words and pictures, Journal of Machine Learning Research, 3:1107-1135, 2003.
111- P. Duygulu, H. Wactlar, Associating video frames with text, Multimedia Information Retrieval Workshop, in conjuction with the 26th annual ACM SIGIR conference on Information Retrieval, August 1, 2003, Toronto, Canada
112- P. Duygulu, Alex Hauptmann, What's news, what's not? Associating News videos with words, The 3rd International Conference on Image and Video Retrieval (CIVR 2004) Ireland, July 21-23, 2004.
113- J.-Y. Pan, H.J. Yang, C. Faloutsos, P. Duygulu, Automatic Multimedia Crossmodal Correlation Discovery, The Tenth ACMSIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2004) Seattle, WA, August 22-25, 2004.
114- T. Miller, A.C. Berg, J. Edwards, M. Maire, R. White, Y.W. Teh, E. Learned- Miller, D.A. Forsyth, Faces and names in the news, International Conference on Computer Vision and pattern recognition, 2004.
115- A. Bobick, A. Wilson, A state-based technique for the summarization and recognition of gesture. Proc. of Intl. Conf. on Computer Vision. Cambridge, 1995, pp. 382-388.
116- K. Takahashi, S. Seki et al., Recognition of dexterous manipulations from time varying images. Proc. of IEEE Workshop on Motion of Non-Rigid and Articulated Objects. Austin, 1994, pp. 23-28.
117- L. Rabinier, A tutorial on hidden Markov models and selected applications in speech recognition. Proc. of IEEE 77 (2) (1989) 257-285.
118- T. Starner, A. Pentland, Real-time American Sign Language recognition from video using hidden Markov models. Proc. of Intl. Symp. on Computer Vision. 1995, pp. 265-270.
119- J. Yamato, J. Ohya, K. Ishii, Recognizing human action in time-sequential images using hidden Markov model. Proc. of IEEE Conf. on Computer Vision and Pattern Recognition. 1992, pp. 379-385.
120- M. Brand, N. Oliver, A. Pentland, Coupled hidden Markov models for complex action recognition. Proc. of IEEE CS Conf. on Computer Vision and Pattern Recognition. 1997, pp. 994-999.
121- Ozer, O.F.; Ozun, O.; Tuzel, C.O.; Atalay, V.; Cetin, A.E.;Vision-based single-stroke character recognition for wearable computing, IEEE Intelligent Systems,Volume: 16 , Issue: 3 , Pages:33 – 37, May-June 2001
122- A. M. Bagci, R. Ansari, A. Khokhar, A. E. Cetin**,** Eye Tracking Using Markov Models, Int. Conf. Pattern Recognition (ICPR), August 23-26,2004.
123- J. Nam, A. E. Cetin, A. Tewfik, "Speaker Identification and Analysis for Hierarchical Video Shot .Classification, Proc. of ICASSP, Munich, 1997, vol. 4, pp. 2597-2600.
124- Bala, E.; Cetin, A.E.; Computationally Efficient Wavelet Affine Invariant Functions for Shape Recognition IEEE Trans. on Pattern Analysis and Machine Intelligence,Volume: 26 , Issue: 8 , Pages:1095 - 1099Aug. 2004
125- H.Ramoser, T.Schlögl, C.Beleznai, M.Winter, H.Bischof; "Shape-based Detection of Humans for Video Surveillance Applications"; Proc. IEEE Int. Conf. on Image Processing; Vol. III; pp. 1013-1016; Barcelona, Spain; 2003
126- C.Beleznai, B.Frühstück, H.Bischof; "Human Detection in Groups Using a Fast Mean Shift Procedure"; Proc. Int. Conf. On Image Processing ICIP, 2004.
127- T. Schlögl, B.Wachmann, H.Bischof, W.Kropatsch; "People Counting in Complex Scenarios"; Proc. Workshop of the AAPR/ÖAGM; pp. 159-166; Graz, Austria; 2002

128- T. Szirányi, J. Zerubia, L. Czúni, D. Geldreich, Z. Kato: ' Image Segmentation Using Markov Random Field Model in Fully Parallel Cellular Network Architectures,' Real-Time Imaging, Vol. 6, No. 3, pp. 195-211, 2000
129- A. Licsár, T. Szirányi, "Hand Gesture Recognition in Camera-Projector System," International Workshop on Human-Computer Interaction, Lecture Notes on Computer Science, Vol. LNCS 3058, pp.83-93, 2004
130- M. Rosenblum, Y. Yacoob, L. Davis, Human emotion recognition from motion using a radial basis function network architecture. Proc. of IEEE Workshop on Motion of Non-Rigid and Articulated Objects. Austin, 1994, pp. 43-49.
131- C-T. Lin, H-W. Nein, W-C. Lin, A space-time delay neural network for motion recognition and its application to lipreading. International Journal of Neural Systems, 9 (4) (1999) 311-334. 159.
132- O. Chomat, J.L. Crowley, Recognizing motion using local appearance. International Symposium on Intelligent Robotic Systems, University of Edinburgh, 1998.
133- Chellappa, Veeraraghavan, Aggarwal, Pattern Recognition in Video. University of Maryland, 2005.
134-Wu, Nevatia, Detection and Tracking of Multiple, Partially Occluded Humans by Bayesian Combination of Edgelet based Part Detectors, University of Southern California, 2006.