



Όραση Υπολογιστών

2η Εργαστηριακή Άσκηση: Εκτίμηση Οπτικής Ροής (Optical Flow) και Εξαγωγή Χαρακτηριστικών σε Βίντεο

8ο Εξάμηνο - Εαρινό εξάμηνο 2018-19 - Ροή Σ

Αντωνιάδης Παναγιώτης (03115009 - e115009@central.ntua.gr)
Μπαζώτης Νικόλαος (03115739 - e115739@central.ntua.gr)

Περιεχόμενα

1 Παρακολούθηση Προσώπου με Χρήση του Πεδίου Οπτικής Ροής (Optical Flow) με τη Μέθοδο των Lucas-Kanade	2
1.1 Ανίχνευση Δέρματος Προσώπου	2
1.2 Τλοποίηση του Αλγόριθμου των Lucas-Kanade	5
1.3 Πολύ-κλιμακωτός υπολογισμός οπτικής ροής	6
1.4 Υπολογισμός της Μετατόπισης του Προσώπου από το Πεδίο Οπτικής Ροής	7
1.5 Συμπεράσματα και Συγκρίσεις Αποτελεσμάτων SingleScale – MultiScale	9
2 Εντοπισμός Χωρο-χρονικών Σημείων Ενδιαφέροντος και Εξαγωγή Χαρακτηριστικών σε Βίντεο Ανθρωπίνων Δράσεων	11
2.1 Χωρο-χρονικά Σημεία Ενδιαφέροντος	11
2.2 Χωρο-χρονικοί Ιστογραφικοί Περιγραφητές	17
2.3 Κατασκευή Δενδρογράμματος για τον Διαχωρισμό των Δράσεων	18
Βιβλιογραφία	23

1 Παρακολούθηση Προσώπου με Χρήση του Πεδίου Οπτικής Ροής (Optical Flow) με τη Μέθοδο των Lucas-Kanade

Σκοπός της εργαστηριακής άσκησης είναι η υλοποίηση ενός συστήματος Παρακολούθησης Προσώπου (Face Tracking) σε μια ακολουθία βίντεο νοηματικής γλώσσας. Το σύστημα αρχικά θα ανιχνεύει στο πρώτο πλαίσιο την περιοχή του προσώπου με χρήση ενός πιθανοτικού ανιχνευτή ανθρώπινου δέρματος. Στη συνέχεια θα μπορεί να παρακολουθεί αυτή την περιοχή του προσώπου χρησιμοποιώντας το διανυσματικό πεδίο οπτικής ροής, υπολογισμένο με τη μέθοδο Lucas-Kanade.



Εικόνα 1: 1o frame του Face Tracking

1.1 Ανίχνευση Δέρματος Προσώπου

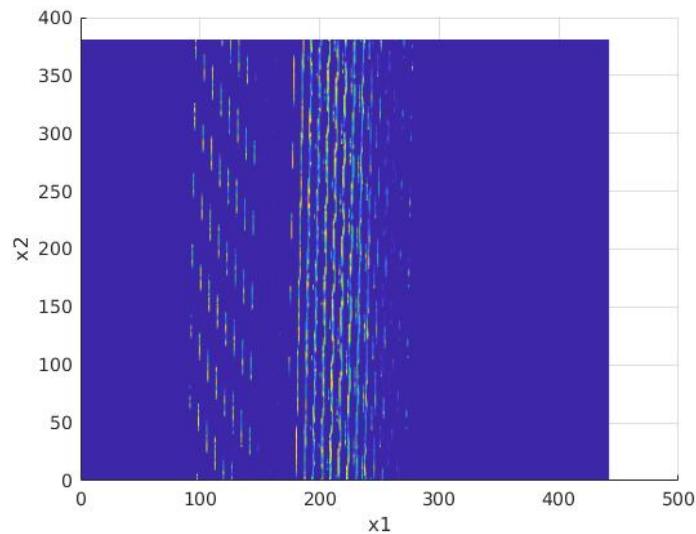
Στο σημείο αυτό θέλουμε να ανιχνεύσουμε την περιοχή του προσώπου. Αρχικά, θα εντοπίσουμε τις περιοχές στις οποίες εμφανίζεται δέρμα και στη συνέχεια θα υποθέσουμε ότι η περιοχή του προσώπου είναι αυτή που έχει την μεγαλύτερη επιφάνεια.

Για την ανίχνευση του δέρματος χρησιμοποιείται ο χρωματικός χώρος $YCbCr$ από το οποίο όμως θα αφαιρέσουμε το πεδίο της φωτεινότητας Y . Το διάνυσμα $c = \begin{vmatrix} C_b \\ C_r \end{vmatrix}$ θα περιγράφει την ταυτότητα του δέρματος. Ακόμη, γνωρίζουμε ότι τα χρώματα του δέρματος μοντελοποιούνται από την παρακάτω 2D-Gaussian:

$$P(c = skin) = \frac{1}{\sqrt{|\Sigma|(2\pi)^2}} e^{-\frac{1}{2}(c-\mu)\Sigma^{-1}(c-\mu)'}$$

όπου $\mu = \begin{vmatrix} \mu_{Cb} \\ \mu_{Cr} \end{vmatrix}$ είναι η μέση τιμή και Σ ο 2×2 πίνακας συνδιαχύμανσης, τα οποία προκύπτουν από τα δείγματα δέρματος που μας δίνονται.

Παρακάτω φαίνεται η επιφάνεια της Γκαουσιανής που προέκυψε μετά από την εύρεση των παραπάνω από τα δείγματα *skinSamplesRGB.mat* που μας δώθηκαν:



Εικόνα 2: Επιφάνεια Γκαουσιανής που εκπειδέυσαμε για την ανίχνευση δέρματος

Επίσης, αξολουθεί η εικόνα πιθανότητας του δέρματος πάνω στην παραπάνω Γκαουσιανή:

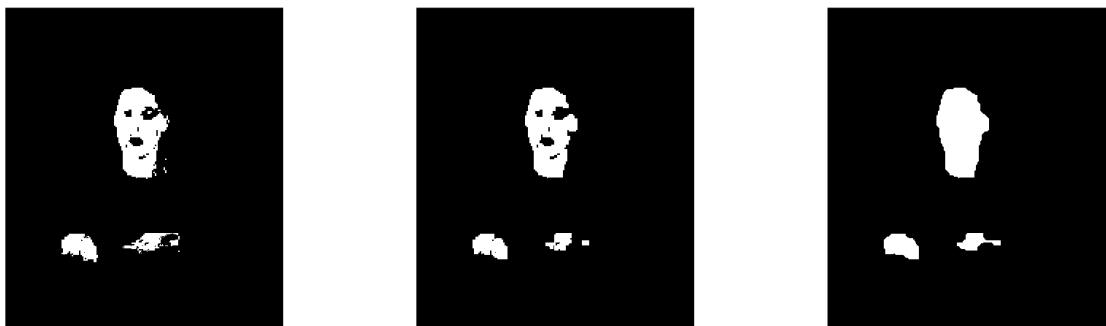


Εικόνα 3: Εικόνα πιθανότητα του δέρματος (πολλαπλασιασμένη με 128 για να γίνει grayscale εικόνα)

Η τελική ανίχνευση της περιοχής δέρματος του προσώπου γίνεται επιλέγοντας την περιοχή με το μεγαλύτερο

εμβαδό από όσες βρέθηκαν. Για το σκοπό αυτό απαιτείται μια μορφολογική επεξεργασία της δυαδικής εικόνας δέρματος και συγκεκριμένα κάλυψη των τρυπών, opening με πολύ μικρό δομικό στοιχείο και closing με μεγάλο δομικό στοιχείο, έτσι ώστε να εξαλειφθούν οι μικρές περιοχές και να αποκτήσουν συνοχή οι περιοχές του προσώπου και των χεριών. Το ορθογώνιο που περιβάλλει την τελική περιοχή δέρματος του προσώπου (bounding box) είναι το παράθυρο της εικόνας που θα χρησιμοποιηθεί στο Μέρος 2 για υπολογισμό του πεδίου Οπτικής Ροής και την τελική παρακολούθηση του προσώπου.

Παρακάτω φαίνονται τα ενδιάμεσα στάδια της επεξεργασίας μέχρι την τελική επιλογή του προσώπου:



Εικόνα 4: Δυαδική εικόνα πρώτου frame: (α) πριν την μορφολογική επεξεργασία, (β) μετά από opening για κάλυψη των τρυπών και τέλος (γ) μετά από closing για να εξαλειφθούν οι μικρές περιοχές.

Αφού έχουμε εντοπίσει τις περιοχές δέρματος χωρίς ασυνέχειες, θεωρούμε πως η περιοχή του προσώπου είναι αυτή που καταλαμβάνει την μεγαλύτερη επιφάνεια από τις υπόλοιπες. Συνεπώς, ορίζουμε αυτήν την περιοχή ως το bounding box. Λαμβάνοντας υπόψιν πλέον την αρχική εικόνα εφαρμόζουμε πάνω της το bounding box που βρήκαμε, όπως βλέπουμε παρακάτω:



Εικόνα 5: Τελική ανίχνευση προσώπου επιλέγοντας την περιοχή με το μεγαλύτερο εμβαδόν.

1.2 Υλοποίηση του Αλγόριθμου των Lucas-Kanade

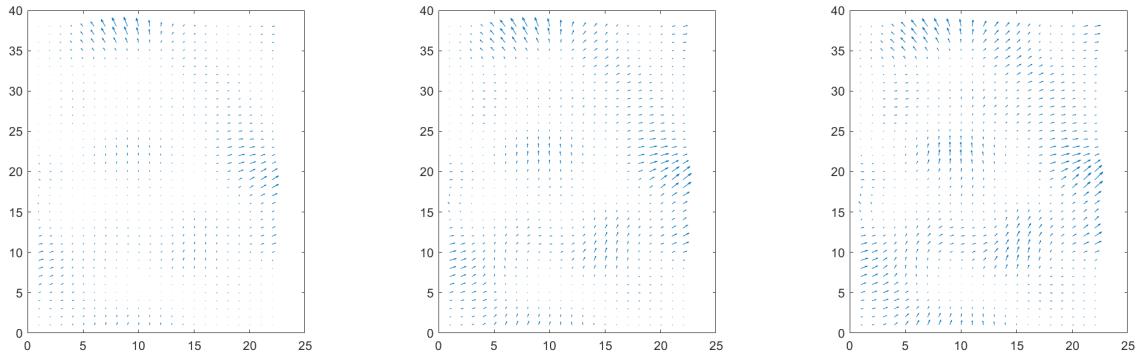
Όπως αναφέρθηκε παραπάνω, σκοπός μας είναι η παρακολούθηση του προσώπου κατά την διάρκεια του βίντεο. Συνεπώς, θέλουμε να εκτιμήσουμε την μετατόπιση του αντικειμένου από το $i - 1$ frame στο i frame, ούτως ώστε να έρθουν σε αντιστοιχία $I_n(x) \approx I_{n-1}(x + d)$. Το μέγεθος αυτό συνιστά την οπτική ροή. Πιο συγκεκριμένα, η οπτική ροή που μας ενδιαιφέρει είναι αυτή του προσώπου. Επομένως δίνουμε ως είσοδο στον αλγόριθμο το bounding box του 1ου ερωτήματος για να υπολογίσει την οπτική ροή. Στην συνέχεια, ο αλγόριθμος επαναληπτικά προσπαθεί να βρει μια προσέγγιση του d , η οποία να ελαχιστοποιεί την διαφορά ανάμεσα στα δύο frames.

Πιο συγκεκριμένα, όσων αφορά την οπτική ροή $d(x) = (d_x, d_y)$ προσπαθούμε να εντοπίσουμε επαναληπτικά το d ελαχιστοποιώντας το τετραγωνικό σφάλμα J_x το οποίο ισούται:

$$J_x(d) = \int_{x' \in R^2} G_p(x - x')[I_n(x') - I_{n-1}(x' + d)]^2 dx'$$

όπου G_p : Gaussian με τυπική απόκλιση p .

Σε κάθε επανάληψη γίνεται προσέγγιση του $d_{i+1} = d_i + u$, όπου i ο γύρος των επαναλήψεων και u το διάνυσμα το οποίο προστίθεται για να βελτιώσει την προσέγγιση του d . Παρακάτω βλέπουμε για τα δύο πρώτα frame πως έχει αλλάξει η οπτική ροή από την 1η επανάληψη έως την 20η.

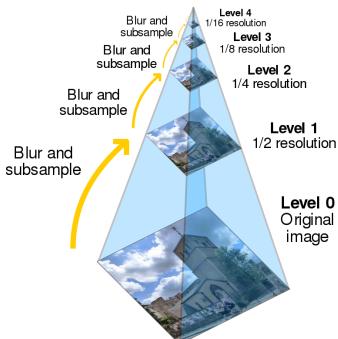


Εικόνα 6: Οπτική ροή από το 1o στο 2o frame (α) στην 1η επανάληψη, (β) στην 10η επανάληψη και (γ) στην 20η επανάληψη.

1.3 Πολύ-χλιμακωτός υπολογισμός οπτικής ροής

Η ιδέα του αλγόριθμου μονής χλίμακας των Lucas-Kanade είναι αποδοτική για μικρές κινήσεις. Ωστόσο, η ανάγκη για υπολογισμό της οπτικής ροής σε μεγαλύτερες κινήσεις που υπερβαίνουν κατά πολύ το 1 pixel είναι μεγάλη και οι παραδοχές που είχαν γίνει στον αλγόριθμο μονής χλίμακας για τους όρους πρώτης τάξης δεν ισχύουν. Η απάντηση στο πρόβλημα αυτό έρχεται από τον πολυχλιμακωτό αλγόριθμό των Lucas-Kanade.

Η νέα εκδοχή του αλγόριθμου χρησιμοποιεί γκαουσιανές πυραμίδες και υπολογίζει την οπτική ροή από τις πιο μεγάλες σε χλίμακα (μικρές σε ανάλυση) εικόνες προς τις πιο μικρές σε χλίμακα (μεγάλες σε ανάλυση εικόνες). Αρχικά υπολογίζει την οπτική ροή στην κορυφαία χλίμακα χρησιμοποιώντας τον κλασικό αλγόριθμο Lucas-Kanade και το αποτέλεσμα του το χρησιμοποιούμε (έπειτα από bilinear interpolation) ως αρχική συνθήκη στον υπολογισμό της οπτικής ροής της επόμενης χλίμακας και συνεχίζεται η διαδικασία έως ότου υπολογίσουμε την οπτική ροή—δια την κανονικής χλίμακας εικόνα.



Γκαουσιανές πυραμίδες: Οι γκαουσιανές πυραμίδες προκύπτουν από την υποδειγματοληψία της εικόνας μειώνοντας την ανάλυση κάθε φορά στο μισό. Ωστόσο για να μετριαστεί η φασματική αναδίπλωση (aliasing) της εικόνας πριν την υποδειγματοληψία φιλτράρουμε την εικόνα με ένα βαθυπερατό φίλτρο, το οποίο στην περίπτωση μας είναι η γκαουσιανή G_p με τυπική απόκλιση 3 pixels.

Πολυπλοκότητα: Αξίζει να σημειωθεί ότι ο αλγόριθμος Multiscale Lucas-Kanade έχει πολυπλοκότητα $O(scales * lk)$ όπου lk ο αλγόριθμος μονής χλίμακας του ερωτήματος 1.2. Ο αλγόριθμος μονής χλίμακας έχει μια

παράμετρο και η οποία προσδιορίζει το πόσες φορές θα προσπαθήσει να κάνει επαναπροσέγγιση της οπτικής ροής. Ενδεικτικά αναφέρεται ότι συγχλίνει γύρω στις 10-15 φορές συνεπώς ο lk έχει πολυπλοκότητα $O(\kappa \text{ φορές } * d)$. Άρα συνολικά ο Multiscale Lucas-Kanade έχει ($scales * \kappa * \text{προσδιορισμός } -d$). Ενώ το μεγάλο και μας δίνει καλύτερες προσεγγίσεις μας αυξάνει πολύ την πολυπλοκότητα.

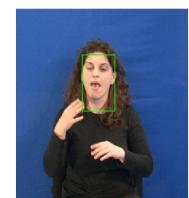
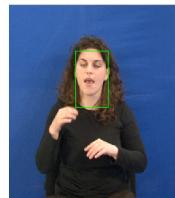
Συνεπώς μια λύση για αυτό το πρόβλημα είναι ο ορισμός μια τιμής κατωφλίου στον απλό lk . Αν η τιμή της νόρμας του διορθωτικού διανύσματος $u = [u_x, u_y]$ είναι μικρότερη της τιμής κατωφλίου τότε σημαίνει ότι ο αλγόριθμος έχει συγχλίνει αρκετά και περαιτέρω επαναλήψεις για καλύτερη προσέγγιση δεν θα βοηθήσουν το αποτέλεσμα. Με αυτόν τον τρόπο γλιτώνουμε αρκετό όγκο περιττών πράξεων διατηρώντας παράλληλα την απόδοση ψηλά.

1.4 Υπολογισμός της Μετατόπισης του Προσώπου από το Πεδίο Οπτικής Ροής

Μέχρι τώρα είχαμε υπολογίσει την οπτική ροή από το ένα frame στο επόμενο, ωστόσο μας ενδιαφέρει η συνολική μετατόπιση του bounding-box και ο υπολογισμός της μετατόπισης αυτής με όσο το δυνατόν μεγαλύτερη ακρίβεια. Γνωρίζουμε ότι τα σημεία τα οποία έχουν έντονη πληροφορία υφής όπως οι ακμές και οι κορυφές έχουν υψηλό επίπεδο ενέργειας οπτικής ροής σε αντίθεση με τα σημεία τα οποία έχουν ομοιόμορφη και επίπεδη υφή που έχουν σχεδόν μηδενική. Επομένως, γίνεται αντιληπτό ότι παίρνοντας την μέση τιμή των όλων διανυσμάτων τα αποτελέσματα παρεκκλίνουν από το επιθυμητό.

Για αυτό θέλοντας να εστιάσουμε κυρίως στα σημαντικά σημεία αυτά δηλαδή που έχουν υψηλή ενέργεια οπτικής ροής επιλέγουμε ένα όριο (threshold) και παίρνουμε την μέση τιμή των διανυσμάτων των οποίων η ενέργεια τους ξεπερνά την τιμή κατωφλίου. Ως ενέργεια διανύσματος ταχύτητας θεωρούμε το $\|d\| = d_x^2 + d_y^2$. Στην συνέχεια, βρίσκουμε την συνολική μετατόπιση του bounding box και συνεχίζουμε όλη την διαδικασία του Lucas-Kanade και την εύρεση του επόμενου bounding box.

Παρακάτω φαίνονται ενδεικτικά τα διανύσματα οπτικής ροής και η αντίστοιχη ενέργεια:



Ενέργεια οπτικής ροής στο frame 12

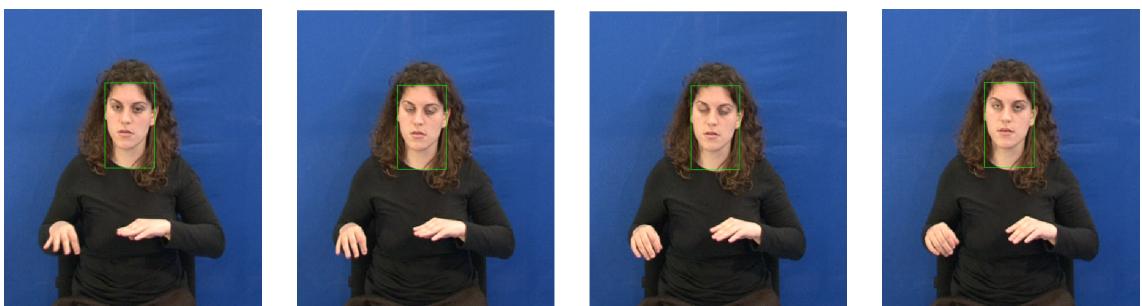
Ενέργεια οπτικής ροής στο frame 13



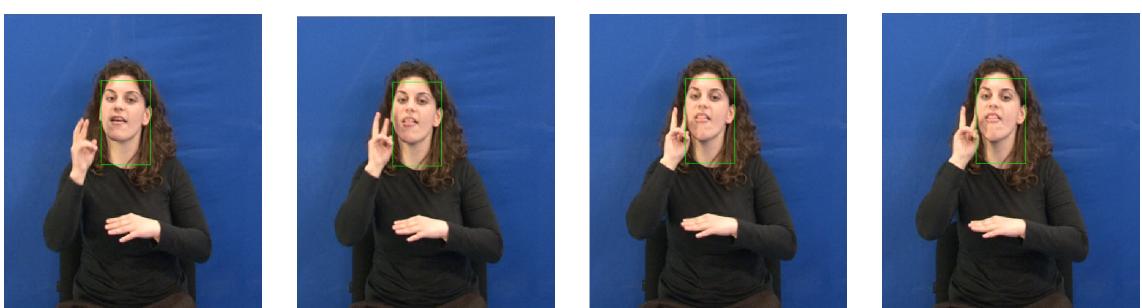
Ενέργεια οπτικής ροής στο frame 14

Ενέργεια οπτικής ροής στο frame 15

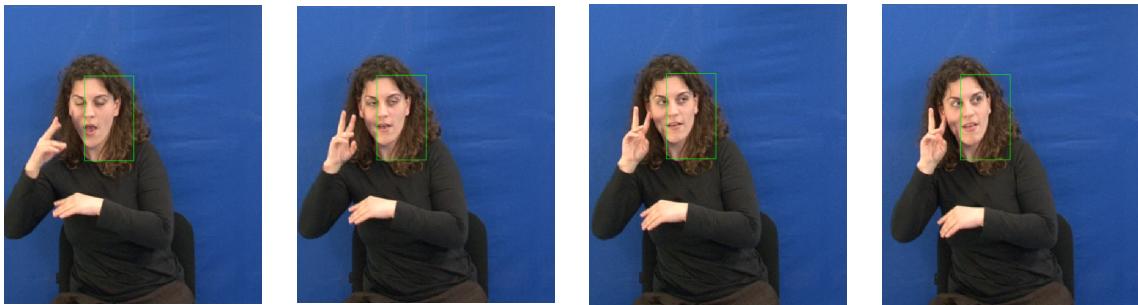
Οι διαδικασίες που αναφέρθηκαν παραπάνω συνοψίζονται στις παρακάτω εικόνες, όπου φαίνεται το tracking του προσώπου κατά την διάρκεια του βίντεο μέσα από τα frames (1,2,3,4), (23,24,25,26) και (64,65,66,67).



Εικόνα 9: Face tracking στα frames 1-4



Εικόνα 10: Face tracking στα frames 23-26



Εικόνα 11: Face tracking στα frames 64-67

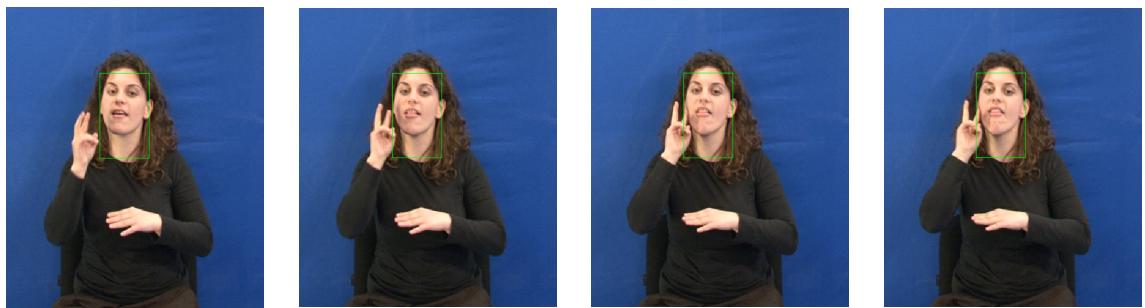
Παρατηρήσεις: Στα τελευταία frames βλέπουμε πως το bounding-box δεν ακολουθεί ακριβώς την περιοχή του προσώπου αλλά ένα μέρος αυτού. Το γεγονός αυτό ήταν αναμενόμενο και οφείλεται στην αδυναμία του αλγόριθμου των Lucas-Kanade να παρακολουθήσεις μεγάλες και απότομες κινήσεις. Στα πρώτα πλαίσια βλέπουμε μικρές σχετικά κινήσεις και τα αποτελέσματα είναι εξαρετικά. Ωστόσο στο σημείο που αρχίζουν οι κινήσεις να γίνονται πιο έντονες διαπιστώνεται η αδυναμία του αλγόριθμου.

1.5 Συμπεράσματα και Συγκρίσεις Αποτελεσμάτων SingleScale – MultiScale

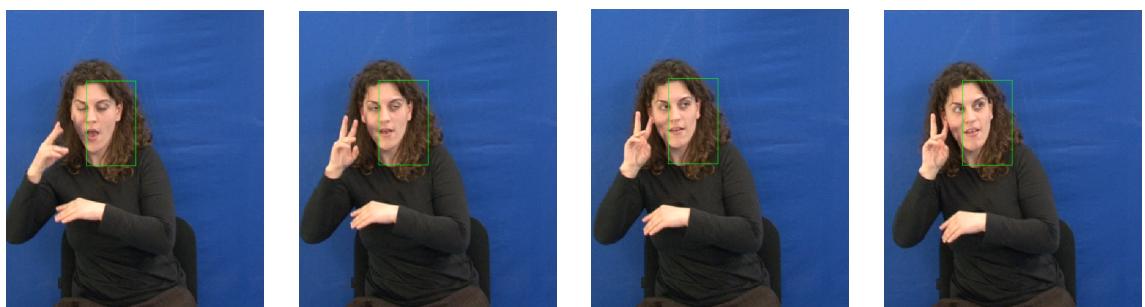
Οι αλγόριθμοι που παρουσιάστηκαν στα 1.2, 1.3 έχουν κάποιες διαφορές οι οποίες αναλύθηκαν παραπάνω. Ωστόσο θέλουμε να δούμε ποιοτικά πως αυτές τους οι διαφορές αποτυπώνονται στην πράξη. Παρακάτω παραθέτουμε κάποια από τα πλαίσια (frames) τα οποία αναφέρθηκαν στο 1.4 για την single scale εκδοχή. Όπως ειπώθηκε και παραπάνω ο αλγόριθμος multiscaling έχει πλεονέκτημα στις μεγάλες κινήσεις. Έτσι, για εξοικόνομηση χώρου παραθέτουμε τις διαφορές των 2 εκδοχών στα πλαίσια που παρουσιάζουν μεγάλη κίνηση για να εντοπίσουμε τις διαφορές των 2.



Εικόνα 12: Single-scale Face tracking στα frame 23-26



Εικόνα 13: Multi-scale Face tracking στα frame 23-26



Εικόνα 14: Single-scale Face tracking στα frame 64-67



Εικόνα 15: Multi-scale Face tracking στα frame 64-67

Παρατηρήσεις: Βλέπουμε παραπάνω πως ειδικά στα πλάσια 64-67 που πραγματοποιείται μεγάλη κίνηση τα αποτελέσματα είναι εμφανώς καλύτερα καθώς προσδιορίζουν την περιοχή του προσώπου στο 100% ενώ οι αντίστοιχες του μονοκλιμακωτού αλγόριθμου αποτυγχάνουν να προσδιορίσουν πλήρως την περιοχή αυτή. Αντίθετα, στις περιοχές που η κίνηση είναι μικρή τα αποτελέσματα δεν έχουν κάποια αισθητή διαφορά καθώς και οι 2 αλγόριθμοι μπορούν να ανταποκριθούν πλήρως.

2 Εντοπισμός Χωρο-χρονικών Σημείων Ενδιαφέροντος και Εξαγωγή Χαρακτηριστικών σε Βίντεο Ανθρωπίνων Δράσεων

Στο μέρος αυτό θα ασχοληθούμε με την εξαγωγή χωρο-χρονικών χαρακτηριστικών με στόχο την εφαρμογή τους στο πρόβλημα κατηγοριοποίησης βίντεο που περιέχουν ανθρώπινες δράσεις. Οπως είδαμε από την 1η εργαστηριακή άσκηση, τα τοπικά χαρακτηριστικά (local features) έχουν δείξει τεράστια επιτυχία σε διάφορα προβλήματα αναγνώρισης της Όρασης Υπολογιστών, όπως η αναγνώριση αντικειμένων. Οι τοπικές αναπαραστάσεις περιγράφουν το προς παρατήρηση αντι- κείμενο με μια σειρά από τοπικούς περιγραφητές που υπολογίζονται σε γειτονιές ανιχνευθέντων σημείων ενδιαφέροντος. Τελικά, η συλλογή των τοπικών χαρακτηριστικών ενσωματώνεται σε μια τελική αναπαράσταση global representation (π.χ. bag of visual words) ικανή να αναπαραστήσει τη στατιστική κατανομή τους και να προχωρήσει στα επόμενα στάδια της αναγνώρισης.

Η αναπαράσταση με χρήση τοπικών χαρακτηριστικών έχει επικρατήσει και στην αναγνώριση ανθρώπινων δράσεων, όπου γίνεται μια επιλογή από δεδομένα που αφ' ενός μειώνουν κατά πολύ τη διάσταση των βίντεο και αφ' ετέρου τα μετασχηματίζονται σε μια αναπαράσταση που τα κάνει κατηγοριοποιήσιμα. Στα πλαίσια αυτής της άσκησης μας δώθηκαν βίντεο από 3 κλάσεις δράσεων (walking,running,boxing) από τα οποία θα εξάγουμε χωρο-χρονικούς περιγραφητές με σκοπό την κατηγοριοποίηση των δράσεων αυτών.

2.1 Χωρο-χρονικά Σημεία Ενδιαφέροντος

Το πρώτο βήμα είναι ο εντοπισμός για κάθε βίντεο χωρο-χρονικών σημείων ενδιαφέροντος. Οι ανιχνευτές τοπικών χαρακτηριστικών αναζητούν χωρο-χρονικά σημεία και κλίμακες ενδιαφέροντος που αντιστοιχούν σε περιοχές που χαρακτηρίζονται από σύνθετη κίνηση ή απότομες μεταβολές στην εμφάνιση του video εισόδου μεγιστοποιώντας μια συνάρτηση οπτικής σημαντικότητας. Στην εργαστηριακή αυτή άσκηση θα ασχοληθούμε με τους παρακάτω 2 ανιχνευτές:

- **Harris Detector**

Ο συγκεκριμένος detector υλοποιήθηκε στη συνάρτηση HarrisDetector.m και η πολυκλιμακωτή εκδοχή του στη MultiscaleHarrisDetector.m. Περιγραφικά, στόχος του Harris είναι να εντοπίσει σημεία όπου υπάρχουν σημαντικές χωρικές ή χρονικές μεταβολές υποδεικνύοντας κάποια γωνία. Τα σημεία αυτά προκύπτουν υπολογίζοντας αρχικά τον πίνακα M:

$$M(x, y, t; \sigma, \tau) = g(x, y, t; s\sigma, s\tau) * (\nabla L(x, y, t; \sigma, \tau)(\nabla L(x, y, t; \sigma, \tau)^T))$$

$$\text{όπου } L(x, y, t; \sigma, \tau) = G(x, y, t; \sigma, \tau) * I(x, y, t) \text{ και } G(x, y, t; \sigma, \tau) = \frac{1}{\sqrt{(2\pi)^3 \sigma^4 \tau^2}} e^{(-\frac{x^2+y^2}{2\sigma^2} - \frac{t^2}{2\tau^2})}$$

Επίσης, για υπολογιστικούς λόγους εκμεταλλευόμαστε το γεγονός ότι τα Gaussian φίλτρα είναι separable, οπότε η 3d συνέλιξη μπορεί να σπάσει σε μονοδιάστατες συνέλιξεις. Ακόμη, οι μερικές παράγωγοι υπολογίστηκαν μέσω συνέλιξης με τον πυρήνα κεντρικών διαφορών $[-101]^T$ (προσαρμοσμένο στην κατάλληλη διάσταση). Αντίστοιχα με τον 2d ανιχνευτή Harris, το 3d κριτήριο γωνιότητας ακολουθεί και αυτό την ίδια λογική:

$$H(x, y, t) = |det(M(x, y, t)) - k * trace(M(x, y, t)^3)|$$

Τα σημεία ενδιαφέροντος προκύπτουν σαν τα τοπικά μέγιστα του κριτηρίου σημαντικότητας. Αφού βρισκόμαστε σε τρεις διαστάσεις, ορίζουμε μία σφράια ακτίνας σ γύρω από κάθε σημείο και θεωρούμε ένα σημείο (x, y, t) ως σημαντικό αν το κριτήριο γωνιότητας έχει τοπικό μέγιστο στο σημείο αυτό και η τιμή του ξεπερνάει ένα ποσοστό της μέγιστης τιμής.

Για να απεικονίσουμε τα αποτελέσματα επιλέγουμε 4 από τα 200 frame σε ομοιόμορφα σημεία μέσα σε κάθε βίντεο και ένα βίντεο από κάθε κατηγορία. Οι τιμές των παραμέτρων που προέκυψαν μετά από ορισμένες δοκιμές:

$$\checkmark \sigma = 1$$

$$\checkmark \tau = 0.7$$

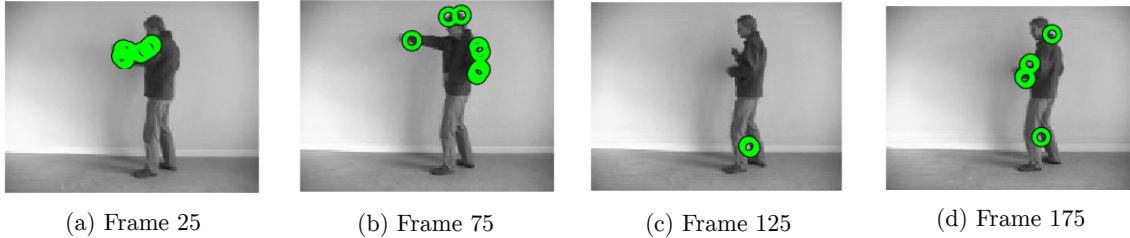
$$\checkmark s = 1.5$$

$$\checkmark k = 0.1$$

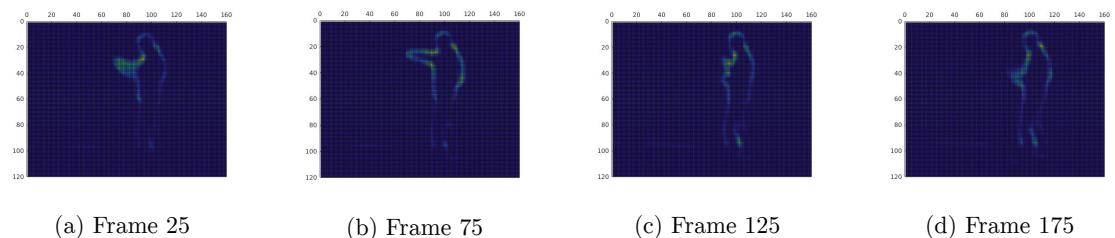
$$\checkmark \text{theta} = 0.1$$

Παρακάτω βλέπουμε για κάθε βίντεο και κάθε frame τα σημεία ενδιαφέροντος με πράσινο κύκλο ανάλογο της κλίμακας όπου εντοπίστηκαν. Επίσης, βλέπουμε και την εικόνα του κριτήριου σημαντικότητας H .

– **person16_boxing_d4_uncomp.avi**

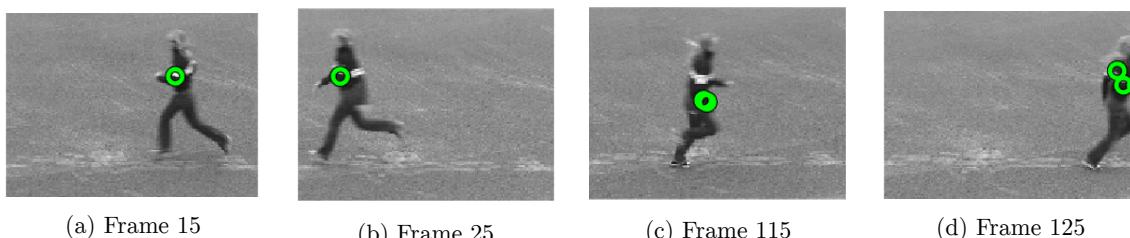


Εικόνα 16: Σημεία ενδιαφέροντος στα frame 25, 75, 125, 175

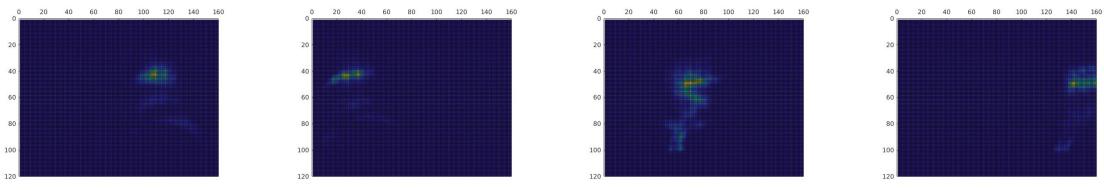


Εικόνα 17: Κριτήριο σημαντικότητας στα frame 25, 75, 125, 175

– **person09_running_d1_uncomp.avi**

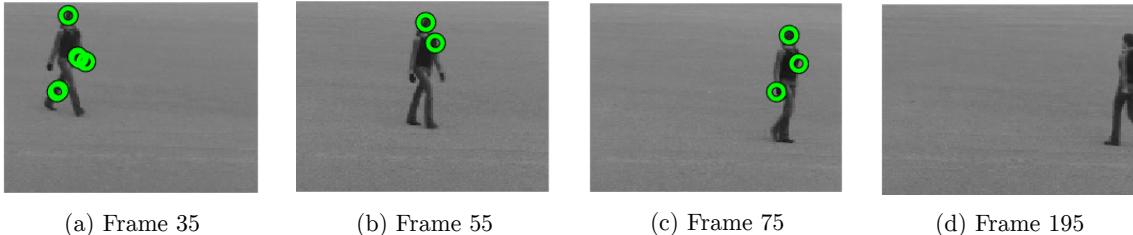


Εικόνα 18: Σημεία ενδιαφέροντος στα frame 15, 25, 115, 125

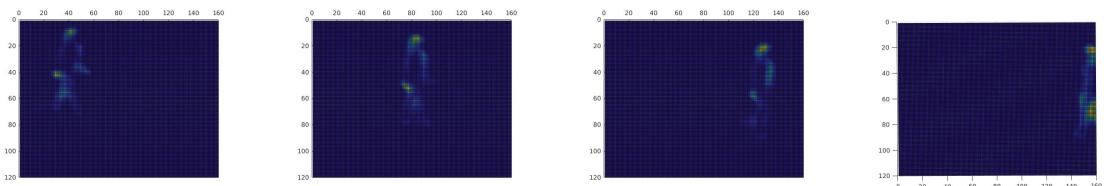


Εικόνα 19: Κριτήριο σημαντικότητας στα frame 15, 25, 115, 125

- person07_walking_d2_uncomp.avi



Εικόνα 20: Σημεία ενδιαφέροντος στα frame 35, 55, 75, 195



Εικόνα 21: Κριτήριο σημαντικότητς στα frame 35, 55, 75, 195

Αντίστοιχα με την πρώτη εργαστηριακή, μπορύμε να εφαρμόσουμε μία πολυκλιμακωτή έκδοση του ανιχνετή Harris, ακολουθώντας ακρις την ίδια λογική. Υπολογίζουμε, δηλαδή, τα σημεία ενδιαφέροντος σε μία ακολουθία από κλίμακες και απορρίπτουμε τα σημεία για τα οποία η κλίμακα που ανιχνεύθηκαν δεν μεγιστοποιεί την IoG μετρική σε μια γειτονιά 2 διαδοχικών κλίμακων. Τα αντίστοιχα αποτελέματα με τα παραπάνω frames φαίνονται παρακάτω:

✓ $\sigma = 1$

✓ s = 1.5

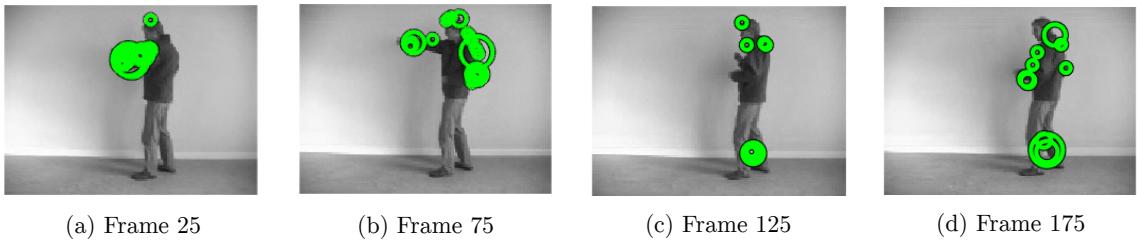
✓ theta = 0.1

$$\checkmark \tau = 0.7$$

✓ k = 0.1

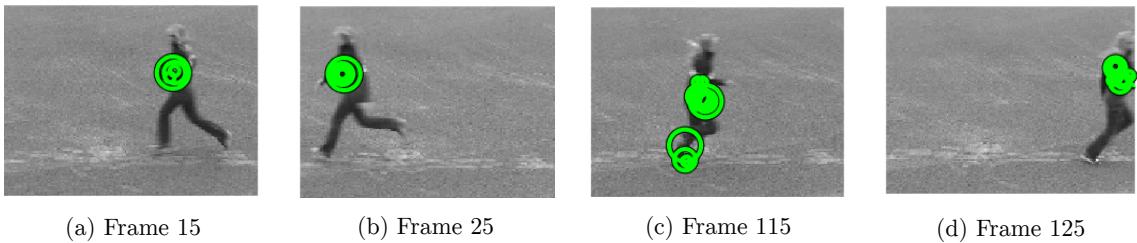
✓ N = 4

- person16_boxing_d4_uncomp.avi



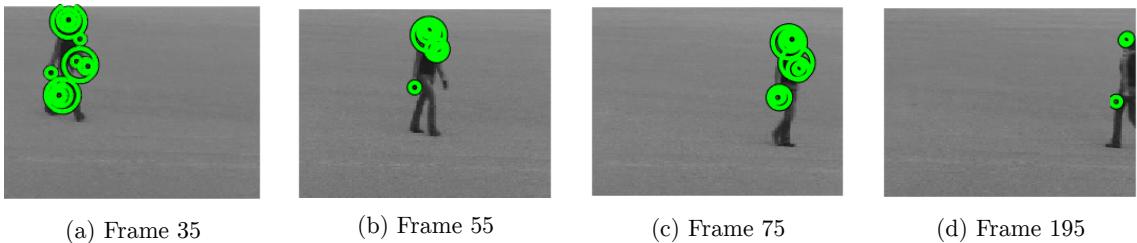
Εικόνα 22: Σημεία ενδιαφέροντος στα frame 25, 75, 125, 175

– **person09_running_d1_uncomp.avi**



Εικόνα 23: Σημεία ενδιαφέροντος στα frame 15, 25, 115, 125

– **person07_walking_d2_uncomp.avi**



Εικόνα 24: Σημεία ενδιαφέροντος στα frame 35, 55, 75, 195

- **Gabor Detector** Ο δεύτερος ανιχνευτής που θα χρησιμοποιήσουμε βασίζεται στο χρονικό φιλτράρισμα του βίντεο με ένα ζεύγος Gabor φίλτρων αφού πρώτα αυτό έχει υποστεί εξομάλυνση στις χωρικές διαστάσεις μέσω ενός 2D γκαουσιανού πυρήνα $g(x, y; \sigma)$ με τυπική απόκλιση σ . Τα Gabor ορίζονται ως:

$$h_{ev}(t; \tau, \omega) = -\cos(2\pi t\omega) e^{\frac{-t^2}{2\tau^2}} h_{od}(t; \tau, \omega) = -\sin(2\pi t\omega) e^{\frac{-t^2}{2\tau^2}}$$

Το κριτήριο σημαντικότητας προκύπτει παίρνοντας την τετραγωνική ενέργεια της εξόδου για το ζεύγος Gabor φίλτρων:

$$H(x, y, t) = (I(x, y, t) * g * h_{ev})^2 + (I(x, y, t) * g * h_{od})^2$$

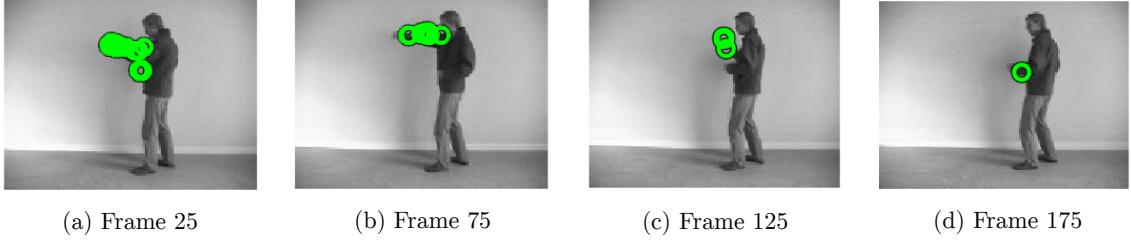
και η επιλογή των σημείων γίνεται αντίστοιχα με τον Harris απλά με το παραπάνω κριτήριο H . Ο συγκεκριμένο detector υλοποιήθηκε στη συνάρτηση `GaborDetector.m` και η πολυκλιμακωτή εκδοχή του στη `MultiscaleGaborDetector.m`. Παρακάτω, βλέπουμε τα σημεία που ανίχνευσε ο Gabor detector με τις εξής παραμέτρους:

✓ $\sigma = 1.6$

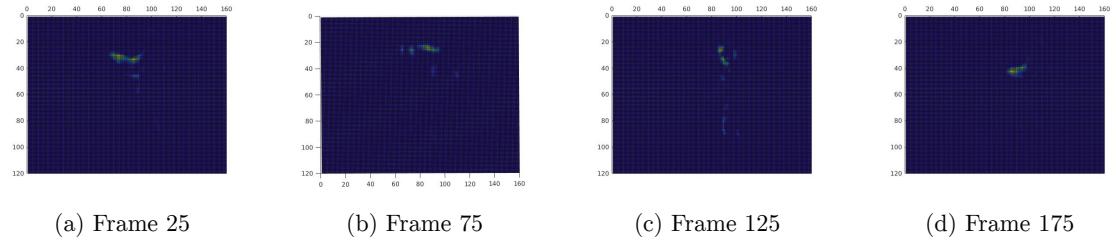
✓ $\tau = 1.5$

✓ theta = 0.1

– person16_boxing_d4_uncomp.avi

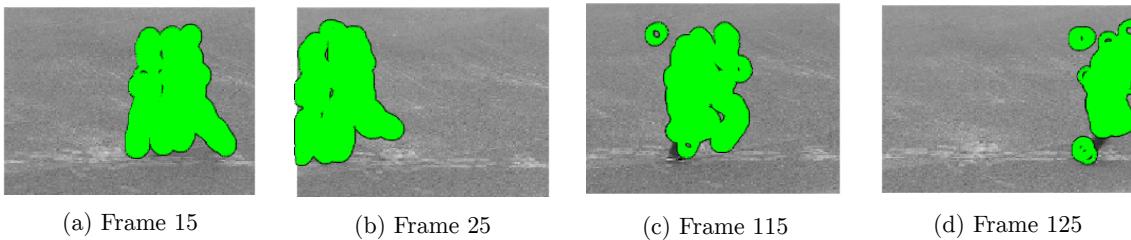


Εικόνα 25: Σημεία ενδιαφέροντος στα frame 25, 75, 125, 175

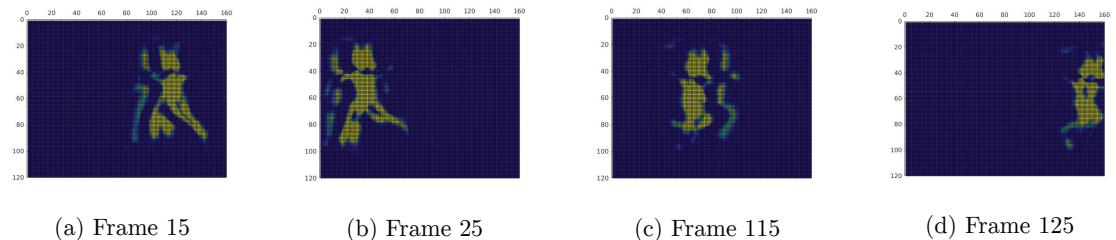


Εικόνα 26: Κριτήριο σημαντικότητας στα frame 25, 75, 125, 175

– person09_running_d1_uncomp.avi

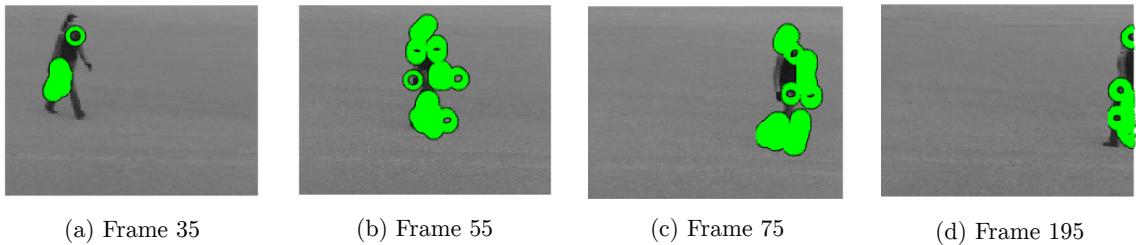


Εικόνα 27: Σημεία ενδιαφέροντος στα frame 15, 25, 115, 125

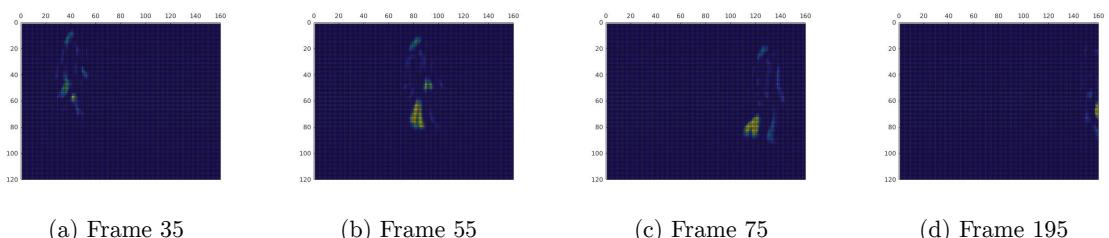


Εικόνα 28: Κριτήριο σημαντικότητας στα frame 15, 25, 115, 125

– **person07_walking_d2_uncomp.avi**



Εικόνα 29: Σημεία ενδιαφέροντος στα frame 35, 55, 75, 195



Εικόνα 30: Κριτήριο σημαντικότητας στα frame 35, 55, 75, 195

Επίσης, η πολυχλιμακωτή ανίχνευση μπορεί να εφαρμοστεί και με τον ανιχνευτή Gabor. Παρακάτω, βλέπουμε τα αποτελέσματα:

$$\checkmark \sigma = 1.6$$

$$\checkmark \tau = 1.5$$

$$\checkmark \theta = 0.5$$

$$\checkmark s = 1.5$$

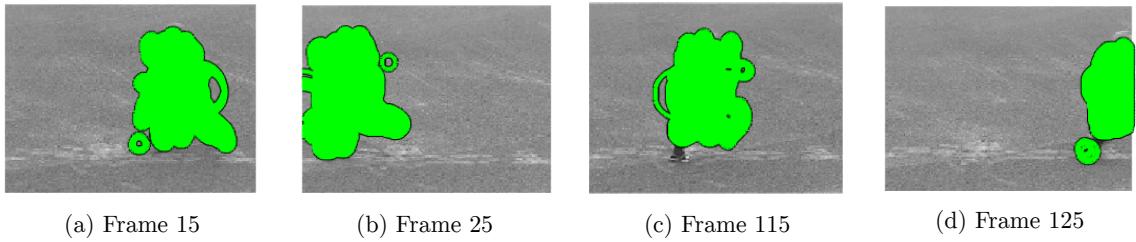
$$\checkmark N = 4$$

- **person16_boxing_d4_uncomp.avi** Εδώ επειδή η κίνηση είναι μικρότερη, το theta μειώθηκε σε 0.4.



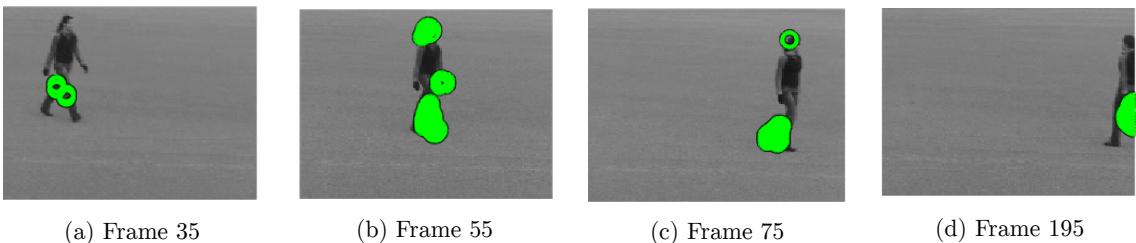
Εικόνα 31: Σημεία ενδιαφέροντος στα frame 25, 75, 125, 175

– **person09_running_d1_uncomp.avi**



Εικόνα 32: Σημεία ενδιαφέροντος στα frame 15, 25, 115, 125

– `person07_walking_d2_uncomp.avi`



Εικόνα 33: Σημεία ενδιαφέροντος στα frame 35, 55, 75, 195

Σύγκριση και Σχολιασμός: Παρατηρούμε ότι ο ανιχνευτής Gabor παρουσιάζει καλύτερα αποτελέσματα, το οποίο οφείλεται στο γεγονός ότι εντοπίζει καλύτερα τις κινήσεις που συμβαίνουν σε αντίθεση με τον Harris, ο οποίος εντοπίζει μόνο δημιουργούμενες γωνίες από frame σε frame. Πιο συγκεκριμένα, στο boxing ο Gabor αναγνωρίζει ορθά μόνο τις κινήσεις των χεριών όταν ο άνθρωπος βίχνει μπουνιά. Αντίθετα, ο Harris εντοπίζει και κάποια λάθος σημεία στα πόδια. Στο running, ο Gabor εντοπίζει πάρα πολλά σημεία στο σώμα του ανθρώπου το οποίο είναι σωστό καθώς έχουμε μία έντονη κίνηση. Από την άλλη ο Harris εντοπίζει πολύ λιγότερα σημεία παρά την έντονη και συνεχή μετατόπιση του ανθρώπου. Τέλος, στο walking, ο Gabor εντοπίζει σώστα κάποια σημεία στα πόδια και τα χέρια ενώ το Harris χάνει πολλές φορές σημεία κίνησης στα πόδια. Η πολυκλιμακωτή εκδοχή τους προφανώς έχει καλύτερα αποτελέσματα καθώς είναι σίγουρα ότι τα σημεία ενδιαφέροντος θα βρίσκονται σε διαφορετικές κλίμακες.

2.2 Χωρο-χρονικοί Ιστογραφικοί Περιγραφητές

Οι χωρο-χρονικοί περιγραφητές που θα χρησιμοποιηθούν βασίζονται στον υπολογισμό ιστογραμάτων της κατευθυντικής παραγώγου (HOG) και της οπικής ροής (HOF - Histograms of Oriented Flow) γύρω από τα σημεία ενδιαφέροντος που υπολογίσαμε. Συγκεκριμένα, χρησιμοποιείται η συνάρτηση *OrientationHistogram.p*, η οποία δέχεται ως είσοδο το διανυσματικό πεδίο (κατευθυντικές παραγώγους είτε κατεύθυνση ροής), το μέγεθος του grid και το πλήθος των bins και επιστρέφει την ιστογραμματική περιγραφή της αντίστοιχης περιοχής. Συγκεκριμένα, στην περίπτωση μας χρησιμοποιούμε μία τετραγωνική περιοχή μεγέθους $4 * scale$ γύρω από κάθε σημείο ενδιαφέροντος (συναρτήσεις *calculateHof.m* και *calculateHog.m*). Ο τελικός περιγραφητής είναι η συνένωση των 2 παραπάνω περιγραφητών. Προφανώς, κάθε εικόνα έχει διαφορετικού μεγέθους περιγραφητή, γεγονός που δεν επιτρέπει να κάνουμε διαχωρισμό δράσεων. Συνεπώς, για την τελική αναπαράσταση ενός βίντεο, χρησιμοποιείται η bag of visual words (BoVW) τεχνική που έχει περιγραφεί και υλοποιηθεί αναλυτικά στην 1η εργαστηριακή άσκηση (χρησιμοποιήθηκαν άμεσα οι αντίστοιχες συναρτήσεις της 1ης εργαστηριακής). Οι αναπαραστάσεις αυτές για διαφορετικούς συνδυασμούς detector και descriptor αποθηκεύτηκαν σε *.mat* μεταβλητές για να χρησιμοποιηθούν στο επόμενο ερώτημα όπου θα τους αξιολογήσουμε.

2.3 Κατασκευή Δενδρογράμματος για τον Διαχωρισμό των Δράσεων

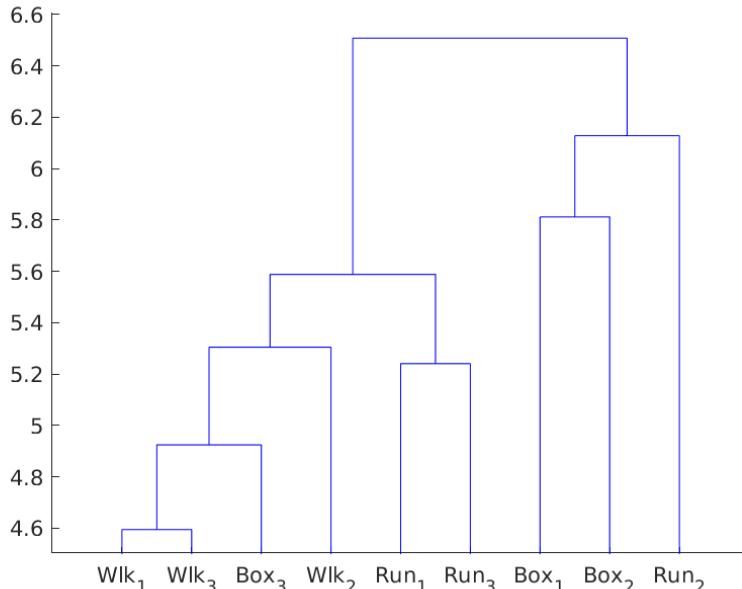
Στο ερώτημα αυτό θα γίνει μια προσπάθεια κατανόησης της ικανότητας κατηγοριοποίησης των βίντεο με τις ανθρώπινες δράσεις σε 3 κατηγορίες/κλάσεις (που η κάθε μία θα αντιπροσωπεύει ένα διαφορετικό είδος δράσης) με χρήση των BoVW αναπαραστάσεων (που βασίζονται σε HOG / HOF χαρακτηριστικά) που υπολογίσαμε στα προηγούμενα ερωτήματα. Αυτό θα επιτευχθεί ποιοτικά με την οπτικοποίηση της απόστασης των διανυσμάτων χαρακτηριστικών μέσω της κατασκευής ενός δενδρογράμματος αποστάσεων που αντιπροσωπεύει την ικανότητα διαχωρισμού των 3 διαφορετικών κατηγοριών.

Ως απόσταση μεταξύ δύο αναπαραστάσεων θα χρησιμοποιηθεί η χ^2 , η οποία θεωρείται κατάλληλη για ιστογράμματα. Η χ^2 για 2 ιστογράμματα $H_i = \{h_{i1}, h_{i2}, \dots, h_{iK}\}$ και $H_j = \{h_{j1}, h_{j2}, \dots, h_{jN}\}$ (όπου K το πλήθος των κέντρων) υπολογίζεται ως $D(H_i, H_j) = \frac{1}{2} \sum_{n=1}^K \frac{(h_{in} - h_{jn})^2}{h_{in} + h_{jn}}$.

Παρακάτω φαίνονται όλα τα δενδρογράμματα που προκέχυψαν με διαφορετικούς συνδυασμούς ανιχνευτών/περιγραφητών:

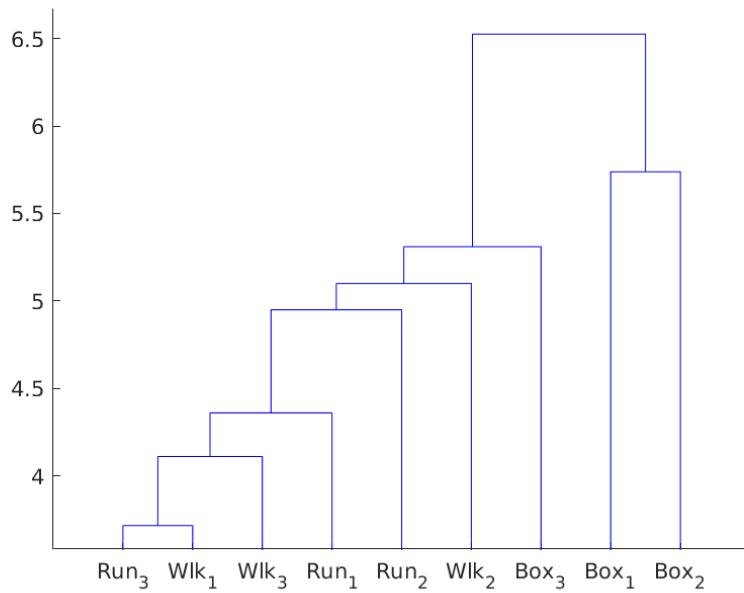
- Harris Detector

- HOG Descriptor



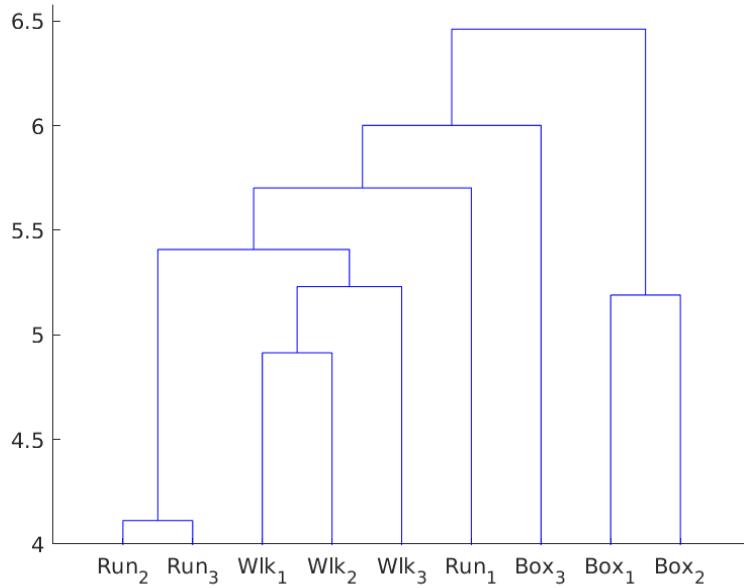
Εικόνα 34: Δενδρόγραμμα με Harris-HOG

- HOF Descriptor



Εικόνα 35: Δενδρογράμμα με Harris-HOF

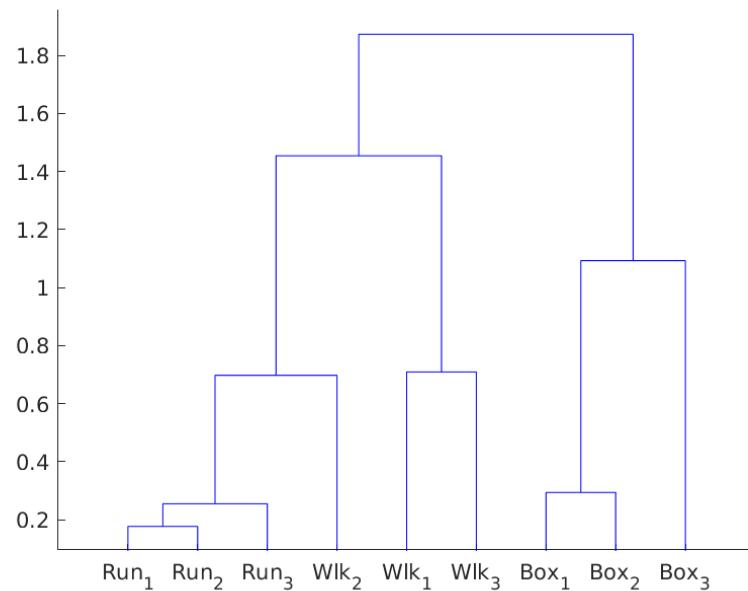
– HOG/HOF Descriptor



Εικόνα 36: Δενδρογράμμα με Harris-HOG/HOF

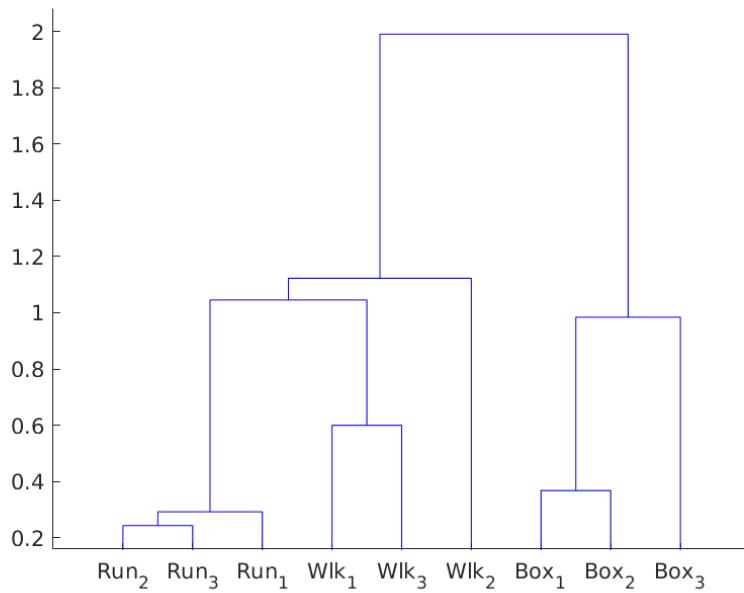
- Gabor Detector

- HOG Descriptor



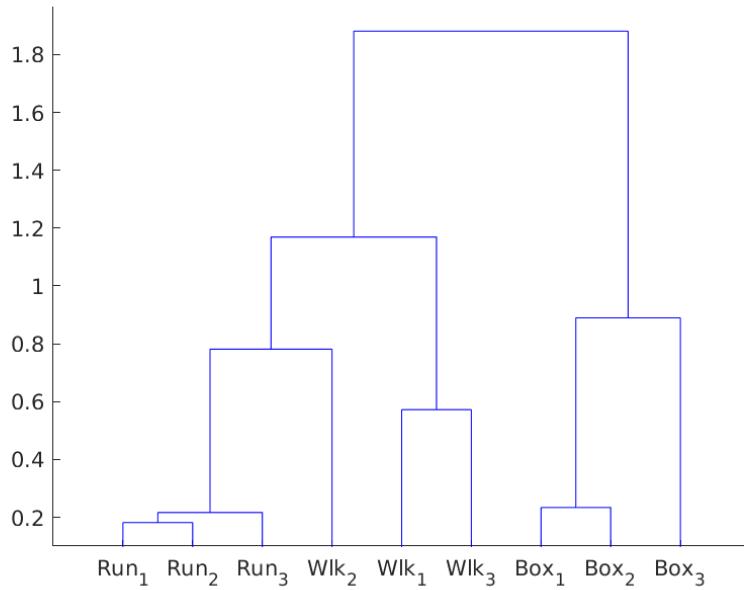
Εικόνα 37: Δενδρογράμμα με Gabor-HOG

- HOF Descriptor



Εικόνα 38: Δενδρόγραμμα με Gabor-HOF

– HOG/HOF Descriptor



Εικόνα 39: Δενδρόγραμμα με Gabor-HOG/HOF

Σύγκριση ως προς ανιχνευτή: Παρατηρούμε ότι σε όλες τις περιπτώσεις ο Gabor detector καταλήγει σε ένα δενδρόγραμμα όπου οι κοινές δράσεις έχουν κοντινή τελική αναπαράσταση. Από την άλλη, ο Harris detector δεν έχει τόσο καλά αποτελέσματα αφού σε συνδυασμό με κανέναν descriptor δεν καταφέρνει να διαχωρίσει τις δράσεις. Δυσκολεύεται κυρίως στον διαχωρισμό του running από το walking το οποίο είναι λογικό αφού αυτές οι δύο δράσεις είναι πιο κοντά σε σχέση με το boxing (στο ένα περιλαμβάνεται μόνο κίνηση χεριών ενώ στα άλλα δύο έχουμε κίνηση και ποδιών και χεριών).

Σύγκριση ως προς περιγραφητή: Εδώ συμβαίνει το αναμενόμενο, δηλαδή ο HOF περιγραφητής είναι καλύτερος από τον HOG. Αυτό συμβαίνει γιατί ο HOG, παρά το γεγονός ότι είναι αναλλοίωτος σε περιστροφές και κλιμακώσεις, δεν μπορεί να συγκρατήσει αρκετή πληροφορία για μία τόσο σύνθετη κίνηση όπως το τρέξιμο, το περπάτημα και το μποξ. Από την άλλη τα ιστογράμματα που προκύπτουν από την οπτική ροή (HOF) εμπεριέχουν πολλή πληροφορία η οποία βοηθάει στον διαχωρισμό των 3 δράσεων όπως φαίνεται στα αποτελέσματα. Τέλος, ο συνδυασμός τους (HOG/HOF) έχει τα καλύτερα αποτελέσματα αφού συνδυάζει τα πλεονεκτήματά τους με μόνο μειονέκτημα ότι θέλουμε διπλάσιο χώρο για την αποθήκευσή.

Επιλογή καλύτερου συνδυασμού: Με βάση όλα τα παραπάνω, ο καλύτερος συνδυασμός είναι Gabor detector με HOG/HOF descriptor αφού απεικονίζει τις κοινές δράσεις σε κοντινές BoW αναπαραστάσεις. Σε περίπτωση που θέλουμε να εξοικονομήσουμε χώρο (και κατ' επέκταση υπολογιστική ισχύ αφού μετά εφαρμόζουμε τον αλγόριθμο kmeans πάνω σε αυτά τα δεδομένα) μπορεί να χρησιμοποιηθεί και ο συνδυασμός Gabor-HOF με εξίσου καλά αποτελέσματα.

Σημείωση: Δεν δημιουργήθηκαν τα δενδρογράμματα για την πολυκλιμακωτή εκδοχή των αλγορίθμων, αλλά υποθέτουμε ότι τα αποτελέσματα θα ήταν καλύτερα.

Βιβλιογραφία

- [1] S. Baker and I. Matthews, “Lucas-kanade 20 years on: A unifying framework,” *International journal of computer vision*, vol. 56, no. 3, pp. 221–255, 2004.
- [2] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, “Behavior recognition via sparse spatio-temporal features,” VS-PETS Beijing, China, 2005.
- [3] C.-S. Fuh and P. Maragos, “Motion displacement estimation using an affine model for image matching,” *Optical Engineering*, vol. 30, no. 7, pp. 881–888, 1991.
- [4] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, “Learning realistic human actions from movies,” 2008.
- [5] B. D. Lucas, T. Kanade, *et al.*, “An iterative image registration technique with an application to stereo vision,” 1981.
- [6] R. Szeliski *et al.*, “Image alignment and stitching: A tutorial,” *Foundations and Trends® in Computer Graphics and Vision*, vol. 2, no. 1, pp. 1–104, 2007.
- [7] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, “Evaluation of local spatio-temporal features for action recognition,” 2009.
- [8] I. Laptev, “On space-time interest points,” *International journal of computer vision*, vol. 64, no. 2-3, pp. 107–123, 2005.
- [9] P. Maragos, *Σημειώσεις Όρασης Υπολογιστών*. 2015.