Όραση Υπολογιστών

2η Εργαστηριακή Άσκηση

ΘΕΜΑ: Εκτίμηση Οπτικής Ροής (Optical Flow) και Εξαγωγή Χαρακτηριστικών σε Βίντεο

Επώνυμο: Διαμάντη
Όνομα: Ιωάννα
ΑΜ: 03115035

Μέρος 1ο : Παρακολούθηση Προσώπου με Χρήση του Πεδίου Οπτικής Ροής (Optical Flow) με τη Μέθοδο Lucas – Kanade.

1.1Ανίχνευση Δέρματος Προσώπου

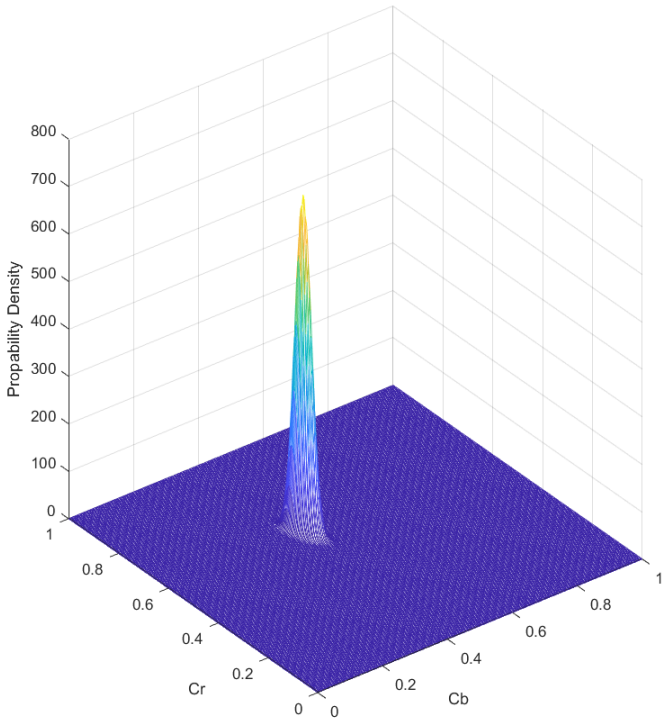
Στο ερώτημα αυτό επιθυμούμε να ανιχνεύσουμε τα pixel του πρώτου frame του βίντεο, τα οποία αντιστοιχούν σε ανθρώπινο δέρμα και στη συνέχεια να ανιχνεύσουμε την περιοχή στην οποία βρίσκεται το πρόσωπο. Για την ανίχνευση των σημείων δέρματος χρησιμοποιείται ο χρωματικός χώρος YCbCr, αφαιρώντας την πληροφορία της φωτεινότητας Y και διατηρώντας τα κανάλια Cb, Cr που περιγράφουν την ταυτότητα του χρώματος. Το χρώμα του δέρματος μοντελοποιείται με δισδιάστατη Γκαουσιανή κατανομή, δηλαδή:

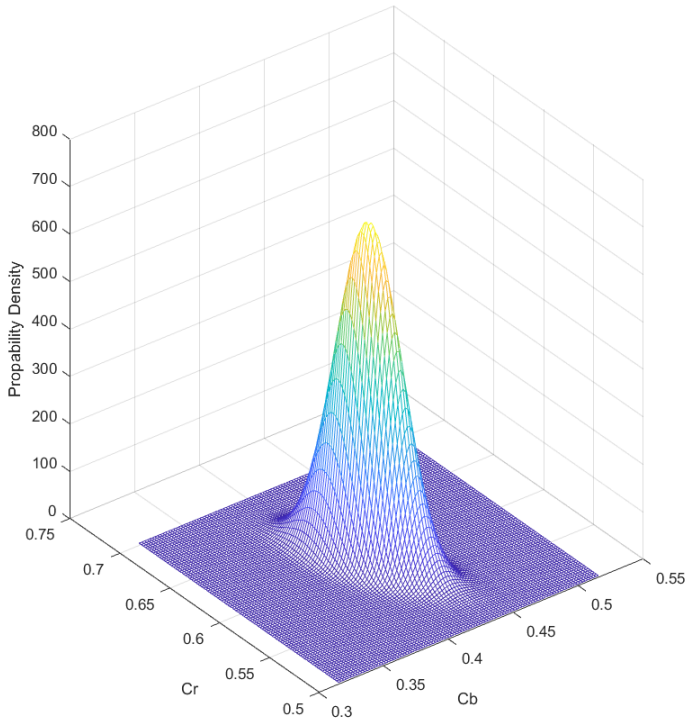
$$P(c = skin) = \frac{1}{\sqrt{|\Sigma|(2\pi)^2}} e^{-\frac{1}{2}(c-\mu)\Sigma^{-1}(c-\mu)'} \quad (1)$$

,όπου c είναι το διάνυσμα τιμών Cb και Cr για κάθε σημείο (x,y) της εικόνας. Η Γκαουσιανή κατανομή εκπαιδεύεται από τα δείγματα δέρματος που δίνονται στο αρχείο skinSamplesRGB.mat (σε μορφή RGB), υπολογίζοντας το 1x2 διάνυσμα μέσης τιμής μ = [μ_{C_b} μ_{C_r}] και τον 2x2 πίνακα συνδιακύμανσης Σ των δειγμάτων. Τα δείγματα δέρματος καθώς και η Γκαουσιανή φαίνονται παρακάτω:

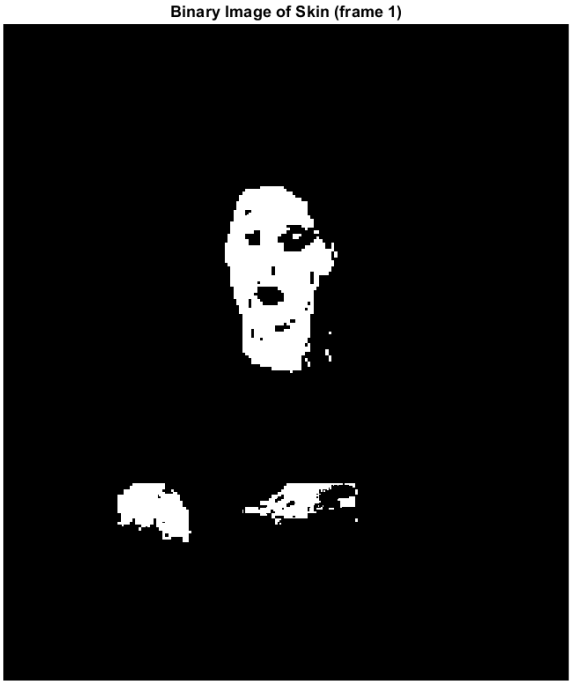
Δείγματα Δέρματος:

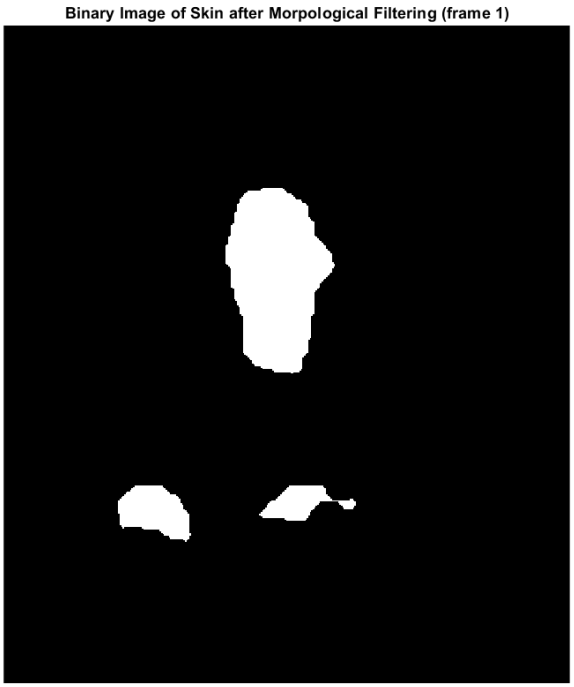
Γκαουσιανή Κατανομή (Συνολική και Κεντραρισμένη):



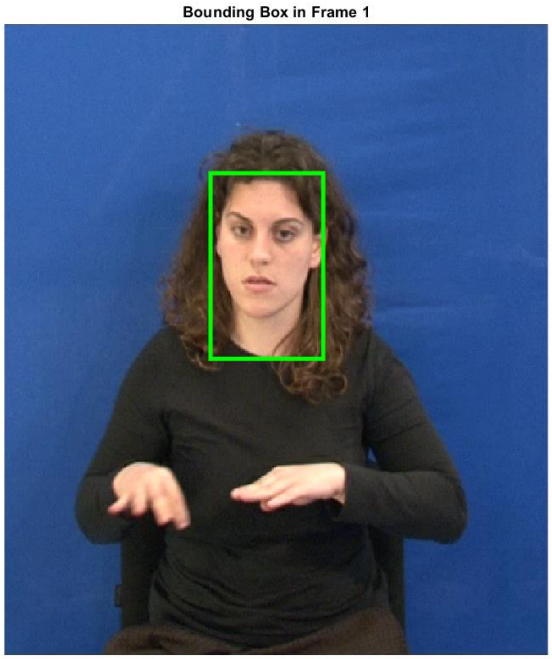


Η δυαδική εικόνα ανίχνευσης δέρματος προκύπτει από την εικόνα πιθανοτήτων P(c(x,y)= skin) > threshold ∀ (x,y), όπου threshold ένα κατώφλι απόφασης. Η τελική ανίχνευση της περιοχής δέρματος του προσώπου γίνεται επιλέγοντας την περιοχή με το μεγαλύτερο εμβαδό από όσες βρέθηκαν. Για το σκοπό αυτό απαιτείται μία μορφολογική επεξεργασία της δυαδικής εικόνας δέρματος και συγκεκριμένα κάλυψη των τρυπών (μάτια, στόμα κλπ.). Αυτό επιτυγχάνεται κάνοντας opening με ένα 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 που παρουσιάζεται το δέρμα. Η παραπάνω διαδικασία υλοποιήθηκε ως αυτόνομη συνάρτηση που δέχεται ως εισόδους μία εικόνα (το πρώτο frame του βίντεο), το διάνυσμα μέσης τιμής μ και τον πίνακα συνδιακύμανσης Σ της Γκαουσιανής και επιστρέφει το πλαίσιο οριοθέτησης προσώπου στη μορφή [x, y, width, height], όπου x, y οι συντεταγμένες του πάνω αριστερά σημείου. Παρακάτω βλέπουμε τις δυαδικές εικόνες δέρματος πριν και μετά την μορφολογική επεξεργασία καθώς και την τελική ανίχνευση προσώπου:





Τελική Ανίχνευση Προσώπου:



1.2 Υλοποίηση του αλγορίθμου Lucas - Kanade

Στο ερώτημα αυτό επιθυμούμε να υπολογίσουμε την οπτική ροή κάθε pixel (x,y) που ανήκει στην περιοχή του προσώπου μεταξύ δύο διαδοχικών frame του βίντεο. Σε μία ακολουθία εικόνων N frames $I_n(\boldsymbol{x})$, όπου n=1,…,N και **x** = (x,y), το πεδίο οπτικλης ροής -d, όπου d(**x**) = (d_x ,d_y) , φέρνει σε αντιστοιχία δύο διαδοχικές εικόνες, έτσι ώστε:

$$I_n(x) \approx I_{n-1}(x+d) \qquad (2)$$

Ο αλγόριθμος Lucas – Kanade υπολογίζει την οπτική ροή σε κάθε σημείο της εικόνας **x** με τη μέθοδο των ελάχιστων τετραγώνων, θεωρώντς ότι το d είναι σταθερό σε ένα μικρό παράθυρο γύρω από το σημείο και ελαχιστοποιώντας το τετραγωνικό σφάλμα:

$$J_x(d) = \int_{x' \in R^2} G_\rho(x-x')[I_n(x') - I_{n-1}(x'+d)]^2 \, dx' \qquad (3)$$

Όπου G_ρ(**x**) είναι μία συνάρτηση παραθύρωσης, π.χ. Γκαουσιανή με τυπική απόκλιση ρ.

Θεωρούμε ότι έχουμε μία εκτίμηση d_i για το d και προσπαθούμε να τη βελτιώσουμε κατά u, δηλαδή $d_{i+1}=d_i + u$. Αναπτύσσοντας κατά Taylor την έκφραση $I_{n-1}(x+d) = I_{n-1}(x+d_i+u)$ γύρω από το σημείο $x + d_i$ προκύπτει ότι:

$$I_{n-1}(x+d) \approx I_{n-1}(x+d_i) + \nabla I_{n-1}(x+d_i)^T u \qquad (4)$$

Βάζοντας αυτή την έκφραση στην Εξ. (3) μπορεί να δειχθεί ότι η λύση των ελάχιστων τετραγώνων για τη βελτίωση του πεδίου οπτικής ροής σε κάθε σημείο είναι:

$$u(x) = \begin{bmatrix} (G_\rho * A_1^2)(x) + \epsilon & (G_\rho * (A_1A_2))(x) \\ (G_\rho * (A_1A_2))(x) & (G_\rho * A_2^2)(x) + \epsilon \end{bmatrix}^{-1} \begin{bmatrix} (G_\rho * (A_1E))(x) \\ (G_\rho * (A_2E))(x) \end{bmatrix} \qquad (5)$$

όπου

$$A(x) = [A_1(x) \;\; A_2(x)] = \left[\frac{\partial I_{n-1}(x+d_i)}{\partial x} \quad \frac{\partial I_{n-1}(x+d_i)}{\partial y}\right] \qquad (6)$$

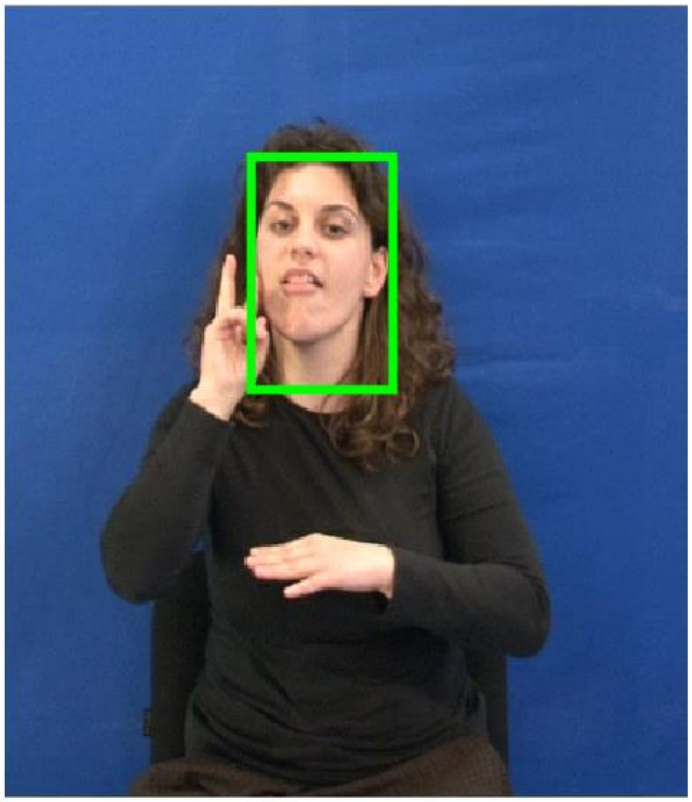
$$E(x) = I_n(x) - I_{n-1}(x + d_i) \quad (7)$$

Η μικρή σταθερά ϵ βελτιώνει το αποτέλεσμα σε επίπεδες περιοχές με μειωμένη υφή και άρα μειωμένη πληροφορία για τον υπολογισμό του πεδίου ροής. Η ανανέωση του πεδίου οπτικής ροής $d_{i+1} = d_i + u$ να υπολογίζεται από την Εξ. (5), επαναλαμβάνεται αρκετές φορές ως τη σύγκλιση.

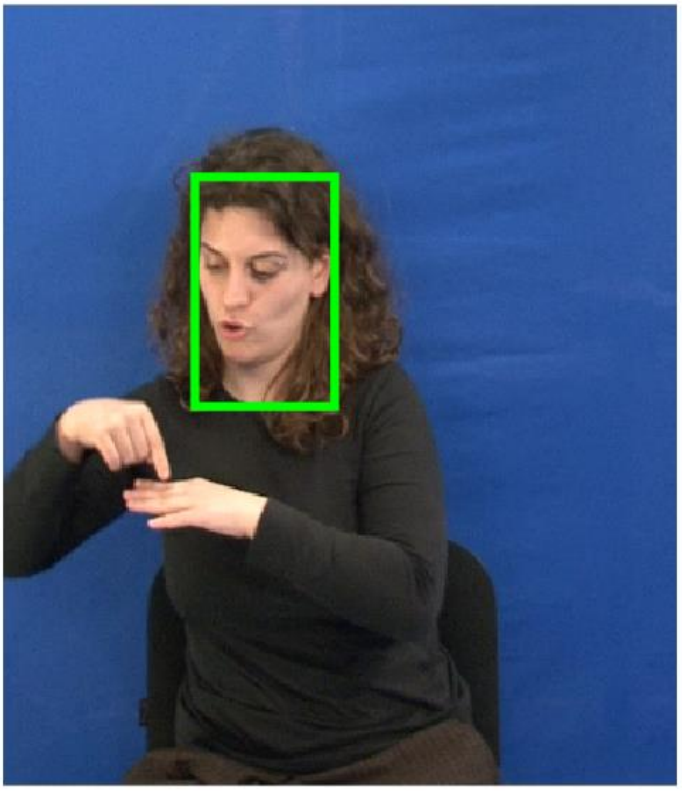
Με βάση τα παραπάνω κατασευάστηκε μία συνάρτηση, η οποία παίρνει ως είσοδο δύο διαδοχικά frame του βίντεο, την κλίμακα ρ ∈ [1,5] του Γκαουσιανού παραθύρου, την σταθερα κανονικοποίσης $\epsilon \in$ [0.01,0.1] και την αρχική εκτίμηση $d_0 = [dx_0\ dy_0]$ για το πεδίο οπτικής ροής (η οποία για κάθε νέο frame αρχικοποείται στο 0) και επιστρέφει το d για κάθε pixel (x,y) στην περιοχή του προσώπου ανάμεσα στα δύο αυτά διαδοχικά frames. Η όλη διαδικασία συνήθως συγκλίνει μετά από 5-10 επαναλήψεις. Εμείς για κάθε δύο frame του βίντεο κάνουμε 7 επαναλήψεις. Για κάθε δύο διαδοχικά frames, κόβουμε την τις δύο εικόνες $I_{n-1}(x), I_n(x)$ με βάση το bounding box που υπολογίσαμε για την $I_{n-1}(x)$ (ξεκινάμε με το bounding box που υπολογίστηκε για το 1ο frame στο 1.1) και υπολογίζουμε την οπτική ροή μόνο για αυτά τα pixel της εικόνας, καθώς μας ενδιαφέρει μόνο η κίνηση του προσώπου και όχι π.χ. των χεριών ή του background το οποίο μάλιστα μετακινείται και ελάχιστα. Συνεχίζοντας, για να υπολογιστεί η οπτική ροή στην περιοχή του προσώπου από την εικόνα $I_n(x)$ στην εικόνα $I_{n+1}(x)$, θα πρέπει μέσω της οπτικής ροής που υπολογίστηκε στο προηγούμενο βήμα για την $I_n(x)$ να υπολογιστεί το συνολικό διάνυσμα μετατόπισης του boudning box από την $I_{n-1}(x)$ στην $I_n(x)$, ώστε να απομονώσουμε σωστά την περιοχή του προσώπου στις $I_n(x)$, $I_{n+1}(x)$. Η διαδικασία αυτή εξηγειται παρακάτω στο μέρος 1.4. Με τον τρόπο αυτό υπολγίζεται το πεδίο οπτικής ροής και η συνολική μετατόπιση του bounding box για κάθε frame. Παρακάτω βλέπουμε ενδεικτικά αποτελέσματα για κάποια frames:

<u>Μετατόπιση Ορθγωνίου:</u>

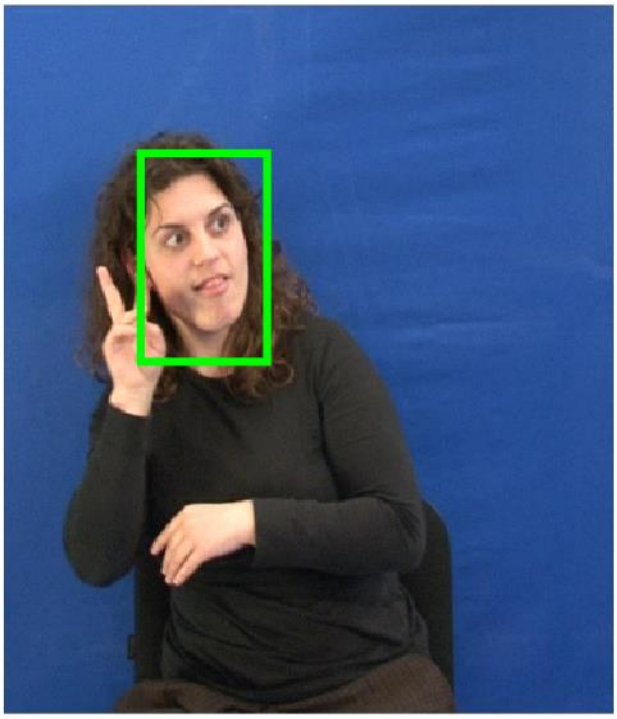
<u>Frame 26:</u>



<u>Frame 60:</u>



Frame 72:



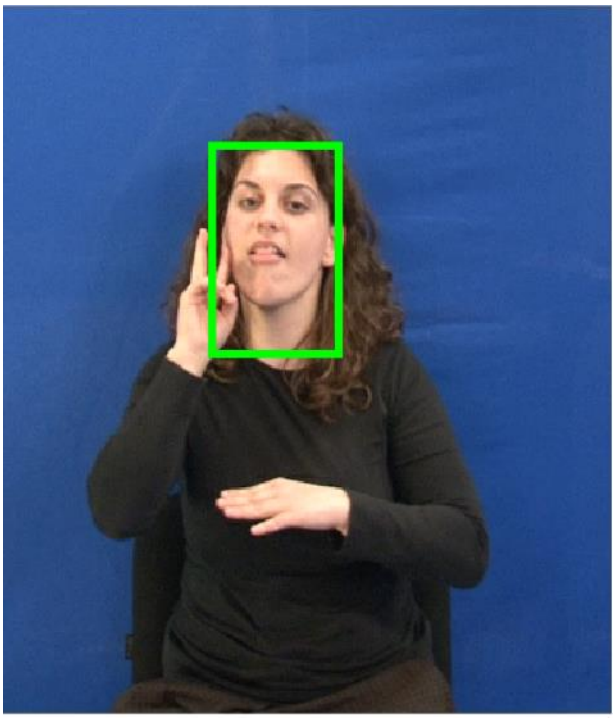
Τα παραπάνω Frames επιλέχθηκαν έτσι ώστε να είναι ευδιάκριτη η αποτελεσματικότητα του αλγορίθμου (σε διαδοχικά frames η μετατόπιση του bounding box είναι λιγότερο ευδιάκριτη).

1.3 Πολύ-Κλιμακωτός Υπολογισμός Οπτικής Ροής

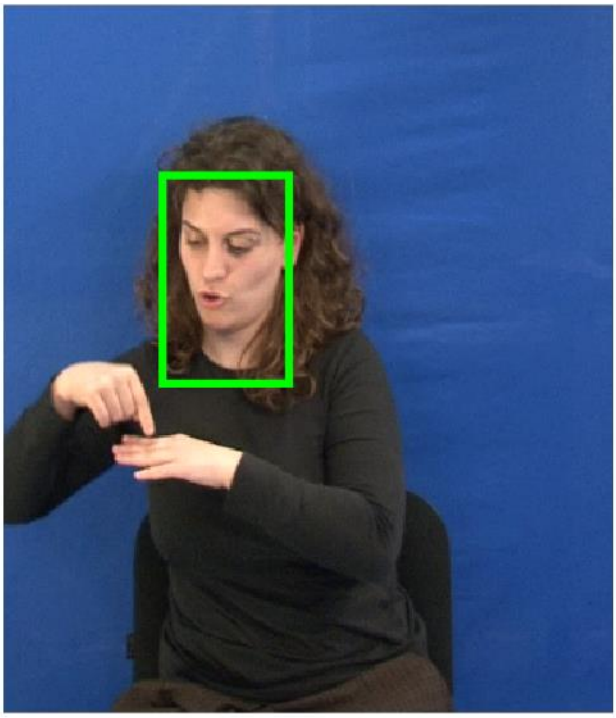
Στο ερώτημα αυτό θα υλοποιήσουμε την πολυκλιμακωτή εκδοχή του αλγόριθμου Lucas – Kanade. Ο αλγόριθμος αρχικά αναλύει τα δύο διαδοχικά frames σε Γκαουσιανές πυραμίδες, πραγματοποιώντας διαδοχικές υποδειγματοληψίες του αρχικού. Για τη μετάβαση από μεγάλες σε μικρές κλίμακες κατά την κατασκευή της Γκαουσιανής πυραμίδας φιλτράρουμε την εικόνα με ένα βαθυπερατό φίλτρο (χρησιμοποιήθηκε γκαουσιανή τυπικής απόκλισης 3 pixel) προκειμένου να μειωθεί η φασματική αναδίπλωση (aliasing) της εικόνας. Έτσι προκύπτουν σύνολα από frames σε ένα εύρος κλιμάκων ως εξής:

Στη συνέχεια ο αλγόριθμος υπολογίζει το πεδίο οπτικής ροής από τις πιο μικρές (τραχείς) στις πιο μεγάλες (λεπτομερείς) κλίμακες – δηλαδή από αριστερά προς τα δεξία με βάση τις παραπάνω εικόνες- , χρησιμοποιώντας το αποτέλεσμα της μικρότερης κλίμακας ως αρχική συνθήκη για την μεγαλύτερη. Τα αποτελέσματα του πολυκλιμακωτού αλγόριθμου για αριθμό κλιμάκων πυραμίδας N = 3 είναι τα εξής:

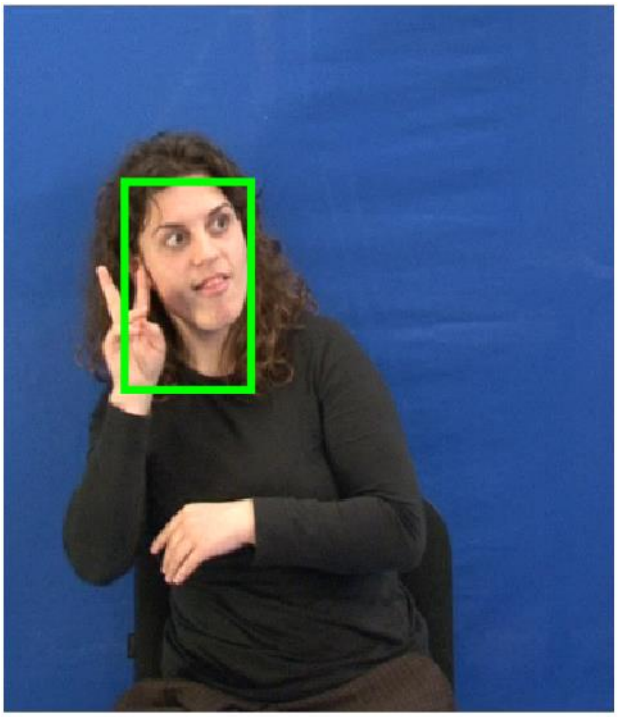
Frame 26:



Frame 60:

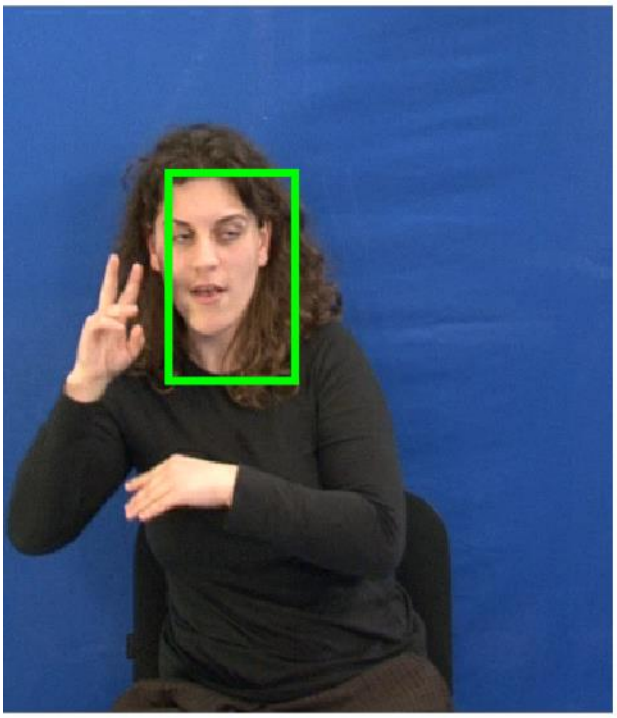


Frame 72:

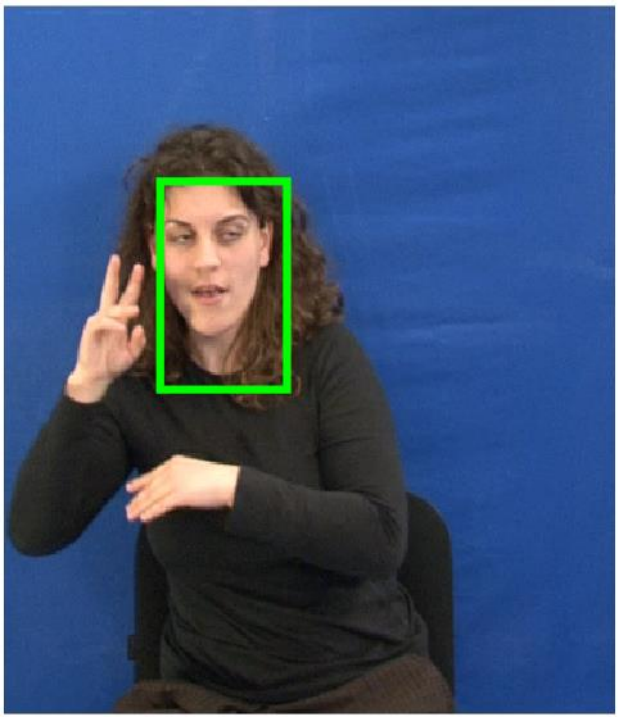


Παρατηρούμε ότι για τον αριθμό επαναλήψεων (7) που θέσαμε στον αλγόριθμο μονής κλίμακας μέχρι την σύγκλιση , τα αποτελέσματα των δύο αλγορίθμων δεν παρουσιάζουν σημαντική διαφορά. Παρόλαυτά τρέχοντας τον ίδιο αλγόριθμο για μικρότερο αριθμό επαναλήψεων, έστω 4, έχουμε τα εξής αποτελέσματα για κάθε μία από τις παραπάνω εκδοχές:

Μονή Κλίμακα:



Τρεις Κλίμακες:

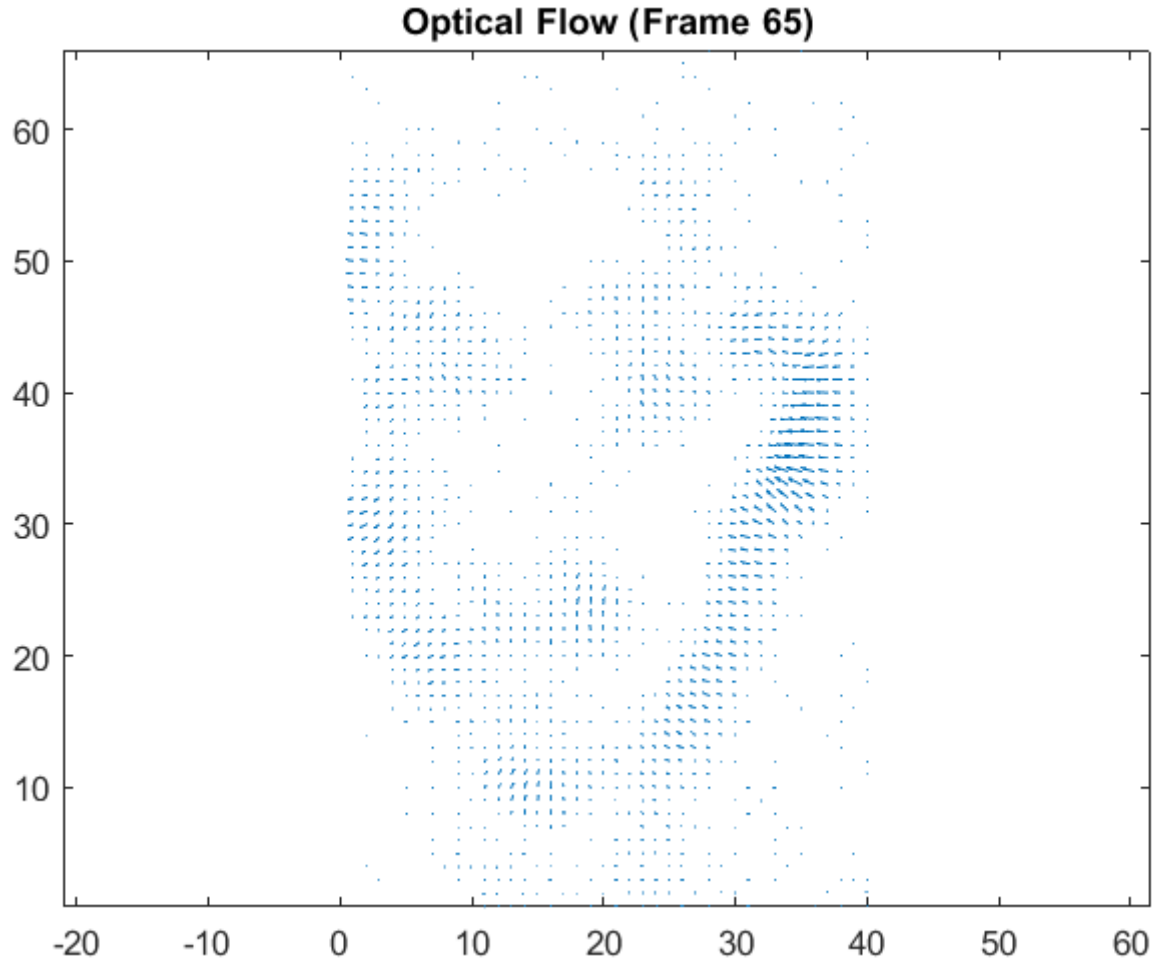


Όπως βλέπουμε υπάρχει εμφανής διαφορά στην αποτελεσματικότητα των αλγορίθμων για μικρότερο αριθμό επαναλήψεων, δηλαδή η πολυκλιμακωτή εκδοχή συγκλίνει ταχύτερα από τον αλγόριθμο μονής κλίμακας, πράγμα το οποίο ήταν και αναμενόμενο, καθώς στην πολυκλιμακωτή εκδοχή εξετάζεται κάθε frame με μεγαλύτερη λεπτομέρεια.

1.4 Υπολογισμός Μετατόπισης του Προσώπου από το Πεδίο Οπτικής Ροής

Όπως αναφέρθηκε παραπάνω, έχοντας υπολογίσει την οπτική ροή της εικόνας I_n στην περιοχή που είχε ορίσει το bounding box της εικόνας I_{n-1}, απομένει να βρούμε το συνολικό διάνυσμα μετατόπισης του bounding box ορθογωνίου, με όσο το δυνατόν μεγαλύτερη ακρίβεια. Για το σκοπό αυτό θα υπολογίσουμε την μέση τιμή των διανυσμάτων μετατόπισης στις 2 κατευθύνσεις. Από την παρακάτω εικόνα του πεδίου οπτικής ροής, παρατηρούμε ότι το μήκος των διανυσμάτων διαφέρει, αναλόγως με την περιοχή. Σε περιοχές με έντονη πληροφορία (ακμές, κορυφές κλπ.) τα διανύσματα έχουν μεγαλύτερο μήκος, ενώ σε περιοχές με ομοιόμορφη και επίπεδη υφή (π.χ. μάγουλα, μέτωπο κλπ.) έχουν σχεδόν μηδενικό μήκος. Για το λόγο αυτό στον υπολογισμό του μέσου όρου των διανυσμάτων μετατόπισης θα ληφθούν υπόψιν μόνο τα διανύσματα με ενέργεια μεγαλύτερη από μία τιμή κατωφλιού. Ως ενέργεια διανύσματος ταχύτητας ορίζουμε $\|d\|^2 = d_x^2 + d_y^2$. Βλέπουμε το πεδίο και την ενέργεια οπτικής ροής για το 65ο frame:

Πεδίο Οπτικής Ροής:

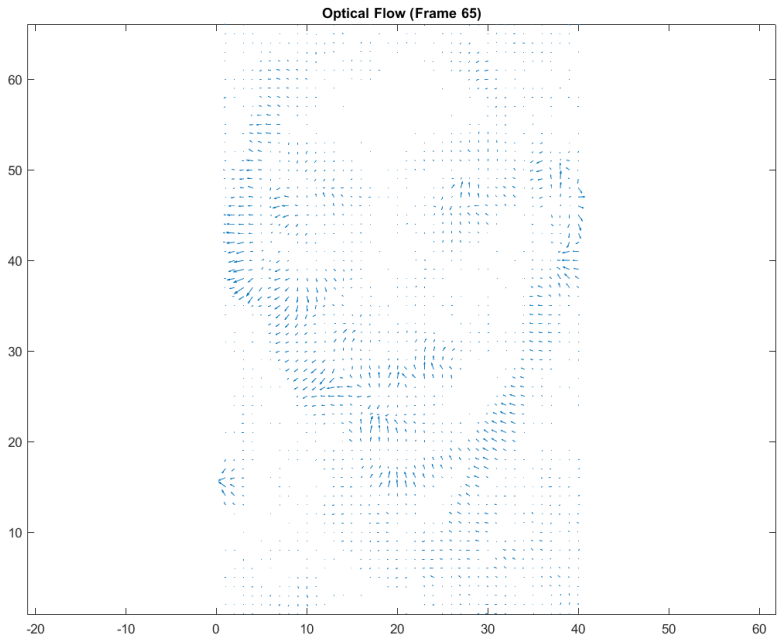


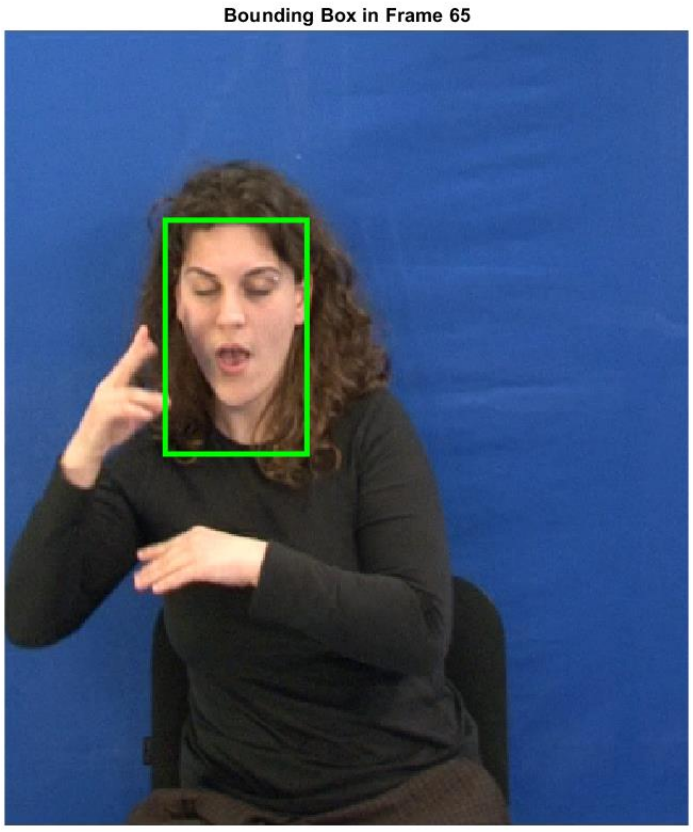
Ενέργεια διανυσμάτων οπτικής ροής:

Πειραματισμός με διαφορετικές τιμές παραμέτρων ρ, ϵ και κατωφλιού ενέργειας:

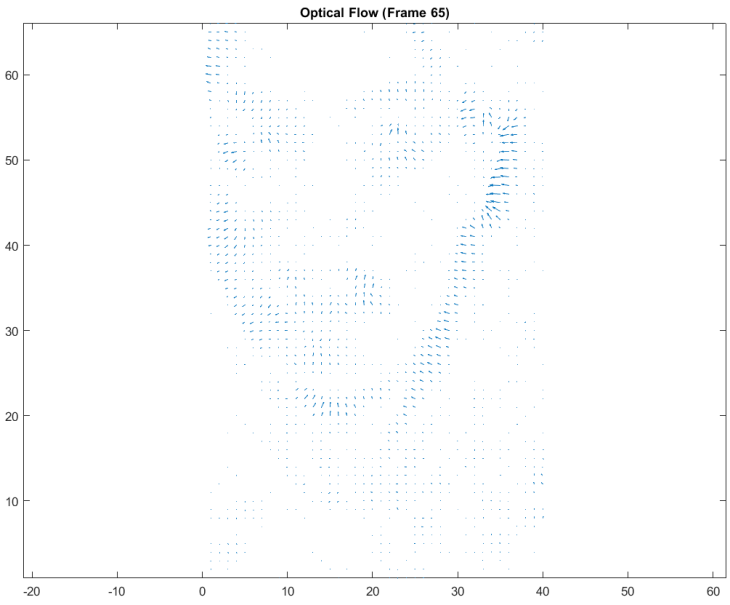
Τα αποτελέσματα που φαίνονται στις παραπάνω εικόνες προέκυψαν για ρ = 3, ϵ = 0.05 και κατώφλι = 0.9. Θα πειραματιστούμε με πολύ διαφορετικές τιμές παραμέτρων.

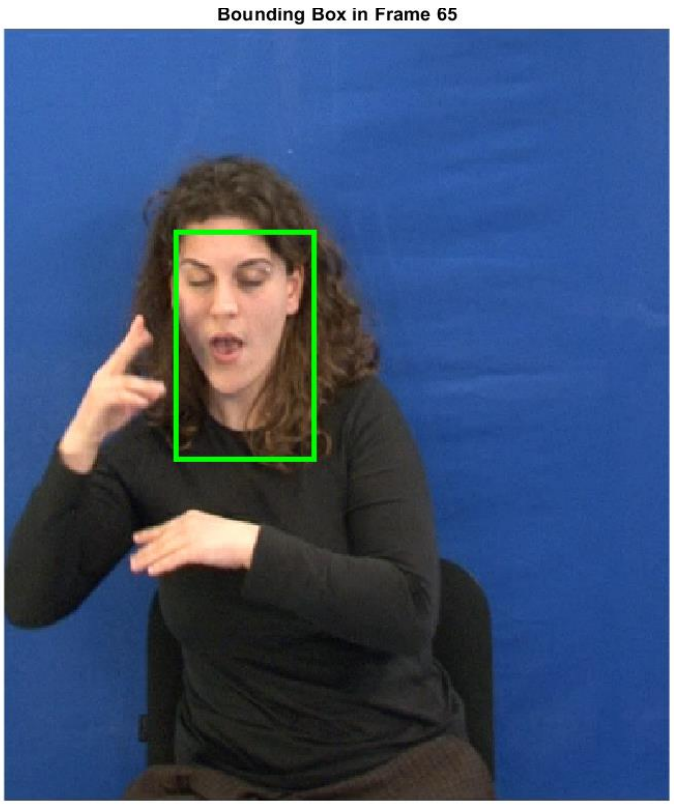
Θέτουμε ρ = 1, ϵ = 0.01, κατώφλι = 0.5 και έχουμε τα εξής αποτελέσματα:



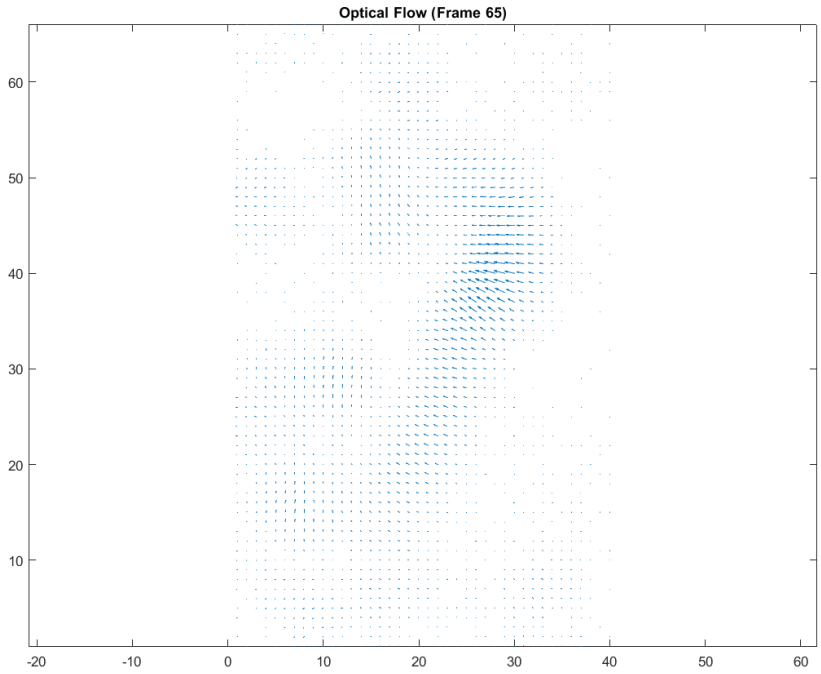


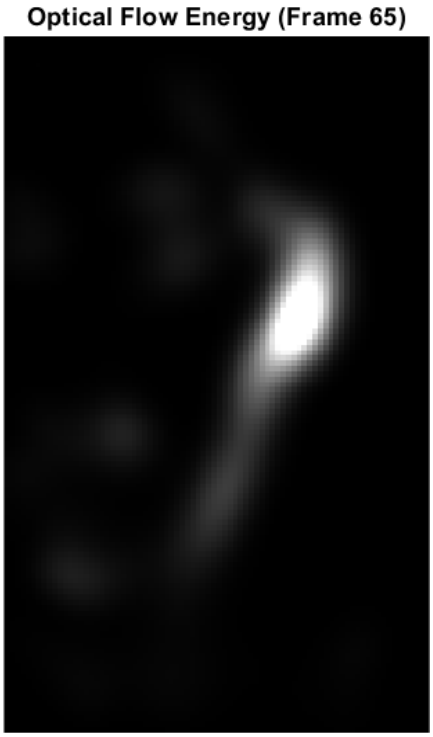
Θέτουμε ρ = 1, ϵ = 0.05, κατώφλι = 0.9 και έχουμε τα εξής αποτελέσματα:

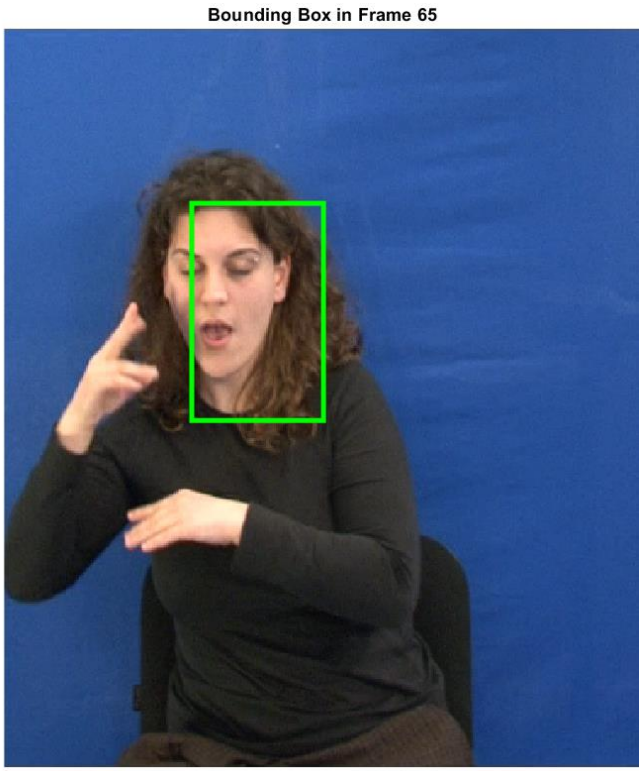




Θέτουμε ρ = 5, ϵ = 0.1, κατώφλι = 0.9 και έχουμε τα εξής αποτελέσματα:







Σύγκριση Αποτελεσμάτων:

Παρατηρούμε γενικά ότι η μεγαλύτερη διαφορά στο αποτέλεσμα προκύπτει από την μεταβολή της παραμέτρου ρ. Συγκεκριμένα το καλύτερο αποτέλεσμα παρουσιάζεται για μικρές τιμές του ρ [1,3] , ενώ για μεγάλες τιμές (π.χ. ρ = 5) παρατηρούμε ότι το αποτέλεσμα δεν είναι το επιθυμητό. Αυτό συμβαίνει διότι για μεγαλύτερες τιμές της παραμέτρου μεγαλώνει το μέγεθος του Γκαουσιανού πυρήνα με αποτέλεσμα να γίνεται υπερβολική εξομάλυνση, να χάνεται πληροφορία και να μειώνεται η ενέργεια ακόμα και σε pixel που αυτό δεν είναι επιθυμητό. Επίσης παρατηρήσαμε ότι αυξάνοντας το κατώφλι ενέργειας και διατηρώντας τις υπόλοιπες παραμέτρους σταθερές παίρνουμε καλύτερα αποτελέσματα, καθώς λαμβάνονται υπόψιν όλο και λιγότερα pixel ομοιόμορφων περιοχών, των οποίων η σχεδόν μηδενική ενέργεια επηρεάζει ανεπιθύμητα το μέσο όρο μετατόπισης. Τέλος για την αλλαγή της σταθεράς ϵ διατηρώντας τις υπόλοιπες παραμέτρους σταθερές παρατηρούμε ότι το καλύτερο αποτέλεσμα προκύπτει για μία μέση τιμή του διαστήματος (π.χ. 0.05), ενώ για ακραίες τιμές 0.01, 0.1 το ορθογώνιο απομακρύνεται από την περιοχή του προσώπου.

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

Στο δεύτερο μέρος της εργαστηριακής άσκησης θα ασχοληθούμε με την εξαγωγή χωρο-χρονικών χαρακτηριστικών με στόχο την εφαρμογή τους στο πρόβλημα κατηγοριοποίησης βίντεο που περιέχουν ανθρώπινες δράσεις. 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 αντικείμενο με μία σειρά από τοπικούς περιγραφητές που υπολογίζονται σε γειτονιές ανιχνευθέντων σημείων ενδιαφέροντος. Τελικά, η συλλογή των τοπικών χαρακτηριστικών ενσωματώνεται σε μία 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.

Για την συγκεκριμένη άσκηση δόθηκαν βίντεο από 3 κλάσεις δράσεων (walking, running, boxing) από τα οποία θα εξαχθούν χωρο-χρονικοί περιγραφητές με σκοπό την κατηγοριοποίηση των δράσεων αυτών.

2.1 Χωρο-χρονικά Σημεία Ενδιαφέροντος

Αρχικά επιθυμούμε να εντοπίσουμε τα χωρο-χρονικά σημεία ενδιαφέροντος. Ο εντοπισμός τους μπορεί να γίνει με χρήση διαφόρων ανιχνευτών. Στην άσκηση θα υλοποιήσουμε τους εξής:

i) Harris Detector και ii) Gabor Detector.

Harris Detector

Ο ανιχνευτής αυτός αποτελεί μία επέκταση τις 3 διαστάσεις του ανιχνευτή γωνιών Harris – Stephens. Πρακτικά στην 3Δ εκδοχή ψάχνουμε για γωνίες όχι μόνο στο χώρο αλλά και στον χρόνο. Για κάθε voxel του βίντεο υπολογίζουμε τον 3x3 πίνακα M(x,y,t) , προσθέτοντας στον δομικό τανυστή J και την χρονική παράγωγο:

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

όπου g(x, y, y; sσ, sτ) ένας 3Δ γκαουσιανός πυρήνας ομαλοποίησης και $\nabla L(x,\ y,\ t;\sigma,\tau)$ οι χωροχρονικές παράγωγοι για την χωρική κλίμακα σ και τη χρονική κλίμακα τ. Τις παραγώγους (χωρικές και χρονικές) τις υπολογίσαμε εφαρμόζοντας συνέλιξη και στις 3 διαστάσεις με τον πυρήνα κεντρικών διαφορών $[-1\ 0\ 1]^T$,

προσαρμοσμένο στην κατάλληλη διάσταση. Η ομαλοποίηση των χωροχρονικών παραγώγων υλοποιήθηκε με 3Δ γκαουσιανό φιλτράρισμα, μπορεί όμως να υλοποιηθεί εύκολα και τρεις 1Δ συνελίξεις, εκμεταλλευόμενοι τη διαχωρισιμότητα (seperability) του φίλτρου. Προκύπτουν έτσι :

$$J_{11}(x,y,t) = g(x, y, t; s\sigma, s\tau) * \left(\frac{\partial L(x,y,t;\sigma,\tau)}{\partial x} \cdot \frac{\partial L(x,y,t;\sigma,\tau)}{\partial x}\right)$$
$$J_{12}(x,y,t) = g(x, y, t; s\sigma, s\tau) * \left(\frac{\partial L(x,y,t;\sigma,\tau)}{\partial x} \cdot \frac{\partial L(x,y,t;\sigma,\tau)}{\partial y}\right)$$
$$J_{13}(x,y,t) = g(x, y, t; s\sigma, s\tau) * \left(\frac{\partial L(x,y,t;\sigma,\tau)}{\partial x} \cdot \frac{\partial L(x,y,t;\sigma,\tau)}{\partial t}\right)$$
$$J_{21}(x,y,t) = J_{12}(x,y,t)$$
$$J_{22}(x,y,t) = g(x, y, t; s\sigma, s\tau) * \left(\frac{\partial L(x,y,t;\sigma,\tau)}{\partial y} \cdot \frac{\partial L(x,y,t;\sigma,\tau)}{\partial y}\right)$$
$$J_{23}(x,y,t) = g(x, y, t; s\sigma, s\tau) * \left(\frac{\partial L(x,y,t;\sigma,\tau)}{\partial y} \cdot \frac{\partial L(x,y,t;\sigma,\tau)}{\partial t}\right)$$
$$J_{31}(x,y,t) = J_{13}(x,y,t)$$
$$J_{32}(x,y,t) = J_{23}(x,y,t)$$
$$J_{33}(x,y,t) = g(x, y, t; s\sigma, s\tau) * \left(\frac{\partial L(x,y,t;\sigma,\tau)}{\partial t} \cdot \frac{\partial L(x,y,t;\sigma,\tau)}{\partial t}\right)$$

Και συνεπώς :

$$M(x, y, t;\sigma, \tau) = \begin{bmatrix} J_{11} & J_{12} & J_{13} \\ J_{21} & J_{22} & J_{23} \\ J_{31} & J_{32} & J_{33} \end{bmatrix}$$

Το 3Δ κριτήριο γωνιότητας ακολουθεί και αυτό την ίδια λογική με τις 2Δ:

$$H(x, y, t) = \det(M(x, y, t)) - k\cdot\text{trace}^3(M(x, y, t))$$

Αντικαθιστώντας και μετά από πράξεις προκύπτει:

$$H(x, y, t) = J_{11}\cdot(J_{22}\cdot J_{33} - J_{23}^2) - J_{12}\cdot(J_{21}\cdot J_{33} - J_{31}\cdot J_{23}) + J_{13}\cdot(J_{21}\cdot J_{32} - J_{31}\cdot J_{22}) - k\cdot((J_{11}+J_{22}+J_{33}))^2.$$

Τέλος, ταξινομούμε τον 3Δ πίνακα H σε φθίνουσα σειρά και αφού «πετάξουμε» τα πρώτα 100 στοιχεία, παίρνουμε τα πρώτα 500 σημεία με τη μέγιστη τιμή. Τα πρώτα 100 στοιχεία τα κάνουμε discard, διότι είναι μεν σημεία με μεγάλη τιμή αλλά στην πραγματικότητα είναι artefacts, έχουν δηλαδή λανθασμένα μεγάλη τιμή απλώς επειδή βρίσκονται στα όρια, είτε χρονικά, είτε χωρικά, χωρίς όμως να είναι όντως σημεία ενδιαφέροντος. Πειραματικά βρέθηκε πως το discard των πρώτων 100 σημείων αντιμετωπίζει αυτό το πρόβλημα.

<u>Gabor Detector</u>

Ο Gabor ανιχνευτής βασίζεται στο χρονικό φιλτράρισμα του βίντεο με ένα ζεύγος Gabor φίλτρων αφού πρώτα αυτό έχει υποστεί εξομάλυνση στις χωρικές διαστάσεις μέσω ενός 2Δ γκαουσιανού πυρήνα g(x,y;σ) με τυπική απόκλιση σ. Τα Gabor ορίζονται ως:

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

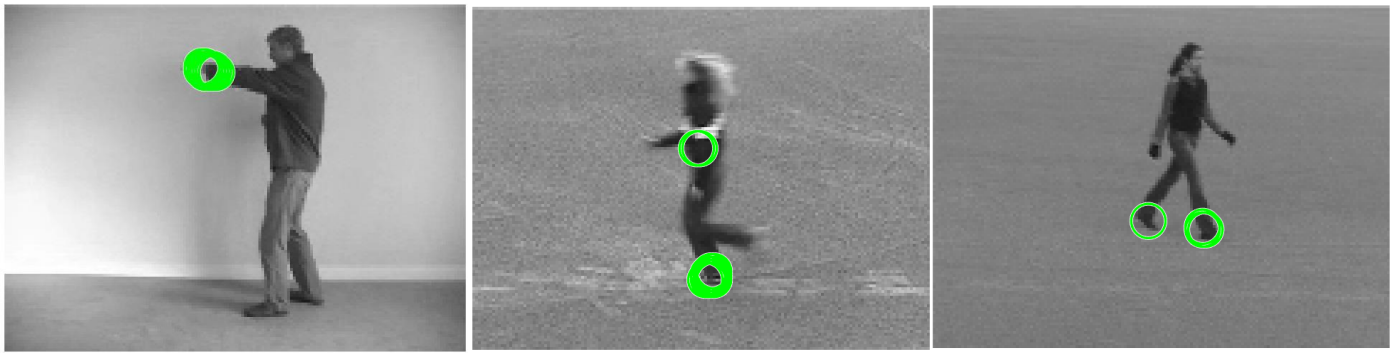
Για τον υπολογισμό της κρουστικής απόκρισης των Gabor θεωρήσαμε μέγεθος παραθύρου [-2τ, 2τ] και κανονικοποιήσαμε με την L1 νόρμα. Η συχνότητα ω του φίλτρου συνδέεται με τη χρονική κλίμακα τ (απόκλιση της Γκαουσιανής συνιστώσας του) μέσω της σχέσης: ω = 4/τ. Το κριτήριο σημαντικότητας προκύπτει τώρα παίρνοντας την τετραγωνική ενέργεια της εξόδου για το ζεύγος Gabor φίλτρων:

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

Τέλος, ταξινομούμε τον 3Δ πίνακα H σε αύξουσα σειρά και αφού «πετάξουμε» τα πρώτα 100 στοιχεία, παίρνουμε τα πρώτα 500 σημεία με την ελάχιστη τιμή. Τα πρώτα 100 στοιχεία τα κάνουμε discard για τον ίδιο λόγο που εξηγήθηκε παραπάνω.

Παρακάτω βλέπουμε τα σημεία ενδιαφέροντος που ανιχνεύθηκαν από τους 2 ανιχνευτές για ορισμένα ενδεικτικά frames:

Harris:



Gabor:

2.2 Χωρο-χρονικοί Ιστογραφικοί Περιγραφητές

Οι χωρο-χρονικού περιγραφητές που θα χρησιμοποιηθούν βασίζονται στον υπολογισμό ιστογραμμάτων της κατευθυντικής παραγώγου (HOG – Histograms of Oriented Gradients) και της οπτικής ροής (HOF – Histograms of Oriented Flow) γύρω από τα σημεία ενδιαφέροντος που ανιχνεύσαμε.

Συγκεκριμένα για κάθε frame υπολογίζουμε αρχικά το διάνυσμα κλίσης (gradient) και για κάθε σημείο ενδιαφέροντος στο frame αυτό υπολογίζουμε την οπτική ροή σε μία τετραγωνική περιοχή 4 x scale γύρω από αυτό, με τον ίδιο τρόπο που εξηγήθηκε στο 1^{ο} μέρος της άσκησης. Στη συνέχεια χρησιμοποιούμε την έτοιμη συνάρτηση OrientationHistogram.p προκειμένου να υπολογίσουμε τους 2 ιστογραφικούς περιγραφητές. Η συνάρτηση αυτή δέχεται ως είσοδο το διανυσματικό πεδίο (είτε κατευθυντικές παραγώγους είτε κατεύθυνση ροής για την περιοχή γύρω από το σημείο ενδιαφέροντος),το μέγεθος του grid, το πλήθος των bins και επιστρέφει την ιστογραμματική περιγραφή της αντίστοιχης περιοχής. Συνεπώς καλούμε την συνάρτηση μία φορά για κάθε σημείο ενδιαφέροντος και για κάθε διανυσματικό πεδίο και συνενώνουμε τους δύο επιμέρους περιγραφητές για να προκύψει ένας ενιαίος χωρο-χρονικός περιγραφητής.

Υπολογίζουμε την τελική αναπαράσταση global representation για κάθε βίντεο υλοποιώντας την bag of visual words BoVW τεχνική ως εξής:

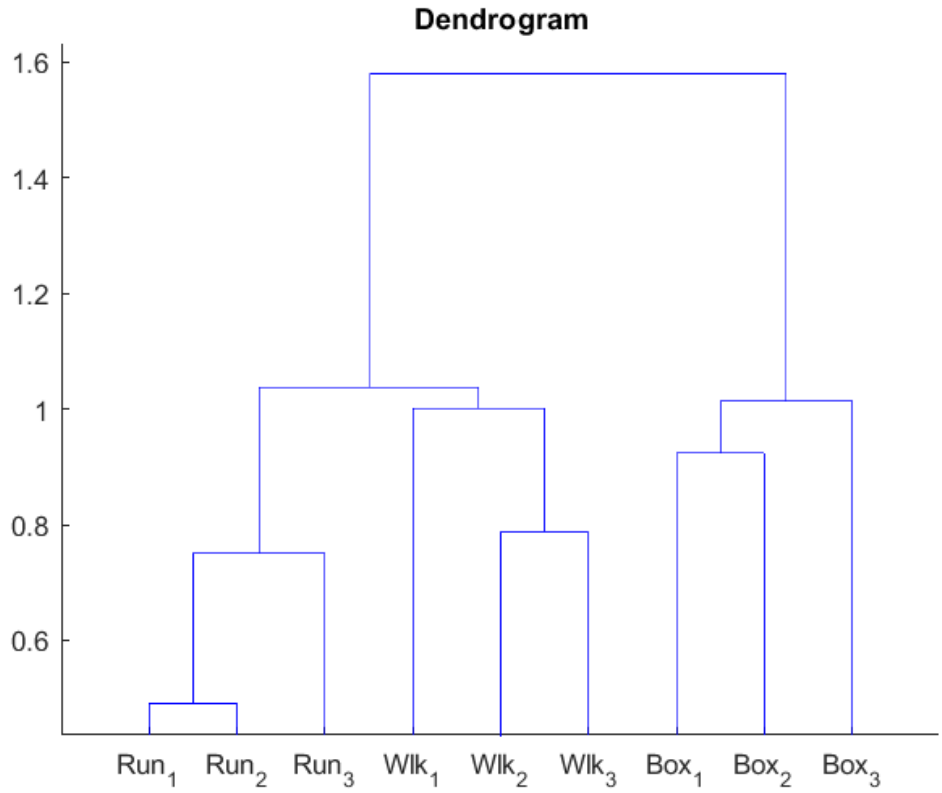
Δημιουργούμε αρχικά το οπτικό λεξικό βάση του οποίου θα υπολογιστούν τα ιστογράμματα. Για τον προσδιορισμό των οπτικών λέξεων του λεξικού θα εφαρμοστεί ο αλγόριθμος συσταδοποίησης kmeans για το σύνολο των περιγραφητών που υπολογίστηκαν στο προηγούμενο βήμα. Στη συνέχεια υπολογίζουμε την ελάχιστη ευκλείδεια απόσταση του κάθε τοπικού περιγραφητή από τα κέντρα που υπολογίστηκαν και με βάση αυτή κατασκευάζουμε για κάθε βίντεο το ιστόγραμμα που προκύπτει από τη συχνότητα εμφάνισης των οπτικών λέξεων του λεξικού. Κανονικοποιούμε κάθε ιστόγραμμα με την L2 νόρμα.

2.3 Κατασκευή Δενδρογράμματος για τον Διαχωρισμό των Δράσεων

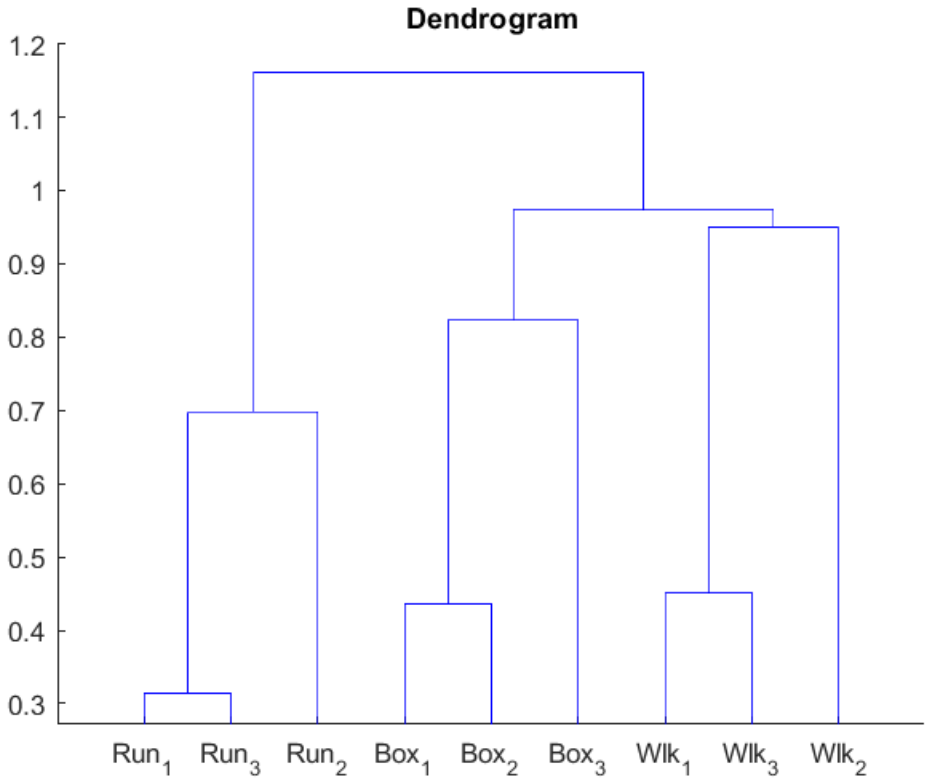
Στο ερώτημα αυτό θα ασχοληθούμε με την ικανότητα κατηγοριοποίησης των βίντεο με τις ανθρώπινες δράσεις σε 3 διαφορετικές κατηγορίες/κλάσεις με χρήση των BoVW ιστογραμμάτων. Θα κατασκευάσουμε ένα δενδρόγραμμα αποστάσεων που αντιπροσωπεύει την ικανότητα διαχωρισμού των τριών διαφορετικών κατηγοριών, προκειμένου να οπτικοποιηθεί η απόσταση των διανυσμάτων χαρακτηριστικών. Για την κατασκευή του δενδρογράμματος χρησιμοποιήθηκε η x^2 απόσταση, η οποία είναι κατάλληλη για ιστογράμματα και ορίζεται ως εξής:

Για 2 ιστογράμματα αν $H_i = \{h_{i1},h_{i2},\dots,h_{iK}\}$ και $H_j = \{h_{j1},h_{j2},\dots,h_{jK}\}$, όπου K το πλήθος των κέντρων, τότε $D(H_i, H_j) = \frac{1}{2}\sum_{n=1}^{K} \frac{(h_{in}-h_{jn})^2}{h_{in}+h_{jn}}$. Γενικά μπορούν να χρησιμοποιηθούν διαφορετικοί συνδυασμοί ανιχνευτών/περιγραφητών/αποστάσεων.

Harris Ανιχνευτής, HOG/HOF περιγραφητής, x^2 απόσταση:



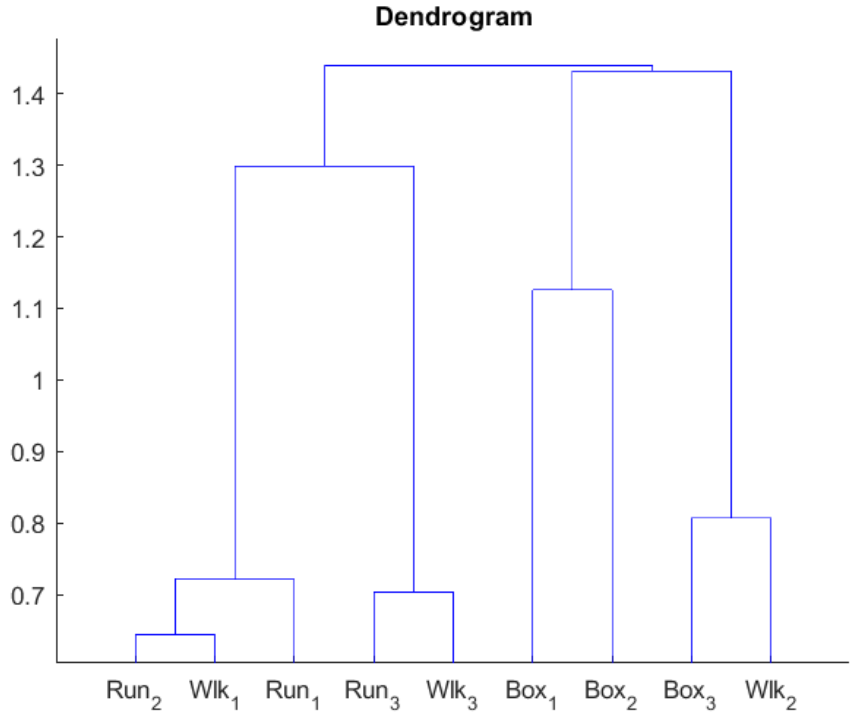
Gabor Ανιχνευτής, HOG/HOF περιγραφητής, x^2 απόσταση:



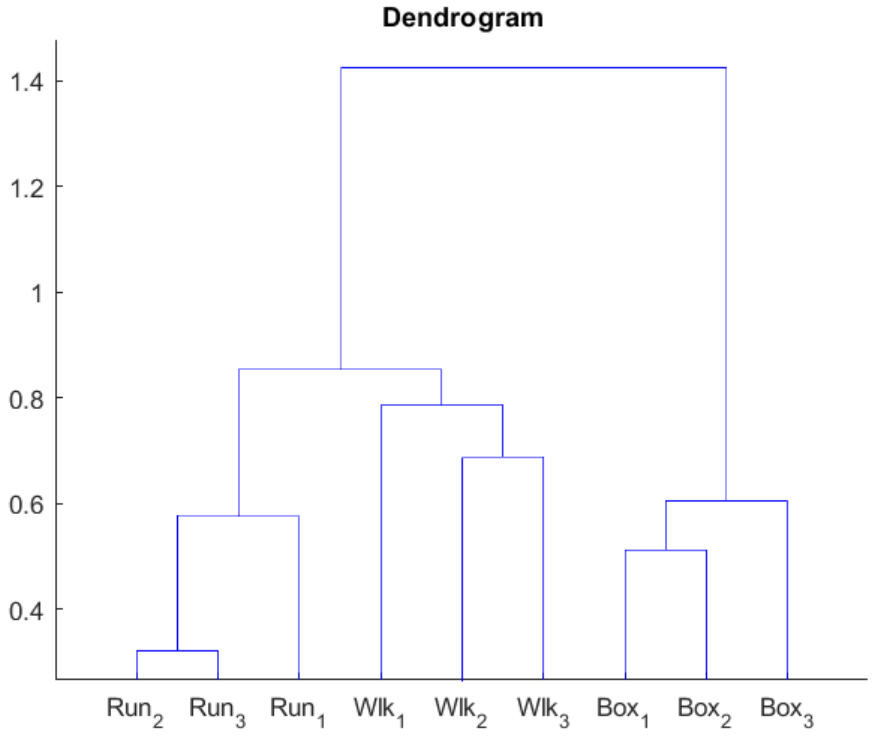
Τα παραπάνω δενδρογράμματα είναι το καλύτερο αποτέλεσμα που πετύχαμε, πειραματιζόμενοι με τις τιμές των παραμέτρων των δύο ανιχνευτών καθώς και με το είδος της απόστασης που θα χρησιμοποιηθεί. Τα δενδρογράμματα, αν και όχι βέλτιστα, είναι αρκετά ικανοποιητικά, καθώς παρατηρούμε ότι η απόσταση των δράσεων που βρίσκονται στην ίδια κλάση είναι αρκετά χαμηλή (0.4 – 1), ενώ η απόσταση των δράσεων που βρίσκονται σε διαφορετικές κλάσεις είναι υψηλότερη (1-1.6). Παρατηρούμε ότι υπάρχουν αποστάσεις μεταξύ διαφορετικών δράσεων, των οποίων η απόσταση δεν είναι τόσο ψηλά, όσο θα περίμενε κανείς. Αυτό συμβαίνει, διότι υπάρχουν δράσεις, οι οποίες έχουν κοινά χαρακτηριστικά (π.χ. περπάτημα και τρέξιμο ανιχνεύουν και οι δύο κίνηση στα πόδια) και συνεπώς συγχέονται ως ένα βαθμό.

Πειραματιζόμενοι με διαφορετικούς συνδυασμούς ανιχνευτών/ περιγραφητών προκύπτουν τα εξής αποτελέσματα:

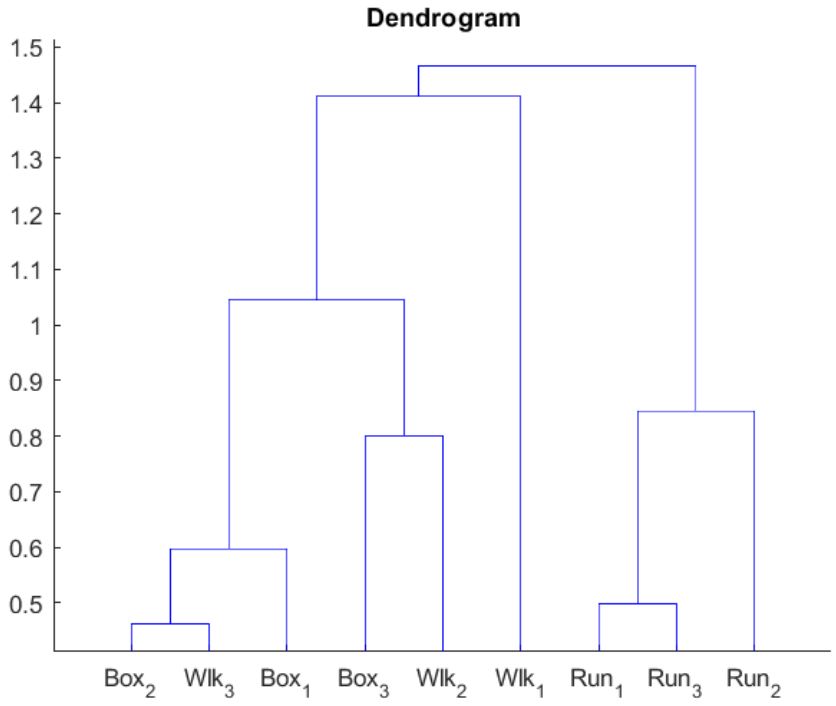
Harris / HOG



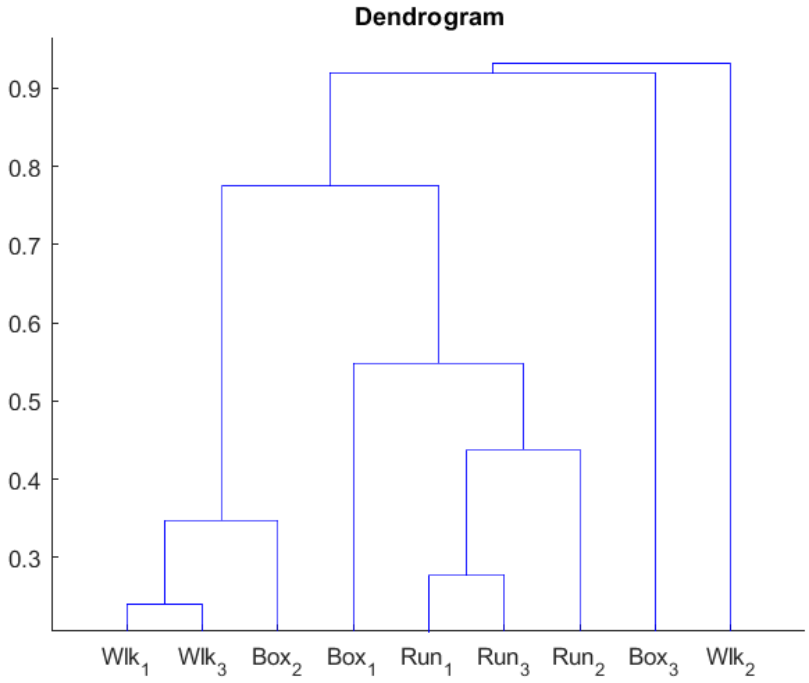
Harris /HOF



Gabor/HOG



Gabor/HOF



Παρατηρούμε όμως ότι σημαντικότερη επίδραση στην αξιοπιστία των αποτελεσμάτων έχει η απώλεια του HOF περιγραφητή, ενώ η αποκλειστική χρήση του είναι πιο κοντά στο βέλτιστο αποτέλεσμα που πετύχαμε. Αυτό ήταν αναμενόμενο, καθώς για τον διαχωρισμό των δράσεων ενδιαφερόμαστε περισσότερο για τα χρονικά δεδομένα και λιγότερο για τα χωρικά. Όπως είναι φανερό, ο διαχωρισμός των δράσεων σε κλάσεις με αποκλειστική χρήση του HOG περιγραφητή δεν είναι καθόλου αποτελεσματικός. Κλάσεις με διαφορετικές δράσεις συγχέονται, ενώ άλλες με ίδια ή παρόμοια δράση παρουσιάζουν μεγάλη απόσταση. Η σύγχυση αυτή υπάρχει, διότι σε αυτά τα δενδρογράμματα χρησιμοποιήθηκε αποκλειστικά χωρικός περιγραφητής και όχι ο συνδυασμός αυτών, με αποτέλεσμα να χάνεται κάθε φορά σημαντική πληροφορία στο πεδίο του χρόνου.