Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών

Όραση Υπολογιστών

8ο Εξάμηνο - Ροή Σ

Δεύτερη Εργαστηριακή Αναφορά

Εκτίμηση Οπτικής Ροής και Εξαγωγή Χαρακτηριστικών
σε Βίντεο για Αναγνώριση Δράσεων

Δημήτρης Δήμος - 031 17 165
dimitris.dimos647@gmail.com

Χρήστος Δημόπουλος - 031 17 037
chrisdim1999@gmail.com

Αθήνα
Άνοιξη, 2021

Περιεχόμενα

Μέρος 1: Παρακολούθηση Προσώπου και Χεριών με Χρήση της Μεθόδου Οπτικής Ροής των Lucas-Kanade

Θέμα

Στο πρώτο μέρος του εργαστηριακού project καλούμαστε να υλοποιήσουμε ένα σύστημα παρακολούθησης προσώπου και χεριών (Face and Hands Tracking) πάνω σε βίντεο ελληνικής νοηματικής γλώσσας γυρισμένο σε στούντιο με ελεγχόμενο φωτισμό. Σε πρώτο στάδιο, ανιχνεύουμε στο πρώτο πλαίσιο την περιοχή του προσώπου και των χεριών (περιοχές ενδιαφέροντος) με χρήση ενός πιθανοτικού ανιχνευτή ανθρώπινου δέρματος. Στη συνέχεια παρακολουθούμε τις περιοχές ενδιαφέροντος χρησιμοποιώντας τα εξαγόμενα διανύσματα οπτικής ροής, υπολογισμένα με τη μέθοδο των Lucas-Kanade.

Ανίχνευση Δέρματος Προσώπου και Χεριών

Στο πρώτο ερώτημα ζητείται η ανίχνευση σημείων δέρματος στο πρώτο πλαίσιο της ακολουθίας και η τελική επιλογή της περιοχής του προσώπου και των χεριών. Για την ανίχνευση των σημείων δέρματος χρησιμοποιείται ο χρωματικός χώρος YCbCr, αφαιρώντας την πληροφορία της φωτεινότητας Y και διατηρώντας τα κανάλια Cb και Cr που περιγράφουν την ταυτότητα του χρώματος. Το χρώμα του δέρματος μοντελοποιείται με μια διδιάστατη Γκαουσιανή κατανομή:

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

όπου $\mathbf{c}$ είναι το διάνυσμα τιμών Cb και Cr για κάθε σημείο (x, y) της εικόνας. Η Γκαουσιανή κατανομή εκπαιδεύεται υπολογίζοντας το 2×1 διάνυσμα μέσης τιμής $\boldsymbol{\mu} = [\mu_{Cb}\mu_{Cr}]^T$ και τον 2×2 πίνακα συνδιακύμανσης Σ από τα δείγματα δέρματος που δίνονται στο αρχείο skinSamplesRGB.mat σε μορφή RGB, ενώ κανονικοποιείται στο διάστημα [0,1]. Στις Εικόνες 1 και 2 φαίνεται η Πυκνότητα Πιθανότητας της εν λόγω Γκαουσσιανής, όπως αυτή εκπαιδεύτηκε από τα δοθέντα δείγματα δέρματος. Ακολούθως, εστιάζοντας στο πρώτο frame του βίντεο νοηματικής γλώσσας, δημιουργούμε τη δυαδική εικόνα ανίχνευσης δέρματος, η οποία προκύπτει από την εικόνα πιθανοτήτων $P(\mathbf{c}(x, y) = \text{skin}), \forall(x, y)$ με κατωφλιοποίηση. Έχοντας κανονικοποιήσει την πυκνότητα πιθανότητας της Γκαουσσιανής, επιλέγουμε ένα κατώφλι πιθανότητας Threshold $= 0.2$. Η δυαδική εικόνα ανίχνευσης δέρματος που αντιστοιχεί στο πρώτο frame φαίνεται στην Εικόνα 3, με τις λευκές περιοχές να αντιστοιχούν σε περιοχές δέρματος.

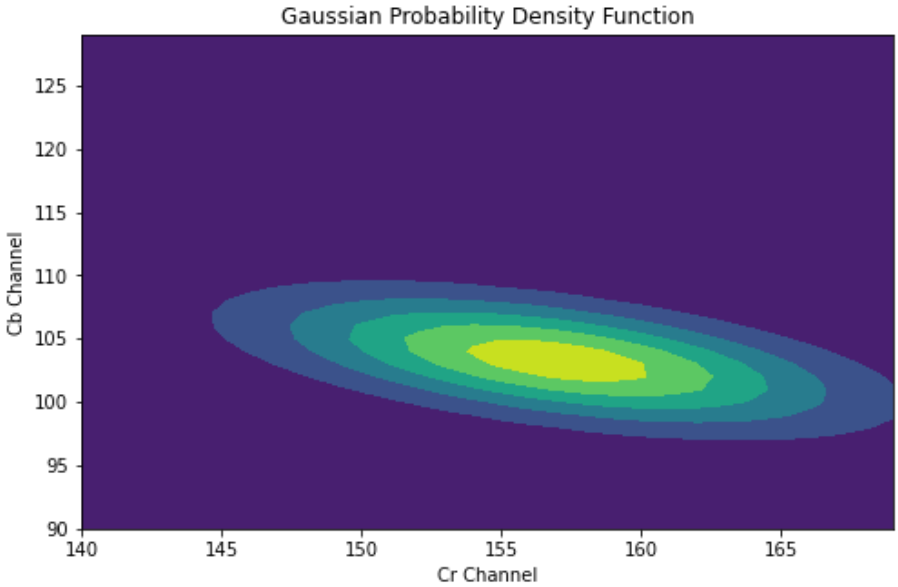


Figure 1: Πυκνότητα Κατανομής Πιθανότητας της Γκαουσσιανής Δέρματος στο επίπεδο Cb-Cr

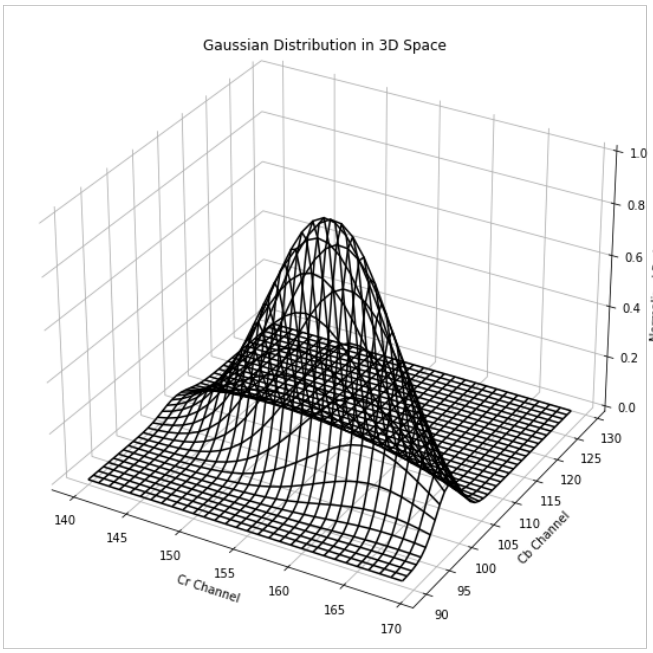


Figure 2: Τρισδιάστατη απεικόνιση της Πυκνότητας Κατανομής Πιθανότητας της Γκαουσσιανής Δέρματος

Figure 3: Δυαδική Εικόνα Ανίχνευσης Δέρματος για το 1ο frame

Στη συνέχεια, προκειμένου να εξαλείψουμε τυχόν συνεκτικές συνιστώσες μικρού εμβαδού και να καλύψουμε προεξοχές, προβαίνουμε σε μια μορφολογική επεξεργασία της δυαδικής εικόνας δέρματος. Συγκεκριμένα, γίνεται κάλυψη των τρυπών που εμφανίζονται, εφαρμόζοντας opening με ένα πολύ μικρό δομικό στοιχείο (διάστασης 2×2 pixels) και closing στο αποτέλεσμα με ένα μεγάλο δομικό στοιχείο (διάστασης 25×25 pixels). Ως εκ τούτου, εξαλείφονται οι μικρές περιοχές και αποκτούν συνοχή οι περιοχές του προσώπου και των χεριών. Τα δύο στάδια της μορφολογικής επεξεργασίας φαίνονται στην Εικόνα 4.

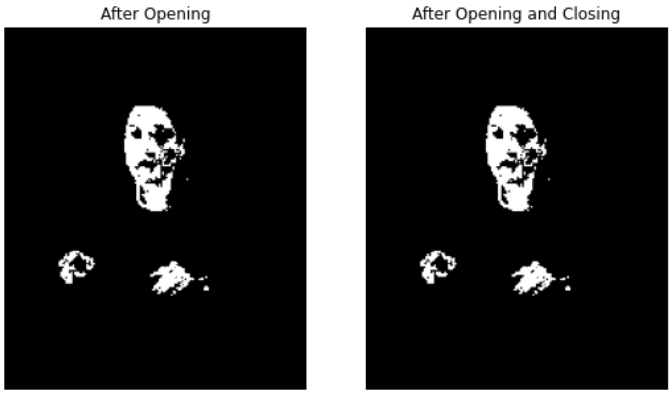


Figure 4: Εφαρμογή φίλτρου opening στη δυαδική εικόνα δέρματος (αριστερά) και φίλτρου closing στην προκύπτουσα εικόνα (δεξιά).

Πλέον οι τρεις βασικές συνεκτικές συνιστώσες δέρματος (πρόσωπο και δύο χέρια) είναι περισσότερο ευδιάκριτες και μπορούν να ανιχνευθούν με ευκολία. Ως αποτέλεσμα, δημιουργούνται τρία ορθογώνια που περιβάλλουν τις περιοχές ενδιαφέροντος-δέρματος (bounding boxes) και εμπρόκειτο να χρησιμοποιηθούν στη συνέχεια για τον υπολογισμό των διανυσμάτων Οπτικής Ροής και την τελική παρακολούθηση του προσώπου και των χεριών.

Figure 5: Οι Τρεις συνεκτικές συνιστώσες Δέρματος (πρόσωπο και δύο χέρια).

Κάθε πλαίσιο οριοθέτησης της περιοχής ενδιαφέροντος υπολογίζεται στη μορφή [x, y, width, height], όπου x, y οι συντεταγμένες του πάνω αριστερά σημείου, width το πλάτος και height το ύψος του bounding box. Τελικώς, ακολουθώντας την παραπάνω διαδικασία ανίχνευσης δέρματος, οριοθετούμε κάθε περιοχή ενδιαφέροντος με τα εξής bounding boxes:

- **Head Component**: x = 139 , y = 91 , Width= 74 , Height= 122
- **Left Hand Component**: x= 63 , y= 259 , Width= 42 , Height= 38
- **Right Hand Component**: x= 169 , y= 271 , Width= 68 ,Height=40

Ώστόσο, προκειμένου να διευκολύνουμε τη διαδικασία της παρακολούθησης κίνησης, εν τέλει χρησιμοποιούμε

Bounding Boxes διευρυμένα, σχετικά με το αποτέλεσμα της ανίχνευσης του δέρματος, όπως αυτά δίνονται, δίχως ωστόσο να αποκλίνουμε κατά πολύ από τη δική μας οριοθέτηση:

- **Head Component**: x= 138, y= 88, Width= 73, Height= 123
- **Left Hand Component**: x = 47, y = 243, Width= 71, Height= 66
- **Right Hand Component**: x = 162, y = 264, Width= 83, Height=48

Παρακολούθηση προσώπου και χεριών

Έχοντας οριοθετήσει τις ανιχνευθείσες περιοχές δέρματος με Bounding Boxes, καλούμαστε να υλοποιήσουμε έναν αλγόριθμο παρακολούθησης οπτικής ροής και να τον εφαρμόσουμε στο βίντεο της νοηματίστριας. Ειδικότερα, υλοποιούμε τον αλγόριθμο των Lucas-Kanade, τόσο στην μονοκλιμακωτή όσο και στην πολυκλιμακωτή εκδοχή του.

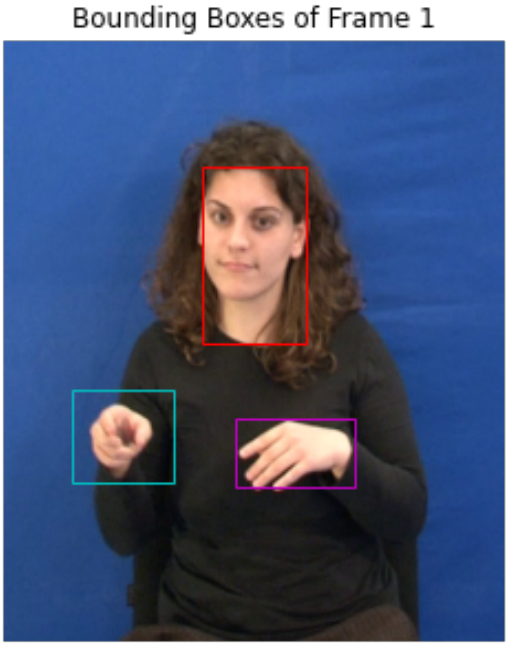


Figure 6: Bounding Boxes για το πρώτο frame του βίντεο

Υλοποίηση του Αλγόριθμου των Lucas-Kanade

Στο μέρος αυτό, υπολογίζουμε το διάνυσμα της οπτικής ροής μεταξύ δύο διαδοχικών frames με χρήση του Αλγορίθμου Lucas-Kanade, υπό την παραδοχή ότι η φωτεινότητα μεταξύ των δύο frames είναι περίπου ίδια. Ο εν λόγω αλγόριθμος υπολογίζει την οπτική ροή σε κάθε pixel x της εικόνας με τη μέθοδο των ελαχίστων τετραγώνων, θεωρώντας ότι το διάνυσμα d παραμένει σταθερό σε ένα μικρό παράθυρο γύρω από το σημείο που εξετάζει και ελαχιστοποιώντας το τετραγωνικό σφάλμα:

$$J_{\mathbf{x}}(\mathbf{d}) = \int_{\mathbf{x}' \in \mathcal{R}^2} G_\rho(\mathbf{x} - \mathbf{x}')[\mathbf{I}_n(\mathbf{x}') - \mathbf{I}_{n-1}(\mathbf{x}' + \mathbf{d})]^2 \, d\mathbf{x}'$$

όπου $G_\rho(\mathbf{x})$ είναι μια Γκαουσιανή συνάρτηση παραθύρωσης, με τυπική απόκλιση ρ. Θεωρούμε ότι έχουμε μια εκτίμηση $\mathbf{d_i}$ για το $\mathbf{d}$ και προσπαθούμε να τη βελτιώσουμε κατά $\mathbf{u}$, δηλαδή $\mathbf{d_{i+1}} = \mathbf{d_i} + \mathbf{u}$. Αναπτύσσοντας κατά Taylor την έκφραση $I_{n-1}(\mathbf{x} + \mathbf{d}) = I_{n-1}(\mathbf{x} + \mathbf{d_i} + \mathbf{u})$ γύρω από το σημείο $\mathbf{x} + \mathbf{d_i}$, προκύπτει ότι:

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

Η λύση ελάχιστων τετραγώνων για τη βελτίωση της εκτίμησης της οπτικής ροής σε κάθε σημείο είναι:

$$\mathbf{u}(\mathbf{x}) = \begin{bmatrix} (G_\rho * A_1^2)(\mathbf{x}) + \epsilon & (G_\rho * A_1 A_2)(\mathbf{x}) \\ (G_\rho * A_1 A_2)(\mathbf{x}) & (G_\rho * A_2^2)(\mathbf{x}) + \epsilon \end{bmatrix}^{-1} \cdot \begin{bmatrix} (G_\rho * A_1 E)(\mathbf{x}) \\ (G_\rho * A_2 E)(\mathbf{x}) \end{bmatrix}$$

όπου:

$$A(\mathbf{x}) = \begin{bmatrix} A_1(\mathbf{x}) & A_2(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} \frac{\partial I_{n-1}(\mathbf{x}+\mathbf{d_i})}{\partial x} & \frac{\partial I_{n-1}(\mathbf{x}+\mathbf{d_i})}{\partial y} \end{bmatrix}$$

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

Η μικρή θετική σταθερά ϵ βελτιώνει το αποτέλεσμα σε επίπεδες περιοχές με μειωμένη υφή και άρα μειωμένη πληροφορία για τον υπολογισμό της οπτικής ροής, ενώ για τον υπολογισμό της συνάρτησης και των μερικών παραγώγων της σε ενδιάμεσα σημεία του πλέγματος της εικόνας, εφαρμόζουμε **γραμμική παρεμβολή**.

Η ανανέωση του διανύσματος οπτικής ροής $\mathbf{d_{i+1}} = \mathbf{d_i} + \mathbf{u}$ επαναλαμβάνεται αρκετές φορές ως τη σύγκλιση (στην υλοποίηση μας 50 φορές). Επιπλέον, προκειμένου να επιταχύνουμε τη διαδικασία σύγκλισης, εφαρμόζουμε ως κριτήριο τερματισμού τον εξής έλεγχο **κατωφλιοποίησης** μεταξύ δύο διαδοχικών επαναλήψεων:

$$\text{if } L2_{norm}(\mathbf{u}) < 0.02 \longrightarrow \text{Terminate}$$

Σημειώνεται, επίσης, ότι ο υπολογισμός της οπτικής ροής δεν πραγματοποιείται για κάθε pixel των δύο διαδοχικών frames, I_1 και I_2, παρά εστιάζουμε την προσοχή μας μόνο σε συγκεκριμένα σημεία ενδιαφέροντος της εικόνας I_2. Για την ανίχνευση των σημείων αυτών – στην προκειμένη περίπτωση γωνίες – γίνεται χρήση ενός **Shi-Tomasi** ανιχνευτή με την εξής παραμετροποίηση:

	Head Bounding Box	Hands Bounding Boxes
max_corners	100	200
quality	0.01	0.01
mindistance	15	5

Έχοντας κανονικοποίησει κάθε εικόνα, ώστε να λαμβάνει τιμές στο διάστημα [0,1], τρέχουμε τον Αλγόριθμο Lucas – Kanade για τιμές παραμέτρων $\epsilon = 0.01 και ρ = 8$. Στις Εικόνες 7 - 8, φαίνεται η οπτική ροή που ανιχνεύθηκε για τα τρία bounding boxes – του προσώπου και των δύο χεριών – ανάμεσα στα πρώτα δύο frames του βίντεο της νοηματίστριας. Παρατηρεί κανείς ότι η επικρατέστερη αρχική κίνηση είναι αυτή του αριστερού χεριού.

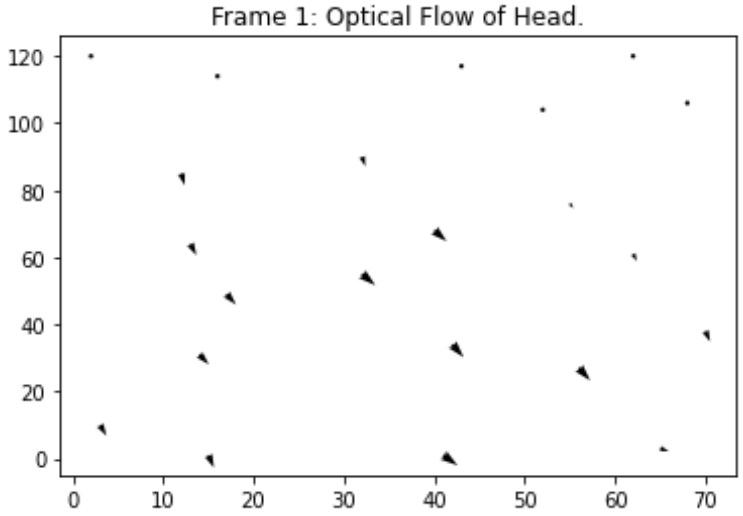
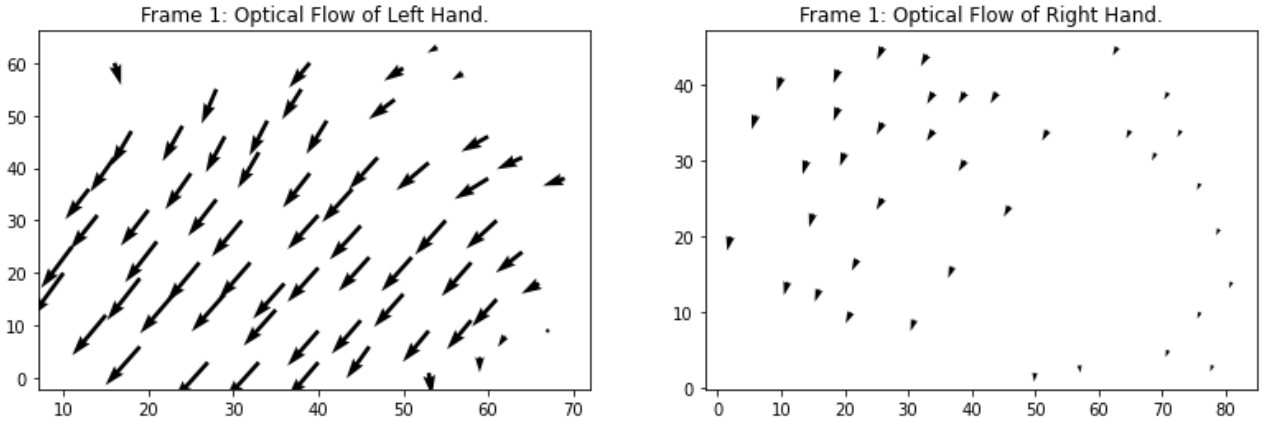


Figure 7: Οπτική Ροή του Bounding Box του προσώπου, ανάμεσα στα πρώτα 2 frames του βίντεο

Προκειμένου να τονιστεί η επίδραση των παραμέτρων ρ και ε στην ανίχνευση της οπτικής ροής, πειραματιζόμαστε για διαφορετικές τιμές τους, εφαρμόζοντας τον αλγόριθμο Lucas-Kanade στο bounding box του αριστερού χεριού για τα δύο πρώτα frames του βίντεο. Παρατηρώντας τις γραφικές παραστάσεις της Εικόνας 9, είναι προφανές ότι για σταθερό ε, καθώς αυξάνεται η παράμετρος ρ το διανυσματικό πεδίο οπτικής ροής **εξομαλύνεται** και αποκτά περισσότερες



(a) Αριστερό Χέρι (b) Δεξί Χέρι

Figure 8: Οπτική Ροή των bounding boxes των δύο χεριών, ανάμεσα στα πρώτα δύο frames του βίντεο

λεπτομέρειες. Κάτι τέτοιο είναι λογικό, καθώς αυξάνεται η τυπική απόκλιση της Γκαουσσιανής συνάρτησης με την οποία συνελλίσονται οι παράγωγοι των εικόνων, με αποτέλεσμα να επιτυγχάνεται μεγαλύτερη εξομάλυνση και κατ'επέκταση ανίχνευση εντονότερης κίνησης. Όσον αφορά την επίδραση της παραμέτρου ε, μολονότι γνωρίζουμε ότι βελτιώνει το αποτέλεσμα υπολογισμού του πεδίου οπτικής ροής σε επίπεδες περιοχές με μειωμένη υφή, παρατηρούμε ότι αυξάνοντας την σε μεγάλο βαθμό, διατηρώντας σταθερή την παράμετρο ρ, το διανυσματικό πεδίο οπτικής ροής φαίνεται να δίνει μικρότερες μετατοπίσεις. Αυτό συμβαίνει διότι, μεγάλες τιμές της παραμέτρου ε σε συνδυασμό με χαμηλές τιμές κατωφλίου του κριτηρίου σύγκλισης, έχουν ως αποτέλεσμα ο αλγόριθμος πολλές φορές να τερματίζει προτού να συγκλίνει, με αποτέλεσμα τα επιστρεφόμενα διανύσματα οπτικής ροής να λαμβάνουν χαμηλές τιμές.

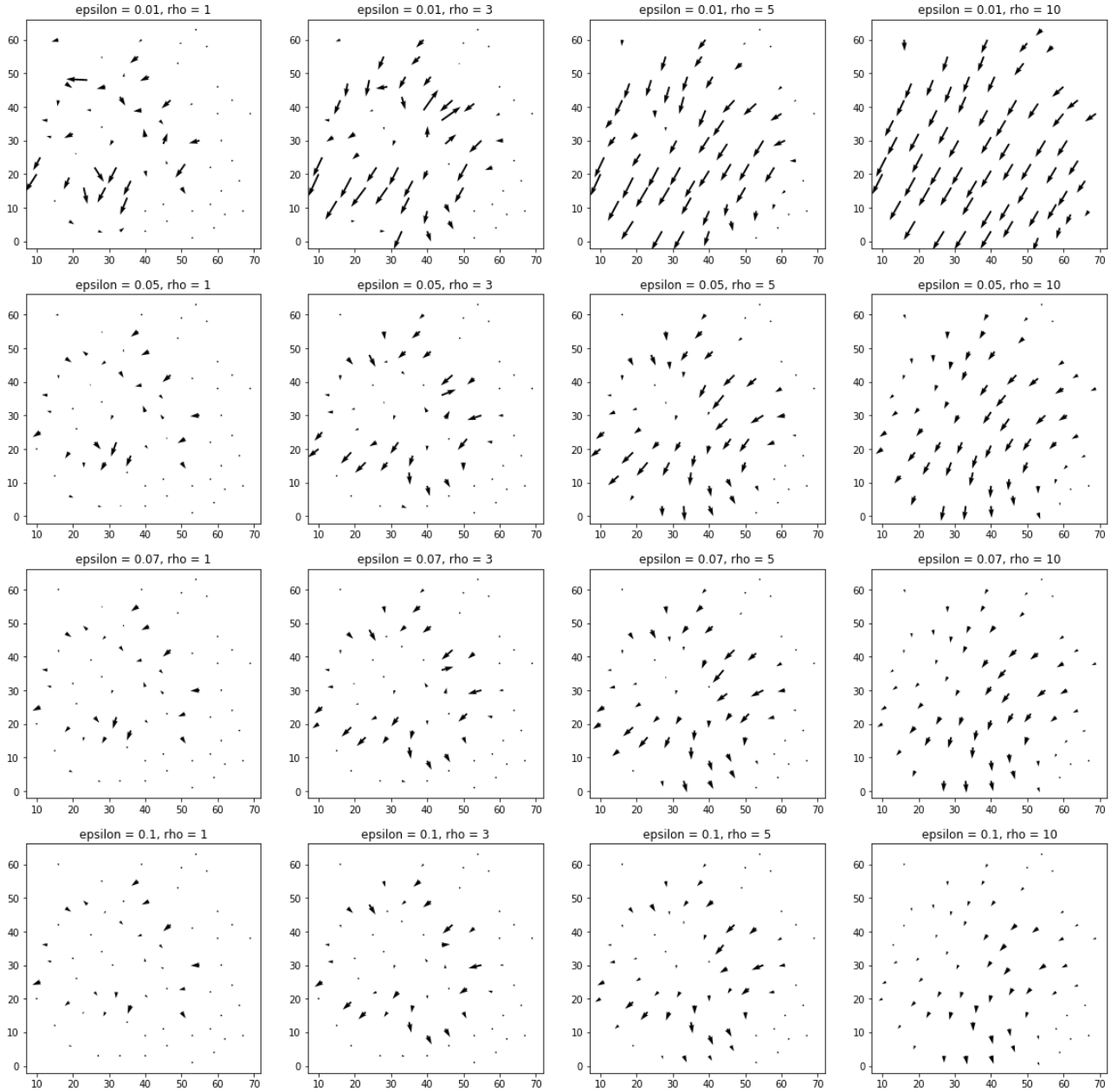


Figure 9: Ανίχνευση Οπτικής Ροής για διάφορες τιμές των παραμέτρων ρ και ε.

Υπολογισμός της Μετατόπισης των Παραθύρων από τα Διανύσματα Οπτικής Ροής

Έχοντας υπολογίσει την οπτική ροή της εικόνας In στα σημεία που ορίζουν τα σημεία ενδιαφέροντος εντός του bounding box της εικόνας In1, απομένει να βρούμε το συνολικό διάνυσμα μετατόπισης του bounding box ορθογωνίου, με όσο το δυνατόν μεγαλύτερη ακρίβεια. Γνωρίζουμε ότι τα διανύσματα οπτικής ροής έχουν κατά κανόνα μεγαλύτερο μήκος σε σημεία που ανήκουν σε περιοχές με έντονη πληροφορία υφής (π.χ. ακμές, κορυφές) και σχεδόν μηδενικό μήκος σε σημεία που ανήκουν σε περιοχές με ομοιόμορφη και επίπεδη υφή. [1] Αρχικά, ως μια πιο αφελής προσέγγιση δοκιμάζουμε να λάβουμε τη συνολική μετατόπιση των bounding boxes ως τη μέση τιμή των διανυσμάτων οπτικής ροής που εμπεριέχει. Παρόλο που η πλειονότητα των ανιχνευθέντων σημείων ενδιαφέροντος (γωνίες) βρίσκονται σε σημεία με έντονη υφή, το εν λόγω κριτήριο δίνει σχετικά ανακριβή αποτελέσματα και επιδέχεται βελτίωση. Παρακάτω φαίνονται οι συνολικές μετατοπίσεις του bounding box που αντιστοιχεί στο αριστερό χέρι για τα πρώτα 2 frames, κάνοντας χρήσης του προαναφερθέντος κριτηρίου και διαφορετικές παραμετροποιήσεις:

epsilon = 0.01, rho = 1: dx = -1.0, dy = -1.0

epsilon = 0.01, rho = 3: dx = -1.0, dy = -3.0
epsilon = 0.01, rho = 5: dx = -1.0, dy = -3.0
epsilon = 0.05, rho = 1: dx = -1.0, dy = -3.0
epsilon = 0.05, rho = 3: dx = -1.0, dy = -3.0
epsilon = 0.05, rho = 5: dx = -1.0, dy = -3.0

epsilon = 0.01, rho = 1: dx = -1.0, dy = -3.0
epsilon = 0.01, rho = 3: dx = -1.0, dy = -3.0
epsilon = 0.01, rho = 5: dx = -1.0, dy = -3.0
epsilon = 0.05, rho = 1: dx = -1.0, dy = -3.0
epsilon = 0.05, rho = 3: dx = -1.0, dy = -3.0
epsilon = 0.05, rho = 5: dx = -1.0, dy = -3.0

epsilon = 0.01, rho = 1: dx = -1.0, dy = -3.0
epsilon = 0.01, rho = 3: dx = -1.0, dy = -3.0
epsilon = 0.01, rho = 5: dx = -1.0, dy = -3.0
epsilon = 0.05, rho = 1: dx = -1.0, dy = -3.0
epsilon = 0.05, rho = 3: dx = -1.0, dy = -3.0
epsilon = 0.05, rho = 5: dx = -1.0, dy = -3.0

epsilon = 0.01, rho = 1: dx = -1.0, dy = -3.0
epsilon = 0.01, rho = 3: dx = -1.0, dy = -3.0
epsilon = 0.01, rho = 5: dx = -1.0, dy = -3.0
epsilon = 0.05, rho = 1: dx = -1.0, dy = -3.0
epsilon = 0.05, rho = 3: dx = -1.0, dy = -3.0
epsilon = 0.05, rho = 5: dx = -1.0, dy = -3.0

Προκειμένου να λάβουμε περισσότερο ακριβή αποτελέσματα και να απορρίψουμε outliers, εφαρμόζουμε ως εναλλακτικό κριτήριο την εξαγωγή της συνολικής μετατόπισης από τη μέση τιμή των διανυσμάτων οπτικής ροής που έχουν ενέργεια μεγαλύτερη από μια τιμή κατωφλίου. Ως ενέργεια διανύσματος ταχύτητας ορίζουμε: $||\mathbf{d}||^2 = d_x^2 + d_y^2$ [1]

Πειραματιζόμενοι με διαφορετικές τιμές των παραμέτρων ρ, ε και κατωφλίου ενέργειας, εξετάζουμε τις συνολικές μετατοπίσεις του bounding box που αντιστοιχεί στο αριστερό χέρι, για τα πρώτας 2 frames:

epsilon = 0.01, rho = 1, Threshold = 0.001: dx = -1, dy = -2
epsilon = 0.01, rho = 3, Threshold = 0.001: dx = -2, dy = -3
epsilon = 0.01, rho = 5, Threshold = 0.001: dx = -2, dy = -3
epsilon = 0.05, rho = 1, Threshold = 0.001: dx = -2, dy = -3
epsilon = 0.05, rho = 3, Threshold = 0.001: dx = -2, dy = -3
epsilon = 0.05, rho = 5, Threshold = 0.001: dx = -2, dy = -3

epsilon = 0.01, rho = 1, Threshold = 0.2: dx = -2, dy = -3
epsilon = 0.01, rho = 3, Threshold = 0.2: dx = -2, dy = -3
epsilon = 0.01, rho = 5, Threshold = 0.2: dx = -2, dy = -3
epsilon = 0.05, rho = 1, Threshold = 0.2: dx = -2, dy = -3
epsilon = 0.05, rho = 3, Threshold = 0.2: dx = -2, dy = -3
epsilon = 0.05, rho = 5, Threshold = 0.2: dx = -2, dy = -3

epsilon = 0.01, rho = 1, Threshold = 0.5: dx = -2, dy = -4
epsilon = 0.01, rho = 3, Threshold = 0.5: dx = -2, dy = -4
epsilon = 0.01, rho = 5, Threshold = 0.5: dx = -2, dy = -4
epsilon = 0.05, rho = 1, Threshold = 0.5: dx = -2, dy = -4
epsilon = 0.05, rho = 3, Threshold = 0.5: dx = -2, dy = -4
epsilon = 0.05, rho = 5, Threshold = 0.5: dx = -2, dy = -4

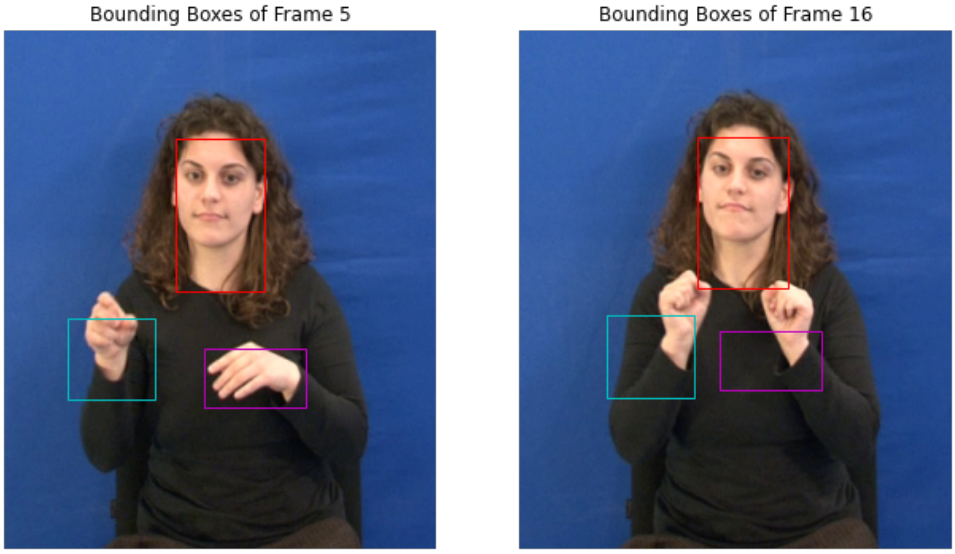
epsilon = 0.01, rho = 1, Threshold = 0.75: dx = -2, dy = -4
epsilon = 0.01, rho = 3, Threshold = 0.75: dx = -2, dy = -4
epsilon = 0.01, rho = 5, Threshold = 0.75: dx = -2, dy = -4
epsilon = 0.05, rho = 1, Threshold = 0.75: dx = -2, dy = -4
epsilon = 0.05, rho = 3, Threshold = 0.75: dx = -2, dy = -4
epsilon = 0.05, rho = 5, Threshold = 0.75: dx = -2, dy = -4

Μπορούμε να εξάγουμε τα εξής συμπεράσματα:

- Όταν η τιμή του κατωφλίου ενέργειας είναι **αρκετά μικρή**, τότε το κριτήριο εξαγωγής συνολικών μετατοπίσεων εκφυλλίζεται στην πρώτη περίπτωση, όπου λαμβάνεται υπόψη η μέση τιμή όλων των διανυσμάτων οπτικής ροής. Όπως αναφέρθηκε κάτι τέτοιο οδηγεί σε λιγότερο ακριβή αποτελέσματα, λόγω της ύπαρξης σημείων που ανήκουν σε περιοχές με ομοιόμορφη και επίπεδη υφή. Η κατάσταση αυτή, μάλιστα, φαίνεται να επιφέρει χειρότερα αποτελέσματα όταν συνοδεύεται με χαμηλές τιμές των παραμέτρων ρ και ε.

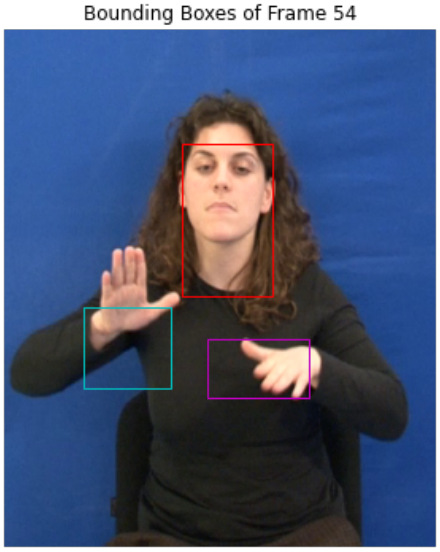
- Στην αντίπερα όχθη, όταν το κατώφλι ενέργειας λαμβάνει **πολύ μεγάλες τιμές**, τότε η συνολική μετατόπιση εξάγεται ως η μέση τιμή διανυσμάτων οπτικής ροής με αρκετά αυξημένη ενέργεια. Κάτι τέτοιο, έχει ως αποτέλεσμα η εξαγόμενη μετατόπιση του bounding box να είναι μεγαλύτερη της πραγματικής και η ανίχνευση κίνησης κατά τη διάρκεια του βίντεο να αποτύχει λόγω μεγάλων μετατοπίσεων των bounding boxes.

- Συμπερασματικά, θα λέγαμε ότι μια ενδεικτική περιοχή τιμών του κατωφλίου ενέργειας θα ήταν το διάστημα [0.2 , 0.5], συνοδευόμενη προφανώς με κατάλληλη παραμετροποίηση των σταθερών ρ και ε. Παρακάτω φαίνονται στιγμιότυπα της ανίχνευσης κίνησης με χρήση του Αλγορίθμου Lucas – Kanade στο βίντεο της νοηματίστριας για τιμές παραμέτρων **$\varepsilon = 0.01$, $\rho = 8$ και Threshold $= 0.5$**:

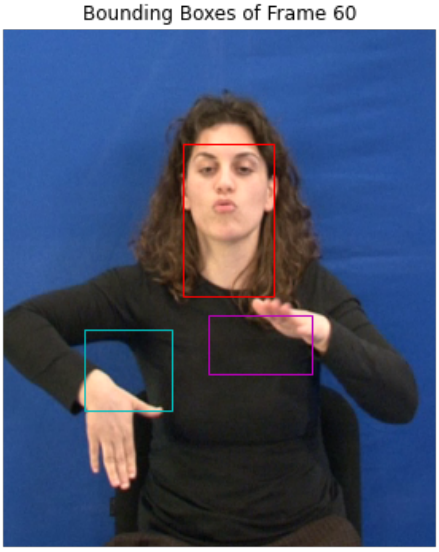


(a)

(b)



(c)



(d)

Figure 10: Στιγμιότυπα Motion Tracking με τον Αλγόριθμο Lucas - Kanade.

Πολυ-Κλιμακωτός Υπολογισμός Οπτικής Ροής

Σε πολλές περιπτώσεις υπάρχει η ανάγκη για τον υπολογισμό της οπτικής ροής σε μεγαλύτερες κινήσεις που υπερβαίνουν κατά πολύ το ένα pixel και επομένως δεν ισχύουν οι παραδοχές για τους όρους πρώτους τάξης του μονοκλιμακωτού αλγορίθμου Lucas – Kanade. Για τον λόγο αυτό, γίνεται χρήση μιας πολυκλιμακωτής εκδοχής του αλγορίθμου, η οποία αναλύει τις αρχικές διαδοχικές εικόνες I_1 και I_2 σε γκαουσιανές πυραμίδες και υπολογίζει το πεδίο οπτικής ροής από τις πιο μεγάλες (τραχείς) στις πιο μικρές (λεπτομερείς) κλίμακες, χρησιμοποιώντας τη λύση της μεγάλης κλίμακας ως αρχική συνθήκη για τη μικρή κλίμακα. Για τη μετάβαση από μεγάλες σε μικρές κλίμακες κατά την κατασκευή της Γκαουσιανής πυραμίδας, εφαρμόζεται φιλτράρισμα της εικόνας με ένα βαθυπεράτο Γκαουσιανό φίλτρο (τυπικής απόκλισης 3 pixel) πριν την υποδειγματοληψία για να μετριάσουμε το aliasing της εικόνας. [1]

Υλοποιώντας τον πολυκλιμακωτό αλγόριθμο Lucas – Kanade, πειραματιζόμαστε ως προς την ανίχνευση οπτικής ροής στο bounding box του αριστερού χεριού για τα 2 πρώτα frames, για διαφορεικές τιμές των παραμέτρων ρ, ε και διαφορετικές κλίμακες της πυραμίδας:

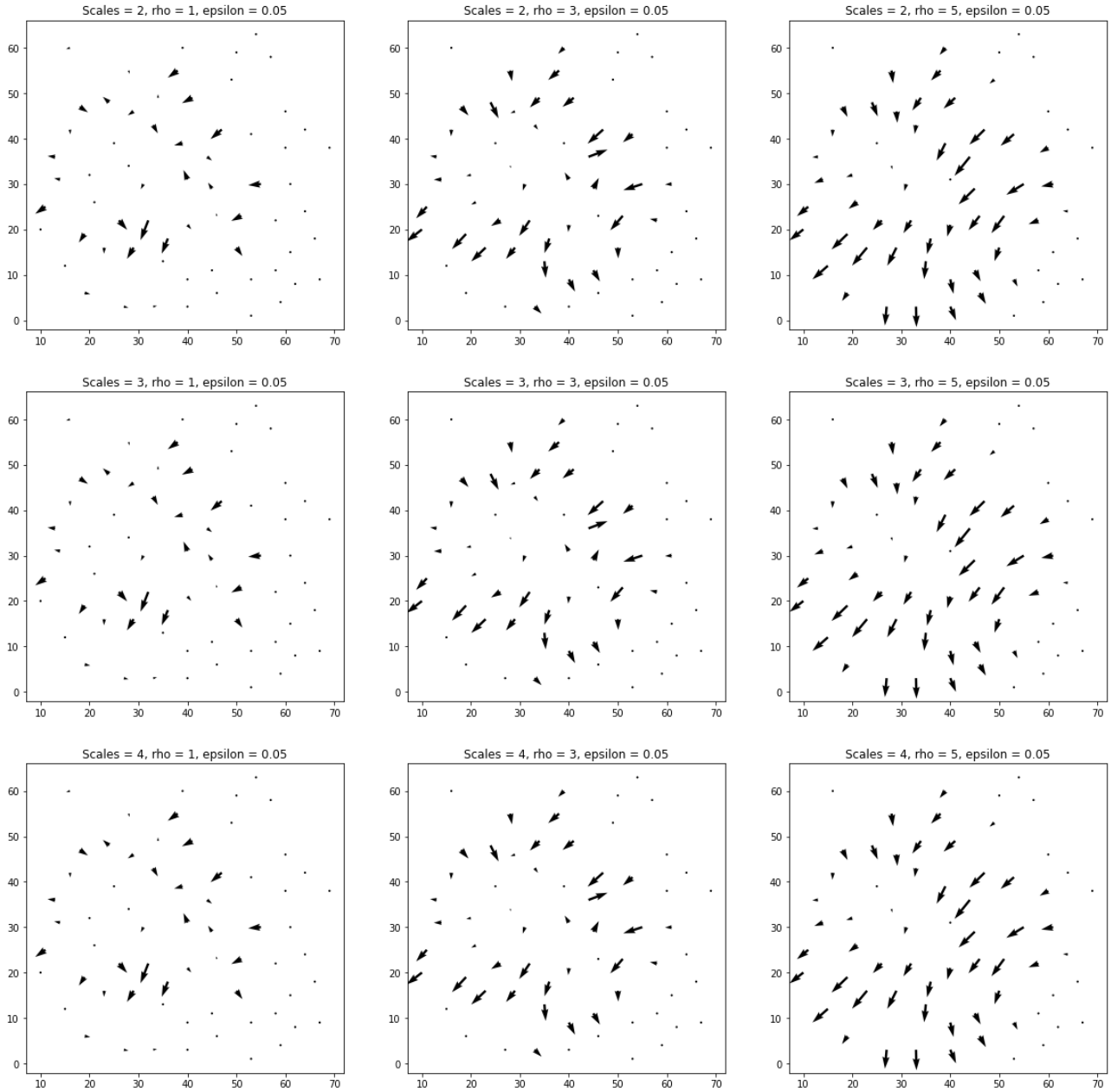


Figure 11: Πολυκλιμακωτή Ανίχνευση Οπτικής Ροής.

Παρατηρώντας τις παραπάνω γραφικές παραστάσεις, και τρέχοντας τον πολυκλιμακωτό αλγόριθμο για ολόκληρο

το βίντεο, οδηγούμαστε στο συμπέρασμα ότι η πολυκλιμακωτή εκδοχή φαίνεται να ανιχνεύει τόσο τις μικρές όσο και τις μεγάλες κινήσεις. Επιπλέον, για κατάλληλη επιλογή κλίμακας παρατηρούμε ότι οι επαναλήψεις που χρειάζονται για τη σύγκλιση της μεθόδου Lucas-Kanade γίνονται λιγότερες, καθώς η προηγούμενη κλίμακα παρέχει μία αρχική εκτίμηση που είναι αντιπροσωπευτική της κίνησης μεταξύ των δύο frames. Ωστόσο, χρησιμοποιώντας μεγάλο αριθμό απο κλίμακες ενδέχεται η συνολική μετατόπιση των bounding boxes να είναι ιδιαίτερα αυξημένη, με αποτέλεσμα να χαθεί το motion tracking. Ως εκ τούτου, μια ασφαλής επιλογή για το μέγεθος της Γκαουσιανής πυραμίδας είναι οι 3-4 κλίμακες.

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

Θέμα

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 αντικείμενο 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 μετασχηματίζουν σε μια αναπαράσταση που τα κάνει κατηγοριοποιήσιμα. Στα πλαίσια αυτής της άσκησης μας δόθηκαν βίντεο από 3 κλάσεις δράσεων (walking,running,boxing) από τα οποία θα εξάγουμε χωρο-χρονικούς περιγραφητές με σκοπό την κατηγοριοποίηση των δράσεων αυτών.

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

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:

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

ο οποίος εκφράζεται σε μητρική μορφής ως εξής:

$$M(x, y, t; \sigma, \tau) = g(x, y, t; s\sigma, s\tau) * \begin{pmatrix} L_x^2 & L_x L_y & L_x L_t \\ L_x L_y & L_y^2 & L_y L_t \\ L_x L_t & L_y L_t & L_t^2 \end{pmatrix}$$

όπου $g(x, y, t; s\sigma, s\tau)$ ένας 3D γκαουσιανός πυρήνας ομαλοποίησης και $\nabla L(x, y, t; \sigma, \tau)$ οι χωρο-χρονικές παράγωγοι για την χωρική κλίμακα σ και τη χρονική κλίμακα τ. Τις παραγώγους (χωρικές και χρονικές) τις υπολογίζουμε εφαρμόζοντας συνέλιξη με τον πυρήνα κεντρικών διαφόρων $\begin{pmatrix} -1 & 0 & 1 \end{pmatrix}$ (προσαρμοσμένο στην κατάλληλη διάσταση).

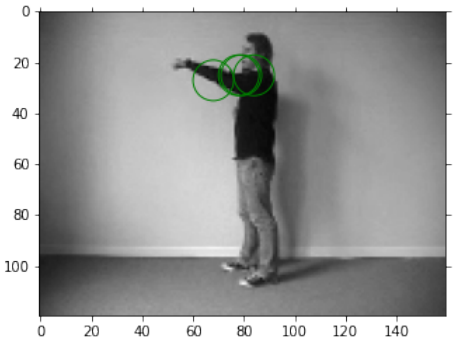
Για την ανίχνευση χρησιμοποιείται το παρακάτω 3D κριτήριο γωνιότητας, το οποίο ακολουθεί και την ίδια λογική με αυτό του ανιχνευτή Harris στις δύο διαστάσεις:

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

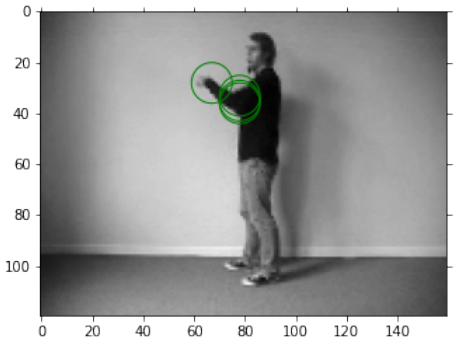
Τα σημεία ενδιαφέροντος προκύπτουν σαν τα τοπικά μέγιστα του κριτηρίου σημαντικότητας. Χρησιμοποιούμε τον παρακάτω συνδυασμό παραμέτρων: $\sigma = 4$, s = 2, $\tau = 1.5$, $\kappa = 0.005$

Στην συνέχεια παραθέτουμε ενδεικτικά frames, για κάθε κατηγορία video, πάνω στα οποία έχουν ανιχνευθεί σημεία ενδιαφέροντος με τη μέθοδο Harris. Επιπλέον, απεικονίζουμε το κριτήριο σημαντικότητας για τα ίδια frames:

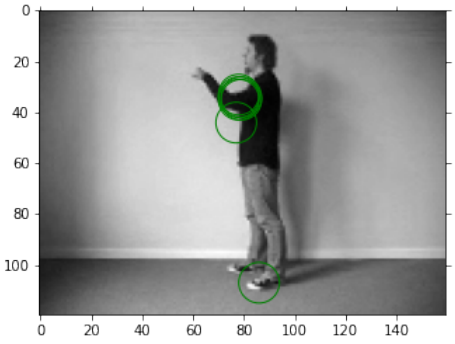
Boxing Video:



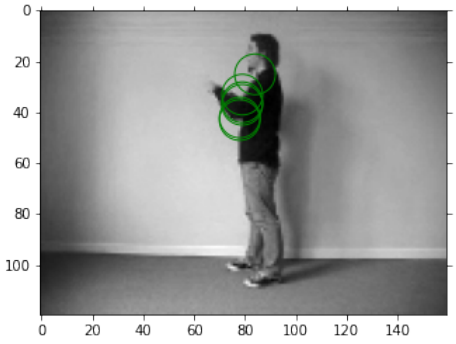
(a) Frame 57



(b) Frame 61

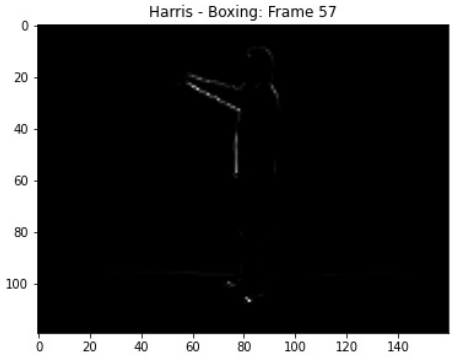


(c) Frame 147

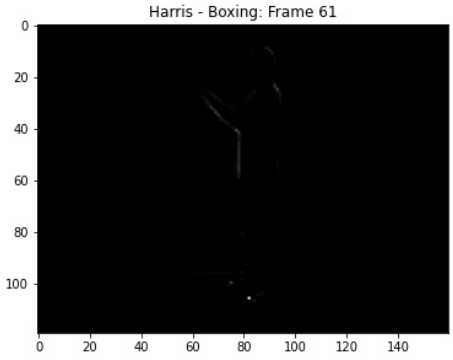


(d) Frame 187

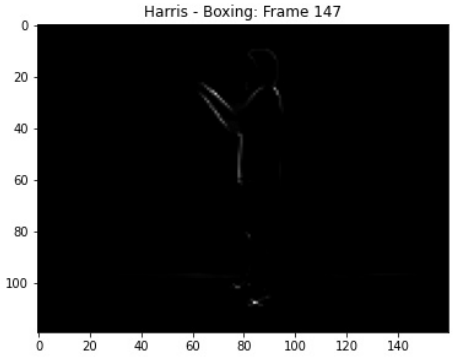
Figure 12: Ανίχνευση σε επιλεγμένα frames της ανίχνευσης με την μέθοδο Harris για boxing video



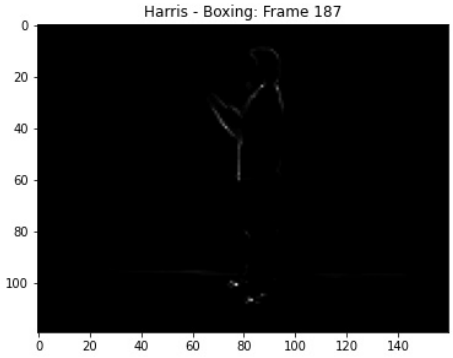
(a) Frame 57



(b) Frame 61



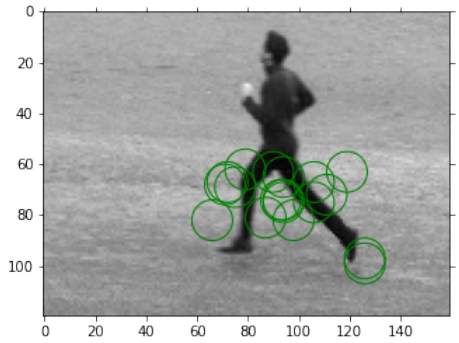
(c) Frame 147



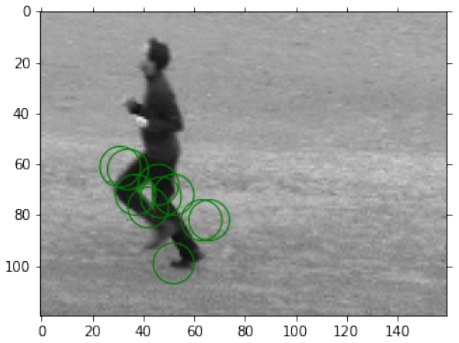
(d) Frame 187

Figure 13: Κριτήριο σημαντικότητας με την μέθοδο Harris για boxing video

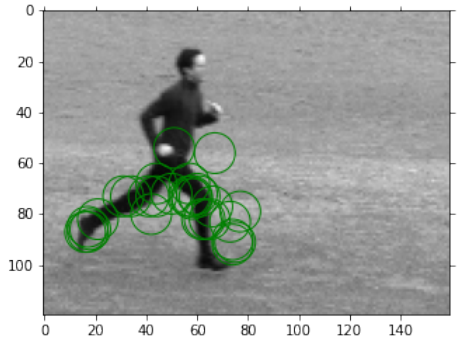
Running Video:



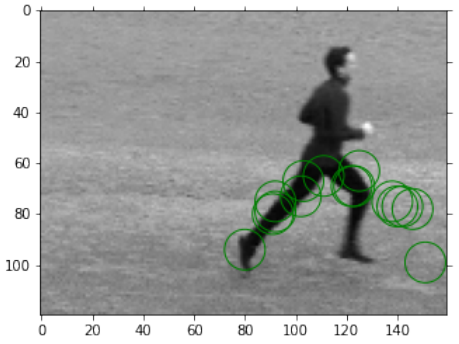
(a) Frame 14



(b) Frame 21

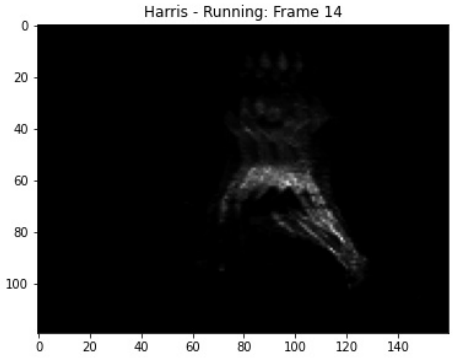


(c) Frame 106

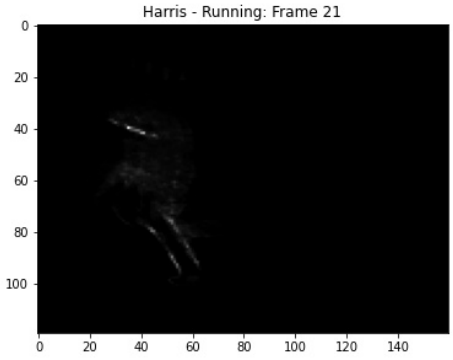


(d) Frame 114

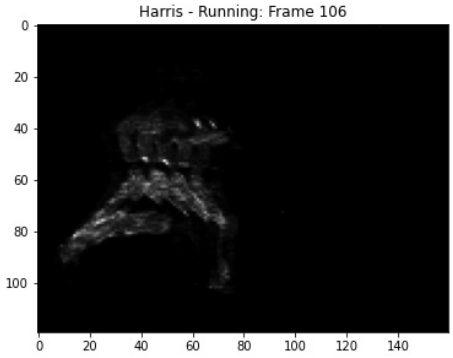
Figure 14: Ανίχνευση σε επιλεγμένα frames της ανίχνευσης με την μέθοδο Harris για running video



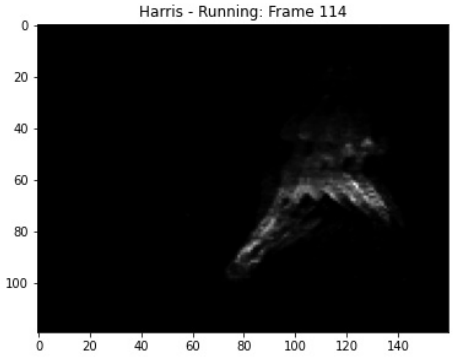
(a) Frame 14



(b) Frame 21



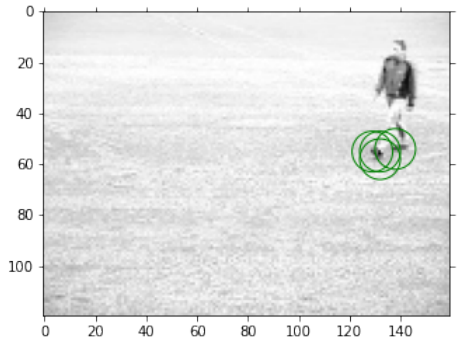
(c) Frame 106



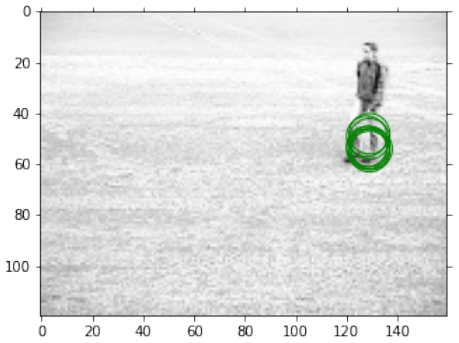
(d) Frame 114

Figure 15: Κριτήριο σημαντικότητας με την μέθοδο Harris για running video

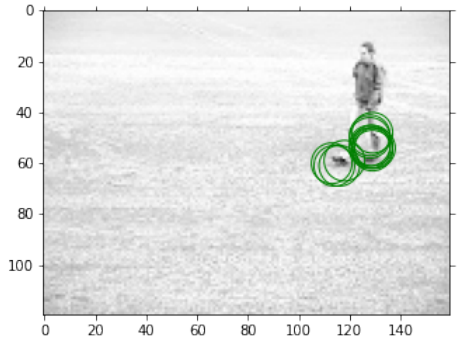
Walking Video:



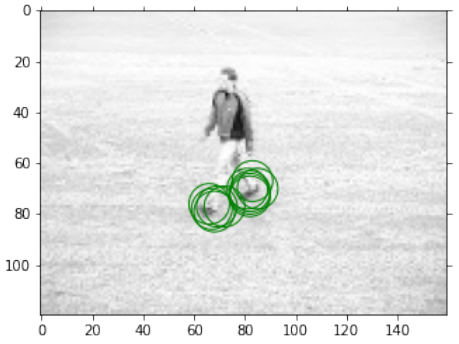
(a) Frame 45



(b) Frame 55

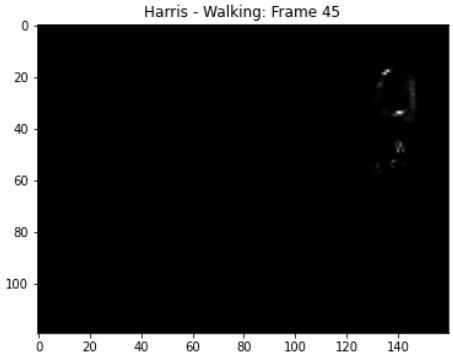


(c) Frame 57

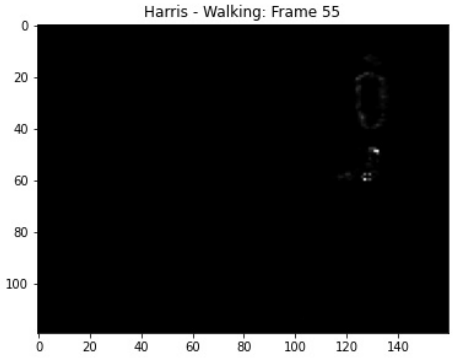


(d) Frame 101

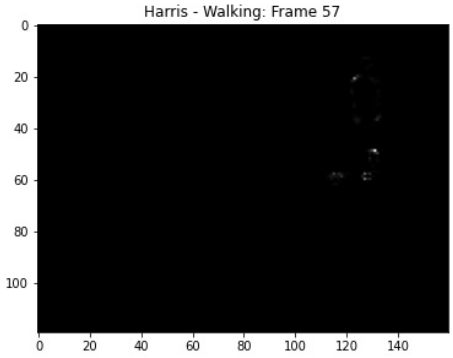
Figure 16: Ανίχνευση σε επιλεγμένα frames της ανίχνευσης με την μέθοδο Harris για walking video



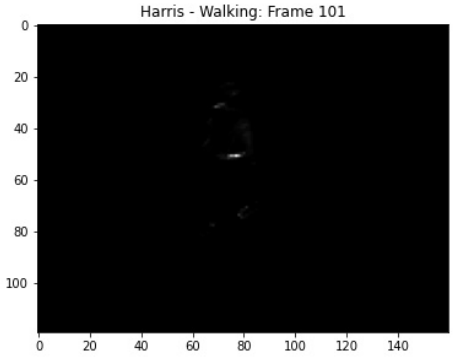
(a) Frame 45



(b) Frame 55



(c) Frame 57



(d) Frame 101

Figure 17: Κριτήριο σημαντικότητας με την μέθοδο Harris για walking video

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:

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

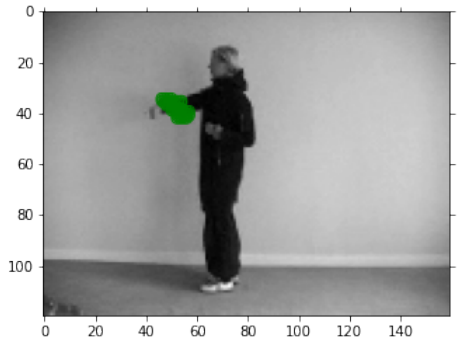
$$h_{od} = -sin(2\pi t\omega)e^{-\frac{t^2}{2\tau}}$$

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

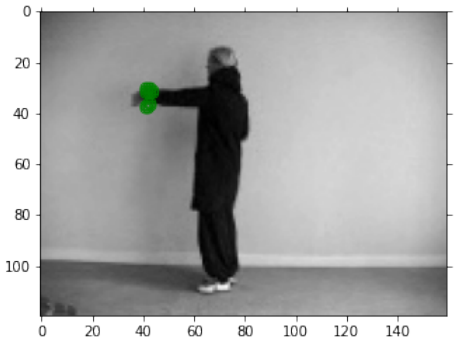
$$H(x, y, t) = (I(x, y, t) * g * h_{ev})^2 + (I(x, y, t) * g * h_{od})^2$$

και η επιλογή των σημείων γίνεται αντίστοιχα με τον Harris, αλλά με το παραπάνω κριτήριο H. Παρακάτω, παρατίθενται επιλεγμένα frames που δείχνουν τα αποτελέσματα ανίχνευσης με μέθοδο Gabor και το κριτήριο σημαντικότητας με παραμέτρους: $\sigma = 1.6$ και $\tau = 1.5$

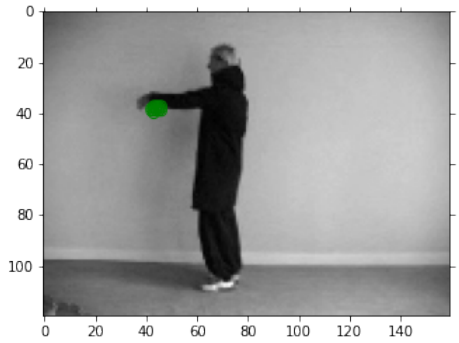
Boxing Video:



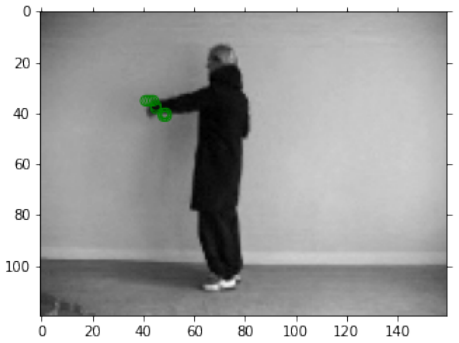
(a) Frame 9



(b) Frame 24

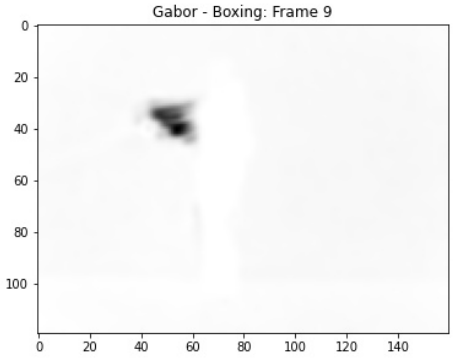


(c) Frame 61

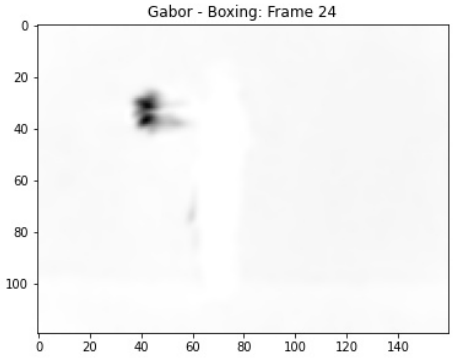


(d) Frame 138

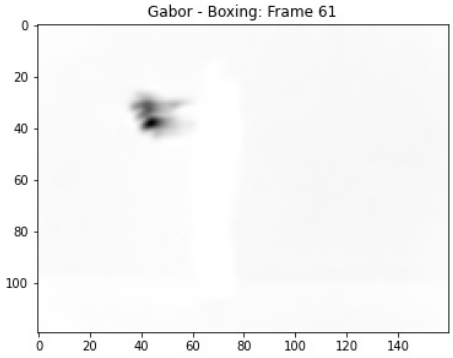
Figure 18: Ανίχνευση σε επιλεγμένα frames της ανίχνευσης με την μέθοδο Gabor για boxing video



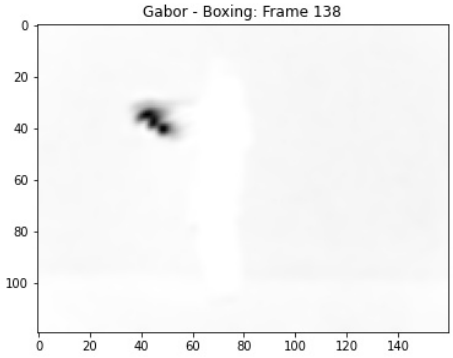
(a) Frame 9



(b) Frame 24



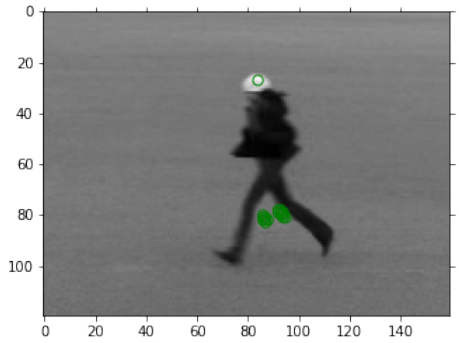
(c) Frame 61



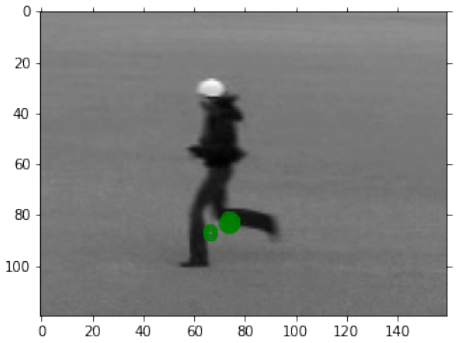
(d) Frame 138

Figure 19: Κριτήριο σημαντικότητας με την μέθοδο Gabor για boxing video

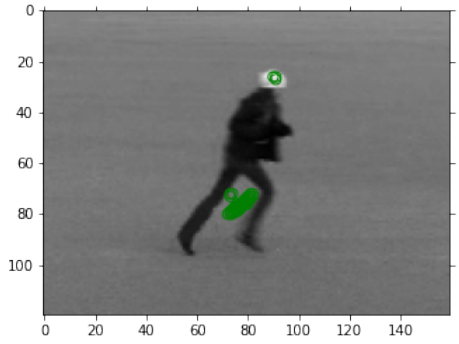
Running Video:



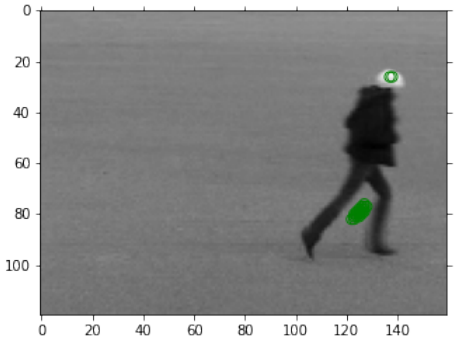
(a) Frame 17



(b) Frame 20

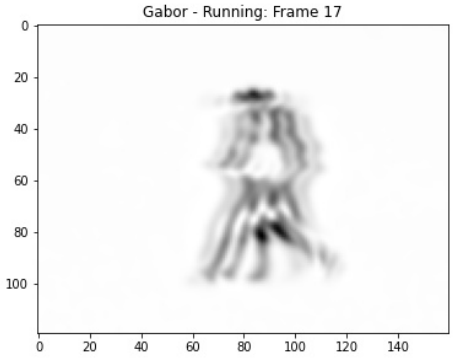


(c) Frame 96

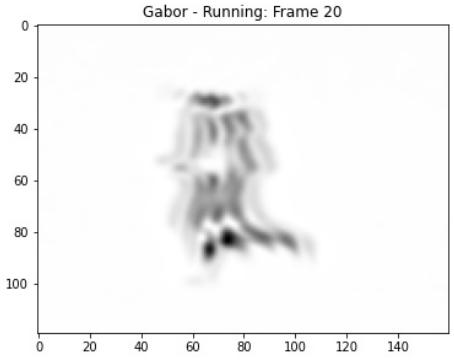


(d) Frame 104

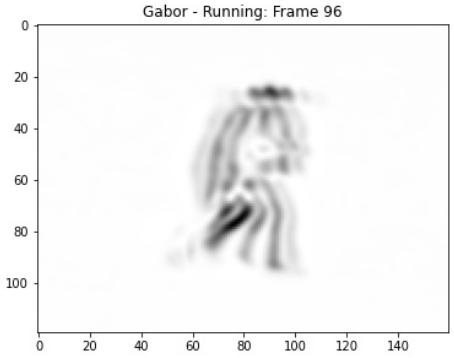
Figure 20: Ανίχνευση σε επιλεγμένα frames της ανίχνευσης με την μέθοδο Gabor για running video



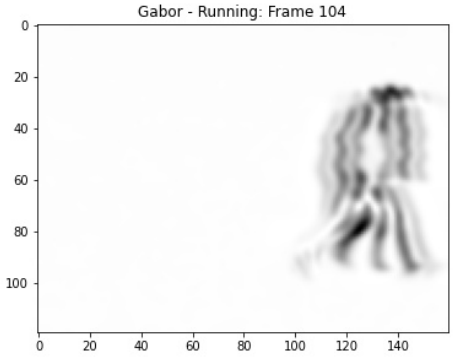
(a) Frame 17



(b) Frame 20



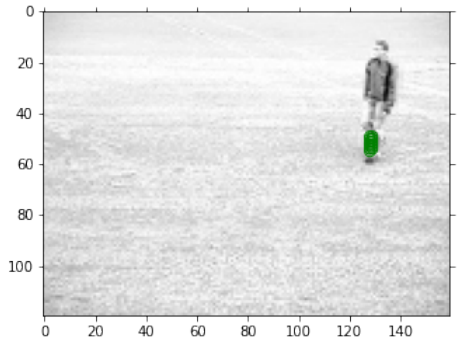
(c) Frame 96



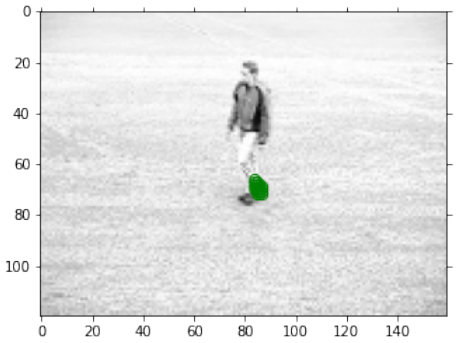
(d) Frame 104

Figure 21: Κριτήριο σημαντικότητας με την μέθοδο Gabor για running video

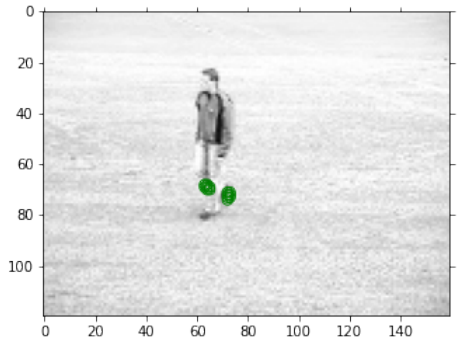
Walking Video:



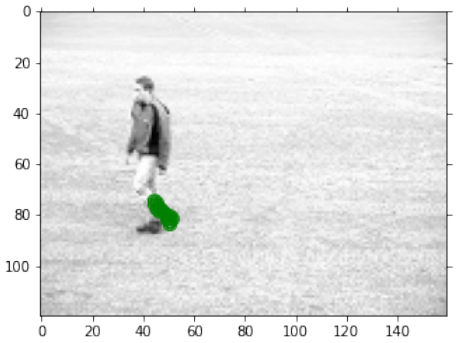
(a) Frame 52



(b) Frame 95

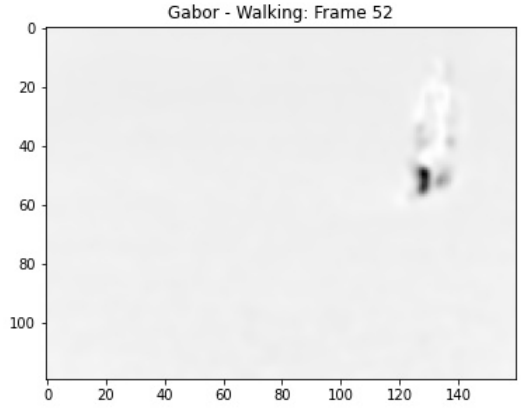


(c) Frame 107

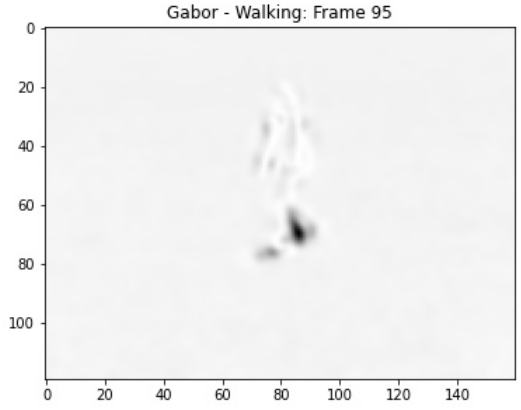


(d) Frame 122

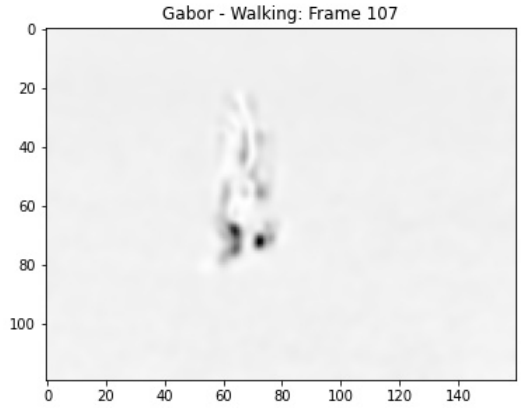
Figure 22: Ανίχνευση σε επιλεγμένα frames της ανίχνευσης με την μέθοδο Gabor για walking video



(a) Frame 52



(b) Frame 95



(c) Frame 107

(d) Frame 122

Figure 23: Κριτήριο σημαντικότητας με την μέθοδο Gabor για walking video

Σχολιασμός Αποτελεσμάτων

Η γενικότερη εικόνα των ανιχνεύσεων οδηγεί στην παρατήρηση ότι η ανίχνευση με τη μέθοδο Gabor είναι καλύτερη από αυτή του Harris και, συγκεκριμένα, πιο ευστοχη. Ο Harris εντοπίζει κυρίως γωνιώδη σημεία που εμφανίζονται ανάμεσα στα frames. Αυτό γίνεται εμφανές κυρίως από τα frames που παρατίθενται για τα boxing video, όπου ο εντοπισμός αφορά κυρίως τους λυγιζμένους αγκώνες του ανθρώπου. Στα βίντεο τρεξίματος παρατουρούμε πως ο εντοπισμός είναι σχετικα "αφηρημένος", δηλαδή τα σημεία που ανιχνεύονται είναι σε μια γενική περιοχή των ποδιών, αλλά όχι συγκεκριμένα στα πόδια. Στα βίντεο περπατήματος παρατηρούμε παρόμοια αποτελέσματα, αν και ο εντοπισμός σε αυτή την περίπτωση είναι σχετικά πιο εύστοχος.

Από την άλλη, η ανίχνευση Gabor είναι αρκετά που συγκεκριμένη. Στα βίντεο boxing παρατηρούμε ότι ανιχεύονται σημεία πολύ συγκεντρωμένα κοντά στην γροθιά του ανθρώπου. Στο τρέξιμο τα σημεία είναι επίσης πολύ συγκεντρωμένα στην περιοχή των ποδιών (ανάμεσα). Όμοιες παρατηρήσεις προκύπτουν για την ανίχνευση των βίντεο με περπάτημα.

Και στους δύο ανιχνευτές εντοπίζονται αραιά και μερικά εσφαλμένα σημεία, ενώ στα frames που δεν υπάρχει κίνηση δεν γίνεται καμία ανίχνευση.

Χωρο-χρονικοί Ιστογραφικοί Περιγραφητές

Οι χωρο-χρονικοί περιγραφητές που θα χρησιμοποιηθούν βασίζονται στον υπολογισμό ιστογραμμάτων της κατευθυντικής παραγώγου (HOG) και της TVL1 οπτικής ροής (HOF - Histograms of Oriented Flow) γύρω από τα σημεία ενδιαφέροντος που εντοπίζουν οι ανιχνευτές μας. Συγκεκριμένα, χρησιμοποιείται η συνάρτηση orientation_histogram, με ορίσματα 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 $4{\cdot}\sigma$ γύρω από κάθε σημείο ενδιαφέροντος. Ο HOG/HOF περιγραφητής προκύπτει με συνένωση των 2 παραπάνω περιγραφητών.

Κατασκευή Bag of Visual Words και χρήση Support Vector Machines για την ταξινόμηση δράσεων

Στην αρχή της διαδικασίας χωρίζουμε τα δεδομένα σε train και test data, ενώ στο τέλος πειραματιζόμαστε και με διαφορετικές διαμερίσεις του συνόλου δεδομένων. Για την τελική αναπαράσταση ενός βίντεο, χρησιμοποιείται η bag of visual words (BoVW) τεχνική που έχει περιγραφεί στην 1η εργαστηριακή άσκηση και για την κατηγοριοποίηση (classification) χρησιμοποιείται ενάς SVM ταξινομητής κατάλληλα προσαρμοσμένος για πολλαπλές κλάσεις.

Παραθέτουμε τα αποτελέσματα του classification για τους διάφορους συνδυασμούς ανιχνευτών-περιγραφητών για τον δοσμένο διαχωρισμό δεδομένων:

- Harris - HOG: 91.67%
- Harris - HOF: 83.33%
- Gabor - HOG: 91.67%
- Gabor - HOF: 83.33%
- Harris - HOG/HOF: 91.67%
- Gabor - HOG/HOF: 91.67%

Ξανατρέχοντας με ακριβώς τις ίδιες παραμέτρους τους ίδιους συνδυασμούς, παρατηρούμε ότι υπάρχει και μια τυχαιότητα στα αποτελέσματα, αφού κάποια ποσοστά διαφοροποιούνται. Τα νέα αποτελέσματα δεν διαφέρουν σημαντικά από τα πρώτα, παρά μόνο στην κατηγοριοποίηση 1 ή δύο βίντεο, ανεβάζονται ή ρίχνοντας το accuracy κατά το αντίστοιχο ποσοστό (όπως φαίνεται στα τρεγμένα κελιά του .ipynb αρχείου).

Ύστερα, για τους συνδυασμούς που δίνουν τα καλύτερα αποτελέσματα ξαναδοκιμάζουμε το classification για διαφορετικές διαμερίσεις:

- Για ένα train_set πολύ μικρό:
  - Gabor - HOG: 87.88%
  - Harris - HOG/HOF: 78.79%
  - Gabor - HOG/HOF: 87.88%
- Για ένα train_set πολύ μεγάλο:
  - Gabor - HOG: 100%
  - Harris - HOG/HOF: 66.67%
  - Gabor - HOG/HOF: 100%
- Για train_set με μόνο 2 βίντεο από την κατηγορία boxing:
  - Gabor - HOG: 50%
  - Harris - HOG/HOF: 72.22%
  - Gabor - HOG/HOF: 50%

Σχολιασμός Αποτελεσμάτων

Για τον δοσμένο διαχωρισμό σε train_set και test_set

Παρατηρούμε ότι ο καλύτερος συνδυασμός περιγραφτή-ανιχνευτή είναι ο Gabor-HOG και Gabor-HOG/HOF. Εκτελώντας πολλαπλές φορές την κατηγοριοποίηση, παρατηρούμε ότι ο καλύτερος ανιχνευτής (όπως ενδεχομένως να παρατηρούσαμε και από την οπτικοποίηση των αποτελεσμάτων ανίχνευσης) είναι ο Gabor. Την μεγαλύτερη διαφοροποίηση στο πρόβλημα του classification την προκαλεί, ωστόσο, η επιλογή του περιγραφητή. Ο HOG υπερτερεί του HOF, καθώς είναι αναλλοίωτος σε περιστροφές και κλιμακώσεις. Ο συνδυασμός τους εν τέλει φαίνεται να παράγει σε όλες τις εκτελέσεις (κατά μέσο όρο) τα καλύτερα αποτελέσματα και για τους δύο τύπους ανιχνευτών.

Στα misclassifications που συμβαίνουν παρατηρούμε ότι η κυριότερη σύγχυση γίνεται μεταξύ των πράξεων run και walk, κάτι που συμφωνεί και με τη διαίσθησή μας, καθώς οι δύο πράξεις είναι πιο σχετικές μεταξύ τους από όσο είναι με το box (τα δύο πρώτα περιέχουν κίνηση ίδιας φύσεως, ενώ το δεύτερο περιέχει στάσιμο άνθρωπο που κινεί μόνο τα χέρια του).

Για τους διαφορετικούς πειραματισμούς

Για τα παρακάτω πειράματα χρησιμοποιούμε τους καλύτερους συνδυασμούς περιγραφητών και ανιχνευτών από τα προηγούμενα.

- Πολύ μικρό train set: Παρατηρούμε χαμηλότερα ποσοστά, με δεδομένο ότι τα δεδομένα εκπαίδευσης είναι πολύ λίγα.
- Πολύ μεγάλο train set: Παρατηρούμε ότι τα ποσοστά είναι τα υψηλότερα δυνατά (σε δύο περιπτώσεις αγγίζεται το 100%). Αυτό δεν συνεπάγεται, ωστόσο, ότι ο μεγαλύτερος αριθμός βίντεο εκπαίδευσης οδηγεί και σε μεγαλύτερα ποσοστά, νομοτελειακά. Θα μπορούσε να οδηγήσει σε φαινόμενα overfitting, όμως τα βίντεο που διαθέτουμε και για τα δύο σύνολο δεδομένων είναι παρεμφερούς φύσης (δηλαδή όλα τα walk videos μοιάζουν μεταξύ τους, όλα τα box videos μοιάζουν μεταξύ τους κοκ) και λίγα σε πλήθος.
- Πολύ λίγα box videos στο train set: Παρατηρούμε τα χαμηλότερα ποσοστά accuracy. Τα misclassifications είναι πολλά, αφού τα run και walk videos είναι εύκολο να μπερδευτούν και για τα box videos δεν έχει αρκετά καλή εκπαίδευση, με αποτέλεσμα να "χάνονται" και αρκετά από αυτά.

References

[1] P.Maragos, *"Image Analysis and Computer Vision"*, Chapter 15: Visual Motion, June 23, 2018.