



Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής

Πολυτροπική Αναγνώριση και Κατάτμηση Δράσεων
Λεπτομέρειας σε Βίντεο

Διπλωματική Εργασία
του

Νικόλαου Χ. Γκανάτσιου

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2017



Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής

Πολυτροπική Αναγνώριση και Κατάτμηση Δράσεων Λεπτομέρειας σε Βίντεο

του

Νικόλαου Χ. Γκανάτσιου

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή στις 26 Οκτωβρίου 2017.

(Υπογραφή)

(Υπογραφή)

(Υπογραφή)

.....
Πέτρος Μαραγκός
Καθηγητής
Ε.Μ.Π

.....
Γεράσιμος Ποταμιάνος
Αναπληρωτής
Καθηγητής
Παν/μίου Θεσσαλίας

.....
Κων/νος Τζαφέστας
Επίκουρος Καθηγητής
Ε.Μ.Π

Αθήνα, Οκτώβριος 2017

(Υπογραφή)

.....
Νικόλαος Χ. Γκανάτσιος

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

© 2017– All rights reserved



Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής

Copyright © Νικόλαος Χ. Γκανάτσιος, 2017.
Με επιφύλαξη παντός δικαιώματος. All rights reserved

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Περίληψη

Ο βασικός στόχος-κίνητρο της παρούσας διπλωματικής εργασίας είναι η εξαγωγή αλγορίθμου δραστηριότητας από βίντεο σύνθετων ανθρώπινων δράσεων. Η πορεία μας εκκινεί από την παρουσίαση μιας γενικής και αφηρημένης μεθοδολογίας σχεδίασης ενός συστήματος που συνδυάζει πολυτροπική πληροφορία σε ένα ενιαίο σύστημα αναγνώρισης και κατάτμησης δράσεων σε βίντεο. Στη συνέχεια, προβαίνουμε στην υλοποίηση ενός τέτοιου συστήματος εστιάζοντας σε δράσεις λεπτομέρειας και πειραματιζόμενοι με την εξαγωγή και τον συνδυασμό χαρακτηριστικών πολλών καναλιών πληροφορίας, από οπτική (Πυκνές Τροχιές) μέχρι ακουστική (πληροφορίες υποτίτλων) και σημασιολογική (σχέσεις αντικειμένων-δράσεων και δράσεων-τύπων λαβής (grasping types)), με την τελευταία να εξάγεται και μέσω ανάλυσης κειμένου. Εξάγομε χαρακτηριστικά από ανάλυση με τη μέθοδο Πυκνών Τροχιών, από ανίχνευση αντικειμένων, τόσο οπτικά, μέσα σε μια δυναμική περιοχή ενδιαφέροντος που παρακολουθούμε με χρήση ανιχνευτή ανθρώπων και προσκηνίου, όσο και μέσω υποτίτλων και από την εξαγωγή τύπων λαβής με χρήση ενός εύρωστου ανιχνευτή χεριών και συνελικτικών χαρακτηριστικών με χρήση ResNet. Εκτελούμε σειρά πειραμάτων σχετικά με την κωδικοποίηση και τις μεθόδους ταξινόμησης αυτών των χαρακτηριστικών και καταλήγουμε στο ενδιαφέρον συμπέρασμα ότι το σχήμα Tf-Idf (ολικής συχνότητας - αντίστροφης συχνότητας κειμένου) ή και η απλή σώρρευση χαρακτηριστικών μπορούν να αντικαταστήσουν τον χ^2 μετασχηματισμό πυρήνων κατά τη σύμμειξη καναλιών διαφορετικής πληροφορίας αυξάνοντας ελαφρά την ακρίβεια αλλά σημαντικά την επίδοση από άποψη ταχύτητας όταν συνδυαστεί με μια γραμμική Μηχανή Διανυσμάτων Στήριξης (SVM). Η ιδιότητα αυτή επιτρέπει στο σχήμα αυτό να χρησιμοποιηθεί αποδοτικά από αλγορίθμους κατάτμησης βίντεο. Η προσέγισή μας στο ζήτημα της κατάτμησης είναι η ελαχιστοποίηση της συνάρτησης κόστους SVM με χρήση πιθανοτήτων και ενός νέου αλγορίθμου δυναμικού προγραμματισμού που είναι αμερόληπτος ως προς το μέγεθος των τελικών τμημάτων. Τελικά, από το αποτέλεσμα της κατάτμησης εξάγομε τον αλγόριθμο της δραστηριότητας κρατώντας τη χρήσιμη πληροφορία. Το σχήμα που χρησιμοποιούμε μας δίνει επιπλέον την πληροφορία αλληλεπίδρασης με τα αντικείμενα στον τελικό αλγόριθμο.

Λέξεις-Κλειδιά

Αναγνώριση Δράσεων, Κατάτμηση Βίντεο, Πολυτροπική Πληροφορία, Σχήμα Tf-Idf

Abstract

The objective of the current Thesis is the extraction of an algorithm describing a complex human activity performed in an observed video. We start by presenting a generic and abstract methodology for designing a joint video action segmentation and classification system, combining multiple modalities. We further present our implementation of such a system, focusing on fine-grained activities and experimenting on efficiently extracting and combining multiple information channels, from Low-Level Visual information (Dense Trajectories) to sound (subtitles) and semantics (action-object relations and grasping type-action relations), with the last category being supported by text analysis. We extract features using Dense Trajectories, object detection and recognition, both visually, searching inside a dynamic region of interest constructed using a combination of human and foreground detection, and via subtitles and lastly using grasping type information. The last type of information is obtained by applying a robust hand detector and then classifying the hand regions using ResNet deep convolutional features. We perform a sequence of experiments regarding feature encoding and classification and reach to an interesting result, that we are able to replace the χ^2 kernel fusion with Tf-Idf encodings or even feature concatenation, slightly increasing classification metrics but especially increase the speed of the classification progress, when a linear SVM is also used. This fact allows this schema to be efficiently used by video segmentation algorithms. Our approach when it comes to video segmentation is minimizing an SVM loss function using probabilities and a novel dynamic programming algorithm, invariant to final segments' length. Our method also returns object handling information in the total extracted algorithm.

Keywords

Action Segmentation, Video Segmentation, Multiple Modalities, Tf-Idf Encoding

Ευχαριστίες

Σε αυτό το σημείο θα ήθελα να ευχαριστήσω τον καθηγητή κ. Πέτρο Μαραγκό, ο οποίος με ενέπνευσε να ασχοληθώ με το αντικείμενο της Όρασης Υπολογιστών.

Πολλές ευχαριστίες οφείλω να δώσω στην οικογένειά μου για τη στήριξή τους τόσο κατά τη διάρκεια αυτής της εργασίας όσο και σε όλα τα στάδια της ζωής μου.

Ακόμα, πρέπει να ευχαριστήσω τα μέλη και συνεργάτες του εργαστηρίου CVSP Αθανασία Ζλατίντση, Πέτρο Κούτρα, Βασίλη Πιτσικάλη και Ισίδωρο Ροδομαγουλάκη για τη συνεισφορά τους στην περαιώση της εργασίας και τα σχόλιά τους πάνω στο κείμενο και τα πειράματα.

Τέλος, ευχαριστώ τους συναδέλφους Γαρούφη Χ., Λεμονίδη Β., Μπουρίτσα Γ., Νικολουδάκη Ε., Πίσσα Θ. και Χάντφιλντ Τ., με τους οποίους εκπονούσαμε παράλληλα τις διπλωματικές μας εργασίες και είχαμε την ευκαιρία να συζητήσουμε συχνά πολλά ενδιαφέροντα ζητήματα σχετικά με τις εργασίες μας.

Περιεχόμενα

Ευχαριστίες

x

Κατάλογος Εικόνων

xv

Κατάλογος Πινάκων

xxiii

1 Εισαγωγή	1
1.1 Η Κατάτμηση Δράσεων Και Οι Επιστήμες Πίσω Από Αυτή	1
1.2 Διάρθρωση Διπλωματικής Εργασίας	4
1.3 Πακέτα Λογισμικού που Χρησιμοποιήθηκαν	5
2 Το Συνολικό Σύστημα	7
2.1 Η Αξία της Σημασιολογίας και Εναλλακτικές Μέθοδοι Αναγνώρισης Δράσεων	9
2.2 Σχετικές Εργασίες	11
2.3 Σύνοψη Ολικού Συστήματος	13
2.3.1 Το Υποσύστημα Ακουστικής Πληροφορίας-Υποτίτλων	15
2.3.2 Το Υποσύστημα Εξαγωγής Χαρακτηριστικών	16
2.3.3 Συνδυασμός Χαρακτηριστικών και Ταξινομητής	18
2.3.4 Ο Αναλυτής Κειμένου	18
2.3.5 Ο Συνδυασμός των Πιθανοτήτων	18
2.3.6 Το Μπλοκ Κατάτμησης	18
2.4 Προτεινόμενη Υλοποίηση	19
2.5 Το Σύνολο Δεδομένων	20
3 Το Υποσύστημα Εξαγωγής Χαρακτηριστικών Όρασης Χαμηλού Επιπέδου	23
3.1 Ιστορικά Στοιχεία	23
3.2 Θεωρητικό Υπόβαθρο	25
3.2.1 Η Μέθοδος των Πυκνών Τροχιών	25
3.2.2 Είδη Περιγραφητών	27
3.2.2.1 Ο Περιγραφητής Τροχιάς	27
3.2.2.2 Ιστογράμματα Κατευθυνόμενων Παραγώγων	28
3.2.2.3 Ιστογράμματα Οπτικής Ροής	29
3.2.2.4 Ιστογράμματα Ορίων Κίνησης	29
3.2.3 Κωδικοποίηση Χαρακτηριστικών	30
3.2.4 Επιπλέον Χαρακτηριστικά: Τα Χαρακτηριστικά Πόζας	32
3.3 Προτεινόμενο Σύστημα	34
4 Σημασιολογικό Υποσύστημα: Αντικείμενα	37

4.1	Ανίχνευση Αντικειμένων σε Εικόνες	37
4.1.1	Γενικά	37
4.1.2	Ιστορικά Στοιχεία	39
4.2	Ανίχνευση Προσκηνίου και Παρακολούθηση Αντικειμένων σε Βίντεο	42
4.2.1	Γενικά	42
4.2.2	Ιστορικά Στοιχεία	44
4.3	Θεωρητικό Υπόβαθρο	47
4.3.1	Τοίριασμα Προτύπων (Template Matching)	47
4.3.2	Εξαγωγή Προσκηνίου με Μοντέλα Μίξης Γκαουσιανών (GMM)	49
4.3.3	Ανίχνευση Ανθρώπων με Χρήση Μοντέλων Παραμορφώσιμων Τμημάτων	50
4.4	Προτεινόμενο Σύστημα	52
4.4.1	Το Σύνολο Δεδομένων για Ανίχνευση Αντικειμένων	52
4.4.2	Η Εισαγωγή της Περιοχής Ενδιαφέροντος	54
4.4.3	Η Ανίχνευση Αντικειμένων	56
4.4.4	Η Αξιοποίηση της Ομιλίας στην Ανίχνευση Αντικειμένων	59
4.4.5	Συνδυασμός Εικόνας και Γλώσσας	60
5	Σημασιολογικό Υποσύστημα: Ο Τύπος Λαβής	63
5.1	Θεωρητικό Υπόβαθρο	64
5.1.1	Νευρωνικά Δίκτυα	64
5.1.2	Το Δίκτυο ResNet	67
5.1.3	Συνελικτικά Χαρακτηριστικά και Μεταβατική Εκμάθηση (Transfer Learning)	68
5.1.4	BING: Binarized Normed Gradients for Objectness Estimation	68
5.2	Προτεινόμενο Σύστημα	69
5.2.1	Ανίχνευση Χεριών	69
5.2.2	Εξαγωγή Τύπου Λαβής	73
6	Πειράματα και Αποτελέσματα Αναγνώρισης Δράσεων	75
6.1	Θεωρητικό Υπόβαθρο	75
6.1.1	Μηχανές Διανυσμάτων Υποστήριξης (SVM)	75
6.1.1.1	Μη Γραμμική SVM: Το Τέχνασμα Πυρήνα	76
6.1.1.2	Πιθανοτική SVM: Η Μέθοδος Platt	78
6.1.2	Το Σχήμα Tf-Idf	79
6.2	Η Δική Μας Προσέγγιση και Τα Πειραματικά αποτελέσματα	79
6.2.1	Οι Ρυθμίσεις του Συνόλου Δεδομένων	79
6.2.2	Αξιοποίηση Πληροφορίας Όρασης Χαμηλού Επιπέδου	80
6.2.3	Αξιοποίηση Σημασιολογικής Πληροφορίας: Αντικείμενα	81
6.2.4	Αξιοποίηση Σημασιολογικής Πληροφορίας: Τύποι Λαβής	82
6.2.5	Επανεκτίμηση Πιθανοτήτων Εξόδου	83
6.2.6	Συνολική Παρουσίαση Αποτελεσμάτων και Σύγκριση με Παγκόσμια Βιβλιογραφία	84
7	Κατάτμηση Δράσεων	87
7.1	Ιδέες και Τεχνικές Κατάτμησης Βίντεο	87
7.2	Ο αλγόριθμος των Hoai et al.	89
7.2.1	Εκπαίδευση SVM	89
7.2.2	Κατάτμηση με Καταπίεση μη-Μεγίστων	90
7.2.3	Δυναμικός Προγραμματισμός	90
7.3	Προτεινόμενος Αλγόριθμος	91

8 Συμπεράσματα-Επίλογος	97
8.1 Συμβολή της Διπλωματικής Εργασίας	97
8.2 Προτάσεις Για Μελλοντική Έρευνα	100
 Παραρτήματα	 103
 A Το Σύνολο Δεδομένων Και Οι Τροποποιήσεις Στις Οποίες Προβήκαμε	 105
A.1 MPII Cooking Activities Dataset [191]	105
A.2 MPII Cooking 2 Dataset [192]	106
A.3 Επιπλέον Επισημειώσεις Στις Οποίες Προβήκαμε	108
A.3.1 Υπότιτλοι	108
A.3.2 Επισημειώσεις Αντικειμένων	108
A.3.3 Επισημειώσεις Χεριών	110
 Βιβλιογραφία	 113

Κατάλογος σχημάτων

- | | | |
|-----|---|----|
| 1.1 | Ενδεικτικές δράσεις από τη βάση KTH [128]. Ένα πρόβλημα αναγνώρισης δράσεων θα αφορούσε το διαχωρισμό των παραπάνω δράσεων. Δηλαδή, δοθέντος ενός τμήματος βίντεο, ζητάμε τη δράση η οποία πραγματοποιείται σε αυτό, επιλέγοντας από τον κατάλογο δράσεων. | 3 |
| 2.1 | Κάποιες από τις εφαρμογές της ανάλυσης βίντεο στην ανθρώπινη ζωή. Είναι φανερό ότι η εξέλιξη της τεχνολογίας επιβάλλει ακόμα ποιοτικότερους αλγορίθμους και ταυτόχρονα προκύπτουν νέα πεδία εφαρμογής. Εικόνα από [48] | 8 |
| 2.2 | Η εμφάνιση δύο δράσεων ίδιου περιεχομένου μπορεί να αλλάζει σημαντικά μεταξύ διαφορετικών δειγμάτων. Άλλαγές στην οπτική της κάμερας, στο φωτισμό, στην εμφάνιση των δραστών, χρονικές μεταβολές στη διάρκεια κάθε δράσης και η ποικιλία στην ανθρώπινη μορφολογία και κίνηση είναι παράγοντες που καθιστούν το πεδίο αναγνώρισης δράσεων ιδιαίτερα απαιτητικό και προκλητικό. Εικόνα από [48] | 8 |
| 2.3 | Η αναγνώριση δράσεων μόνο οπτικά είναι συχνά δύσκολη σε περιπτώσεις όπου η μεταβλητότητα μεταξύ των κλάσεων είναι μικρή οπτικά. Χαρακτηριστικά όπως η στάση του σώματος δεν είναι εύρωστα στο μεταβλητό περιεχόμενο της δράσης. Για το λόγο αυτό, η ενσωμάτωση σημασιολογίας και πρότερης γνώσης μπορεί να συνεισφέρει στη διάκριση όπου η οπτική αντίληψη δεν επαρκεί. Εικόνα από [48] | 9 |
| 2.4 | Ένα περιγραφικό σύστημα αναγνώρισης δράσεων μελετά τα βίντεο εισόδου σε πολλαπλά επίπεδα και συνδυάζει την πληροφορία για να εξάγει αποτέλεσμα. Εκτός από την κίνηση, τα συστήματα αυτά συγκεντρώνουν και σημασιολογικές περιγραφές της δράσης. Ταυτόχρονα, η οπτική ανάλυση δε γίνεται με παραδοσιακές μεθόδους, αλλά με περιγραφική μοντελοποίηση σε διάφορες κλίμακες. Η προσέγγιση αυτή εξελίχθηκε στο συνδυασμό χαρακτηριστικών χαμηλού επιπέδου οπτικής πληροφορίας με σημασιολογία χωρίς να μεσολαβεί η μοντελοποίηση με αυστηρούς κανόνες. Εικόνα από [202] | 10 |
| 2.5 | Το σύστημα του [259]. Δύο νευρωνικά δίκτυα για τύπο λαβής και αντικείμενα δίνουν μια κατανομή πιθανοτήτων η οποία κωδικοποιείται με μια γραμματική χωρίς συμφραζόμενα. Παράλληλα, το μοντέλο γραμματικής ενισχύεται από πληροφορία κειμένου. Το αποτέλεσμα είναι ένα συντακτικό δέντρο που δίνει τον τύπο δράσης, τα συσχετιζόμενα αντικείμενα και τον τύπο λαβής τους. Εικόνα από [259] | 12 |

- 2.6 Το συνολικό σύστημα αναγνώρισης-κατάτμησης δράσεων που προτείνουμε σε μορφή μπλοκ διαγράμματος. Οι δομικές του μονάδες αλληλεπιδρούν αλλά σχεδιάζονται και υλοποιούνται ανεξάρτητα ώστε να μπορούν να ενημερώνονται αυτόνομα. Σε κάθε τμήμα βίντεο γίνεται συνδυασμός χαρακτηριστικών πολλαπλών καναλιών και εξάγεται η κατανομή πιθανοτήτων των δράσεων. Ο αλγόριθμος κατάτμησης χρησιμοποιεί δυναμικό προγραμματισμό για να προτείνει νέο τμήμα βίντεο προς εξέταση. Η ταξινόμηση στα τελικά τμήματα γίνεται σύμφωνα με τη μέγιστη εκ των υστέρων (posterior) πιθανότητα. 14
- 2.7 Η εσωτερική δομή του Υποσυστήματος Ακουστικής Πληροφορίας. Μετά τη μετατροπή των ακουστικών υποτίτλων σε κείμενο γίνεται η συντακτική ανάλυση και η χρονική ευθυγράμμιση με την εικόνα. Οι κατανομές πιθανοτήτων που εξάγονται από την ανάλυση αυτή θα συνδυαστούν με τις αντίστοιχες που θα προκύψουν από τα υπόλοιπα κανάλια πληροφορίας. 15
- 2.8 Μια τραπέζοειδής ασαφής συνάρτηση. Τέτοιας μορφής συναρτήσεις χρησιμοποιούνται για να αποδόσουν μια ένδειξη χαλαρών άκρων συνάρτησης. Η χρονική συσχέτιση λόγου και εικόνας στο σύστημά μας γίνεται με τέτοιες συναρτήσεις γύρω από τη διάρκεια του κάθε υποτίτλου. 16
- 2.9 Το Υποσύστημα Εξαγωγής Χαρακτηριστικών αναλαμβάνει το καθήκον εξαγωγής οπτικών και σημασιολογικών χαρακτηριστικών από την παρακολούθηση της οπτικής πληροφορίας του βίντεο, ενώ η πληροφορία υποτίτλων μπορεί να ενισχύει την πεποίθηση σημασιολογικών χαρακτηριστικών. Τα χαρακτηριστικά κωδικοποιούνται σε μορφή ιστογραμμάτων σε αυτό το στάδιο, οπότε η πληροφορία των καναλιών είναι συγκρίσιμη και να μπορεί να γίνει σύζευξη πριν το στάδιο του ταξινομητή. 17
- 2.10 Το Υποσύστημα Σημασιολογικής Πληροφορίας αναμιγνύει την οπτική ανίχνευση με την αντίληψη του συστήματος για τη δομή των δράσεων. Παρότι στο σχήμα απεικονίζεται η δομή που τελικά ακολουθήσαμε, συνδυάζοντας πληροφορία αντικειμένων και τύπων λαβής (grasping type), είναι φανερό ότι πολλαπλά κανάλια σημασιολογίας μπορούν να εισαχθούν ανεξάρτητα και τελικά να παραχθούν τα αντίστοιχα ιστογράμματα. 17
- 2.11 Το σύνολο δεδομένων που χρησιμοποιούμε (MPII Cooking Activities Dataset) αποτελείται από βίντεο στα οποία διαφορετικοί άνθρωποι εκτελούν συνταγές μαγειρικής και ετοιμάζουν διαφορετικά πιάτα. Το σύνολο δεδομένων προσφέρεται τόσο για αναγνώριση δράσεων (action) όσο και για αναγνώριση δραστηριοτήτων (activity). Εμείς κινούμαστε προς την αναγνώριση δράσεων. 21
- 3.1 Η εξέλιξη της Αναγνώρισης δράσεων με τεχνικές που στηρίζονται αποκλειστικά στην οπτική αντίληψη, από τις ισχυρές υποθέσεις στην αφηρημένη δομή δικτύου. **Αριστερά:** ένα παράδειγμα σειριακής προσέγγισης που κάνει χρήση Κρυφού Μαρκοβιανού Μοντέλου για να αναπαραστήσει την απλή δράση "Σπρώχνω" με καταστάσεις που αντιστοιχούν σε ενδιάμεσες στάσεις σώματος. Εικόνα από [48]. **Δεξιά:** Δομή Συνελικτικού Νευρωνικού Δικτύου για ανάλυση αθλητικών αγώνων από την εργασία [116]. 24

- 3.2 Απεικόνιση της μεθόδου των Πυκνών Τροχιών. **Αριστερά:** Τα σημεία παρακολούθησης δειγματοληπτούνται πυκνά για πολλαπλές χωρικές κλίμακες. **Κέντρο:** Η παρακολούθηση γίνεται στην αντίστοιχη χωρική κλίμακα και για L frames. **Δεξιά:** Οι περιγραφητές τροχιάς βασίζονται στο σχήμα της, που αντιπροσωπεύεται από σχετικές συντεταγμένες σημείων, αλλά και σε πληροφορίες κίνησης και εμφάνισης σε γειτονιές $N \times N$ pixels κατά μήκος της τροχιάς. Προκειμένου να ενσωματωθεί δομική πληροφορία, η κάθε γειτονιά χωρίζεται επιπλέον σε ένα χωρο-χρονικό πλέγμα διαστάσεων $n_\sigma \times n_\sigma \times n_\tau$. Εικόνα από το [248]. 26
- 3.3 Απεικόνιση της πληροφορίας που περιέχεται στους περιγραφητές HOG, HOF και MBH. Για κάθε εικόνα, τα διανύσματα κλίσης/οπτικής ροής διακρίνονται με χρώμα (hue) και τα αντίστοιχα μέτρα τους με τη χρωματική συνιστώσα κορεσμού (saturation). Τα όρια κίνησης υπολογίζονται ως παράγωγοι των συνιστωσών x και y της οπτικής ροής ξεχωριστά. Σε σύγκριση με την οπτική ροή, τα όρια κίνησης συμπλέζουν το μεγαλύτερο μέρος της κίνησης του υποβάθρου λόγω κινήσεων της κάμερας και τονίζουν την κίνηση προσκηνίου. Σε αντίθεση με την πληροφορία που λαμβάνεται από τις παραγώγους, τα όρια κίνησης σβήνουν περισσότερη πληροφορία υφής από στατικά παρασκήνια. Εικόνα από το [248]. 28
- 3.4 Απεικόνιση της μεθόδου του Σάκου Οπτικών Λέξεων (Bag of Visual Words, εν συντομίᾳ BoW). Οι διαφορετικοί τύποι χαρακτηριστικών εξάγονται σε σημεία ενδιαφέροντος στο βίντεο. Στη συνέχεια κωδικοποιούνται για να σχηματίσουν ένα λεξιλόγιο, τον Σάκο Λέξεων. Ένα βίντεο τώρα αναπαρίσταται σαν μια πρόταση από οπτικές λέξεις και ταξινομείται σύμφωνα με το ιστόγραμμα των λέξεων αυτών. Μια απλή περίπτωση είναι η ταξινόμηση απευθείας σύμφωνα με τα ιστογράμματα αυτά. Στην πορεία αυτής της εργασίας θα δούμε αρκετά πιο σύνθετες μεθόδους. Εικόνα από το [47]. 31
- 3.5 Μερικά παραδείγματα επιτυχούς ανίχνευσης στάσης του άνω ήμισυ του σώματος. Η στάση μπορεί να σχετίζεται άμεσα με την δράση. Ιδιαίτερα η θέση και ο σχηματισμός των χεριών αποτελούν σημαντικό διακριτικό στοιχείο των λεπτομερών δράστηριοτήτων. Εικόνα από το [5]. 33
- 3.6 Μερικά παραδείγματα εξαγωγής εμπροσθόδρομων και οπισθόδρομων τροχιών των 10 τμημάτων του άνω μισού μέρους του σώματος κατά την παρακολούθησή τους στην εξαγωγή των χαρακτηριστικών πόζας. Με πράσινο απεικονίζονται οι οπισθόδρομες μετακινήσεις και με κόκκινο οι εμπροσθόδρομες. Με κυανό σημειώνεται η αρχική θέση του κάθε συνδέσμου. Εικόνα από το [191]. 35
- 4.1 Παράδειγμα εξόδου συστήματος ανίχνευσης αντικειμένων. Βλέπουμε ότι το σύστημα μας επιστρέφει το ορθογώνιο στο οποίο εντοπίζεται το ζητούμενο αντικείμενο. Στα ορθογώνια επισημειώνεται επιπλέον το σκορ πεποιθησης του συστήματος για την ανίχνευση και ταυτοπίηση στην οποία προέβη. Εικόνα από το [186]. 38
- 4.2 Η αμεταβλητότητα αποτέλεσε πρωταρχικό στόχο των εργασιών ανίχνευσης και αναγνώρισης αντικειμένων. Είναι πρώτης σημασίας οι διαφορετικές όψεις του αντικειμένου να έχουν παρόμοια αναπαράσταση στο χώρο των χαρακτηριστικών που χρησιμοποιούνται για την ταυτοποίησή τους. Εδώ για παράδειγμα απεικονίζονται τρεις διαφορετικές όψεις ενός εργαλείου και η συνάρτηση αναπαράστασής τους. Παρατηρούμε την επιθυμητή ομοιότητα μεταξύ των αναπαραστάσεων αυτών. Εικόνα από το [200]. 40

- 4.3 Η εργασία των Viola-Jones αποτέλεσε σταθμό στην έρευνα πάνω στην ανίχνευση αντικειμένων χάρη στην απλότητα, ταχύτητα και αποτελεσματικότητά της. Στην εικόνα φαίνονται τα αποτελέσματα της εκτέλεσης του αλγορίθμου ανίχνευσης προσώπου πάνω στις δοθείσες εικόνες. Παρότι η αρχική σχεδίαση έγινε για ανίχνευση προσώπου, η μέθοδος γενικεύεται και δίνει ικανοποιητικά αποτελέσματα για μια ποικιλία αντικειμένων υπό ορισμένες προϋποθέσεις και περιορισμούς στους μετασχηματισμούς που μπορούν να υφίστανται τα ανιχνευόμενα αντικείμενα και απαιτεί μικρό αριθμό θετικών δειγμάτων εκπαίδευσης. Εικόνα από το [245]. 41
- 4.4 Ζήτημα της ανίχνευσης προσκηνίου είναι η εύρεση όχι μόνο των κινούμενων αντικειμένων αλλά και των κινήσεων ενδιαφέροντος. Παρότι ο ορισμός αυτός δεν είναι σαφής, διαισθητικά ζητάμε την κίνηση που αφορά την απεικονιζόμενη δράση στην οποία και εστιάζουμε και όχι πιθανές άλλες δράσεις που μπορεί να συμβαίνουν στο περιβάλλον, κάτι που ονομάζουμε παρασκήνιο. Στο εικονιζόμενο παράδειγμα, προσκήνιο αποτελούν οι αθλήτριες και η μπάλα, αλλά όχι κάποια κίνηση που μπορεί να πραγματοποιείται στην εξέδρα. Εικόνα από το [54]. 43
- 4.5 Παρακολούθηση κινούμενου αυτοκινήτου σε πραγματικό χρόνο με τη βοήθεια Κάρτας Γραφικών Υπολογιστή (GPU). Παρατηρούμε ότι η κάμερα κινείται και ως εκ τούτου η σχετική θέση κάμερας και περιβάλλοντος μεταβάλλεται διαρκώς. Εν τούτοις το σύστημα κατορθώνει να απομονώνει το αντικείμενο ενδιαφέροντος και να κατατάσσει την υπόλοιπη εικόνα ως παρασκήνιο. Οι κάρτες γραφικών συνέβαλλαν δραστικά στην αύξηση της ταχύτητας υπολογισμών για τις απαιτητικές εφαρμογές της σημερινής εποχής. Εικόνα από το [159]. 46
- 4.6 Η λογική της μεθόδου Ταιριάσματος Προτύπων: Δοθεντος ενός προτύπου, αναζητούμε τη βέλτιστη τοποθέτησή του στην εικόνα έτσι ώστε να μεγιστοποιείται μια συνάρτηση ομοιότητας μεταξύ προτύπου και τμημάτος εικόνας. Η συνάρτηση ομοιότητας μπορεί να επιλεχθεί κατάλληλα σύμφωνα με το πρόβλημα και τις απαιτήσεις του σχεδιαστή. Συχνά, χρησιμοποιείται κατώφλι στη συνάρτηση ομοιότητας και τιμές μικρότερες αυτού σημαίνουν μη ταίριασμα και απουσία του προτύπου από την εικόνα. 47
- 4.7 Ενδιάμεσα στάδια κατά την εφαρμογή της μεθόδου των Μοντέλων Παραμορφώσιμων Τμημάτων. **Αριστερά:** Το μοντέλο για την ανίχνευση ανθρώπων. Στο σχήμα φαίνονται τα σταθερά μέρη του ανθρώπου (όκρα και σώμα) τα οποία μπορούν να κινούνται σχετικά και να παραμένουν ενωμένα με χαλαρούς συνδέσμους. Ακόμα, βλέπουμε την απεικόνιση του μοντέλου στο πεδίο των περιγραφητών HOG και τα μέρη του όταν αναλυθούν. **Δεξιά:** Η κάθε εικόνα αναλύεται σε πολλαπλές κλίμακες έτσι ώστε να ενσωματώσει την πληροφορία του μεγέθους. Βλέπουμε πώς η πυραμίδα στο επίπεδο των κλιμάκων εικόνων μεταφέρεται στο χώρο χαρακτηριστικών HOG. Εικόνα από το [69]. 51
- 4.8 Παρά την αλλαγή στη δομή και στη φύση τους, τα αντικείμενα συχνά εκπροσωπούνται από την ίδια ετικέτα κλάσης. Για παράδειγμα, μια πατάτα μπορεί να κοπεί σε κομμάτια, τα οποία διατηρούν την ιδιότητα της πατάτας. Ομοίως και μια σαλάτα καρότων, εντάσσεται στην κατηγορία καρότων. Τέτοιες διενέξεις στην εμφάνιση και την ετικέτα συμβαίνουν συχνά στο σύνολο δεδομένων που χρησιμοποιείται σε αυτή την εργασία, γεγονός που δυσχεραίνει σημαντικά τον εντοπισμό και την παρακολούθηση των αντικειμένων στη διάρκεια του βίντεο. 53

- 4.9 Ένα παράδειγμα εξαγωγής περιοχής ενδιαφέροντος. Σε πρώτη φάση τρέχουμε ανίχνευση ανθρώπων βασισμένο σε DPM. Σε δεύτερη φάση, τρέχουμε ανίχνευση προσκηνίου με GMM. Συνενώνουμε το πλήθος των περιοχών που προκύπτουν σε όλη τους την έκταση. Επεκτείνουμε την περιοχή ενδιαφέροντος κατά ένα μικρό αριθμό πίξελ δεξιά και αριστερά, ενώ την αφήνουμε να λάβει όλο το ύψος της εικόνας. Έτσι μπορούμε να συλλάβουμε τμήματα αντικειμένων χώρου τα οποία μας βοηθούν στην ανίχνευσή τους όταν χρειάζεται, χωρίς όμως να κείνται εντός της περιοχής προσκηνίου ή εύρεσης ανθρώπου. 55
- 4.10 Η οπτική ανίχνευση αντικειμένων μπορεί εύκολα να λειτουργήσει αποτελεσματικά, με τη μέθοδο Ταιριάσματος Προτύπων ή άλλη, σε εικόνες όπου γνωστά αντικείμενα κείνται ακίνητα και πρακτικά χωρίς επικαλύψεις. Εν τούτοις, μόνο η οπτική πληροφορία εισάγει με μεγάλη πιθανότητα σφάλματα μη χρήσης. Στην εικόνα, μπορούμε εύκολα να διακρίνουμε αντικείμενα όπως φρούτα και μαχαίρι πάνω στον πάγκο, ωστόσο ο δράστης δεν τα χρησιμοποιεί. Η πληροφορία ανίχνευσης αυτών των αντικειμένων είναι θόρυβος που επιβαρύνει τον τελικό ταξινομητή σημασιολογίας. Το πρόβλημα λύνεται με την εισαγωγή της περιοχής ενδιαφέροντος, χάρη στην οποία διαπιστώνουμε ότι τα εντοπιζόμενα αντικείμενα δεν βρίσκονται στην περιοχή προσκηνίου και άρα δε λαμβάνονται υπόψιν. 57
- 4.11 Πολλές φορές η ανίχνευση αντικειμένων σε μια εικόνα είναι δυσεπίλυτο πρόβλημα ακόμα και για τον άνθρωπο. Δεν αρκεί μια εικόνα, κατά πάσα πιθανότητα, έτσι ώστε να συμπεράνουμε με ασφάλεια τι κρατάει στα χέρια του ο άνθρωπος της εικόνας. Σε αυτό μπορεί να συμβάλλει η εισαγωγή πρότερης γνώσης ή η ομιλία. Σε τέτοια βίντεο είναι αρκετά συχνή η συμπερίληψη των χρησιμοποιούμενων αντικειμένων στον λόγο. Συνδυάζοντας κατάλληλα λόγο και εικόνα λαμβάνουμε αρκετά πιο ισχυρές ενδείξεις για την εμφάνιση ενός αντικειμένου. 60
- 4.12 Θεωρούμε αυξημένη την πιθανότητα ένα αντικείμενο να εμφανίζεται ”όσο διαρκεί ο υπότιτλος και λίγο μετά”. Η φράση αυτή ερμηνεύεται μαθηματικά με μια ασαφή συνάρτηση μετοχής, η οποία έχει τιμή 1 στα χρονικά όρια του υποτίτλου και φθίνει γραμμικά μέχρι μηδενισμού από το πέρας του υποτίτλου μέχρι και 10% της διάρκειάς του έπειτα από αυτόν. Για παράδειγμα, η συνάρτηση μετοχής για ένα αντικείμενο που αναφέρεται στον υπότιτλο στο χρονικό παράθυρο μεταξύ 20 και 30 δευτερολέπτων έχει τη μορφή της εικόνας. 61
- 5.1 Ο σχηματισμός των χεριών στο χειρισμό των αντικειμένων εμπεριέχει σημασιολογική πληροφορία η οποία μπορεί να υποδεικνύει την πρόθεση ή τη δράση αυτή καθαυτή. Στην εικόνα αριστερά για παράδειγμα, ο τρόπος λαβής του μαχαιριού είναι προσανατολισμένος στη δύναμη και μπορούμε να υποθέσουμε ότι το μαχαίρι πρόκειται να χρησιμοποιηθεί σε κάποια εργασία. Αντίθετα στη μεσαία εικόνα, ο τύπος λαβής είναι προσανατολισμένος στην ακρίβεια. Η φυσική απόκριση στη λαβή αυτή είναι να δεχτούμε το μαχαίρι που μας αποδίδεται, όπως στην εικόνα δεξιά. Εικόνα από [255]. 63

5.2	Η εξέλιξη των νευρωνικών στην πορεία τους. Εκκινώντας εμπνεόμενα από τα βιολογικά νευρικά κύτταρα, τα πρώτα νευρωνικά δίκτυα αποτελούνταν από μια γραμμική συνδεσμολογία απλών νευρώνων σε επίπεδα. Στο Neocognitron [79] εμφανίζεται η οργάνωση σε επίπεδα με διαφορετικό συναπτικό πεδίο που απεικονίζεται σε διαφορετικά τμήματα εισόδου. Οι νευρώνες και κατά συνέπεια τα εκφραζόμενα χαρακτηριστικά αποκτούν μια τοπικότητα η φαίνεται να χρησιμεύει ιδιαίτερα στην ανάλυση εικόνων και βίντεο. Στα τελευταία χρόνια, η χρήση ταχύτερου υλικού (hardware) και η ανάπτυξη προγραμματιστικών τεχνικών (παραλληλία) επέτρεψαν τη σχεδίαση και εφαρμογή βαθιών συνελικτικών δικτύων, όπως το εικονιζόμενο GoogLeNet [233], τα οποία έγιναν ασυναγώνιστα σε πολλά πεδία της Μηχανικής Μάθησης.	65
5.3	Ένα συνελικτικό νευρωνικό δίκτυο οργανώνει τους νευρώνες του σε τρεις διαστάσεις, πλάτος, ύψος και βάθος, όπως οπτικοποιούνται εδώ σε ένα από τα επίπεδά του. Κάθε επίπεδο ενός συνελικτικού νευρωνικού δικτύου μετασχηματίζει έναν τρισδιάστατο όγκο εισόδου σε έναν τρισδιάτατο χώρο ενεργοποιήσεων νευρώνων.	65
5.4	Η δράση της συνέλιξης. Μια σειρά από φίλτρα κυλίονται επί της εισόδου και το αποτέλεσμα του τοπικού εσωτερικού γινομένου απεικονίζεται απευθείας στην έξοδο. Η έξοδος θα έχει διάσταση βάθους όσο και το πλήθος των φίλτρων. Οι χωρικές διαστάσεις προκύπτουν από το πλήθος των σημείων όπου υπολογίζονται τοπικά τα εσωτερικά γινόμενα καθώς το φίλτρο κυλίεται. Συνήθως, βήμα ολίσθησης μοναδιαίο σε συνδυασμό με κατάλληλη αρχική προσαύξηση με μηδενικά αφήνουν αναλλοίωτες τις χωρικές διαστάσεις εξόδου ως προς την είσοδο, ενώ η μεταβλητή είναι η διάσταση βάθους.	66
5.5	Η μορφή των residual συναρτήσεων. Ένας εμπροσθόδρομος βρόχος παρακάμπτει ενδιάμεσα στάδια και αθροιζεται στην έξοδό τους. Το σχήμα αυτό κρύβει την πεμπτουσία των δικτύων ResNet, τα οποία βελτιστοποιούνται και πετυχαίνουν κορυφαίες αποδόσεις, ακόμα και συγκρινόμενα με άλλα συστήματα. Εικόνα από [89].	67
5.6	Παρόλο που ο χώρος των παραθύρων αντικειμένων (κόκκινο) και του παρασκηνίου (πράσινο) παρουσιάζει τεράστιες διακυμάνσεις στο χώρο των εικόνων, σε κατάλληλες κλίμακες, τα κανονικοποιημένα διανύσματα παραγγώνεις εμφανίζουν μεγάλη συσχέτιση. Με ένα απλό μοντέλο 64 διαστάσεων, μπορούμε να λάβουμε αξιόπιστες προτάσεις για την ύπαρξη ή μη αντικειμένου. Εικόνα από [33].	69
5.7	Παραδείγματα επιτυχούς ανίχνευσης χεριών. Τα προκύπτοντα παράθυρα μπορούν να εισέλθουν στο επόμενο στάδιο εξαγωγής χαρακτηριστικών και τελικά να ταξινομηθούν ως τύποι λαβής.	70
5.8	Περιπτώσεις όπου η ανίχνευση χεριών αποτυγχάνει. Είναι πιθανό, περιοχές κοντά σε χέρια να ταξινομηθούν ως χέρια λόγω του χρωματικού κατωφλίου και των περιορισμών έκτασης που εφαρμόζονται στο τικό στάδιο του αλγορίθμου ανίχνευσης. Εν τούτοις, προτιμήθηκε η ύπαρξη τέτοιων σφαλμάτων μπροστά στα οφέλη μιας τέτοιας σχεδίασης. Κατά την εξαγωγή του τύπου λαβής, τα σφάλματα εισάγουν θόρυβο που αντισταθμίζουμε με την εισαγωγή επιπλέον κατηγοριών στο στάδιο μη επιβλεπόμενου διαχωρισμού.	71

- 5.9 Βήματα του αλγορίθμου ανίχνευσης χεριών. Η αρίθμηση εφαρμόζεται από πάνω αριστερά προς τα κάτω δεξιά. **Εικόνα 1:** Η αρχική εικόνα. **Εικόνα 2:** Η περιοχή που καλύπτουν τα παράθυρα με μηδενική πιθανότητα συμπερίληψης χεριού. Τα παράθυρα αυτά προκύπτουν από την κύλιση ενός παραθύρου πάνω στην εικόνα, την εξαγωγή τοπικών χαρακτηριστικών και την ταξινόμηση αυτών. **Εικόνα 3:** Ο χάρτης πιθανοτήτων ύπαρξης χεριού. Ο χάρτης αυτός προκύπτει μετά από κανονικοποίηση της εικόνας 2. **Εικόνα 4:** Τα superpixels. Ο χάρτης των superpixels προκύπτει από την κατωφλίωση του χάρτη πιθανοτήτων. Βλέπουμε ότι υπάρχουν δύο συμπαγείς περιοχές πλέον. Καθώς στο άξονα γ δεν υπάρχει επικάλυψη, η πάνω περιοχή αποκόπτεται. Η κάτω περιοχή χρειάζεται επιπλέον ανάλυση για διαχωρισμό των χεριών. **Εικόνα 5:** Με την εφαρμογή χρωματικών κατωφλίων και μορφολογικών τελεστών απομονώνουμε τις περιοχές των χεριών. **Εικόνα 6:** Η τελική εικόνα ανίχνευσης. 72
- 5.10 Οι τύποι λαβής στους οποίους στηριζόμαστε ως χρήσιμη πληροφορία για την αναγνώριση δράσεων. Λαβές που δεν ταξινομούνται σε αυτές τις κατηγορίες θα είναι λαβές έκτασης και ξεκούρασης. Εικόνα από [255]. 74
- 6.1 Είναι φανερό ότι σε ένα πρόβλημα ταξινόμησης δύο γραμμικά διαχωρίσιμων κλάσεων υπάρχουν άπειρες λύσεις. Ποια από αυτές είναι η προτιμότερη; Αν το κριτήριο είναι η ικανότητα γενίκευσης, τότε βέλτιστη λύση, για τα υπάρχοντα δεδομένα, είναι η γραμμή που αφήνει μέγιστο περιθώριο και για τις δύο κλάσεις, έτσι ώστε να διατηρείται ισορροπία μεταξύ των χώρων σφαλμάτων των κλάσεων. Ο αλγόριθμος SVM αναζητά τη βέλτιστη αυτή γραμμή, μέσα από ένα πρόβλημα ελαχιστοποίησης κόστους. Η αναζήτηση βέλτιστης λύσης και όχι απλά λύσης διαχωρίζει τον ταξινομητή SVM από τους υπόλοιπους γραμμικούς ταξινομητές. Εικόνα από [171] 76
- 6.2 Η μέθοδος ταξινόμησης SVM αναζητά τους βέλτιστους εκπροσώπους των ορίων των κλάσεων, τα λεγόμενα Διανύσματα Υποστήριξης (Support Vectors). Σύμφωνα με αυτά τα διανύσματα, οριοθετεί γραμμικά τους χώρους των κλάσεων και λαμβάνει τη μεσοπαράλληλο των γραμμικών συνόρων ως τη βέλτιστη διαχωριστική γραμμή μεταξύ των κλάσεων. Σε παραπάνω διαστάσεις, η διαχωριστική επιφάνεια είναι ένα υπερεπίπεδο. Με μετασχηματισμό πυρήνα, το υπερεπίπεδο αυτό μπορεί να αναπαριστά μια υπερεπιφάνεια η οποία προκύπτει από μια μη γραμμική απεικόνιση του χώρου χαρακτηριστικών σε έναν άλλο, διαφορετικής πιθανόν διάστασης. Εικόνα από [171] 77
- 6.3 Ο ίδιος χώρος δεδομένων μπορεί να διαχωριστεί αρκετά διαφορετικά αν μεσολαβήσει μη γραμμική απεικόνιση σε έναν άλλο χώρο όπου μια μη γραμμική διαχωριστική επιφάνεια μπορεί να γίνει διαχωριστική. Για παράδειγμα, σε έναν χώρο που προκύπτει από έναν πολυωνυμικό μετασχηματισμού βαθμού n , ένα πολυώνυμο βαθμού n μπορεί να αναπαριστά μια ευθεία. Μετασχηματισμοί αυτού του είδους χρησιμοποιούνται για να διευκολύνουν τη γραμμική διαχωρισμότητα και την αποδοτική χρήση της μεθόδου SVM. Εικόνα από [171] 78
- 7.1 Κατά το σπάσιμο ενός τμήματος AB έχουμε πολλαπλές επιλογές ως το πλήθος και το μήκος των επιμέρους τμημάτων. Ποια είναι όμως η βέλτιστη; Σύμφωνα με τους Hoai et al. [92], αυτό προκύπτει αναζητώντας το μέγιστο περιθώριο κέρδους μεταξύ της πιο πιθανής κλάσης και της αμέσως λιγότερο πιθανής. Οπότε στην εικόνα του σχήματος, όπου έχουμε δύο κλάσεις με σκορ που φαίνονται στο σχήμα για κάθε τμήμα, προτιμότερο είναι το σπάσιμο στο σημείο N. Εικόνα από [92] 89

- 7.2 Αποτελέσματα της κατάτμησης του βίντεο s19-d01 για το δυαδικό πρόβλημα δράσεων υποβάθρου εναντίον δράσεων μη υποβάθρου. Ποιοτικά (εικόνα) αλλά και ποσοτικά (80.5% mAP) τα αποτελέσματα είναι ικανοποιητικά. Η παρούσα εικόνα δείχνει με μαύρο τις περιοχές δράσεων υποβάθρου και με λευκό τις δράσεις προσκηνίου. Όπως φαίνεται, το σχήμα περικλείει δύο γραμμες αποτελεσμάτων. Στην πάνω γραμμή φαίνεται η πραγματική (ground truth) κατάτμηση και στην κάτω η υπολογισμένη κατάτμηση με χρήση του αλγορίθμου που προτείνουμε. 95
- A.1 Παράδειγμα δημιουργίας επισημειώσεων με το λογισμικό trainingImageLabeler του MATLAB. Εισάγουμε τα ορθογώνια και την ετικέτα και το λογισμικό εξάγει τις συντεταγμένες. Η χρήση του πακέτου αυτού διευκολύνει τη μαζική δημιουργία θετικών δειγμάτων εκπαίδευσης για ανιχνευτές αντικειμένων. 109
- A.2 Παράδειγμα αναπαράστασης σταθερού αντικειμένου με το τμήμα του το οποίο εμφανίζεται μόνο όταν το αντικείμενο χρησιμοποιείται. Έτσι, εξασφαλίζεται ότι το αντικείμενο θα εντοπίζεται μόνο σε περιπτώσεις ενδιαφέροντος και όχι σε όλη τη διάρκεια του βίντεο. 110
- A.3 Παραδείγματα επισημειώσεων πόζας. Με μπλε απεικονίζονται τα σημεία που χαρακτηρίζονται ως χέρια μέσω των επισημειώσεων πόζας. Δείχνουμε τις περιοχές ως τετράγωνα 10×10 γύρω από τα σημεία επισημείωσης για να είναι εμφανείς οπτικά στον αναγνώστη. Βλέπουμε ότι ενώ στη δεξιά εικόνα οι επισημειώσεις ταυτίζονται με τη θέση των χεριών, τα οποία είναι εμφανή οπτικά, στην εικόνα αριστερά οι επισημειώσεις δείχνουν τη θέση που βρίσκεται το χέρι πίσω από το σώμα του ανθρώπου. Είναι φανερό ότι εικόνες σαν την αριστερή δεν μπορούν να χρησιμεύσουν σαν θετικά δείγματα για έναν ανιχνευτή χεριών και έτσι αποκλείονται από τα θετικά δεδομένα εκπαίδευσης και χρησιμοποιούνται μόνο για την εξαγωγή αρνητικών παραδειγμάτων. . . 111

Κατάλογος πινάκων

4.1	Αποτελέσματα 4 μετρικών για την ανίχνευση αντικειμένων πάνω σε όλα τα βίντεο του συνόλου δεδομένων test που χρησιμοποιήσαμε. Παρατηρούμε ότι με σύζευξη λόγου και εικόνας μπορούμε να πετύχουμε ικανοποιητικά αποτελέσματα, πολύ πιο εύρωστα από ό,τι με χρήση μόνο ενός από τους δύο τρόπους.	62
6.1	Αποτελέσματα των μεθόδων που αξιοποιούν μόνο πληροφορία χαμηλού επιπέδου Όρασης Υπολογιστών. Η προτεινόμενη μέθοδος φαίνεται ότι υπερισχύει των υπολοίπων και ότι η χρήση χ^2 πυρήνα είναι απαραίτητη.	81
6.2	Αποτελέσματα των μεθόδων που αξιοποιούν και πληροφορία αντικειμένων σε σύγκριση. Βλέπουμε ότι τα ποσοστά ανεβήκαν και ότι η ανάγκη για χ^2 μετασχηματισμό υποχωρεί.	82
6.3	Αποτελέσματα των μεθόδων που αξιοποιούν και πληροφορία τύπων λαβής. Υπάρχει μικρή βελτίωση στην απόδοση, ενώ η ανάγκη για χ^2 μετασχηματισμό εξαλείφεται.	82
6.4	Αποτελέσματα των μεθόδων που προσαρμόζουν τις πιθανότητες στην έξοδο. Η υπόθεση στατιστικής ανεξαρτησίας των δύο καναλιών, οπτικής και γλωσσικής πληροφορίας ενσωματώνει σημασιολογία υψηλότερου επιπέδου η οποία συνεισφέρει σημαντικά στην αποτελεσματική ταξινόμηση.	84
6.5	Συνολικά αποτελέσματα όλων των μεθόδων που δοκιμάσαμε. Βλέπουμε ότι το εύρος είναι 15%, μια αξιόλογη βελτίωση από τη χρήση μόνο οπτικής πληροφορίας. Επιπλέον, η απλή σώρρευση χαρακτηριστικών είναι αρκετά ταχύτερη του υπολογισμού χ^2 πυρήνα.	85
6.6	Σύγκριση αποτελεσμάτων της παγκόσμιας βιβλιογραφίας για το σύνολο δεδομένων MPII Cooking Activities. Επισημειώνονται με έντονα γράμματα η καλύτερη επίδοση [35] και η δική μας καλύτερη επίδοση, η οποία καταλαμβάνει τη δεύτερη θέση.	85
A.1	Οι 65 δράσεις του συνόλου δεδομένων MPII Cooking Activities Dataset. Στα πειράματά μας, οι δράσεις put on cutting-board, read και smell ενσωματώνονται στην κατηγορία background activity και η δράση wash hands συνενώνεται με τη δράση wash hands.	106
A.2	Τα 44 βίντεο του συνόλου δεδομένων MPII Cooking Activities Dataset. Από αυτά, χρησιμοποιούμε για εκπαίδευση όσα αφορούν τους δράστες 11 ως 15 και τα s08-d02 και s10-d10.	106
A.3	Οι 92 κατηγορίες αντικειμένων που εμφανίζονται στα βίντεο του συνόλου MPII Cooking Activities Dataset και που χρησιμοποιούμε στα πειράματά μας.	107
A.4	Παράδειγμα υποτίτλων για το βίντεο s08-d04. Όλοι οι χρόνοι αναφέρονται σε δευτερόλεπτα από την αρχή του βίντεο.	108

Κεφάλαιο 1

Εισαγωγή

1.1 Η Κατάτμηση Δράσεων Και Οι Επιστήμες Πίσω Από Αυτή

Η ανάπτυξη του διαδικτύου έχει ανοίξει νέους ορίζοντες στη δυνατότητα απόκτησης γνώσης: Online how-to tutorials, εκπαιδευτικά βίντεο και εκατομμύρια χρήστες που μοιράζονται τις γνώσεις και τις εμπειρίες τους σε έναν καθημερινά αυξανόμενο ιστό. Αν το διαδίκτυο μπορεί να έχεις τέτοιες επιπτώσεις στην ανθρώπινη γνώση, το να συνεισφέρει στην τεχνητή νοημοσύνη είναι πολύ μεγαλύτερη πρόκληση. Οι μηχανές «αντιλαμβάνονται» με μαθηματικά καθορισμένους αλγορίθμους και κριτήρια, αντίθετα με την αφηρημένη σύλληψη και συσχέτιση εννοιών και αισθήσεων των ανθρώπων. Με λίγα λόγια, αν δούμε τις μηχανές σαν έμβια όντα, απαιτείται ένας εκπαιδευτής (ο σχεδιαστής και προγραμματιστής) ώστε η μηχανή να «μάθει» τελικά ακριβώς αυτό το οποίο της διδάσκεται. Η εκμάθηση από το διαδίκτυο μπορεί να διευκολύνει αυτή τη διαδικασία. Με online δεδομένα, μια μηχανή μπορεί αυτόματα να συλλέξει γνώσεις και να «εκπαιδευτεί». Τελικά η σχεδίαση εστιάζει στη γενική μεθοδολογία απόδοσης της δυνατότητας εκμάθησης παρά στην καθαυτή διδασκαλία κάθε καθήκοντος ξεχωριστά. Η επίτευξη αυτού του στόχου είναι ένα βήμα προς την ενσωμάτωση των ευφυών συστημάτων (π.χ. ρομπότ) στην καθημερινή ζωή με στόχο την απλοποίησή της.

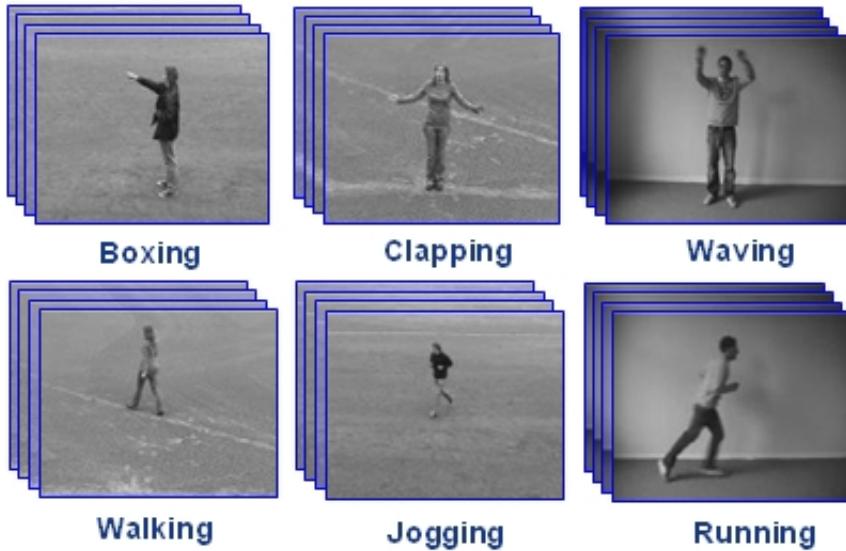
Ένα βήμα προς την παραπάνω κατεύθυνση είναι η εστίαση στην απόκτηση γνώσης μέσω βίντεο. Τα βίντεο περιέχουν συμπαγή οπτικοακουστική πληροφορία και μπορούν να χρησιμεύσουν για εξαγωγή αλγορίθμων, εκμάθηση αντικειμένων αλλά και χειρισμού αυτών για την ολοκλήρωση της εργασίας. Πώς ωστόσο μπορούμε να εκμεταλλευτούμε με αυτόματο τρόπο όλη αυτή την πληροφορία; Πρώτα επιχειρούμε να μοντελοποιήσουμε τη διαδικασία με την οποία ένας άνθρωπος μπορεί να εξάγει πληροφορία από βίντεο. Αρχικά, βλέπει μια αλληλουχία κινήσεων και αντιλαμβάνεται ότι διαφορετικές ενέργειες εκτελούνται με την αίσθηση της όρασης. Με την ίδια αίσθηση αντιλαμβάνεται, ταυτοποιεί και συσχετίζει αντικείμενα με την κάθε ενέργεια. Ωστόσο η τελική του απόφαση για το τι παρακολουθεί και τελικά η κατανόηση του βίντεο δεν στηρίζονται μόνο στην αίσθηση αλλά και στη σημασιολογία: η ταυτοποίηση των αντικειμένων ενισχύεται από την πρότερη γνώση για τον τρόπο και λόγο χρήσης ενός αντικειμένου. Έτσι σε ένα βίντεο στο οποίο ένα υποκείμενο κόβει ένα χαρτόνι με ψαλίδι, η a priori γνώση της χρήσης του ψαλιδιού για κοπή μπορεί να σβήσει τις αμφιβολίες για τη φύση του αντικειμένου που μπορεί να προκύψουν

από δυσκολίες στην οπτική αναγνώριση στο βίντεο, όπως πιθανές επικαλύψεις με άλλα αντικείμενα ή τα χέρια, θόρυβο λόγω κακής ανάλυσης κ.α. Τέλος, η ακουστική πληροφορία είναι επεξηγηματική για αυτά τα οποία προβάλλονται, προβλήθηκαν ή θα προβληθούν στο βίντεο. Στο παραπάνω παράδειγμα, η ακουστική πληροφορία «Τώρα κόβουμε το χαρτόνι με ένα ψαλίδι» μπορεί να ενισχύσει την αντίληψη του παρατηρητή ώστε να αντιληφθεί το ψαλίδι παρά τις προαναφερθείσες δυσκολίες από την απευθείας αντίληψή του με όραση.

Επομένως, βήμα-κλειδί στην εκμάθηση μέσω βίντεο για τις μηχανές είναι η δυνατότητα κατάτμησης των ενεργειών του βίντεο και η εξαγωγή επιπλέον χρήσιμων πληροφοριών για τα εργαλεία και το χειρισμό τους. Με άλλα λόγια, χρειάζεται μια αναπαράσταση των δράσεων με τρόπο τέτοιο ώστε μια μηχανή να είναι σε θέση να τις ταυτοποιήσει αλλά και να τις διακρίνει, εντοπίζοντάς τες μέσα σε μια χρονική ακολουθία δράσεων. Τι είναι όμως μια δράση; Αναφερθήκαμε στον όρο με γενικότητα και φυσικότητα, ωστόσο είναι δύσκολο να ορίσουμε τη «δράση», παρότι το περιεχόμενο του όρου είναι πολύ εύκολα αισθητό από τους ανθρώπους. Θα μπορούσαμε να πούμε ότι δράση είναι «το γεγονός ή η διαδικασία του να κάνεις κάτι, τυπικά με κάποιο σκοπό» (από το λεξικό της Google). Ο ορισμός πάσχει από μικρή αυτοαναφορά αλλά και περιλαμβάνει το σκοπό. Πολλές φορές σε μια δράση δεν υπάρχει τέτοιο κίνητρο. Το [56] ορίζει τη δράση ως «την ενέργεια που εκτελείται από συνειδητή θέληση και μπορεί να χαρακτηριστεί από φυσική ή πνευματική δραστηριότητα». Μια άλλη προσέγγιση είναι αυτή της ακολουθίας πρωταρχικών ενεργειών, οι οποίες αποτελούν μικρές κινήσεις που δεν μπορούν αυτόνομα να χαρακτηριστούν ως δράση. Ο ορισμός αυτός είναι βέβαια ασαφής ως προς το όριο διάκρισης αυτό, συν ότι γίνεται αυτοαναφορικός, αφού ορίζει τη δράση μέσω των πρωταρχικών ενεργειών, οι οποίες με τη σειρά τους ορίζονται σε συσχέτιση με τη δράση. Τελικά, δεν έχει νόημα ένας καθολικός ορισμός για τη δράση, τουλάχιστον σε αυτή την εργασία, αλλά θα κρατήσουμε τη διαίσθηση για τον όρο, οπότε θα αναμένουμε ότι και η εκπαίδευση των μηχανών για αναγνώριση της δράσης όπως προκύπτει μέσα από την ανθρώπινη αντίληψη θα είναι επιτυχής, αν η διαδικασία εκπαίδευσης περιλαμβάνει δείγματα που συμβαδίζουν με την αντίληψη αυτή.

Γύρω από τις δράσεις έχουν μελετηθεί τα εξής τρία προβλήματα που μας ενδιαφέρουν: Το πρώτο είναι η αναγνώριση δράσεων. Ο σκοπός εδώ είναι η κατηγοριοποίηση ενός τμήματος βίντεο σε κάποια δράση, έχοντας πάντα προηγηθεί εκπαίδευση της μηχανής. Το δεύτερο πρόβλημα είναι η ανίχνευση δράσης, με στόχο τον εντοπισμό, σε ένα βίντεο που περιέχει πολλές δράσεις, εκείνων των τμημάτων στα οποία συμβαίνει η δράση ενδιαφέροντος. Τέλος, η κατάτμηση δράσεων αφορά το σπάσιμο ενός βίντεο σε τμήματα στα οποία εκτυλίσσεται ακριβώς μία δράση. Στα ανωτέρω θίξαμε κυρίως το ζήτημα της αυτόματης εξαγωγής γνώσης μέσα από βίντεο ως σημαντική εφαρμογή της κατάτμησης δράσεων σε βίντεο, δεδομένου ότι σχετίζεται όμεσα με το θέμα της παρούσας εργασίας. Εν τούτοις είναι προφανές ότι η αναγνώριση, η ανίχνευση και η κατάτμηση δράσεων έχουν πολλαπλές εφαρμογές και σε άλλα πεδία. Ταυτόχρονα, είναι αρκετές και οι προκλήσεις των εργασιών αυτών. Θα αφήσουμε όμως την ανάλυση των προκλήσεων και των εφαρμογών για τα κεφάλαια που αφορούν πιο συγκεκριμένα την αναγνώριση και την κατάτμηση δράσεων.

Ως αντικείμενα, η αναγνώριση και κατάτμηση δράσεων εμπίπτουν στο χώρο πολλών επιστημονικών πεδίων. Πρώτιστο είναι το πεδίο της Όρασης Υπολογιστών, αλλά και της Μηχανικής Μάθησης, της Επεξεργασίας Σημάτων και της Επεξεργασίας Φυσικής Γλώσσας.



Σχήμα 1.1: Ενδεικτικές δράσεις από τη βάση KTH [128]. Ένα πρόβλημα αναγνώρισης δράσεων θα αφορούσε το διαχωρισμό των παραπάνω δράσεων. Δηλαδή, δοθέντος ενός τμήματος βίντεο, ζητάμε τη δράση η οποία πραγματοποιείται σε αυτό, επιλέγοντας από τον κατάλογο δράσεων.

Η Όραση Υπολογιστών είναι ο τομέας της επιστήμης και της τεχνολογίας που μελετά μεθόδους για την ανάλυση και κατανόηση μιας εικόνας ή ακολουθίας εικόνων με στόχο την εξαγωγή συμβολικής πληροφορίας. Σκοπός είναι η εξαγωγή συμπερασμάτων από τις εικόνες και τις ακολουθίες, σε ένα εύρος που εκτείνεται από «χαμηλού επιπέδου» πληροφορία (γωνίες, ακμές, υφή) μέχρι «μεσαίου» (χρώμα, υφή, σχήμα, συνεκτικότητα) και «ψηλού επιπέδου» πληροφορία (ταυτοποίηση αντικειμένων, εντοπισμός αντικειμένων, αναγνώριση δράσης, κίνηση). Το πεδίο έχει λιγότερα από 60 χρόνια ζωής και αναπτύσσεται συνεχώς, καθώς βρίσκει εφαρμογή σε πολλές πτυχές της καθημερινής ζωής, όπως στη ρομποτική, βιοϊατρική τεχνολογία και ιατρική απεικόνιση, την επικοινωνία ανθρώπου-υπολογιστή, την τηλεπισκόπηση, τα ευφυή συστήματα, τον κινηματογράφο και την τεχνολογία βίντεο.

Η Μηχανική Μάθηση είναι το σύνολο των μεθόδων και τεχνικών που σύμφωνα με τον Arthur Samuel το 1959, «αποδίδει στους υπολογιστές τη δυνατότητα να μαθαίνουν χωρίς να προγραμματίζονται σαφώς». Το πεδίο αποτελεί εξέλιξη της Αναγνώρισης Προτύπων και της Τεχνητής Νοημοσύνης και μελετά την κατασκευή αλγορίθμων για συστήματα που μπορούν να εκπαιδεύονται και να κάνουν προβλέψεις για δεδομένα που δεν έχουν ξανασυναντήσει. Η Μηχανική Μάθηση δίνει λύσεις σε περιπτώσεις όπου ο ρητός προγραμματισμός είναι πρακτικά αδύνατος. Για το λόγο αυτό συναντά ευρεία εφαρμογή σε ιστοσελίδες, ηλεκτρονικό ταχυδρομείο, στην ανάλυση δεδομένων, στη λήψη αποφάσεων και στην υπολογιστική όραση και τη ρομποτική.

Η Επεξεργασία Σημάτων είναι πιο παλιός και γενικός τομέας από την Όραση Υπολογιστών και την Αναγνώριση Προτύπων αφορά την ανάλυση, σύνθεση και τροποποίηση σημάτων, τα οποία ορίζονται ευρέως σαν συναρτήσεις που περιέχουν πληροφορία για τις ιδιότητες ή τη συμπεριφορά ενός φαινομένου. Οι απαρχές του τομέα αυτού φαίνεται να ξεκινάνε τον 17ο αιώνα, αλλά η ψηφιακή επεξεργασία ξεκινάει τη δεκαετία του 1940. Ο γενικός ορισμός των σημάτων δίνει τεράστιο εύρος εφαρμογών, από ακουστικά και ηλεκτρικά σήματα μέχρι ιατρικά σήματα και εικόνες. Για παράδειγμα, η Επεξεργασία Σημάτων χρησιμοποιείται για τη βελτίωση μετάδοσης σημάτων, την

επάρκεια αποθηκευτικού χώρου στους υπολογιστές, την αποθορυβοποίηση και την ανίχνευση σημάτων ενδιαφέροντος.

Τέλος, η Επεξεργασία Φυσικής Γλώσσας εστιάζει στον λόγο, παρά στις ιδιότητες του σήματος, ώστε να δώσει τη δυνατότητα σε μηχανές να επεξεργάζονται την ανθρώπινη γλώσσα. Σημαντικά προβλήματα της Επεξεργασίας Φυσικής Γλώσσας είναι η κατανόηση φυσικής γλώσσας, η σύνθεση λόγου, η συντακτική ανάλυση και συμπερασματολογία, τα διαλογικά συστήματα και η σύνδεση γλώσσας με την αντίληψη των μηχανών. Στη σημερινή εποχή χρησιμοποιείται κατά κόρον σε συστήματα αυτόματης μετάφρασης, ανακατασκευής κειμένου, συντακτικής σημασιολογίας και σε ευφυή μοντέλα αλληλεπίδρασης ανθρώπου-μηχανής, όπως οι αυτόματοι ηλεκτρονικοί βοηθοί.

1.2 Διάρθρωση Διπλωματικής Εργασίας

Στην εργασία αυτή προτείνεται μια μέθοδος σχεδίασης αλλά γίνεται και η υλοποίηση ενός ενιαίου συστήματος αναγνώρισης και κατάτμησης δράσεων σε βίντεο, το οποίο λαμβάνει υπόψιν την οπτική αντίληψη και πληροφορία από κείμενο και υπότιτλους για αποφασίσει, με ενδείξεις αλλά και χρήση σημασιολογίας, για την κατηγοριοποίηση κάθε τμήματος βίντεο και το βέλτιστο σπάσιμο του βίντεο σε συμπαγή τμήματα. Η σχεδίαση είναι αρχικά γενική και προχωράει σε βαθύτερα στάδια αφαιρετικά, ενώ στη συνέχεια παρουσιάζεται η υλοποίηση που επιλέχθηκε. Τελικά η εργασία μας συνεισφέρει σε πολλαπλά επίπεδα: σε επίπεδο σχεδίασης, προτείνουμε μια μέθοδο που συνδυάζει πολλαπλά κανάλια πληροφορίας και διατηρεί την αυτονομία κάθε υποσυστήματος για να επιτρέπει την ενημέρωσή του σύμφωνα με τις τελευταίες state-of-the-art εξελίξεις. Σε επίπεδο υλοποίησης, η μέθοδος μας εφαρμόζεται σε ένα απαιτητικό σύνολο δεδομένων με δράσεις λεπτομέρειας και επιδεικνύει απόδοση που συγκρίνεται με τις τρέχουσες state-of-the-art επιδόσεις για αυτό το σύνολο. Τέλος, προτείνουμε και υλοποιούμε έναν νέο αλγόριθμο κατάτμησης βίντεο που στηρίζεται στις μεταβολές των παρατηρούμενων αντικειμένων.

Η εργασία ακολουθεί την εξής πορεία:

- Στο κεφάλαιο 2 παρουσιάζεται η σχεδίαση του συνολικού συστήματος και των υπομονάδων του, καθώς και το σύνολο δεδομένων στο οποίο εργαζόμαστε. Γίνεται συσχέτιση με άλλες μεθόδους οι οποίες έχουν σχέση με την εργασία μας, ενώ αφήνουμε τις λεπτομέρειες κάθε υποσυστήματος για να παρουσιαστούν ξεχωριστά.
- Στο κεφάλαιο 3 παρουσιάζεται το ζήτημα της αναγνώρισης δράσεων με αξιοποίηση μόνο οπτικής πληροφορίας και εξηγείται η λειτουργία του υποσυστήματος Όρασης Χαμηλού Επιπέδου. Αναλύεται η μέθοδος των Πυκνών Τροχιών [248] και μια επέκτασή της που λαμβάνει υπόψιν της επιπλέον χαρακτηριστικά πόζας [191]. Τέλος, περιγράφουμε τη σύνδεση αυτών των εργασιών στη δική μας σχεδιαστική επιλογή.
- Στο κεφάλαιο 4 παρουσιάζεται το ζήτημα της ανίχνευσης αντικειμένων και η σύνδεσή του με το σημασιολογικό υποσύστημα. Στη σημασιολογία εντάσσουμε και το υποσύστημα ακουστικής πληροφορίας, οπότε δεν αφιερώνουμε για αυτό

ξεχωριστό κεφάλαιο. Γίνεται μια ιστορική επισκόπηση της ανίχνευσης αντικειμένων και προσκηνίου, παρουσιάζεται η μέθοδος των Μοντέλων Παραμορφώσιμων Τμημάτων [69], των Μοντέλων Μίξης Γκαουσιανών για εξαγωγή προσκηνίου και η Μέθοδος Ταιριάσματος στην ανίχνευση αντικειμένου οπτικά και γλωσσικά. Δικαιολογούμε τις σχεδιαστικές μας επιλογές και παρουσιάζουμε αποτελέσματα ανίχνευσης συνδυάζοντας γλωσσική και οπτική πληροφορία.

- Στο κεφάλαιο 5 παρουσιάζεται η διαδικασία εξαγωγής πληροφορίας για τον τύπο λαβής (grasping type) στα frames του βίντεο ως επιπλέον αναγνωριστικό χαρακτηριστικό. Παρουσιάζονται θέματα από την περιοχή των Συνελικτικών Νευρωνικών Δικτύων και η από αρχής υλοποίηση ενός συστήματος ανίχνευσης χεριών.
- Στο κεφάλαιο 6 γίνεται ανάλυση των αλγορίθμων και των αποτελεσμάτων ταξινόμησης. Εδώ συγκρίνουμε όλες τις μεθόδους και τα αποτελέσματά μας με άλλες εργασίες.
- Στο κεφάλαιο 7 αναπτύσσουμε μια νέα μέθοδο κατάτμησης δράσεων και δείχνουμε τα πειραματικά αποτελέσματα για αυτή πάνω σε βίντεο του συνόλου δεδομένων μας.
- Στο κεφάλαιο 8, κλείνουμε αυτή την εργασία παρουσιάζοντας τις συνεισφορές μας και προτείνοντας μελλοντικές ερευνητικές κατευθύνσεις.

1.3 Πακέτα Λογισμικού που Χρησιμοποιήθηκαν

Ο πηγαίος κώδικας για την παρούσα εργασία είναι κατά βάση δικό μας έργο και γράφτηκε ως επί το πλείστον σε MATLAB, ενώ κάποια τμήματα έχουν γραφεί σε Python. Ωστόσο, στην έκταση της εργασίας μας χρησιμοποιούνται και επιπλέον πακέτα λογισμικού από άλλες πηγές τα οποία αναφέρουμε εδώ:

- Scikit-Learn [171]: Το πακέτο αυτό λογισμικού είναι μια ανοιχτού κώδικα βιβλιοθήκη υλοποίησης αρκετών μεθόδων (ταξινόμησης ή επεξεργασίας δεδομένων) που σχετίζονται με τη Μηχανική Μάθηση για τη γλώσσα Python. Όλοι οι ταξινομητές επιβλεπόμενης μάθησης που χρησιμοποιήθηκαν ή δοκιμάστηκαν, εκτός από τα Συνελικτικά Νευρωνικά Δίκτυα, βασίστηκαν στη βιβλιοθήκη Scikit-Learn. Χρήση της ίδιας βιβλιοθήκης γίνεται και στην εξαγωγή μετρικών για αποτελέσματα ταξινόμησης αλλά και για την κωδικοποίηση των χαρακτηριστικών με το σχήμα Tf-Idf.
- MatConvNet [241]: Το MatConvNet είναι μια εργαλειοθήκη Συνελικτικών Νευρωνικών Δικτύων για MATLAB. Όπου στην εργασία αυτή γίνεται χρήση τέτοιων δικτύων, γίνεται μέσω αυτής της εργαλειοθήκης.
- Voc-Release 5 [68]: Το πακέτο αυτό περιέχει τον πηγαίο κώδικα για την εργασία των [68] πάνω στα Μοντέλα Παραμορφώσιμων Τμημάτων. Χρησιμοποιούμε τέτοια μοντέλα για ανίχνευση ανθρώπων στη φάση εξαγωγής της περιοχής ενδιαφέροντος.

Κεφάλαιο 2

Το Συνολικό Σύστημα

Στο εισαγωγικό κεφάλαιο εστιάσαμε στη σημασία της αναγνώρισης και της κατάτμησης δράσεων στην εκμάθηση των μηχανών. Είναι φανερό ότι η οπτική αντίληψη είναι ουσιαστική για την εξέλιξη των ευφυών συστημάτων. Ταυτόχρονα, πολλαπλές εφαρμογές βασίζονται στην αναγνώριση δράσεων. Συστήματα επίβλεψης με κάμερες μπορούν αυτόματα να ανιχνεύουν και να προειδοποιούν για ασυνήθιστη ή επικίνδυνη δραστηριότητα. Τα έξυπνα οικιακά συστήματα χρησιμοποιούν αναγνώριση και ανίχνευση δράσεων για να προσαρμόσουν τη λειτουργία τους στην κίνηση και τη βούληση του χρήστη. Παράλληλα, συστήματα υποβοήθησης για ηλικιωμένους ή άτομα με ειδικές ανάγκες μπορούν να βρίσκονται σε εγρήγορση και εύκολα να προβούν σε εξυπηρέτηση. Η ανάλυση τύπων βίντεο μπορεί να γίνει αυτόματα, όπως για παράδειγμα ανάλυση φάσεων ενός αθλητικού αγώνα. Πέρα από τον τύπο, αυτόματα μπορεί να εξαχθεί και το περιεχόμενο, εξέλιξη που βοηθά την κατηγοριοποίηση βίντεο παγκόσμιου ιστού, με στόχο την αποτελεσματικότερη εύρεση σχετικών βίντεο αλλά και βίντεο απαγορευμένου περιεχομένου. Κλείνοντας, θέλουμε να αποσαφηνίσουμε ότι οι παραπάνω είναι μόνο ελάχιστες από τις δυνατές χρήσεις της τεχνολογίας αναγνώρισης δράσεων κι ότι υπάρχει πληθώρα εφαρμογών στους τομείς της ιατρικής, της κοινωνιολογίας, των τεχνικών προβλέψεων και στον κινηματογράφο.

Από την άλλη μεριά, πολλές είναι και οι προκλήσεις του αντικειμένου της αναγνώρισης δράσεων. Η βασική δυσκολία είναι η μεγάλη ενδομεταβλητότητα μεταξύ των βίντεο δράσεων ίδιας κλάσης. Με άλλα λόγια, η ποικιλομορφία των παραγόντων που συνδέονται με βίντεο ίδιας κλάσης είναι τεράστια. Ξεκινώντας από τις διαφορές στο περιβάλλον του βίντεο, το υπόβαθρο μπορεί να αλλάζει σημαντικά. Ταυτόχρονα, το παρασκήνιο μπορεί να είναι δυναμικό: υπάρχει κίνηση η οποία δεν πρέπει να ερμηνευθεί ως προσκήνιο και να επηρεάσει την ταξινόμηση. Η κίνηση μπορεί επιπλέον να οφείλεται σε μη σταθερή κάμερα. Ακόμα και στην περίπτωση σταθερής κάμερας όμως, η οπτική γωνία προσδίδει σημαντική διαφοροποίηση στην παρατηρούμενη δράση. Έπειτα, η καθαυτή μεταβλητότητα έγκειται στην εκτέλεση της δράσης. Κάθε άνθρωπος εκτελεί την ίδια δράση διαφορετικά. Ακόμα και το ίδιο άτομα δρα διαφορετικά υπό διαφορετικές συνθήκες. Η κίνηση δεν είναι ο μόνος παράγοντας που μπορεί να αλλάζει ωστόσο. Άλλαγές παρατηρούνται και στη διάρκεια της εκτέλεσης. Τέλος, οι επικαλύψεις, οι διαφορετικές ενδυμασίες και οι μη συμπαγείς κινήσεις είναι παράγοντες που αυξάνουν επιπλέον τη διαφοροποίηση αυτή των εκτελέσεων. Παράγοντες όμως που δυσχεραίνουν την επιτυχή αναγνώριση είναι και η έλλειψη επαρκών δεδομένων εκπαίδευσης, όπως και η σύγχυση της γλωσσικής ετικέτας, όταν

Why video analysis?

Applications:



First appearance of N. Sarkozy on TV



Sociology research:
Influence of character smoking in movies



Education: How do I make a pizza?



Where is my cat?



Predicting crowd behavior
Counting people



Motion capture and animation

Σχήμα 2.1: Κάποιες από τις εφαρμογές της ανάλυσης βίντεο στην ανθρώπινη ζωή. Είναι φανερό ότι η εξέλιξη της τεχνολογίας επιβάλλει ακόμα ποιοτικότερους αλγορίθμους και ταυτόχρονα προκύπτουν νέα πεδία εφαρμογής. Εικόνα από [48]



Σχήμα 2.2: Η εμφάνιση δύο δράσεων ίδιου περιεχομένου μπορεί να αλλάζει σημαντικά μεταξύ διαφορετικών δειγμάτων. Αλλαγές στην οπτική της κάμερας, στο φωτισμό, στην εμφάνιση των δραστών, χρονικές μεταβολές στη διάρκεια κάθε δράσης και η ποικιλία στην ανθρώπινη μορφολογία και κίνηση είναι παράγοντες που καθιστούν το πεδίο αναγνώρισης δράσεων ιδιαίτερα απαιτητικό και προκλητικό. Εικόνα από [48]



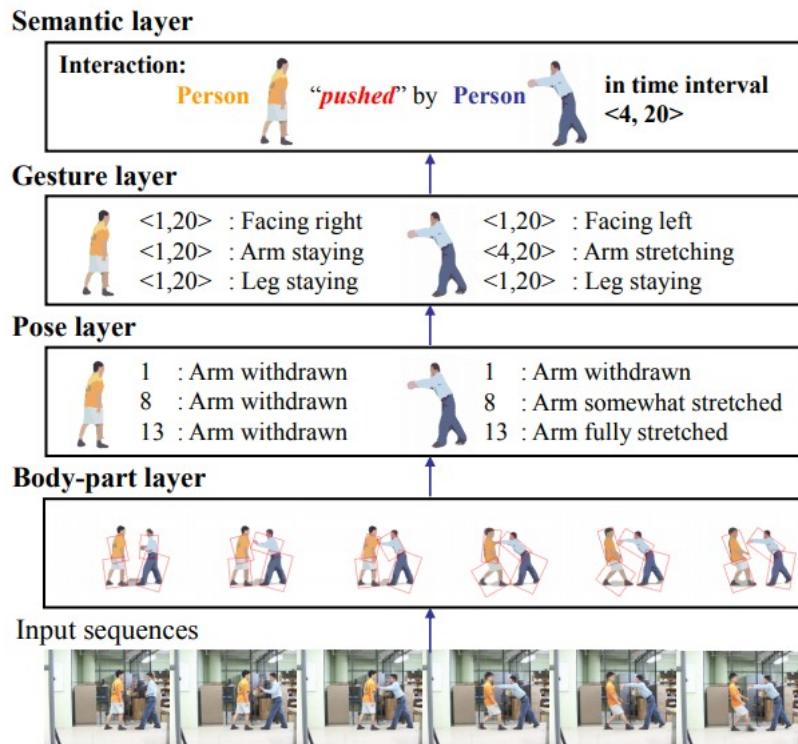
Σχήμα 2.3: Η αναγνώριση δράσεων μόνο οπτικά είναι συχνά δύσκολη σε περιπτώσεις όπου η μεταβλητή τα μεταξύ των κλάσεων είναι μικρή οπτικά. Χαρακτηριστικά όπως η στάση του σώματος δεν είναι εύρωστα στο μεταβλητό περιεχόμενο της δράσης. Για το λόγο αυτό, η ενσωμάτωση σημασιολογίας και πρότερης γνώσης μπορεί να συνεισφέρει στη διάκριση όπου η οπτική αντίληψη δεν επαρκεί. Εικόνα από [48]

η δράση προσδιορίζεται μόνο από το ρήμα (π.χ. ανοίγω την πόρτα αλλά και ανοίγω την τηλεόραση).

2.1 Η Αξία της Σημασιολογίας και Εναλλακτικές Μέθοδοι Αναγνώρισης Δράσεων

Οι προσεγγίσεις για το ζήτημα της αναγνώρισης είναι αρκετές και με διαφορετική σκοπιά η καθεμία. Η πρώτη προσέγγιση ήταν με χρήση γεωμετρικής μοντελοποίησης και χαμηλού επιπέδου χαρακτηριστικών όρασης υπολογιστών. Η προσέγγιση αυτή εισχωρεί βαθιά στο πρόβλημα της οπτικής αντίληψης και εξακολουθεί να αναπτύσσεται μέχρι σήμερα σε πολλές παραλλαγές της. Αναλύουμε περισσότερο αυτές τις προσεγγίσεις στο Κεφάλαιο 3. Μία άλλη επαφή με το πρόβλημα εστίασε στη σημασιολογία. Οι iεραρχικές προσεγγίσεις, όπως ονομάσθηκαν, αντιμετωπίζουν τη δράση ως μια ακολουθία υπογεγονότων. Επιπλέον, αξιοποιούν υψηλού επιπέδου σημασιολογική πληροφορία, όπως πληροφορίες που σχετίζονται με τον άνθρωπο και τα αντικείμενα που συμμετέχουν στη δράση, για να συνθέσουν μια συντακτική αναπάρασταση των δράσεων. Κίνητρο είναι ότι το περιβάλλον μιας δράσης μπορεί να επηρεάσει την κατηγοριοποίησή της (εικόνα 2.3). Θα κάνουμε εδώ μια μικρή ιστορική αναφορά σε αυτές τις μεθόδους.

Η πρώτη κατηγορία iεραρχικών προσεγγίσεων είναι η συντακτική και βασίζεται σε χρήση γραμματικών για την εξαγωγή υψηλού επιπέδου πληροφορίας γύρω από τη δράση. Η χρήση γραμματικών στην Όραση Υπολογιστών είχε ήδη ξεκινήσει από αρκετά νωρίς [239] για τη γεωμετρική αναπαράσταση προτύπων με συντακτικό τρόπο. Η οπτική πληροφορία απεικονίζεται σαν συμβολοσειρά η οποία αναλύεται συντακτικά από έναν στοχαστικό αναλυτή. Στην αναγνώριση δράσεων η εμφάνιση πιθανοτικών γραμματικών και συντακτικών αναλύσεων εμφανίστηκε στα τέλη της δεκαετίας του '90 [18], [161]. Η μοντελοποίηση του θορύβου αποτέλεσε κύριο πρόβλημα αυτών των μεθόδων, αλλά μοντελοποιήθηκε σταδιακά. Με χρήση θεωρημάτων από τη Θεωρία Πληροφορίας, οι [119] κατορθώνουν να πετύχουν ευρωστία σε μορφές



Σχήμα 2.4: Ένα περιγραφικό σύστημα αναγνώρισης δράσεων μελετά τα βίντεο εισόδου σε πολλαπλά επίπεδα και συνδυάζει την πληροφορία για να εξάγει αποτέλεσμα. Εκτός από την κίνηση, τα συστήματα αυτά συγκεντρώνουν και σημασιολογικές περιγραφές της δράσης. Ταυτόχρονα, η οπτική ανάλυση δε γίνεται με παραδοσιακές μεθόδους, αλλά με περιγραφική μοντελοποίηση σε διάφορες κλίμακες. Η προσέγγιση αυτή εξελίχθηκε στο συνδυασμό χαρακτηριστικών χαμηλού επιπέδου οπτικής πληροφορίας με σημασιολογία χωρίς να μεσολαβεί η μοντελοποίηση με αυστηρούς κανόνες. Εικόνα από [202]

Θορύβου όπως η εισαγωγή, η διαγραφή και η αντικατάσταση χαρακτήρων. Ωστόσο απαιτείται ευρύ σύνολο δεδομένων εκπαίδευσης και το υπολογιστικό κόστος αυξάνει. Οι προσεγγίσεις αυτές δίνουν ποιοτικά αποτελέσματα σε περιπτώσεις όπου οι δράσεις εμφανίζουν βαθιά ιεραρχική δομή ή κυκλικές επαναλήψεις (π.χ. κοπή ενός φρούτου). Σε περιπτώσεις μεγάλης αβεβαιότητας οι μέθοδοι αυτές δεν είναι λειτουργικές.

Μια άλλη προσέγγιση είναι η στατιστική και κάνει χρήση μαρκοβιανών μοντέλων και γραμματικών χωρίς συμφραζόμενα αλλά με διαφορετική φιλοσοφία. Οι μέθοδοι αυτές βασίζονται στην υπόθεση ισχυρής μαρκοβιανής ιδιότητας και στην πρότερη γνώση των δυναμικών των δράσεων [167]. Καθώς οι δράσεις μπορούν να συμβαίνουν ταυτόχρονα, πρέπει να παρακολουθούμε ανεξάρτητες ακολουθίες παράλληλα. Τη λύση σε αυτό επιχειρούν οι [219] με χρήση δρομολογητικών δικτύων (propagation networks). Οι στατιστικές προσεγγίσεις εφαρμόζονται με επιτυχία συμπληρωματικά των συντακτικών, σε περιπτώσεις μεγάλης αβεβαιότητας και θορύβου, όταν ισχύουν οι υποθέσεις των σειριακών μοντέλων (μαρκοβιανή ιδιότητα) και μπορούν να ενσωματωθούν οι δυναμικές των δράσεων. Αντίθετα, δεν αποδίδουν καλά σε περιπτώσεις σύνθετων δράσεων με βαθιά ιεραρχική δομή που δύσκολα αναπαρίσταται με δενδρικές δομές.

Πιο κοντά στην έννοια της σημασιολογίας όπως τη χρησιμοποιούμε σε αυτή την εργασία είναι οι μέθοδοι βασιζόμενες σε περιγραφή, οι οποίες άρχισαν να αναπτύσσονται στις αρχές του 21ου αιώνα. Η ιδέα είναι η αναπαράσταση των δράσεων σε σημασιολογικούς χώρους και η αναγνώρισή τους μέσω σημασιολογικού ταιριάσματος. Για παράδειγμα η ανάλυση μπορεί να γίνεται σε πολλαπλά επίπεδα (παρακολούθηση κίνησης και στάσης διαφορετικών μερών του σώματος, αλληλεπίδραση με αντικείμενα και ανθρώπους κλπ) και οι πληροφορίες να συνθέτουν την αναπαράσταση της δράσης [202]. Οι [165] προχωρούν σε υλοποίηση γλωσσών αναπαράστασης. Οι [238] ισχυροποιούν τις αναπαραστάσεις προσθέτοντας στοχαστικό χαρακτήρα μέσω των δικτύων μαρκοβιανής λογικής που χρησιμοποιούν. Η γενίκευση σε δραστηριότητες που αφορούν ένα σύνολο ατόμων γίνεται στο [203]. Πιο σύνθετες προσεγγίσεις και αναπαραστάσεις συνεχίζουν να βγαίνουν τα τελευταία χρόνια [204], [104], [257], [174], [256], [254].

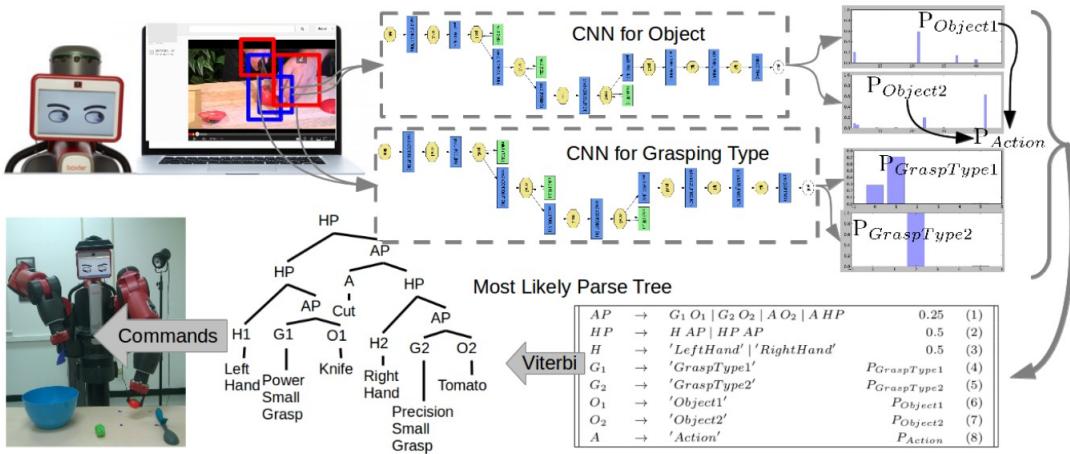
2.2 Σχετικές Εργασίες

Εδώ αναφέρουμε εργασίες που μας ενέπνευσαν ή αξιοποιήθηκαν κατά τη φάση σχεδίασης και υλοποίησης αυτής της εργασίας.

Ξεκινάμε από το [259], όπου παρουσιάζεται ένα ολοκληρωμένο σύστημα αξιοποίησης σημασιολογικής πληροφορίας για αναγνώριση δράσεων. Αφού γίνει ανίχνευση χεριών και αντικειμένων, χρησιμοποιούνται δύο συνελικτικά νευρωνικά δίκτυα, ένα για αναγνώριση τύπου λαβής και ένα για αναγνώριση αντικειμένων. Με τη βοήθεια λεξικού υπολογίζεται μια κατανομή πιθανοτήτων για τις δράσεις ως προς τα αντικείμενα. Με αυτά τα δεδομένα, σχηματίζεται η οπτική πρόταση, μια επτάδα της μορφής [Αριστερό Χέρι, Λαβή Α.Χ., Αντικείμενο Α.Χ., Δράση, Αντικείμενο Δ.Χ., Λαβή Δ.Χ., Δεξί Χέρι], με τα Α.Χ. και Δ.Χ. να δηλώνουν το αριστερό και το δεξί χέρι αντίστοιχα. Η επτάδα εξάγεται από τα υπολογισμένα χαρακτηριστικά πιθανοτικά και μια γραμματική χωρίς συμφραζόμενα αναλαμβάνει να βρει την πιο πιθανή οπτική λέξη και να αναγνωρίσει τελικά τη δράση. Η επιτυχία της εργασίας αυτής ήταν ένα κίνητρο για τη χρήση σημασιολογίας και τύπων λαβής.

Αρκετές εργασίες προχώρησαν σε πολυτροπικές μεθόδους αναγνώρισης δράσεων. Από αυτές, το [152] αξιοποιεί εικόνα, ήχο και κείμενο και ευθυγραμμίζει τους διαφορετικούς τρόπους με κρυφά μαρκοβιανά μοντέλα. Το [3] χρησιμοποιεί τεχνικές μη επιβλεπόμενης μάθησης συνδυάζοντας υπότιτλους και εικόνα. Από μια ομάδα βίντεο παρόμοιου περιεχομένου εξάγει με συντακτική ανάλυση των υποτίτλων τα απαραίτητα βήματα κάθε αλγορίθμου. Η ευθυγράμμιση με το βίντεο γίνεται με αναζήτηση σε ένα παράθυρο γύρω από το χρονικό σημείο τοποθέτησης του βήματος. Τα βήματα έχουν επιλεχθεί λύνοντας ένα πρόβλημα βελτιστοποίησης στον χώρο όλων των βίντεο εκπαίδευσης. Τέλος, το [211] επίσης χρησιμοποιεί μη επιβλεπόμενη μάθηση και πάλι προηγείται η γλωσσική ομαδοποίηση των βίντεο τόσο για εύρεση ομοιότητας και ευθυγράμμιση μεταξύ τους όσο και για αποκοπή outliers. Η τελευταία εργασία μάλιστα εξάγει εικόνες αντικειμένων συσχετίζοντάς τες με τη γλωσσική πληροφορία, προσβλέποντας μελλοντικά σε μια αυτοματοποιημένη εξαγωγή συνόλου δεδομένων από τον παγκόσμιο ιστό.

Στο [105] παρουσιάζονται πειράματα με ανιχνευτές αντικειμένων βασισμένους σε βαθιά νευρωνικά δίκτυα σε κάθε frame του βίντεο και διαπιστώνεται ότι η γνώση για τα αντικείμενα μπορεί να συνεισφέρει κατά μεγάλο περιθώριο στο κέρδος. Προς αυτή



Σχήμα 2.5: Το σύστημα του [259]. Δύο νευρωνικά δίκτυα για τύπο λαβής και αντικείμενα δίνουν μια κατανομή πιθανοτήτων η οποία κωδικοποιείται με μια γραμματική χωρίς συμφραζόμενα. Παράλληλα, το μοντέλο γραμματικής ενισχύεται από πληροφορία κειμένου. Το αποτέλεσμα είναι ένα συντακτικό δέντρο που δίνει τον τύπο δράσης, τα συσχετιζόμενα αντικείμενα και τον τύπο λαβής τους. Εικόνα από [259]

την κατεύθυνση το [192] χρησιμοποιεί επισημειώσεις αντικειμένων που εξάγονται από δεδομένα κειμένου (ανθρώπινες περιγραφές για τα βίντεο) ώστε να εξάγει πληροφορία αντικειμένων και να τη χρησιμοποιήσει στην αναγνώριση. Καθώς οι δράσεις που μελετώνται εστιάζουν στα χέρια, η εργασία αυτή εξάγει τοπικές τροχιές γύρω από τα χέρια αντί για πυκνές τροχιές σε όλο την έκταση του frame. Τα αποτελέσματα δείχνουν ότι η περιοχή παρακολούθησης μπορεί να συρρικνωθεί σημαντικά αν υπάρχει διαθέσιμος ένας εύρωστος ανιχνευτής χεριών. Σε παρόμοιο κλίμα, το [125] αναζητά τη σημασιολογική θραύση των δράσεων, εστιάζοντας όμως περισσότερο στη δομή τους.

Στο συνδυασμό των πιθανοτήτων διαδοχικών συστημάτων μια σχετική εργασία είναι η [142] η οποία εξάγει οπτικά χαρακτηριστικά για τα αντικείμενα και τις αλληλεπιδράσεις τους σε μια εικόνα και προσδιορίζει σημασιολογικά τη δράση, παρέχοντας μια περιγραφή των εικόνων. Οι σημασιολογικές πληροφορίες ενσωματώνονται ως πρότερη γνώση στο σύστημα, επαναπροσδιορίζοντας τις αρχικές πιθανότητες από την οπτική αντίληψη. Μια διαφορετική αντίληψη της ίδιας ιδέας παρουσιάζεται στο [93], όπου ο επαναπροσδιορισμός πιθανοτήτων γίνεται με χρήση μετα-ταξινομητών που δρουν πάνω στα σκορ και τις πιθανότητες των αρχικών εκτιμητών.

Στην κατάτμηση των βίντεο, η πιο σχετική εργασία είναι η [92], η οποία θα αναλυθεί στο κεφάλαιο της κατάτμησης δράσεων.

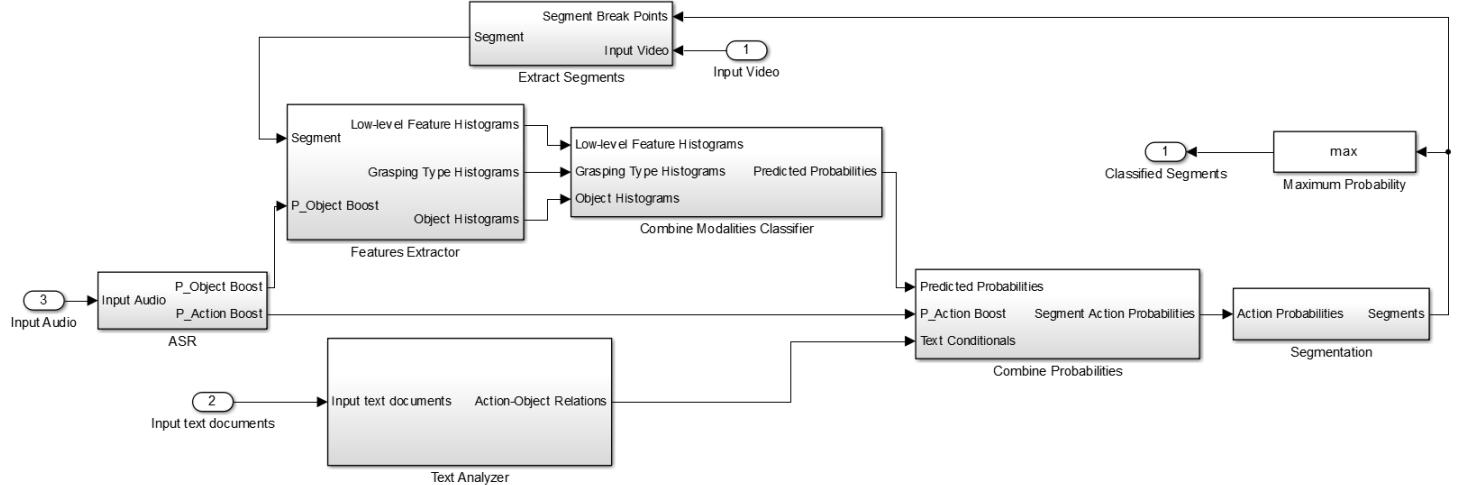
Σχετικές με την εργασία μας όχι από άποψη προσέγγισης απαραίτητα αλλά διότι χρησιμοποιούν το ίδιο σύνολο δεδομένων είναι και οι εργασίες [36], [268], [73], [34] και [35]. Το [36] συνδυάζει τη βελτιωμένη μέθοδο πυκνών τροχιών [247] με χαρακτηριστικά πόζας. Η καινοτομία του είναι ότι εισάγει μια νέα δομή συνελικτικού νευρωνικού δικτύου, ονομαζόμενη PCNN (Pose Convolutional Neural Network), η οποία μπορεί να εξάγει αυτόματα και με μεγάλη ακρίβεια τα χαρακτηριστικά πόζας ενσωματώνοντας και χρονική πληροφορία. Κοντινή με τη δική μας η εργασία η [268], όπου όμως τα αντικείμενα δεν ανιχνεύονται σύμφωνα με εκ των προτέρων γνωστά μοντέλα, αλλά προτείνονται θέσεις αντικειμένων με μέτρα objectness (ο όρος θα αναλυθεί σε

επόμενα κεφάλαια) οι οποίες παρακολουθούνται με χρήση πυκνών τροχιών στα επόμενα frames. Σύμφωνα με την κατανομή αυτών εκπαιδεύονται ταξινομητές. Το [73] εφαρμόζει μια πρωτοποριακή μέθοδο που προτείνει αυτόματα τη διάταξη των frames στο βίντεο. Οι συγγραφείς βρήκαν ότι μια τέτοια συνάρτηση είναι σε θέση να ταξινομήσει αποτελεσματικά τα τμήματα του βίντεο. Το [34] κάνει χρήση νευρωνικών δικτύων απευθείας στα βίντεο προτείνοντας μια βελτιωμένη αρχιτεκτονική που επιδρά στα επίπεδα pooling των δικτύων. Η ιδέα εξελίσσεται στο [35] το οποίο προτείνει μια νέα αρχιτεκτονική που ονομάζεται generalized rank pooling (GRP), η οποία δέχεται σαν είσοδο ενδιάμεσα χαρακτηριστικά από πρότερα στάδια ενός δικτύου που έχει εκπαιδευτεί με μικρές υπακολουθίες βίντεο και παράγει σαν έξοδο παραμέτρους ενός υποχώρου που παρέχει στα χαρακτηριστικά μικρή τάξη διατηρώντας τη χρονική τους σειρά. Ως τον Σεπτέμβριο του 2017, η μέθοδος αυτή είναι η state-of-the-art. Θα συγκρίνουμε τα αποτελέσματα των παραπάνω μεθόδων και της δικής μας σε επόμενο κεφάλαιο.

Τέλος, κοντινές αλλά όχι άμεσα σχετικές με τη δική μας είναι οι εργασίες που αξιοποιούν ημιεπιβλεπόμενη μάθηση. Στις ημιεπιβλεπόμενης μάθησης μεθόδους το ζήτημα έγκειται πρακτικά στην ευθυγράμμιση της εικόνας με το κείμενο. Για παράδειγμα στο [20] γίνεται κατάτμηση που υπακούει σε δεδομένη σειρά ενεργειών. Το πρόβλημα γενικεύεται στο [19] όπου ζητάμε να προσδιορίσουμε χρονικά το τμήμα βίντεο στο οποίο πραγματοποιείται μια ενέργεια που αναγράφεται στο κείμενο. Στο ίδιο κλίμα τα [95], [126] χρησιμοποιούν ασθενή επίβλεψη (weakly supervised). Τέλος το [221] αντιμετωπίζει το ζήτημα της ευθυγράμμισης με το κείμενο με μη επιβλεπόμενο τρόπο.

2.3 Σύνοψη Ολικού Συστήματος

Στην εργασία μας αυτή προτείνουμε μια νέα ενοποιημένη αρχιτεκτονική που συνδυάζει πολλαπλά κανάλια πληροφορίας για την επιτυχή ταυτόχρονη αναγνώριση και κατάτμηση δράσεων στο βίντεο. Ο αναγνώστης μπορεί έχει μία όψη του συνολικού συστήματος στο σχήμα 2.6. Συνολικά, το σύστημά μας δέχεται ως είσοδο ένα βίντεο και παράγει στην έξοδο τα ταξινομημένα τμήματα και τους χρόνους εκκίνησής τους. Το σύστημα όπως απεικονίζεται προϋποθέτει ότι οι δομικές του μονάδες έχουν εκπαιδευτεί κατάλληλα. Δηλαδή το σχήμα δείχνει το σύστημα στη φάση ελέγχου (test). Η παρουσίαση εδώ είναι αφηρημένη εσκεμμένα, ώστε να δείξει τη γενικότητα της σχεδίασης αλλά και τη δυνατότητα αυτόνομης αντικατάστασης ενός τμήματος με κάποιο αντίστοιχης λειτουργίας αλλά μεγαλύτερων δυνατοτήτων.



Σχήμα 2.6: Το συνολικό σύστημα αναγνώρισης-κατάτμησης δράσεων που προτείνουμε σε μορφή μπλοκ διαγράμματος. Οι δομικές του μονάδες αλληλεπιδρούν αλλά σχεδιάζονται και υλοποιούνται ανεξάρτητα ώστε να μπορούν να ενημερώνονται αυτόνομα. Σε κάθε τμήμα βίντεο γίνεται συνδυασμός χαρακτηριστικών πολλαπλών καναλιών και εξάγεται η κατανομή πιθανοτήτων των δράσεων. Ο αλγόριθμος κατάτμησης χρησιμοποιεί δυναμικό προγραμματισμό για να προτείνει νέο τμήμα βίντεο προς εξέταση. Η ταξινόμηση στα τελικά τμήματα γίνεται σύμφωνα με τη μέγιστη εκ των υστέρων (posterior) πιθανότητα.

Συνολικά, το βίντεο διαχωρίζεται σε εικόνα και ήχο, αν αυτός υπάρχει. Ο ήχος διέρχεται από το σύστημα Ανάλυσης Ήχου σε Κείμενο (ASR) για να μετατραπεί σε υπότιτλους (αν δεν διαθέτουμε ήδη υπότιτλους) και από εκεί λαμβάνουμε τις συναρτήσεις ενίσχυσης πιθανοτήτων για αντικείμενα και δράσεις. Οι συναρτήσεις αυτές ονομάζονται έτσι διότι θα χρησιμοποιηθούν έτσι ώστε να ενισχύσουν τις αντίστοιχες πιθανότητες που θα προκύψουν από την οπτική ανάλυση του βίντεο. Η εικόνα διέρχεται από το σύστημα εξαγωγής χαρακτηριστικών (Features Extractor), το οποίο περιλαμβάνει οπτική ανάλυση χαμηλού επιπέδου (Low-level Vision) και σημασιολογικού επιπέδου (Semantics). Τα χαρακτηριστικά από τα διάφορα υποσυστήματα συνδυάζονται μεταξύ τους στο στάδιο του ταξινομητή (Combine Modalities Classifier), ο οποίος εξάγει τις πιθανότητες για κάθε δράση στο συγκεκριμένο τμήμα. Οι πιθανότητες αυτές συνδυάζονται με τη συνάρτηση ενίσχυσης πιθανοτήτων δράσεων από το ακουστικό υποσύστημα και τις δεσμευμένες πιθανότητες δράσεων-αντικειμένων που προκύπτουν από το υποσύστημα επεξεργασίας κειμένου (Text Analyzer). Οι τρεις τύποι πιθανοτήτων αποτελούν τις εισόδους του συστήματος Combine Probabilities και από εκεί εξάγεται η τελική κατανομή πιθανοτήτων για το τρέχον τμήμα βίντεο. Λαμβάνοντας αυτές τις πιθανότητες, το σύστημα κατάτμησης (Segmentation) προτείνει νέα τμήματα βίντεο για εξέταση, σύμφωνα με έναν αλγόριθμο δυναμικού προγραμματισμού. Σε κάθε τμήμα που διατηρείται τελικά, λαμβάνουμε την κλάση του ως αυτή με τη μέγιστη συνολική πιθανότητα. Να σημειωθεί ότι το υποσύστημα επεξεργασίας κειμένου εξάγει τις δεσμευμένες πιθανότητες στη φάση εκπαίδευσης αλλά εικονίζεται εδώ για πληρότητα στη ροή της πληροφορίας. Συνολικά, τα διαφορετικά κανάλια πληροφορίας που συνδυάζονται είναι:

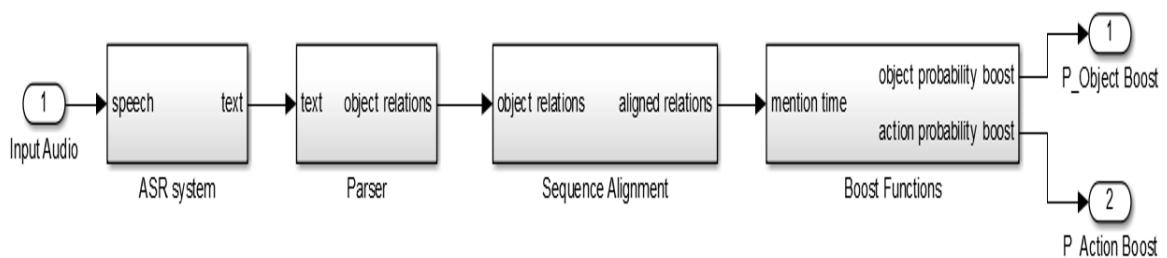
- Το κανάλι οπτικής πληροφορίας, που χρησιμοποιείται για την εξαγωγή χαρακτηριστικών χαμηλού επιπέδου αλλά και υψηλότερου επιπέδου (αντικείμενα, χέρια)

- Το κανάλι υποτίτλων (πιθανά μέσω ακουστικής πληροφορίας, αν αυτή υπάρχει).
- Το κανάλι κειμένου, το οποίο χρησιμοποιείται για την εξαγωγή των σχέσεων δράσεων και λεξιλογίου (αντικειμένων).
- Το σημασιολογικό κανάλι, το οποίο συνδυάζει πληροφορίες υψηλού επιπέδου (εμφάνιση αντικειμένων, τύπων λαβής κλπ) από διαφορετικά κανάλια για συνθέσει μια απεικόνιση της δράσης συνδέοντάς τη με έννοιες (όπως τα αντικείμενα που εμφανίζονται) και όχι χαρακτηριστικά αυτά καθαυτά.

Προχωράμε σε ανάλυση κάθε υποσυστήματος ξεχωριστά.

2.3.1 Το Υποσύστημα Ακουστικής Πληροφορίας-Υποτίτλων

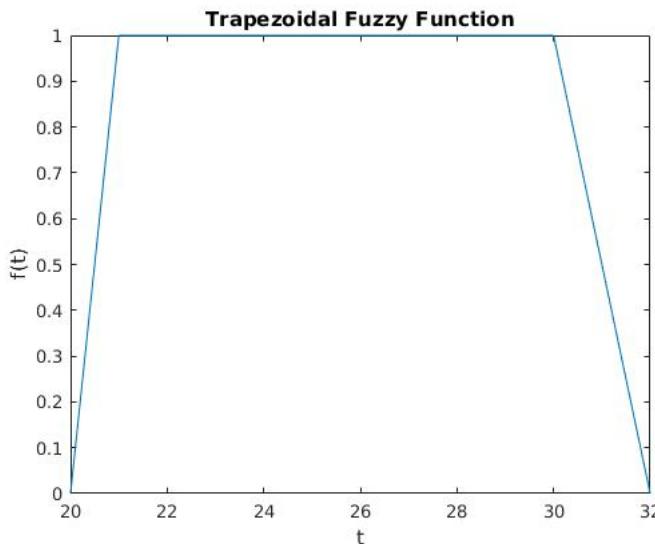
Το υποσύστημα ακουστικής πληροφορίας χρησιμοποιείται για την εξαγωγή υποτίτλων από το βίντεο σε περίπτωση που αυτοί δεν υπάρχουν και την περαιτέρω ανάλυσή τους. Σε πρώτη φάση, η ηχητική πληροφορία διέρχεται από σύστημα αυτόματης μετατροπής σε κείμενο, οπότε λαμβάνουμε τους υπότιτλους και τους σχετικούς χρόνους σε μορφή κειμένου. Προφανώς το βήμα αυτό παραλείπεται αν διαθέτουμε υπότιτλους εκ των προτέρων. Ακολουθεί συντακτικός αναλυτής ο οποίος εξάγει τις εμφανίσεις των λέξεων του λεξιλογίου μας. Ως λεξιλόγιο ορίζουμε εδώ το σύνολο των δράσεων (συνήθως ένα σύνολο ρημάτων) ενωμένο με το σύνολο των σημασιολογικών παραγόντων (συνήθως σύνολα ουσιαστικών). Για παράδειγμα, σε μια πρόσεγγιση όπου σημασιολογικοί παράγοντες είναι τα εμφανιζόμενα αντικείμενα, τότε το λεξιλόγιο θα αποτελείται από τους τίτλους δράσεων και τα ονόματα των αντικειμένων. Όπως θα δούμε στη συνέχεια όταν αναλύσουμε την υλοποίησή μας για το προτεινόμενο σύστημα, το δικό μας λεξιλόγιο αποτελείται από ένα σύνολο δράσεων, αντικειμένων και τύπων λαβής.



Σχήμα 2.7: Η εσωτερική δομή του Υποσυστήματος Ακουστικής Πληροφορίας. Μετά τη μετατροπή των ακουστικών υποτίτλων σε κείμενο γίνεται η συντακτική ανάλυση και η χρονική ευθυγράμμιση με την εικόνα. Οι κατανομές πιθανοτήτων που εξάγονται από την ανάλυση αυτή θα συνδυαστούν με τις αντίστοιχες που θα προκύψουν από τα υπόλοιπα κανάλια πληροφορίας.

Η εμφάνιση των λέξεων του λεξιλογίου δεν έχει κάποια χρηστική αξία αν δεν ευθυγραμμιστεί με το συνολικό βίντεο. Για το σκοπό αυτό, η χρονική πληροφορία εμφάνισης μιας λέξης μεταφράζεται σε frames με μια συνάρτηση ασαφούς λογικής.

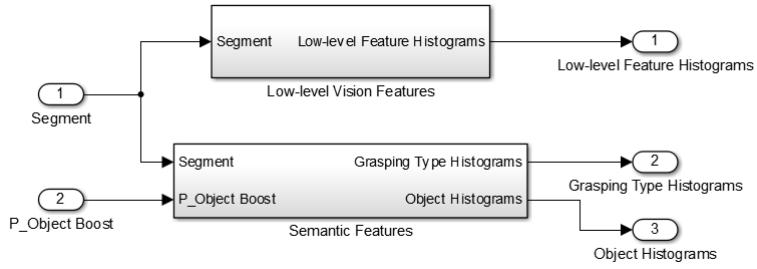
Επομένως, θεωρούμε ότι η δράση που αναφέρεται όντως πραγματοποιείται ή το αντικείμενο όντως εμφανίζεται σε ένα παράθυρο γύρω από τη χρονική στιγμή αναφοράς της σχετικής λέξης. Το σχήμα 2.8 απεικονίζει ένα παράδειγμα μιας τέτοιας συνάρτησης. Η συνάρτηση αυτή δηλώνει την πιθανότητα εμφάνισης του αντικειμένου ή της πραγματοποίησης της δράσης και μια συστοιχία τέτοιων συναρτήσεων (μία για κάθε λέξη του λεξιλογίου) αποτελεί την έξοδο του υποσυστήματος ακουστικής πληροφορίας, τις λεγόμενες, στη γενική περιγραφή, συναρτήσεις ενίσχυσης πιθανοτήτων. Κλείνοντας, στεκόμαστε στη μορφή αυτών των συναρτήσεων, που η καθεμία αποτελεί την επαλληλία της μηδενικής συνάρτησης στο εύρος του βίντεο με τις ασαφείς συναρτήσεις που προκύπτουν για την αντίστοιχη λέξη.



Σχήμα 2.8: Μια τραπεζοειδής ασαφής συνάρτηση. Τέτοιας μορφής συναρτήσεις χρησιμοποιούνται για να αποδόσουν μια ένδειξη χαλαρών άκρων συνάρτησης. Η χρονική συσχέτιση λόγου και εικόνας στο σύστημά μας γίνεται με τέτοιες συναρτήσεις γύρω από τη διάρκεια του κάθε υποτίτλου.

2.3.2 Το Υποσύστημα Εξαγωγής Χαρακτηριστικών

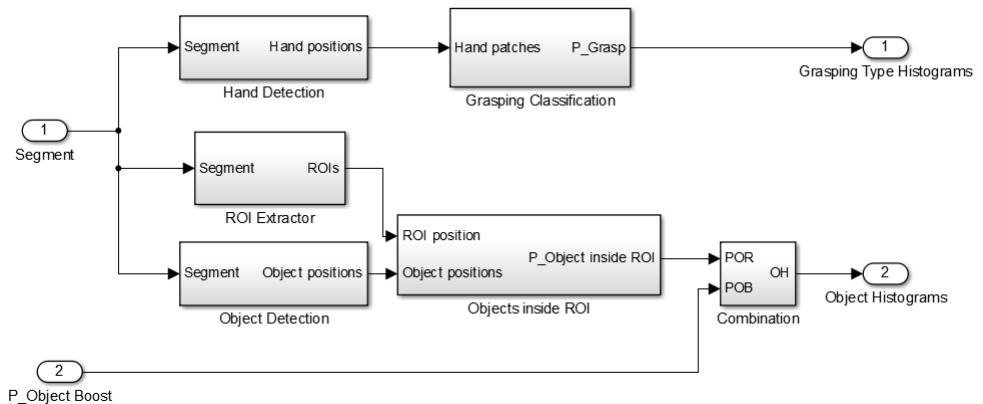
Το υποσύστημα εξαγωγής χαρακτηριστικών δέχεται ως είσοδο ένα τμήμα βίντεο και τις συναρτήσεις ενίσχυσης πιθανοτήτων του λεξιλογίου από το υποσύστημα ακουστικής πληροφορίας (για απλότητα θα τις αναφέρουμε ως συναρτήσεις ενίσχυσης πιθανοτήτων αντικειμένων). Στο εσωτερικό του υποσυστήματος στεγάζονται δύο μικρότερες υπομονάδες εξαγωγής χαρακτηριστικών, το υποσύστημα χαρακτηριστικών χαμηλού επιπέδου όρασης (Low-level Vision Features) και το σημασιολογικό υποσύστημα (Semantic features). Η έξοδος του συνολικού συστήματος προκύπτει από το σύνολο των εξόδων των επιμέρους υπομονάδων του. Θα αναλύσουμε τώρα κάθε υποσύστημα ξεχωριστά.



Σχήμα 2.9: Το Υποσύστημα Εξαγωγής Χαρακτηριστικών αναλαμβάνει το καθήκον εξαγωγής οπτικών και σημασιολογικών χαρακτηριστικών από την παρακολούθηση της οπτικής πληροφορίας του βίντεο, ενώ η πληροφορία υποτίτλων μπορεί να ενισχύει την πεποίθηση σημασιολογικών χαρακτηριστικών. Τα χαρακτηριστικά κωδικοποιούνται σε μορφή ιστογραμμάτων σε αυτό το στάδιο, οπότε η πληροφορία των καναλιών είναι συγκρίσιμη και να μπορεί να γίνει σύζευξη πριν το στάδιο του ταξινομητή.

Πρώτο και απλούστερο είναι το υποσύστημα εξαγωγής χαμηλού επιπέδου οπτικών χαρακτηριστικών. Το σύστημα περιλαμβάνει μία μέθοδο εξαγωγής χαρακτηριστικών και ένα στάδιο κωδικοποίησής τους για τη δημιουργία ιστογραμμάτων. Η μέθοδος εξαγωγής μπορεί να επιλεχθεί ανεξάρτητα από το υπόλοιπο σύστημα, με δυνατότητα να ανανεώνεται σύμφωνα με το εκάστοτε state-of-the-art.

Δεύτερο και πιο σύνθετο το υποσύστημα σημασιολογικής πληροφορίας. Εδώ γίνεται η οπτική εξαγωγή χαρακτηριστικών για όλα τα ουσιαστικά του λεξιλογίου μας (πιο γενικά, για όλες τις λέξεις πέραν των τίτλων των δράσεων). Οι πιθανότητες αντικειμένων συνδυάζονται με τις συναρτήσεις ενίσχυσης πιθανοτήτων και προκύπτουν οι κατανομές λεξιλογίου. Στην εικόνα 2.10 φαίνονται και οι επιπρόσθετοι περιορισμοί που μπορεί να εισέλθουν σε ένα σύστημα ανίχνευσης. Για παράδειγμα η ανίχνευση αντικειμένων γίνεται μέσα σε μια περιοχή ενδιαφέροντος. Η κωδικοποίηση των σημασιολογικών χαρακτηριστικών για την εξαγωγή ιστογραμμάτων γίνεται επίσης μέσα στο υποσύστημα αυτό.



Σχήμα 2.10: Το Υποσύστημα Σημασιολογικής Πληροφορίας αναμιγνύει την οπτική ανίχνευση με την αντίληψη του συστήματος για τη δομή των δράσεων. Παρότι στο σχήμα απεικονίζεται η δομή που τελικά ακολουθήσαμε, συνδυάζοντας πληροφορία αντικειμένων και τύπων λαβής (grasping type), είναι φανερό ότι πολλαπλά κανάλια σημασιολογίας μπορούν να εισαχθούν ανεξάρτητα και τελικά να παραχθούν τα αντίστοιχα ιστογράμματα.

2.3.3 Συνδυασμός Χαρακτηριστικών και Ταξινομητής

Στο στάδιο του ταξινομητή γίνεται η σύζευξη οπτικών χαμηλού-επιπέδου και σημασιολογικών χαρακτηριστικών και τελικά η ταξινόμησή τους. Η επιλογή του ταξινομητή αλλά και της μεθόδου σύζευξης δεν επηρεάζουν το υπόλοιπο σύστημα. Αν ο ταξινομητής δεν επιστρέφει εγγενώς πιθανότητες, μπορούμε να τις εξάγουμε ποιοτικά με κάποια προσέγγιση (calibration).

2.3.4 Ο Αναλυτής Κειμένου

Ο αναλυτής κειμένου είναι ένας συντακτικός αναλυτής ο οποίος εξάγει τις σχέσεις μεταξύ δράσεων και υπολοίπου λεξιλογίου, όπου αυτό είναι δυνατό. Στο υποσύστημα αυτό παρέχονται στη φάση εκπαίδευσης δεδομένα από κείμενα σχετικά με τις δραστηριότητες που αναλύει το σύστημα. Στόχος του υποσυστήματος είναι η εξαγωγή των δεσμευμένων πιθανοτήτων των δράσεων όταν εμφανίζονται γλωσσικά μαζί με τις υπόλοιπες λέξεις του λεξιλογίου. Αυτό μπορεί να υλοποιηθεί με οποιαδήποτε μορφής ανάλυση, είτε με συντακτικές σχέσεις ρήματος-αντικειμένου-υποκειμένου είτε με απλή παράθεση. Οι δεσμευμένες πιθανότητες εκφράζουν τη σχέση των δομικών λίθων του λεξιλογίου και χρησιμοποιούνται στο φιλτράρισμα των πιθανοτήτων των άλλων δομών.

2.3.5 Ο Συνδυασμός των Πιθανοτήτων

Μπορούμε να συνδυάσουμε τις πιθανότητες δράσεων που προκύπτουν από τα τρία υποσυστήματα (οπτικό, ακουστικό και γλωσσικό) για να εξάγουμε την τελική κατανομή πιθανοτήτων δράσεων. Πρακτικά ζητάμε τη δράση α που μεγιστοποιεί την ποσότητα $P(\alpha|vocabulary, features)$, δηλαδή την πιθανότητα της δράσης α δεδομένης της εμφάνισης των χαρακτηριστικών λεξιλογίου *vocabulary* και των λοιπών χαρακτηριστικών *features* τα οποία εμφανίστηκαν. Υπό διάφορες στατιστικές υποθέσεις, ο χειρισμός αυτού του προβλήματος βελτιστοποίησης μπορεί να γίνει ποικιλοτρόπως. Αποτελεί σχεδιαστική επιλογή το σε ποιες υποθέσεις θα καταλήξουμε.

2.3.6 Το Μπλοκ Κατάτμησης

Το υποσύστημα κατάτμησης εφαρμόζει έναν αλγόριθμο δυναμικού προγραμματισμού για να προτείνει τμήματα προς εξέταση. Παρότι στο σχήμα φαίνεται σαν μια αναδρομή που επαναλαμβάνει τον υπολογισμό των χαρακτηριστικών σε κάθε κύκλο, εν τούτοις η εξαγωγή μπορεί να επιτευχθεί σε ένα κύκλο και να εξετάζονται μόνο τα κατάλληλα χαρακτηριστικά σε κάθε επανάληψη. Όταν το τμήμα κατάτμησης αποφασίσει εύρεση ορίου τότε το αντίστοιχο τμήμα μπορεί να ταξινομηθεί σύμφωνα με τη μέγιστη πιθανότητα δράσης για αυτό.

2.4 Προτεινόμενη Υλοποίηση

Θα δείξουμε τώρα τη δική μας σχεδίαση πάνω στο γενικό σύστημα που προαναφέραμε. Καταρχήν το λεξιλόγιο που χρησιμοποιούμε αποτελείται από ένα σύνολο 61 ρηματικών συνόλων (δράσεις), ένα πλήθος ουσιαστικών χώρου και αντικειμένων και 8 τύπους λαβής (grasping types). Στο Παράρτημα A παραθέτουμε αναλυτικούς καταλόγους με τις δράσεις και τα ουσιαστικά που χρησιμοποιούμε, ενώ οι τύποι λαβής αναλύονται στο Κεφάλαιο 5. Ας δούμε όμως κάθε υποσύστημα ξεχωριστά.

- Για το υποσύστημα ακουστικής πληροφορίας θεωρήσαμε ότι διαθέτουμε τους υπότιτλους ώστε να γλιτώσουμε από άχρηστο θόρυβο λόγω σφαλμάτων του συστήματος αναγνώρισης φωνής. Για τη συντακτική ανάλυση επιλέξαμε γλωσσικό ταίριασμα προτύπων λαμβάνοντας υπόψιν τις διαφορετικές μορφές που μπορεί να λάβει το κάθε ουσιαστικό. Πληροφορία για τύπο λαβής δε βρίσκουμε στους υποτίτλους. Επιπλέον, αναζητούμε μόνο αντικείμενα και όχι δράσεις, οπότε δεν αξιοποιούμε τη συνάρτηση ενίσχυσης πιθανοτήτων των δράσεων. Για τη συνάρτηση πιθανοτήτων των αντικειμένων, θεωρούμε ότι η εμφάνιση κάθε ενός εξαπλώνεται στη διάρκεια του υποτίτλου και λίγο παρακάτω χρονικά, το οποία μεταφράζεται σε επιπλέον 10% του υποτίτλου, με μια φθίνουσα γραμμική συνάρτηση που μηδενίζεται στο σημείο εκείνο.
- Στο σύστημα εξαγωγής οπτικών και σημασιολογικών χαρακτηριστικών έχουμε να επιλέξουμε αρκετές παραμέτρους. Αρχικά, για τα χαμηλού-επιπέδου χαρακτηριστικά επιλέγουμε τη μέθοδο των Πυκνών Τροχιών [248] και κωδικοποιούμε τα χαρακτηριστικά με χρήση k-μέσων. Στο υποσύστημα σημασιολογικής πληροφορίας, σχεδιάζουμε ανιχνευτή χεριών με χρήση HOG [49], BING [33], ιστογραμμάτων χρώματος και ταξινομητή Τυχαίου Δάσους (Random Forest) [91] και για τις εικόνες των ανευρεθέντων χεριών εξάγουμε συνελικτικά χαρακτηριστικά μέσω ResNet τα οποία τοποθετούμε σε έναν χώρο 2048 διαστάσεων με ομαδοποίηση k-μέσων. Για την ανιχνευση αντικειμένων τρέχουμε ταίριασμα προτύπων με δειγματοληπτημένες εικόνες τους από τα διάφορα βίντεο. Η περιοχή ενδιαφέροντος σχηματίζεται από έναν ανιχνευτή προσκηνίου σε συνδυασμό με έναν ανιχνευτή ανθρώπων. Τα αντικείμενα που ανιχνεύονται συνδυάζονται με τις συναρτήσεις ενίσχυσης από το ακουστικό υποσύστημα και με χρήση κατωφλίου προκύπτουν τα ιστογράμματα αντικειμένων.
- Στο στάδιο του ταξινομητή δοκιμάζονται διάφοροι τρόποι σύζευξης χαρακτηριστικών και ταξινομητές. Γενικά καλύτερα αποτελέσματα δίνει η απλή συνέννωση με σχήμα Tf-Idf και χρήση γραμμικής Μηχανής Διανυσμάτων Στήριξης (SVM).
- Ο αναλυτής κειμένου δέχεται ένα σύνολο από λεπτομερείς περιγραφές βίντεο μαγειρικής και εξάγει τις κοινές εμφανίσεις αντικειμένων και δράσεων στην ίδια πρόταση. Μετά από μια κανονικοποίηση Laplace των εμφανίσεων, προκύπτουν οι πιθανότητες των δράσεων σε σχέση με την εμφάνιση των αντικειμένων. Χρησιμοποιούμε αυτές τις πιθανότητες για να προσαρμόσουμε την έξοδο του ταξινομητή. Η στατιστική υπόθεση είναι ότι αντικείμενα και χαρακτηριστικά χαμηλού-επιπέδου όρασης εμφανίζονται ανεξάρτητα, άρα

$$P(\alpha|vocabulary, features) = P(\alpha|vocabulary) \times P(\alpha|features) \quad (2.1)$$

- Για την κατάτμηση χρησιμοποιούμε μια νέα παραλλαγή του αλγορίθμου των [92], η οποία μοιράζει το βάρος μεταξύ των τμημάτων στα οποία σπάει κάθε μεγαλύτερο τμήμα, παρά αθροίζοντας το κόστος κι έτσι επιβραβεύοντας μεγάλα τμήματα.

Θα αναλύσουμε κάθε δομική υπομονάδα, σχεδιαστική επιλογή και αποτέλεσμα σε ξεχωριστά κεφάλαια και με λεπτομέρεια. Ωστόσο αξίζει να σταθούμε εδώ για να σχολιάσουμε τη δύναμη που προσφέρει η συμπληρωματικότητα των μεθόδων που χρησιμοποιούνται στο σύστημά μας. Από διαισθητική και μόνο πλευρά, το σύστημά μας αξιοποιεί όλες τις αισθήσεις στις οποίες βασίζεται ένας άνθρωπος προκειμένου να αποκωδικοποιήσει ένα βίντεο. Αφενός η οπτική πληροφορία είναι η πιο σημαντική, αφού μπορούμε να διακρίνουμε τι συμβαίνει (κίνηση και χαμηλού-επιπέδου πληροφορία), πού συμβαίνει και αν έχει σχέση με τη δράση (περιοχή ενδιαφέροντος) και πώς συμβαίνει (με τι αντικείμενα και τι χειρισμό). Ταυτόχρονα, εκμεταλλευόμαστε την ηχητική πληροφορία για να διαλευκάνουμε σημεία που η οπτική αντίληψη δυσκολεύεται να διακρίνει. Η ακουστική πληροφορία έχει μία τοπικότητα και προσαρμόζεται στη δυναμική του βίντεο. Τέλος, η πληροφορία κειμένου αφορά το “διάβασμα” του ακροατή. Ο όρος διάβασμα δεν αφορά την ανάγνωση αλλά τη μελέτη και τη συσσώρευση εμπειρίας. Με άλλα λόγια, οι πληροφορίες κειμένου εκφράζουν τη σύνδεση των αντίληψης με την ανθρώπινη λογική, προσφέροντας μια ταξινόμηση υψηλού επιπέδου.

2.5 Το Σύνολο Δεδομένων

Το σύνολο δεδομένων που χρησιμοποιείται στα πειράματα είναι το MPII Cooking Activities Dataset [191]. Το dataset αυτό περιλαμβάνει 44 βίντεο μαγειρικής. Συγκεκριμένα, 13 διαφορετικοί άνθρωποι ετοιμάζουν μια σειρά από διαφορετικά πιάτα. Παρέχονται επισημειώσεις για τη θέση και το είδος κάθε μιας από τις 65 διαφορετικές κατηγορίες δράσεων στα βίντεο. Στις δράσεις αυτές περιλαμβάνεται και η κλάση background activity, η οποία αφορά δραστηριότητες οι οποίες δεν έχουν επισημανθεί σε κάποια από τις 65 κατηγορίες. Σημειώνεται ότι για να επισημανθεί ένα τμήμα βίντεο ως δράση υποβάθρου, πρέπει να έχει διαρκεία τουλάχιστον 1 δευτερόλεπτο. Επιπλέον το dataset παρέχει επισημάνσεις πόζας οι οποίες χρησιμοποιούνται στην εξαγωγή των χαρακτηριστικών πόζας. Τέλος, στην ιστοσελίδα των ερευνητών παρέχονται προϋπολογισμένα τα χαρακτηριστικά που εξάγονται κατά τη μέθοδο Πυκνών Τροχιών [248] ανά frame, σε μορφή ολοκληρωτικών ιστογραμμάτων.

Τροποποιούμε το σύνολο δεδομένων ελαφρώς ώστε να διευκολύνει σε περισσότερα πειράματα. Αρχικά, δημιουργούμε αρχεία υποτίτλων για ένα σύνολο βίντεο που χρησιμοποιούνται στον έλεγχο του συνολικού συστήματος. Για μεγαλύτερη αξιοπιστία οι υπότιτλοι κατασκευάστηκαν από ανθρώπους επισημειώτες στους οποίους παρέχονταν το βίντεο, ένας κατάλογος λεξιλογίου δράσεων και αντικειμένων που εμφανίζονται μέσα στο βίντεο και ένας χάρτης των δράσεων στη χρονική διάρκεια του βίντεο. Επιβάλλαμε τον περιορισμό της χρήσης των λέξεων του λεξιλογίου μας αντί άλλων συνώνυμων λέξεων όταν γίνεται αναφορά σε αντικείμενα που υπάρχουν στο λεξιλόγιο.

Επιπλέον, προβαίνουμε σε object annotations. Η χρησιμότητα των επισημειώσεων αυτών αφορά την εκπαίδευση συστημάτων ανίχνευσης αντικειμένων μέσα στα βίντεο. Αντί να εκπαιδεύσουμε γενικούς ταξινομητές για όλα τα είδη αντικειμένων και όλα



Σχήμα 2.11: Το σύνολο δεδομένων που χρησιμοποιούμε (MPII Cooking Activities Dataset) αποτελείται από βίντεο στα οποία διαφορετικοί άνθρωποι εκτελούν συνταγές μαγειρικής και ετοιμάζουν διαφορετικά πιάτα. Το σύνολο δεδομένων προσφέρεται τόσο για αναγνώριση δράσεων (action) όσο και για αναγνώριση δραστηριοτήτων (activity). Εμείς κινούμαστε προς την αναγνώριση δράσεων.

τα βίντεο, επιλέγουμε μόνο τα αντικείμενα με μικρή οπτική μεταβλητότητα. Η ιδέα είναι ότι συνήθως τα αντικείμενα που παραμένουν συμπαγή και εντοπίζονται εύκολα είναι τα εργαλεία και οι χώροι υποβάθρου ενώ τα αντικείμενα μεγάλης μεταβλητότητας που εντοπίζονται δύσκολα οπτικά είναι τα σημαντικά εργαλεία χειρισμού και τα υλικά, τα οποία αναφέρονται στους υπότιτλους. Συν τα παραπάνω, εκπαιδεύουμε για κάθε βίντεο ελέγχου ξεχωριστά συστήματα εντοπισμού. Ουσιαστικά χώρου, όπως “πάγκος”, δε μας παρέχουν σημαντική πληροφορία για τη δράση κι έτσι αμελούνται. Αποφεύγουμε να συνενώσουμε ποικιλίες παρόμοιων αντικειμένων γιατί αυτές προσφέρουν γνώση για τη λεπτομέρεια της δράσης, κάτι απαραίτητο στο διαχωρισμό λεπτομερών δράσεων όπως στην περίπτωσή μας. Διαισθητικά, αντί να δημιουργήσουμε ένα σύστημα που έχει κατασκευάσει ένα μοντέλο για το τι είναι μήλο για παράδειγμα, ρωτάμε σε κάθε βίντεο τι μορφή θα έχει το μήλο ώστε να το εντοπίσουμε. Λεπτομέρειες για το σύνολο δεδομένων και τις επισημειώσεις στις οποίες προβήκαμε παρέχονται στο Παράρτημα A.

Τέλος, χρησιμοποιούμε ως πρόσθετο υλικό μια νεότερη μορφή dataset, την MPII Cooking 2 Dataset [192], [193], [191], πιο ευρεία σε πλήθος βίντεο αλλά με διαφορετικές επισημειώσεις. Εδώ υπάρχουν κειμενικές περιγραφές για τις δράσεις του βίντεο. Οι περιγραφές αυτές έχουν αναλυθεί και έχει εξαχθεί ground truth συνδυασμός ρήματος και αντικειμένων στην πρόταση. Χρησιμοποιούμε τα κειμενικά αυτά δεδομένα για στατιστικά σχέσεων αντικειμένου και με αυτά αναπαριστούμε τα διανύσματα στο σημασιολογικό υποσύστημα. Στο Παράρτημα A περιγράφουμε διεξοδικότερα τη μορφή του συνόλου δεδομένων και τις επεμβάσεις μας σε αυτό.

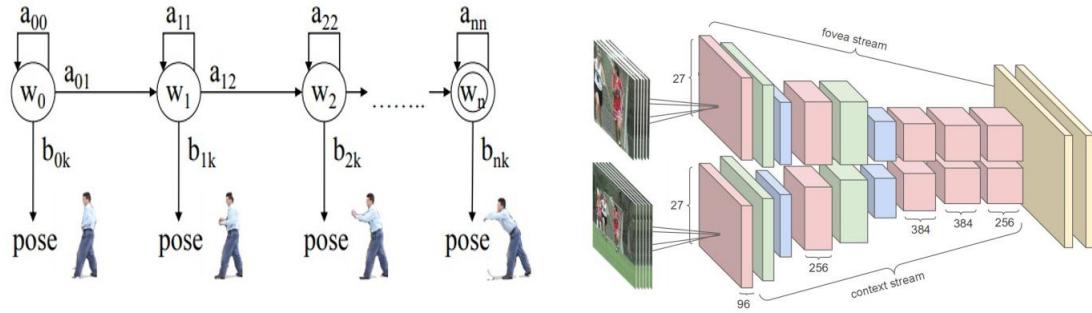
Κεφάλαιο 3

Το Υποσύστημα Εξαγωγής Χαρακτηριστικών Όρασης Χαμηλού Επιπέδου

Στο προηγούμενο κεφάλαιο είδαμε τη δομή του συνολικού συστήματος. Ακόμα νωρίτερα, στο ίδιο κεφάλαιο, γνωρίσαμε διάφορες προσεγγίσεις αναγνώρισης δράσεων που χρησιμοποιούν σημασιολογία ή οπτικά χαρακτηριστικά υψηλού επιπέδου. Η αναγνώριση δράσεων με χαρακτηριστικά χαμηλού επιπέδου είναι εξίσου σημαντική και πιθανόν ισχυρή. Διαισθητικά, η πληροφορία κίνησης εξασφαλίζει την ύπαρξη δράσης, την ώρα που απλές επισημειώσεις περιβάλλοντος αδυνατούν να εγγυηθούν κάτι παρόμοιο. Ακόμα περισσότερο, η αναπαράσταση των δράσεων στις iεραρχικές μεθόδους συχνά προϋποθέτει ισχυρά μοντέλα ταξινομητών χαμηλού επιπέδου. Το παρόν κεφάλαιο οργανώνεται ως εξής: αρχικά γίνεται μια ιστορική ανασκόπηση της αναγνώρισης δράσεων με οπτικά χαρακτηριστικά χαμηλού επιπέδου. Στη συνέχεια παρουσιάζεται η θεωρία της μεθόδου των Πυκνών Τροχιών και των χαρακτηριστικών που χρησιμοποιεί. Αναλύονται επίσης τα χαρακτηριστικά πόζας και η κωδικοποίηση των χαρακτηριστικών για δημιουργία ιστογραμμάτων. Τέλος, παρουσιάζουμε τη δική μας μέθοδο η οποία συνδυάζει όλα τα παραπάνω.

3.1 Ιστορικά Στοιχεία

Η πρώτη απόπειρα μοντελοποίησης των δράσεων σε βίντεο μπορεί να θεωρηθεί η σειρά πειραμάτων του Johansson [110] το 1973. Στην εργασία αυτή, συγκεντρώνεται ένα σύνολο δεδομένων με ανθρώπους σε μαύρη ενδυμασία στους κύριους συνδέσμους των οποίων έχουν τοποθετηθεί μικρά φώτα. Σκοπός του πειράματος ήταν η εξέταση του αν η πληροφορία αυτή είναι αρκετή ώστε ένας άνθρωπος να αποφανθεί για την δράση του υποκειμένου. Το 1982 το [250] είναι η πρώτη εργασία που μελετά την εύρεση της δομής των αντικειμένων από την κίνηση και θεωρείται ότι γέννησε την αναγνώριση δράσεων μέσω συνδέσμων. Η έρευνα εντείνεται από τη δεκαετία του '90 μέχρι και σήμερα, επιτυγχάνοντας αρκετά αξιόλογα αποτελέσματα και εξελίσσοντας πολλαπλές μεθόδους.



Σχήμα 3.1: Η εξέλιξη της Αναγνώρισης δράσεων με τεχνικές που στηρίζονται αποκλειστικά στην οπτική αντίληψη, από τις ισχυρές υποθέσεις στην αφηρημένη δομή δικτύου. **Αριστερά:** ένα παράδειγμα σειριακής προσέγγισης που κάνει χρήση Κρυφού Μαρκοβιανού Μοντέλου για να αναπαραστήσει την απλή δράση "Σπρώχνω" με καταστάσεις που αντιστοιχούν σε ενδιάμεσες στάσεις σώματος. Εικόνα από [48]. **Δεξιά:** Δομή Συνελικτικού Νευρωνικού Δικτύου για ανάλυση αθλητικών αγώνων από την εργασία [116].

Η πρώτη προσέγγιση ήταν η λεγόμενη σειριακή (sequential). Η ιδέα είναι ότι μία δράση μπορεί να εκφραστεί ως ακολουθία καταστάσεων των μερών του σώματος. Κάθε frame επομένως περιγράφει μία συγκεκριμένη διάταξη αυτών των μερών. Η αναγνώριση μπορεί να γίνει με Κρυφά Μαρκοβιανά Μοντέλα (Hidden Markov Models, HMM) όπου μια κατάσταση αντιστοιχεί σε μία συγκεκριμένη πόζα σώματος [253], [226]. Μια εναλλακτική είναι η χρήση δυναμικού προγραμματισμού για το ταίριασμα δύο ακολουθιών [80]. Η προσέγγιση αυτή εξελίχθηκε με αρκετές παραλλαγές τα επόμενα χρόνια, όπως χρήση συζευγμένων Κρυφών Μαρκοβιανών [168] και Ημι-Μαρκοβιανών Μοντέλων [164], καθώς και Δυναμικών Μπεϋζιανών Δικτύων (Dynamic Bayesian Networks) [170]. Ωστόσο, βασικό μειονέκτημα αυτών των προσεγγίσεων είναι η εξάρτηση από ισχυρά χαρακτηριστικά (features) για τις δράσεις και η αδυναμία διαχωρισμού σύνθετων δράσεων χωρίς πολύ μεγάλο μεγεθος συνόλου δεδομένων εκπαίδευσης.

Μια επόμενη προσέγγιση ήταν η χωροχρονική (space-time), η οποία αντιμετωπίζει τα βίντεο σαν τρισδιάστατους όγκους, με την τρίτη διάσταση να αποτελεί τον χρόνο. Εδώ το πρόβλημα είναι το ταίριασμα όγκων ως λύση της ταξινόμησης. Μία λύση σε αυτό ήταν το απευθείας ταίριασμα προτύπων (template matching) όπως στο [17] όπου εισάγεται η έννοια των Εικόνων Ιστορίας Κίνησης (Motion History Images) ως μια σταθμισμένη προβολή ενός χωροχρονικού όγκου προσκηνίου. Στο [117] το ταίριασμα γίνεται σε υποόγκους του βίντεο και τα επιμέρους σκορ συνδυάζονται. Η ιδέα εξελίχθηκε στην εξαγωγή χωροχρονικών χαρακτηριστικών από τα βίντεο, καθολικά (global), όπως οπτική ροή [63], αλλά και τοπικά, όπως στο [210] και αραιά, όπως τα κυβοειδή του [58]. Οι προσεγγίσεις αυτές εξελίχθηκαν σημαντικά στα επόμενα χρόνια [128], ώστε να λαμβάνουν υπόψιν τους τις χρονικές εξαρτήσεις των χαρακτηριστικών [208], [251].

Η εξέλιξη των μεθόδων χωροχρονικών χαρακτηριστικών ήταν συνυφασμένη με την απομάκρυνση από τις αναλύσεις σχήματος και πόζας. Η κίνηση θεωρήθηκε το κύριο στοιχείο που καθορίζει τη δράση. Από τους τύπους χαρακτηριστικών, αυτά που επικράτησαν ήταν τα τοπικά, καθώς περιγράφουν με ευρωστία τα κύρια συστατικά της κίνησης. Ένα πολύ σημαντικό βήμα προς αυτή την κατεύθυνση ήταν η εργασία των [248] που εισήγαγε τις Πυκνές Τροχιές. Η μέθοδος αυτή βελτιώθηκε και εξελίχθηκε [247] επιτυγχάνοντας ποσοστά ακριβείας βελτιωμένα κατά μεγάλο περιθώριο σε πολλές βάσεις βίντεο. Άξια αναφοράς η εργασία των [115] η οποία εστιάζει

στην αποδοτικότητα των αλγορίθμων αναγνώρισης δράσεων και παρέχει αλγορίθμους τάξεις μεγέθους γρηγορότερους από τις Πυκνές Τροχιές με κάποιο κόστος στην ακρίβεια. Άλλες μέθοδοι αξιοποιούν και δεδομένα βάθους [29], ενώ, τελευταία, τα Συνελικτικά Νευρωνικά Δίκτυα (Convolutional Neural Networks, CNN) μπήκαν επίσης στο πεδίο της αναγνώρισης δράσεων με το [116] και πρωταγωνιστούν στην εξέλιξη και τα state-of-the-art αποτελέσματα [35].

3.2 Θεωρητικό Υπόβαθρο

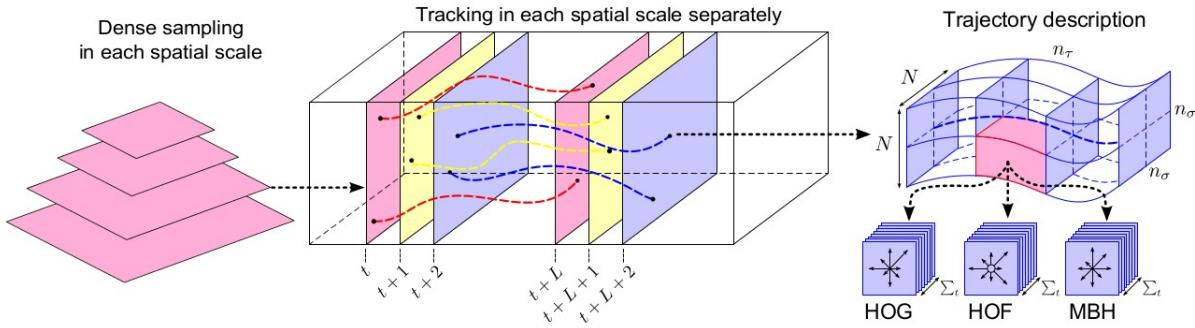
3.2.1 Η Μέθοδος των Πυκνών Τροχιών

Η μέθοδος των Πυκνών Τροχιών αποτελεί μια state-of-the-art μέθοδο για αναγνώριση δράσης σε βίντεο αξιοποιώντας μόνο οπτική πληροφορία. Ως Πυκνές Τροχιές ορίζουμε τη χωροχρονική τροχιά των σημείων ενδιαφέροντος πάνω σε ένα πυκνά δειγματοληπτημένο πλέγμα. Η ιδέα βασίζεται στην επιτυχία της πυκνής δειγματοληψίας [249] στην εξαγωγή σημείων ενδιαφέροντος και των μεθόδων ανίχνευσης κίνησης και προσκηνίου και προτάθηκε στο [248]. Η μέθοδος απέκτησε δημοφιλία λόγω της απλότητας, της ταχύτητας αλλά και της αποτελεσματικότητάς της, καθώς βελτίωνε τα αποτελέσματα αναγνώρισης σε όλα τα σύνολα δεδομένων. Επιπλέον, έχει τη δυνατότητα να αναιρεί την υπόθεση της σταθερής κάμερας, μοντελοποιώντας την κίνησή της. Μάλιστα, στο [247] επεκτείνεται το μοντέλο και πετυχαίνει καλύτερη αντιστάθμιση της κίνησης της κάμερας με αντίστοιχη βελτίωση στην ακρίβεια του συστήματος. Στη συνέχεια θα παρουσιάσουμε συνοπτικά τα κυριότερα σημεία της μεθόδου.

Οι πυκνές τροχιές εξάγονται σε πολλαπλές χωρικές κλίμακες, μέχρι 8 το πλήθος, οι οποίες προκύπτουν από υποδειγματοληψία της προηγούμενής τους κλίμακας (με αρχική κλίμακα 1, δηλαδή αρχική εικόνα την αυθεντική) κατά έναν παράγοντα $1/\sqrt{2}$. Τα σημεία ενδιαφέροντος εξάγονται από ένα χωρικό πλέγμα με κατάλληλο βήμα W pixels, έτσι ώστε να μπορεί να καλυφθεί όλη η εικόνα. Παρακολουθούμε τα σημεία ενδιαφέροντος αυτά σε κάθε κλίμακα ανεξάρτητα, με την κίνηση κάθε σημείου $P_t = (x_t, y_t)$ από το frame t στο frame $t + 1$ να προσεγγίζεται από την εξίσωση:

$$P_{t+1} = P_t + (M * \omega)|_{(\bar{x}_t, \bar{y}_t)} \quad (3.1)$$

όπου ένα φίλτρο διαμέσου (median) διάστασης 3×3 , (\bar{x}_t, \bar{y}_t) η θέση του σημείου στο frame t κατόπιν στρογγυλοποίησης των συντεταγμένων και ω το πεδίο πυκνής οπτικής ροής, υπολογισμένο με τη μέθοδο του Farneback [65]. Τα σημεία (P_t, P_{t+1}, \dots) συνενώνονται για να σχηματίσουν μια ακολουθία που ονομάζεται τροχιά. Καθώς οι τροχιές έχουν την τάση να ολισθαίνουν κατά τη διάρκεια ενός βίντεο, το μήκος τους περιορίζεται σε $L = 15$ σημεία και ένα σημείο παύει να παρακολουθείται όταν η τροχιά του ξεπεράσει το προκαθορισμένο αυτό μήκος. Ταυτόχρονα, για να εξασφαλιστεί η πυκνή κάλυψη του βίντεο, σε κάθε frame γίνεται έλεγχος αν παρακολουθείται κάποιο σημείο εντός μιας γειτονιάς $W \times W$ σημείων από τις αρχικές. Αν δεν βρεθεί τέτοιο σημείο, δειγματοληπτείται ένα νέο και προστίθεται στη διαδικασία παρακολούθησης. Τροχιές που είναι στατικές αφαιρούνται, μιας και δεν περιέχουν καμία πληροφορία που σχετίζεται με κίνηση εντός του βίντεο, ενώ τροχιές που παρουσιάζουν απότομες μετατοπίσεις μεταξύ δύο διαδοχικών frame επίσης εκτοπίζονται. Θα



Σχήμα 3.2: Απεικόνιση της μεθόδου των Πυκνών Τροχιών. **Αριστερά:** Τα σημεία παρακολούθησης δειγματοληπτούνται πυκνά για πολλαπλές χωρικές κλίμακες. **Κέντρο:** Η παρακολούθηση γίνεται στην αντίστοιχη χωρική κλίμακα και για L frames. **Δεξιά:** Οι περιγραφήτες τροχιάς βασίζονται στο σχήμα της, που αντιπροσωπεύεται από σχετικές συντεταγμένες σημείων, αλλά και σε πληροφορίες κίνησης και εμφάνισης σε γειτονιές $N \times N$ pixels κατά μήκος της τροχιάς. Προκειμένου να ενσωματωθεί δομική πληροφορία, η κάθε γειτονιά χωρίζεται επιπλέον σε ένα χωρο-χρονικό πλέγμα διαστάσεων $n_\sigma \times n_\sigma \times n_\tau$. Εικόνα από το [248].

σταθούμε σε αυτό το σημείο για να παρουσιάσουμε κάποια βήματα υπολογισμού των τροχιών και σχεδιαστικές επιλογές με μεγαλύτερη λεπτομέρεια.

Ξεκινάμε από την πυκνή δειγματοληψία, η οποία εφαρμόζεται σε 8 κλίμακες ταυτόχρονα και σε ένα τετράγωνο χωρικό πλαίσιο $W \times W$ (δηλαδή με βήμα W σε αμφότερες τις χωρικές διαστάσεις). Πειράματα στο [248] δείχνουν ότι τιμή $W = 5$ είναι επαρκής για την πυκνή δειγματοληψία και ότι περαιτέρω μείωση του W δε συνεισφέρει σημαντικά στην απόδοση του αλγορίθμου παρά επιβαρύνει υπολογιστικά. Σε κάθε πλέγμα απορρίπτουμε τα σημεία που ανήκουν σε ομοιογενείς περιοχές της εικόνας, καθώς είναι αδύνατο να τα παρακολουθήσουμε με χρήση οπτικής ροής. Για το σκοπό αυτό, χρησιμοποιούμε το κριτήριο των Shi-Tomasi [217], το οποίο βασίζεται στον ανιχνευτή γωνιών Harris [88] προκειμένου να εξάγει έναν πίνακα αυτοσυσχέτισης $M(x, y)$, ο οποίος αποτελεί ένα μέτρο της μεταβολής της εικόνας σε κάθε pixel ως προς τις δύο κατευθύνσεις. Αν τώρα λ_1 και λ_2 οι ιδιοτιμές του πίνακα αυτού, το κριτήριο γωνιότητας των Shi-Tomasi ελέγχει αν η μικρότερη των δύο ιδιοτιμών αυτών είναι μεγαλύτερη από ένα κατώφλι, το οποίο συχνά επιλέγεται ίσο με

$$\text{threshold} = 0.001 \max_{x,y} (\min(\lambda_1, \lambda_2))$$

Τα σημεία τα οποία δεν ικανοποιούν το κριτήριο Shi-Tomasi απορρίπτονται από την διαδικασία παρακολούθησης.

Συνεχίζουμε με την παρακολούθηση των σημείων ενδιαφέροντος. Βασικό συστατικό στην εξίσωση (3.1) είναι το πεδίο οπτικής ροής. Η οπτική ροή εκφράζει την σχετική κίνηση κάμερας και σκηνής σε ένα βίντεο. Διαισθητικά εκφράζει τη στιγμιαία ταχύτητα των επιφανειών, ενώ ένα πιο μαθηματικό ανάλογο θα μπορούσαν να αποτελούν τα διανύσματα κατευθυντήριων παραγώγων, κάθετων στις επιφάνειες κίνησης. Η χρήση της οπτικής ροής εδώ έχει σημαντικά πλεονεκτήματα: Αφενός, η μέθοδος παρακολούθησης γίνεται αρκετά πιο εύρωστη από μεθόδους παρεμβολής και αφετέρου, διευκολύνεται ο πυκνός υπολογισμός, αφού μόλις υπολογιστεί η πυκνή οπτική ροή, κάθε σημείο μπορεί να παρακολουθείται χωρίς επιπλέον υπολογιστικό κόστος. Για

τον υπολογισμό της οπτικής ροής υπάρχουν αρκετοί αλγόριθμοι [144], αλλά επιλέγεται ο αλγόριθμος του Farneback ως μια συμβιβαστική λύση στο δίλημμα ταχύτητας-ακρίβειας. Συνοπτικά, ο αλγόριθμος αυτός στηρίζεται στην υπόθεση αναπαράστασης της φωτεινότητας εντός πίξελ με ένα πολυωνυμικό ανάπτυγμα του οποίου οι όροι υπολογίζονται με τη μέθοδο των Ελαχίστων Τετραγώνων. Στη συνέχεια, εφαρμόζεται ένας επαναληπτικός αλγόριθμος ο οποίος ελαχιστοποιεί μια συνάρτηση κόστους. Το υπολογιστικό πλεονέκτημα αυτού του αλγορίθμου είναι η γρήγορη σύγκλιση και η αποδοτική υλοποίηση, οπότε αρκούν λίγες επαναλήψεις και η οπτική ροή υπολογίζεται ανά δύο frames. Ωστόσο, δεν αξιοποιείται βαθύτερη χωροχρονική πληροφορία κι έτσι θυσιάζεται μικρό μέρος της ακρίβειας.

Έχοντας παρακολουθήσει τα σημεία ενδιαφέροντος, τα συνενώνουμε, σχηματίζοντας μια τροχιά. Πειράματα [248] για το μήκος της τροχιάς δείχνουν καλύτερη λειτουργία για τιμή 15 με 20 frames, καθώς μεγαλύτερο μήκος συνεπάγεται μεγαλύτερη πιθανότητα απόκλισης. Το επόμενο βήμα είναι η εξαγωγή των χαρακτηριστικών. Τοπικοί περιγραφητές υπολογίζονται σε έναν τρισδιάστατο χωροχρονικό όγκο γύρω από τα σημεία της τροχιάς, μεγέθους $N \times N$ pixels και L frames, ο οποίος δειγματοληπτείται σε ένα χωροχρονικό πλέγμα $n_\sigma \times n_\sigma \times n_\tau$ προκειμένου να ενσωματώσει δομική πληροφορία στην αναπαράσταση με χαρακτηριστικά. Καλή επιλογή για την τιμή αυτών των παραμέτρων είναι $N = 32$, $n_\sigma = 2$, $n_\tau = 3$, όπως προκύπτει πειραματικά [248]. Οι περιγραφητές που υπολογίζονται είναι οι HOF [128], HOG [49], MBH [50] και ένας Περιγραφητής Τροχιάς. Ο HOG περιγράφει το σχήμα και την εμφάνιση γύρω από κάθε τροχιά, ενώ οι HOF και MBH κωδικοποιούν την κίνηση, με τον τελευταίο μάλιστα να αντισταθμίζει εν μέρει και την κίνηση της κάμερας. Ο Περιγραφητής Τροχιάς κωδικοποιεί απλώς το σχήμα της τροχιάς. Γίνεται επομένως φανερή η συμπληρωματικότητα των περιγραφητών, οι οποίοι κωδικοποιούν διαφορετικά κανάλια πληροφορίας που χαρακτηρίζουν μια δράση. Θα δούμε πώς εκμεταλλευόμαστε το φανινόμενο αυτό στο στάδιο του ταξινομητή, όπου και θα συνδυάσουμε τα διαφορετικά κανάλια. Προς το παρόν, θα αναλύσουμε διεξοδικότερα τη φύση των χαρακτηριστικών που εξάγονται για κάθε τροχιά.

3.2.2 Είδη Περιγραφητών

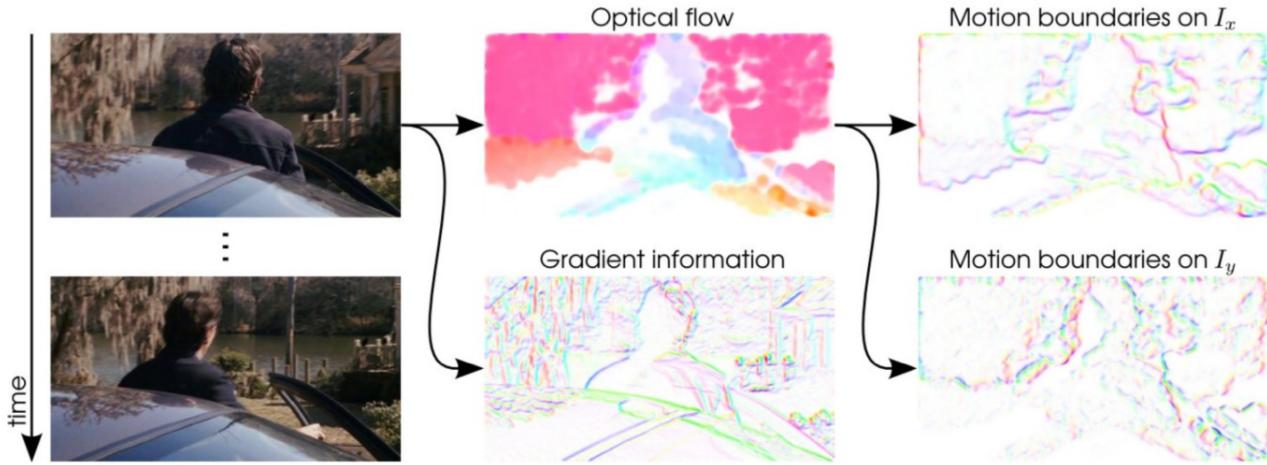
3.2.2.1 Ο Περιγραφητής Τροχιάς

Ο περιγραφητής τροχιάς ορίσθηκε στην αυθεντική εργασία των Πυκνών Τροχιών [248]. Έστω μια τροχιά μήκους L που εκκινεί από το frame t . Είναι φανερό ότι μπορούμε να περιγράψουμε το σχήμα της με την ακολουθία των διανυσμάτων μετατόπισης $\Delta P_t = P_{t+1} - P_t$, η οποία θα έχει τη μορφή:

$$S = (\Delta P_{t+1}, \dots, \Delta P_{t+L-1}) \quad (3.2)$$

Ορίζουμε ως περιγραφητή τροχιάς το διάνυσμα που προκύπτει από την κανονικοποίηση της παραπάνω ακολουθίας με το άθροισμα των μέτρων των διανυσμάτων μετατόπισης:

$$S' = \frac{(\Delta P_{t+1}, \dots, \Delta P_{t+L-1})}{\sum_{j=t}^{t+L-1} \|\Delta P_j\|} \quad (3.3)$$



Σχήμα 3.3: Απεικόνιση της πληροφορίας που περιέχεται στους περιγραφητές HOG, HOF και MBH. Για κάθε εικόνα, τα διανύσματα κλίσης/οπτικής ροής διακρίνονται με χρώμα (hue) και τα αντίστοιχα μέτρα τους με τη χρωματική συνιστώσα κορεσμού (saturation). Τα όρια κίνησης υπολογίζονται ως παράγωγοι των συνιστωσών x και y της οπτικής ροής ξεχωριστά. Σε σύγκριση με την οπτική ροή, τα όρια κίνησης συμπλέζουν το μεγαλύτερο μέρος της κίνησης του υποβάθρου λόγω κινήσεων της κάμερας και τονίζουν την κίνηση προσκηνίου. Σε αντίθεση με την πληροφορία που λαμβάνεται από τις παραγώγους, τα όρια κίνησης σβήνουν περισσότερη πληροφορία υφής από στατικά παρασκήνια. Εικόνα από το [248].

Η χρήση του περιγραφητή τροχιάς είναι σημαντική καθώς μοντελοποιεί το σχήμα, το οποίο κωδικοποιεί χαρακτηριστικά της κίνησης. Μάλιστα, στην ίδια εργασία διαπιστώνεται πειραματικά ότι η χρήση μεταβλητού L , παρότι σαν ιδέα μπορεί να εκφράσει δράσεις με διαφορετικές ταχύτητες, πρακτικά δε βελτιώνει τα αποτελέσματα, οπότε προτιμάται το σταθερό μήκος τροχιάς.

3.2.2.2 Ιστογράμματα Κατευθυνόμενων Παραγώγων

Τα Ιστογράμματα Κατευθυνόμενων Παραγώγων (Histograms of Oriented Gradients, εν συντομίᾳ HOG) εισήχθησαν στο [49] χρησιμοποιούμενα για την αναπαράσταση αντικειμένων στο πρόβλημα της αναγνώρισης αντικειμένων. Η δύναμη του περιγραφητή HOG έγκειται στην αμεταβλητότητά του σε φωτομετρικές και γεωμετρικές μεταβολές, χάρη στην τοπικότητα του υπολογισμού του και των κανονικοποιήσεων που υφίσταται. Η βασική ιδέα είναι η δυνατότητα ικανοποιητικής αναπαράστασης του σχήματος και της εμφάνισης των αντικειμένων σε μια μικρή περιοχή της εικόνας από την κατανομή των τοπικών κλίσεων, δηλαδή των κατευθύνσεων των ακμών στην περιοχή αυτή. Ο υπολογισμός του HOG σε μια γειτονιά περιλαμβάνει τον υπολογισμό τοπικών διακριτών παραγώγων και την κατασκευή των ιστογραμμάτων τους. Περιγράφουμε τη διαδικασία υπολογισμού για εικόνας, ωστόσο είναι φανερό ότι η μέθοδος γενικεύεται σε βίντεο, αν τα θεωρήσουμε ως ένα χωροχρονικό όγκο και μελετήσουμε γειτονιές τριών διαστάσεων.

Το πρώτο βήμα της μεθόδου είναι η εφαρμογή του εκθετικού κανόνα (γάμμα διόρθωση) για κωδικοποίηση της φωτεινότητας. Ακολουθεί ο υπολογισμός των παραγώγων της εικόνας με συνέλιξη της εικόνας με τους τελεστές $[-1, 0, 1]$ και $[-1, 0, 1]^T$, μια εύρωστη παραλλαγή της μεθόδου πεπερασμένων διαφορών. Η εικόνα χωρίζεται

σε κελιά διάστασης 8×8 pixels συνήθως και υπολογίζεται το ιστόγραμμα των διεύθυνσεων της κλίσης, σταθμισμένο με το αντίστοιχο μέτρο της κλίσης, για τα pixel εντός κάθε κελιού. Κάθε θέση του ιστογράμματος αντιστοιχεί σε μια στάθμη κβάντισης των διευθύνσεων κλίσεων κι έτσι σε ένα μικρό εύρος διευθύνσεων, ανάλογα με το πλήθος των στάθμεων. Τα κελιά ομαδοποιούνται σε μπλοκ (συνήθως 2×2 κελιών) και τα ιστογράμματα των επιμέρους κελιών συνενώνονται και κανονικοποιούνται σε έναν ενιαίο περιγραφητή. Η συνένωση γίνεται για τη μετρίαση ή μεταφορά της τοπικής σημαντικότητας των pixels σε επίπεδο κελιού στη συνολική σημαντικότητα σε επίπεδο μπλοκ εικόνας, ενώ η κανονικοποίηση γίνεται ώστε να αντισταθμιστούν αλλαγές στη φωτεινότητα και την αντίθεση της εικόνας. Τελικά μια εικόνα περιγράφεται από τη συνένωση των περιγραφητών των μπλοκ που την αποτελούν. Αντίστοιχα σε ένα βίντεο, τα κελιά και τα μπλοκ εκτείνονται σε τρεις διαστάσεις και η λογική είναι παρόμοια.

3.2.2.3 Ιστογράμματα Οπτικής Ροής

Τα ιστογράμματα οπτικής ροής (Histograms of Optical Flow, εν συντομίᾳ HOF) εισήχθησαν στο [128] ως χαρακτηριστικά χρήσιμα στην αναγνώριση δράσεων. Ο περιγραφητής HOF μοντελοποιεί την κίνηση σε ένα μικρό χωροχρονικό όγκο βίντεο με φιλοσοφία παρόμοια με του HOG και χρησιμοποιείται ευρέως σε προβλήματα ανάλυσης και κωδικοποίησης της κίνησης σε βίντεο. Ωστόσο, η διακριτική του ικανότητα εξαρτάται άμεσα από την ακρίβεια της οπτικής ροής. Στην περίπτωση των Πυκνών Τροχιών, ο υπολογισμός των περιγραφητών συνδυάζεται κομψά με τον υπολογισμό της καθαυτής τροχιάς: η οπτική ροή έχει ήδη υπολογιστεί, οπότε ο ίδιος υπολογισμός χρησιμοποιείται στην εξαγωγή των HOF χαρακτηριστικών.

Από άποψη υλοποίησης, η πορεία μοιάζει αρκετά με αυτή για τον υπολογισμό του HOG. Συγκεκριμένα, υπολογίζεται αρχικά η οπτική ροή ως προς τους δύο άξονες και κατόπιν το μέτρο και η διεύθυνσή της. Η διεύθυνση κβαντίζεται σε στάθμες με την τελευταία από αυτές να προορίζεται για τα pixels εκείνα στα οποία η οπτική ροή είναι κατά μέτρο μικρότερη από κάποιο κατώφλι. Ακολουθεί, σε πλήρη αντιστοιχία με τον περιγραφητή HOG, η υποδιαίρεση της περιοχής ενδιαφέροντος σε κελιά, για κάθε ένα από τα οποία κατασκευάζεται το ιστόγραμμα των διεύθυνσεων της οπτικής ροής. Τα επιμέρους ιστογράμματα ενώνονται και κανονικοποιούνται σχηματίζοντας το τελικό διάνυσμα που περιγράφει την κίνηση εντός του αρχικού όγκου. Προφανώς εδώ γίνεται αντιληπτό το γεγονός ότι ο HOF είναι εγγενώς τρισδιάστατος περιγραφητής, αφού κατασκευάζεται με μεγέθη που αφορούν βίντεο (οπτική ροή).

3.2.2.4 Ιστογράμματα Ορίων Κίνησης

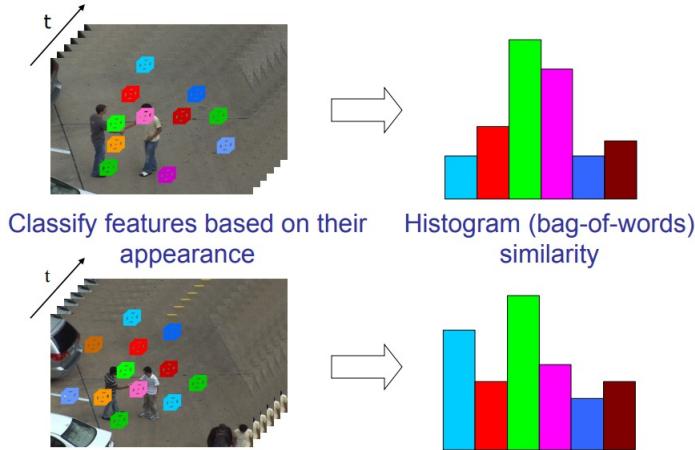
Τα ιστογράμματα ορίων κίνησης (Motion Boundary Histograms, εν συντομίᾳ MBH), εισήχθησαν στο [50] με βλέψεις αντιστάθμισης της επίπτωσης της κίνησης της κάμερας στη διακριτική ικανότητα άλλων περιγραφητών που χρησιμοποιούνται για την περιγραφή της κίνησης, όπως ο μεταγενέστερος HOF, οι οποίοι κωδικοποιούν την κίνηση του παρασκηνίου εισάγοντας σημαντικό θόρυβο στην αποτύπωση της πραγματικής κίνησης ενδιαφέροντος. Η ιδέα του MBH είναι η περιγραφή της κίνησης μέσω της παραγώγου της οπτικής ροής προκειμένου να αντισταθμιστεί η κίνηση της κάμερας, εκμεταλλευόμενος την ομοιομορφία της κίνησης λόγω των μεταθέσεων κάμερας ένοντι στην τυχαιότητα και τις απότομες μεταβολές που εισάγει μια κίνηση προσκηνίου.

Ο υπολογισμός του MBH ξεκινά και πάλι από τον υπολογισμό της οπτικής ροής, που πάλι είναι ήδη υπολογισμένη στην περίπτωση της μεθόδου Πυκνών Τροχιών. Το πεδίο οπτικής ροής χωρίζεται σε οριζόντιο και κατακόρυφο μέρος και υπολογίζεται η παράγωγος για καθένα από αυτά. Η συνέχεια είναι παρόμοια με τη διαδικασία υπολογισμού των HOG και HOF: η κλίση κβαντίζεται σε στάθμες και εντός ενός κελιού κατασκευάζεται το ιστόγραμμα των κλίσεων σταθμισμένων με το μέτρο τους. Οι δύο περιγραφητές που προκύπτουν (MBH_x και MBH_y) κανονικοποιούνται ξεχωριστά και συχνά χρησιμοποιούνται ξεχωριστά, αλλά και σε ενοποιημένη μορφή. Ο MBH είναι αρκετά εύρωστος στην κίνηση της κάμερας, αφού όταν η κίνηση είναι ομαλή, η οπτική ροή περιέχει μια σταθερή συνιστώσα, η οποία αφαιρείται κατά τον υπολογισμό της παραγώγου. Έτσι, κωδικοποιούνται μόνο οι μεταβολές στο πεδίο της οπτικής ροής.

3.2.3 Κωδικοποίηση Χαρακτηριστικών

Η χρήση των σημείων ενδιαφέροντος εισάγει μια ιδιαιτερότητα στην αναπαράσταση των εικόνων-βίντεο με χαρακτηριστικά: κάθε εικόνα έχει διαφορετικό πλήθος σημείων ενδιαφέροντος κι έτσι διαφορετικό μήκος διανύσματος αναπαράστασης. Τα πράγματα γίνονται ακόμα χειρότερα όταν εισάγονται πολλαπλοί ταξινομητές, καθώς αφενός το μήκος των διανυσμάτων αναπαράστασης ενδέχεται να αλλάζει ακόμα και για την ίδια εικόνα λόγω της κατασκευής του κάθε περιγραφητή, αφετέρου οι τιμές των χαρακτηριστικών περιέχουν τελείως διαφορετική και ανεξάρτητη πληροφορία μεταξύ τους, οπότε η ενιαία αναπαράσταση και σύγκριση δεν έχει κάποια ερμηνεία. Έτσι, παρόλο που η συμπληρωματική φύση των περιγραφητών ωφελεί τα χαρακτηριστικά και αυξάνει τη διαχωρισμό την των κλάσεων, απαιτείται κωδικοποίηση για την αποτελεσματική τους χρήση. Μια επιτυχής λύση στο πρόβλημα αυτό είναι η κωδικοποίηση με τη μέθοδο του «σάκου λέξεων» (bag-of-words). Η ιδέα είναι εμπνευσμένη από την κατηγοριοποίηση κεμένου, ενός χώρου όπου το πρόβλημα είχε εμφανιστεί εγγενώς στο παρελθόν. Εκεί το πρόβλημα είναι η μοντελοποίηση προτάσεων διαφορετικού μήκους με έναν ενιαίο τρόπο. Τελικά, αντί να αναπαριστούμε τις προτάσεις με τις λέξεις τους, τις αναπαριστούμε με την κατανομή των λέξεων και μάλιστα κατασκευάζουμε ένα λεξικό μέτις τις λέξεις των προτάσεων του συνόλου εκπαίδευσης και μετράμε την εμφάνιση ή μη αυτών των λέξεων σε κάθε νέα πρόταση. Φυσικά υπάρχουν και πιο σύνθετες αναπαραστάσεις [111], αλλά αυτή που περιγράφηκε είναι η πιο σχετική με τη λύση στα προβλήματα Όρασης Υπολογιστών. Θα δούμε τώρα πώς, συσχετίζοντας τα προβλήματα αναπαράστασης γλωσσικών προτάσεων και βίντεο, μπορούμε να εξάγουμε ανάλογη λύση.

Μπορούμε να σκεφτούμε μια εικόνα ή ένα βίντεο σαν μια οπτική πρόταση, σε αναλογία με την γλωσσική πρόταση. Η αναλογία αυτή διατηρεί την ποικιλομορφία και το μεταβλητό μέγεθος, ωστόσο υπάρχει ένα λεπτό σημείο που πρέπει να διευκρινιστεί για να είναι η αναλογία έγκυρη: το σύνολο των λέξεων που σχηματίζουν κάθε πρόταση. Από τη μία μεριά το σύνολο των γλωσσικών λέξεων είναι αριθμήσιμο (μπορούμε να σκεφτούμε ότι κάθε λέξη είναι μια μετάθεση ενός συνδυασμού των γραμμάτων του αλφαριθμητού), ενώ, από την άλλη, δεν έχουμε ορίσει σύνολο οπτικών λέξεων. Η ιδέα να ορίσουμε ως λέξη ένα ιστόγραμμα χαρακτηριστικών απορρίπτεται καθώς οι τιμές των οπτικών χαρακτηριστικών απεικονίζονται σε έναν συνεχή χώρο, την στιγμή που ζητάμε ένα αριθμήσιμο σύνολο. Το πρώτο βήμα επομένως, είναι η κβάντιση του συνόλου οπτικών χαρακτηριστικών για τη δημιουργία ενός οπτικού λεξικού. Συνήθως αυτή η διαδικασία εφαρμόζεται χωριστά σε κάθε περιγραφητή, αφού, όπως



Σχήμα 3.4: Απεικόνιση της μεθόδου του Σάκου Οπτικών Λέξεων (Bag of Visual Words, εν συντομίᾳ BoW). Οι διαφορετικοί τύποι χαρακτηριστικών εξάγονται σε σημεία ενδιαφέροντος στο βίντεο. Στη συνέχεια κωδικοποιούνται για να σχηματίσουν ένα λεξιλόγιο, τον Σάκο Λέξεων. Ένα βίντεο τώρα αναπαρίσταται σαν μια πρόταση από οπτικές λέξεις και ταξινομείται σύμφωνα με το ιστογραμμα των λέξεων αυτών. Μια απλή περίπτωση είναι η ταξινόμηση απευθείας σύμφωνα με τα ιστογράμματα αυτά. Στην πορεία αυτής της εργασίας θα δούμε αρκετά πιο σύνθετες μεθόδους. Εικόνα από το [47].

εξηγήσαμε, η σύγκριση τιμών διαφορετικών περιγραφητών δεν έχει καμία φυσική έννοια.

Το οπτικό λεξικό για κάθε περιγραφητή θα προκύψει από έναν αλγόριθμο ομαδοποίησης. Η μέθοδος των Πυκνών Τροχιών, όπως προτάθηκε στην αυθεντική της μορφή, υπολογίζει λεξικά μεγέθους 4000 λέξεων για κάθε περιγραφητή που χρησιμοποιεί, τρέχοντας τον αλγόριθμο k-μέσων (k-means) [228]. Ο k-means είναι ίσως ο δημοφιλέστερος αλγόριθμος διανυσματικής κβάντισης και χρησιμοποιείται για να διαιρέσει τον αρχικό χώρο X των χαρακτηριστικών σε k περιοχές, κάθε μια από τις οποίες αντιπροσωπεύεται από ένα διάνυσμα $d_i \in D = d_1, \dots, d_k$, όπου D τα κεντροειδή του k-means, δηλαδή το οπτικό λεξικό. Για τον υπολογισμό των κέντρων D , ο k-means εκκινεί από αρχικοποιημένες τιμές k κέντρων $m_1^{(1)}, \dots, m_k^{(1)}$ και επαναλαμβάνει τα εξής δύο βήματα:

$$\text{Assignment Step: } S_i^{(t)} = \{x_p : \|x_p - m_i^{(t)}\| \leq \|x_p - m_j^{(t)}\| \forall j, 1 \leq j \leq k\} \quad (3.4)$$

και

$$\text{Update Step: } m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j \quad (3.5)$$

Οι επαναλήψεις σταματούν όταν δεν υπάρχει άλλη ανανέωση. Με άλλα λόγια, λύνεται ένα πρόβλημα ομαδοποίησης όπου κάθε διάνυσμα αντιστοιχίζεται στην κλάση του κοντινότερού του μέσου. Η προσθήκη ενός νέου διάνυσματος σε μια κλάση επηρεάζει το μέσο διάνυσμα της κλάσης αυτής (και της κλάσης από την οποία αφαιρείται), οπότε επανεκτιμούμε την ανάθεση των υπολοίπων διανυσμάτων σε αυτή τη νέα

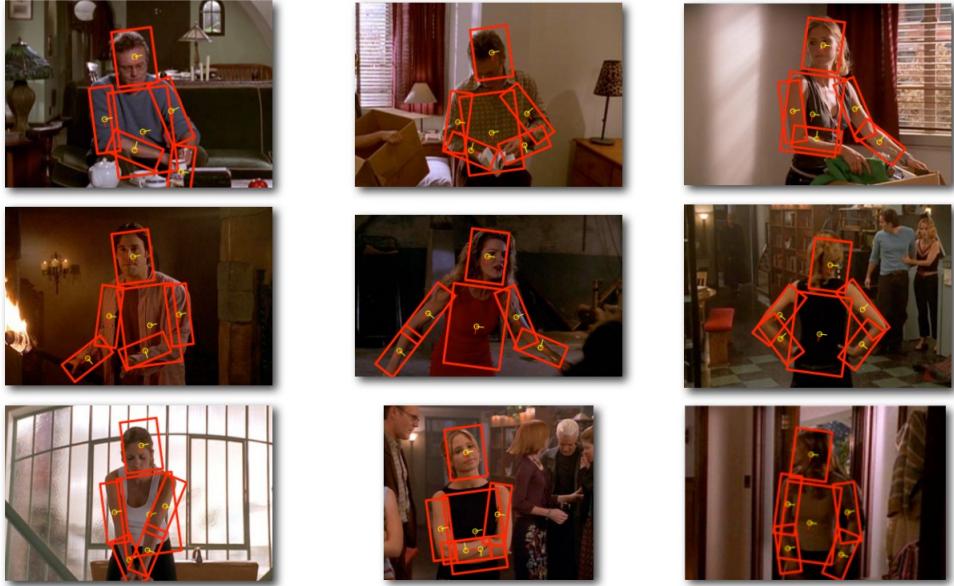
κλάση. Αναφέρεται βέβαια ότι ο αλγόριθμος δεν εγγυάται βέλτιστη διαμέριση του χώρου.

Έχοντας κατασκευάσει το οπτικό λεξικό, πρέπει να κωδικοποιήσουμε τα οπτικά χαρακτηριστικά ώστε να τα μετατρέψουμε σε οπτικές λέξεις. Η διαδικασία εδώ είναι απλή: κάθε χαρακτηριστικό αντιστοιχίζεται στην λέξη που αντιστοιχεί στο κοντινότερο (υπό κάποια μετρική απόστασης, στη μέθοδο Πυκνών Τροχιών χρησιμοποιείται η Ευκλείδεια απόσταση) κεντροειδές του οπτικού λεξικού. Έτσι το πρόβλημα διακριτότητας των λέξεων έχει λυθεί. Όπως στην περίπτωση των γλωσσικών προτάσεων, έτσι και μια οπτική πρόταση αντιπροσωπεύεται από την κατανομή (εμφάνιση ή μη και συχνότητα) των λέξεων που την αποτελούν, πάνω στο, κατασκευασμένο κατά τη διάρκεια εκπαίδευσης, οπτικό λεξικό, δηλαδή λαμβάνουμε το ιστόγραμμα των λέξεων της πρότασης. Στην περίπτωση των βίντεο, αυτή η αναπαράσταση συμβαίνει σε κάθε frame και για την αναπαράσταση όλου του τμήματος βίντεο λαμβάνουμε το αθροιστικό ιστόγραμμα πάνω στα επιμέρους frames που το αποτελούν. Ωστόσο αυτή η επιλογή παρουσιάζει δύο ακόμα προκλήσεις. Αρχικά, μια πρόταση έχει μεταβλητό μέγεθος. Επομένως τα ιστογράμματα θα εμφανίζουν μεγάλες διακυμάνσεις στις τιμές τους ακόμα και στην αναπαράσταση της ίδιας δράσης. Μας ενδιαφέρει η κατανομή κι όχι το μήκος της πρότασης, οπότε λαμβάνουμε τα κανονικοποιημένα ιστογράμματα οπτικών λέξεων. Τώρα η αναπαράσταση είναι ενιαία και ανεξάρτητη του μήκους βίντεο ή γλωσσικής πρότασης. Η δεύτερη πρόκληση αφορά την απαλοιφή της χωροχρονικής συσχέτισης μεταξύ των λέξεων. Πράγματι, τόσο στην περίπτωση των γλωσσικών όσο και των οπτικών προτάσεων, η εξαγωγή του ιστογράμματος βασίζεται στην εμφάνιση και μόνο μιας λέξης κι όχι τη χωρική της σχέση με τις υπόλοιπες λέξεις. Παρότι υπάρχουν μέθοδοι που μπορούν να αποδώσουν κάποια χωρική σχέση (όπως τα n-grams στην περίπτωση των γλωσσικών προτάσεων), η απόδοση των μέθοδων BoW είναι υψηλή, οπότε συνεχίζουν να χρησιμοποιούνται.

3.2.4 Επιπλέον Χαρακτηριστικά: Τα Χαρακτηριστικά Πόζας

Στο [191] εισάγεται μια επέκταση της μεθόδου Πυκνών Τροχιών, η οποία λαμβάνει υπόψιν της περισσότερα χαρακτηριστικά. Οι συγγραφείς χρησιμοποιούν επιπλέον χαρακτηριστικά της στάσης του άνω τμήματος του σώματος. Το κίνητρο πίσω από αυτό το εγχείρημα είναι η αίσθηση ότι οι λεπτομερείς δραστηριότητες αφορούν κυρίως τα χέρια και τις κινήσεις των άνω μερών του σώματος. Σε πρώτη φάση εξάγεται η στάση του σώματος και υπολογίζονται οι τροχιές στάσης στα βίντεο. Γύρω από τις τροχιές εξάγονται χαρακτηριστικά πόζας τα οποία κωδικοποιούνται και μπορούν να αναπαραστήσουν τη δράση. Περιγράφουμε τώρα αναλυτικότερα τα επιμέρους βήματα του αλγορίθμου.

Το πρώτο βήμα είναι η εκτίμηση της στάσης του σώματος. Για το σκοπό αυτό, οι συγγραφείς χρησιμοποιούν τον εκτιμητή πόζας του [5]. Η εργασία αυτή βασίζεται στις Εικονικές Δομές (Pictorial Structures, εν συντομίᾳ PS) [71] και σε ισχυρούς ανιχνευτές τμημάτων, όπως των [156] και [244]. Πιο συγκεκριμένα, γίνεται χρήση περιγραφητών σχήματος [157] για την εξαγωγή πυκνών αναπαραστάσεων και εκπαιδεύονται ταξινομητές παραμορφώσιμων τμημάτων με τον αλγόριθμο AdaBoost [77]. Οι έξοδοι των ταξινομητών αυτών συνδυάζονται με ένα μοντέλο PS, σύμφωνα με τα σκορ του ταξινομητή. Καθώς ο αλγόριθμος είναι γενικός και απευθύνεται σε οποιοδήποτε σύνολο δεδομένων, οι συγγραφείς του [191] εκπαιδεύουν ένα μοντέλο αποκλειστικά για το



Σχήμα 3.5: Μερικά παραδείγματα επιτυχούς ανίχνευσης στάσης του άνω ήμισυ του σώματος. Η στάση μπορεί να σχετίζεται άμεσα με την δράση. Ιδιαίτερα η θέση και ο σχηματισμός των χεριών αποτελούν σημαντικό διακριτικό στοιχείο των λεπτομερών δράστηριοτήτων. Εικόνα από το [5].

δικό τους σύνολο δεδομένων. Ο αλγόριθμος τροποποιείται ελαφρά ώστε να χρησιμοποιεί 10 αντί για 6 τμήματα του σώματος. Τα αποτελέσματα εκτίμησης πόζας βελτιώνονται αισθητά μετά τη νέα εκπαίδευση, καθώς το μοντέλο προσαρμόζεται στο σύνολο δεδομένων.

Έχοντας υπολογίσει την πόζα με τον παραπάνω αλγόριθμο, χρειαζόμαστε χρονική πληροφορία, δηλαδή την τροχιά των συνδέσμων του σώματος. Για εξοικονόμηση υπολογισμών, εκμεταλλεύμενοι την πυκνότητα του συνόλου δεδομένων μπορούμε να υπολογίσουμε την στάση του σώματος ανά 10 frames και να την παρακολουθήσουμε σε ένα χρονικό πλαίσιο 100 frames με κέντρο το κάθε ένα από τα frames στα οποία έχει γίνει η εκτίμηση. Για την παρακολούθηση (tracking), χρησιμοποιείται ο αλγόριθμος SIFT [140]. Η παρακολούθηση γίνεται για κάθε σύνδεσμο ξεχωριστά και τα αποτελέσματα των διαφορετικών κέντρων παρακολούθησης συνεκτιμώνται. Μάλιστα γίνεται επιλογή των χαρακτηριστικών που παρουσιάζουν τη μέγιστη συνεκτική κίνηση για κάθε σύνδεσμο και η θέση ανανεώνεται μόνο σύμφωνα με αυτά τα χαρακτηριστικά για την αποκοπή outliers. Το πλεονέκτημα αυτής της προσέγγισης είναι ο συνδυασμός ενός γενικού μοντέλου εμφάνισης με ένα μοντέλο ειδικής εμφάνισης όπως αυτό που παρακολουθεί ο αλγόριθμος SIFT.

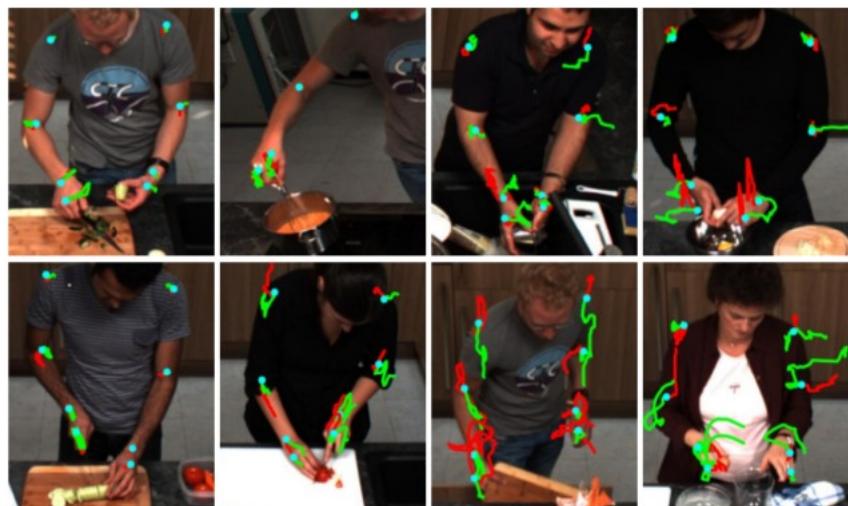
Η εξαγωγή των τροχιών στάσης σώματος ανοίγει το δρόμο για τον υπολογισμό των χαρακτηριστικών πόζας. Πρώτα χαρακτηριστικά είναι αυτά του Μοντέλου Σώματος (Body Model, εν συντομίᾳ BM). Η διαδικασία υπολογισμού αυτών των χαρακτηριστικών εκκινεί με τον υπολογισμό των ταχυτήτων και των επιταχύνσεων για όλους τους συνδέσμους και την κβάντισή τους σε ιστογράμματα 8 θέσεων. Ο διαχωρισμός γίνεται με βάση την κατεύθυνση του διανύσματος σταθμισμένη με τον αριθμό των pixels ανά frame. Επιπλέον, υπολογίζονται οι αποστάσεις μεταξύ των αντίστοιχων συνδέσμων (αριστερού-δεξιού) και των διαφόρων κατηγοριών συνδέσμων του άνω μέρους του σώματος μεταξύ τους. Για κάθε μία από τις τροχιές αυτών των αποστάσεων υπολογίζονται στατιστικά (μέση τιμή, διάμεσο, τυπική απόκλιση, μέγιστο και

ελάχιστο) και ιστογράμματα παρόμοια με αυτά της ταχύτητας και της επιτάχυνσης. Τέλος, υπολογίζονται οι τροχιές και γωνιών και των γωνιακών ταχυτήτων όλων των εσωτερικών σημείων (καρπών, αγκώνων και ώμων) και γίνεται εξαγωγή στατιστικών για αυτές. Δεύτερα χαρακτηριστικά που εξάγονται γύρω από τις τροχιές πόζας είναι τα χαρακτηριστικά FFT, εμπνευσμένα από την επιτυχία τους στο [272]. Τα χαρακτηριστικά αυτά περιέχουν 4 εκθετικές ζώνες, 10 συντελεστές Cepstrum και την φοσματική εντροπία και ενέργεια για κάθε συντεταγμένη x και y των τροχιών όλων των συνδέσμων.

Παρότι η αυθεντική μέθοδος των Πυκνών Τροχιών απορρίπτει το μεταβλητό μήκος τροχιάς λόγω της χαμηλής του προσφοράς στο τελικό αποτέλεσμα, η ιδέα εφαρμόζεται τελικά στα χαρακτηριστικά πόζας. Συγκεκριμένα, κάθε είδος χαρακτηριστικού (BM, FFT) κωδικοποιείται χειχωριστά. Για τρία διαφορετικά μήκη τροχιάς (20, 50, 100 frames) με κέντρο το frame όπου έγινε η ανίχνευση εξάγονται το χαρακτηριστικά και συνενώνονται για να αναπαραστήσουν με μεγαλύτερη επιτυχία δράσεις με διαφορετικό μήκος. Η κωδικοποίηση γίνεται με μέγεθος λεξιλογίου διπλάσιο του μήκους κάθε διανύσματος χαρακτηριστικών για κάθε μία από τις δύο κατηγορίες χαρακτηριστικών. Με αυτόν τον τρόπο, έχουμε αναπαραστήσει κάθε δράση και με κίνηση συνδέσμων και στάση σώματος.

3.3 Προτεινόμενο Σύστημα

Συνοψίζοντας όσα παρουσιάσαμε παραπάνω, η προσέγγισή μας συνδυάζει όλη την πορεία από την παρακολούθηση του βίντεο και την εξαγωγή χαρακτηριστικών μέχρι την κωδικοποίησή τους σε ιστογράμματα. Διατηρήσαμε τις ήδη ορισμένες (default) τιμές παραμέτρων όπου χρειάστηκε. Συγκεκριμένα, σε πρώτη φάση εξαγάγαμε τα χαρακτηριστικά HOG, HOF, MBH (σε ενιαία, μετρική μορφή) και Περιγραφητή Τροχιάς, καθώς και τα χαρακτηριστικά πόζας BM και FFT. Στη φάση της κωδικοποίησης, τρέξαμε τον αλγόριθμο k-μέσων για τη δημιουργία λεξιλογίων (codebooks) πλήθους 4000 για τα χαρακτηριστικά των τροχιών (HOG, HOF, MBH, Περιγραφητή Τροχιάς), 3336 για τα χαρακτηριστικά BM και 1536 για τα χαρακτηριστικά 1536 FFT. Αφού λάβαμε τα ιστογράμματα των χαρακτηριστικών ελέγχαμε ποικίλους τρόπους συνδυασμού πριν την ταξινόηση, ώστε να αποκομίσουμε τα μέγιστα οφέλη από τη συμπληρωματική φύση των χαρακτηριστικών. Αφήνουμε όμως τα πειράματα αυτά για το κεφάλαιο 6, αφού θα έχουμε εξάγει τα ιστογράμματα από τα υπόλοιπα κανάλια πληροφορίας.



Σχήμα 3.6: Μερικά παραδείγματα εξαγωγής εμπροσθόδρομων και οπισθόδρομων τροχιών των 10 τμημάτων του άνω μισού μέρους του σώματος κατά την παρακολούθησή τους στην εξαγωγή των χαρακτηριστικών πόζας. Με πράσινο απεικονίζονται οι οπισθόδρομες μετακινήσεις και με κόκκινο οι εμπροσθόδρομες. Με κυανό σημειώνεται η αρχική θέση του κάθε συνδέσμου. Εικόνα από το [191].

Κεφάλαιο 4

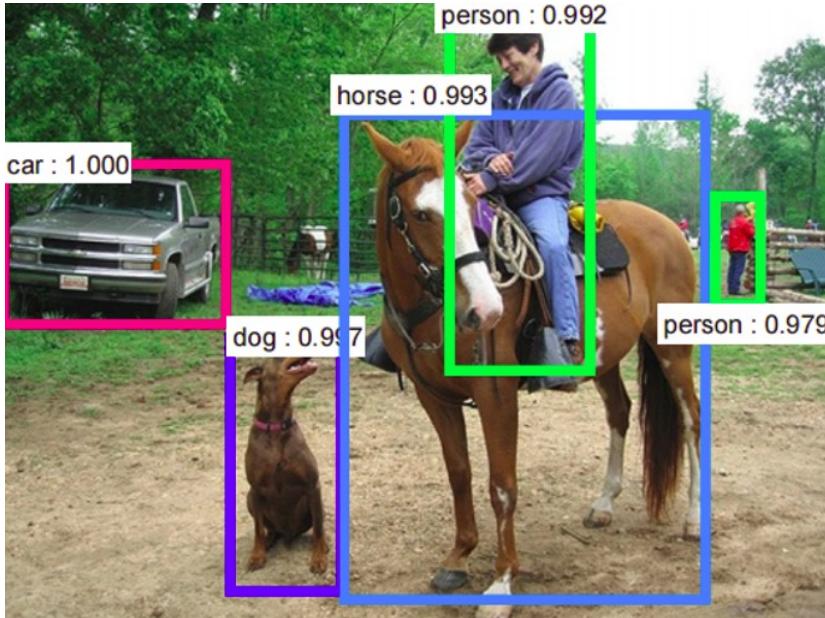
Σημασιολογικό Υποσύστημα: Αντικείμενα

Στο κεφάλαιο 2 είδαμε τη σημασία της σημασιολογίας στην αναγνώριση δράσεων. Παρότι οι προσεγγίσεις που βασίζονται σε αυτή είναι συχνά εξαρτημένες από συνύπαρξη με μεθόδους χαρακτηριστικών χαμηλού επιπέδου, η συνεισφορά τους με την υψηλού επιπέδου αναπαράσταση των δράσεων στην εκτίμηση αυτών είναι καίρια. Σε αυτή την εργασία αναζητάμε σημασιολογικά χαρακτηριστικά μέσω αντικειμένων και τύπων λαβής. Σε αυτό το κεφάλαιο ασχολούμαστε μόνο με τη σημασία των αντικειμένων και αφήνουμε για επόμενο κεφάλαιο την ανάλυση των τύπων λαβής. Η οργάνωση του κεφαλαίου έχει ως εξής: αρχίζουμε με ιστορική αναφορά στις μεθόδους ανίχνευσης αντικειμένων και εύρεσης προσκηνίου. Ακολουθεί το θεωρητικό υπόβαθρο των μεθόδων που αξιοποιούμε (Ταίριασμα Προτύπων, εξαγωγή προσκηνίου με Γκαουσιανά Μοντέλα Μίξης, ανιχνευτής ανθρώπων με χρήση Μοντέλων Παραμορφώσιμων Τμημάτων) και αναλύουμε εκτενώς τις σχεδιαστικές μας επιλογές. Κλείνουμε με ένα σχόλιο πάνω στη θαυμαστή συνθετότητα της ανθρώπινης αντίληψης στο πρόβλημα της ανίχνευσης και ταυτοποίησης αντικειμένων.

4.1 Ανίχνευση Αντικειμένων σε Εικόνες

4.1.1 Γενικά

Η ανίχνευση αντικειμένων είναι ένα σημαντικό και δύσκολο πρόβλημα της Όρασης Υπολογιστών. Όπως δηλώνει το όνομά της, το ζήτημα είναι ο χωρικός εντοπισμός της εμφάνισης μιας κλάσης μέσα σε μια εικόνα και μπορεί να διατυπωθεί μαθηματικά ως εξής: Δοθείσας μιας εικόνας και μιας κατηγορίας αντικειμένων, να βρεθεί η ελάχιστη σε εμβαδόν περιοχή, η οποία περικλείει όλο το αντικείμενο. Η περιοχή αυτή μπορεί να έχει αυθαίρετο σχήμα, να ταυτίζεται με το σχήμα του αντικειμένου, όπως στην περίπτωση της κατάτμησης, αλλά συχνά επιλέγεται το ελάχιστο ορθογώνιο με την παραπάνω ιδιότητα. Οπότε ένα σύστημα ανίχνευσης αντικειμένου πρέπει να λάβει δύο αποφάσεις: αφενός να αποφασίσει για την εμφάνιση ή μη του ζητούμενου αντικειμένου στην εικόνα και αφετέρου να επιστρέψει την περιοχή εμφάνισης.



Σχήμα 4.1: Παράδειγμα εξόδου συστήματος ανίχνευσης αντικειμένων. Βλέπουμε ότι το σύστημα μας επιστρέφει το ορθογώνιο στο οποίο εντοπίζεται το ζητούμενο αντικείμενο. Στα ορθογώνια επισημειώνεται επιπλέον το σκορ πεποίθησης του συστήματος για την ανίχνευση και ταυτοποίηση στην οποία προέβη. Εικόνα από το [186].

Η τελευταία είναι και η ειδοποιός διαφορά από τα συστήματα ταυτοποίησης αντικειμένων τα οποία επιστρέφουν ως έξοδο μια προκαθορισμένη αριθμητική ή λογική τιμή ή ένα κατηγόρημα.

Η σημασία της δυνατότητας ανίχνευσης αντικειμένων είναι σπουδαία, τόσο για ένα φυσικό ον, όπως ο άνθρωπος, αλλά και για εφαρμογές Όρασης Υπολογιστών. Θα σταθούμε κυρίως στη δεύτερη κατηγορία αλλά θα δούμε και περιπτώσεις που η ανθρώπινη ανάγκη εμπνέει την Τεχνητή Νοημοσύνη. Ξεκινάμε από εφαρμογές όπου η ανίχνευση αντικειμένων είναι αυτοσκοπός, όπως τα συστήματα επίβλεψης και ασφαλείας χώρων ή αυτόματου εντοπισμού αστεροειδών σε εικόνες διαστημικών τηλεσκοπίων και γενικά βιομηχανικά συστήματα [90]. Περαιτέρω, η ανίχνευση αντικειμένων μπορεί να υποβοηθήσει την κατηγοριοποίησή τους, ενώ μπορεί ακόμα να συνεισφέρει στη σημασιολογική αντίληψη ενός χώρου. Τέλος, ο άνθρωπος χρειάζεται να εντοπίζει αντικείμενα για να αλληλεπιδράσει μαζί τους ή να λάβει αποφάσεις. Ένα ρομπότ που θα μπορεί ιδανικά να εντοπίζει αντικείμενα στο χώρο κίνησής του θα είναι σε θέση να υποβοηθήσει τον άνθρωπο σε πτυχές της καθημερινότητάς του [135].

Από την άλλη, η ανίχνευση αντικειμένων είναι ένα δύσκολο καθήκον για τους υπολογιστές, παρότι ταυτόχρονα είναι κάτι τετριμμένο για τους ανθρώπους [173]. Υπάρχουν πολλαπλά αίτια για αυτό το φαινόμενο. Θέματα όπως η αμεταβλητότητα στους γεωμετρικούς μετασχηματισμούς, στον φωτισμό, στις διαφορετικές όψεις του αντικειμένου, στις διαφορετικές αποστάσεις κάμερας και στις μερικές επικαλύψεις έχουν σε ικανοποιητικό βαθμό προβλεφθεί και μοντελοποιηθεί. Υπάρχουν όμως εγγενείς δυσκολίες στην ανίχνευση αντικειμένων, όπως η σημαντική ενδομεταβλητότητα μεταξύ αντικειμένων της ίδιας κλάσης, οι παραμορφώσεις των αντικειμένων και η αστάθεια στη φυσική τους υπόσταση (π.χ. ένα κρεμμύδι μπορεί να είναι ωμό, ακέραιο, μαγειρεμένο, τεμαχισμένο, τσιγαρισμένο και πάλι αντιπροσωπεύει την ίδια κλάση). Τέλος, ένα σύστημα οπτικής ανίχνευσης αντικειμένων χάνει τη σημασιολογική πληροφορία που οι άνθρωποι χρησιμοποιούν για να απαντήσουν στο ερώτημα εντοπισμού.

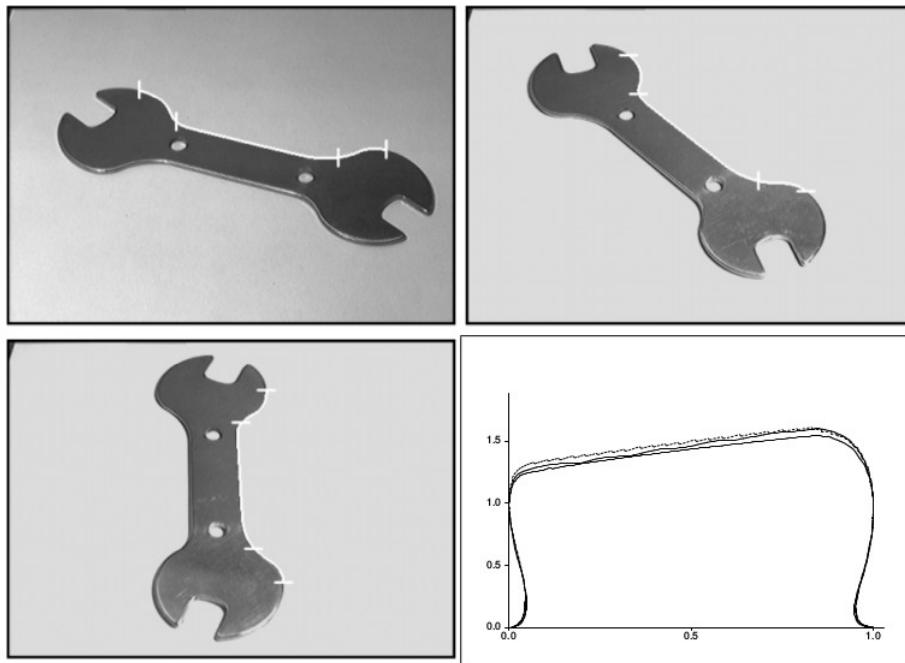
Το ζήτημα αυτό αντιμετωπίζεται μερικώς με την αποδόμηση του αντικειμένου σε μικρότερες οντότητες, ωστόσο, συχνά το περιβάλλον και η χρήση είναι καταλληλότεροι παράγοντες απάντησης αλλά μοντελοποιούνται εξαιρετικά δύσκολα. Το σύνολο δεδομένων που χρησιμοποιούμε για τα πειράματά μας θα μας επιτρέψει στη συνέχεια να διαπιστώσουμε ορισμένους παράγοντες ανθρώπινης αντίληψης που σχετίζονται με την ανίχνευση αντικειμένων.

4.1.2 Ιστορικά Στοιχεία

Η σημασία του προβλήματος της ανίχνευσης αντικειμένων, σε συνδυασμό με τη δύσκολία επίλυσής του, ώθησαν πολυάριθμες και μακροχρόνιες ενασχολήσεις με διαφορετικές προσεγγίσεις και αποτελέσματα. Η ιστορία αυτών των προσπαθειών ξεκινά ήδη από το 1965, όταν ο Roberts [187] δημοσιεύει την πρωτοπόρα εργασία του, η οποία σηματοδοτεί την εποχή των γεωμετρικών προσεγγίσεων και της ευθυγράμμισης. Η γενική φιλοσοφία είναι η εύρεση σημείων τα οποία μας βοηθούν να αποφασίσουμε τον βέλτιστο μετασχηματισμό, συνήθως περιστροφή, μετάθεση και κλιμάκωση, ο οποίος θα ταυτίσει το πρότυπο που έχουμε για το αντικείμενο με αυτό που περιέχεται στην εικόνα [99]. Η καθοριστική παραδοχή που γίνεται είναι η αμεταβλητότητα του αντικειμένου, το οποίο μπορεί να αναπαραστεί επιτυχώς από ένα προκαθορισμένο πρότυπο. Στο [83] χρησιμοποιούνται γεωμετρικά μοντέλα για την εύρεση μερικώς επικαλυπτόμενων αντικειμένων. Μάλιστα δίνεται έμφαση στη δύναμη των μοντέλων αυτών για την συρρίκνωση της περιοχής αναζήτησης. Το [141] χρησιμοποιεί επιπλέον παραδοχές για το μετασχηματισμό ευθυγράμμισης έτσι ώστε με προεπεξεργασία να ελαττώσει το χρόνο ανίχνευσης. Μια ακόμα παραδοχή της εποχής είναι η ακαμψία των αντικειμένων, η οποία μπορεί να γίνει εκμεταλλεύσιμη στην γεωμετρική αναπαράσταση των αντικειμένων με γραμμές και επίπεδα [66]. Η ανάγκη για αναισθησία στην οπτική γωνία παρακολούθησης του αντικειμένου εμφανίζεται αυτή την εποχή στο [139], το οποίο θίγει επίσης πιθανοτικά μοντέλα για ταχύτερο εντοπισμό περιοχών ενδιαφέροντος. Πριν κλείσουμε με το “ρεύμα” της ευθυγράμμισης και των γεωμετρικών αναπαραστάσεων, αξίζει να αναφέρουμε ότι στην εποχή εκείνη, δεν υπάρχει σαφής διάκριση ανίχνευσης και ταυτοποίησης αντικειμένων. Μάλιστα, τρισδιάστατα και δισδιάστατα μοντέλα εναλλάσσονται κατά μήκος της ίδιας εργασίας, ενώ ορισμένοι αλγόριθμοι περιορίζονται φανερά από το hardware της εποχής.

Η γεωμετρική προσέγγιση είναι φανερό ότι δεν έχει γενική χρήση. Ήδη το 1972, το [62] επιχειρεί να προσεγγίσει το ζήτημα της εύρεσης γραμμών με ένα μετασχηματισμό που γενικεύει για κυρτές καμπύλες. Η ιδέα της αμεταβλητότητας βρέθηκε στο επίκεντρο της μελέτης στις αρχές της δεκαετίας του '90. Στο [162] παρουσιάζεται μία μέθοδος τρισδιάστατης αναγνώρισης αντικειμένων η οποία βασίζεται σε γεωμετρικά χαρακτηριστικά τα οποία παραμένουν αναλλοίωτα σε μετασχηματισμούς, προοπτική κάμερας και αλλαγή φωτισμού, όπως η συμμετρία και η ανάλυση σε πολύεδρα. Το [200] επιδεικνύει τη χρησιμότητα τοπικά αναλλοίωτων χαρακτηριστικών με τρόπο που να πετυχαίνει κάποια επιπλέον αναισθησία σε επικαλύψεις αντικειμένων.

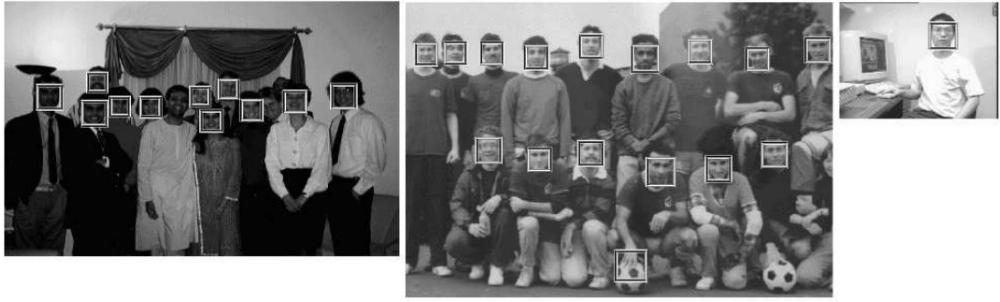
Παράλληλα με τα παραπάνω, δημιουργείται μια άλλη προσέγγιση από τη δεκαετία του '70, η οποία εξελίσσεται με τα χρόνια. Αντί για ταύτιση γεωμετρικών μοντέλων, γίνεται εστίαση στην αναπαράσταση του αντικειμένου ώστε να μπορεί να αναγνωρισθεί ως πρότυπο. Στο [166] γίνεται απόπειρα ανάλυσης αντικειμένων σε απλούστερες δομές, με σκοπό να αυξηθεί η συνθετότητα των αντικειμένων προς αναγνώριση.



Σχήμα 4.2: Η αμεταβλητότητα αποτέλεσε πρωταρχικό στόχο των εργασιών ανίχνευσης και αναγνώρισης αντικειμένων. Είναι πρώτηστης σημασίας οι διαφορετικές όψεις του αντικειμένου να έχουν παρόμοια αναπαράσταση στο χώρο των χαρακτηριστικών που χρησιμοποιούνται για την ταυτοποίησή τους. Εδώ για παράδειγμα απεικονίζονται τρεις διαφορετικές όψεις ενός εργαλείου και η συνάρτηση αναπαράστασής τους. Παρατηρούμε την επιθυμητή ομοιότητα μεταξύ των αναπαραστάσεων αυτών. Εικόνα από το [200].

Στο [154] επιχειρείται η τρισδιάστατη μοντελοποίηση των αντικειμένων σε δομικά συστατικά. Ωστόσο η δυσκολία αυτής της αναπαράστασης αποτρέπει την καρποφορία των προσεγγίσεων αυτών. Το 1987 το [14] επαναφέρει το πρόβλημα με έναν ενδιαφέροντα τρόπο: αναλύει κάθε αντικείμενο σε μονάδες που ονομάζει geons, οι οποίες είναι τρισδιάστατα γεωμετρικά σχήματα με επιπλέον ιδιότητες αμεταβλητότητας σε μετασχηματισμούς. Η θεωρία αυτή έχει ψυχοφυσικό υπόβαθρο [86], [64] και είναι εύρωστη σε γεωμετρικούς μετασχηματισμούς και αλλαγή οπτικής, δημιουργώντας παράλληλα ένα τεράστιο λεξιλόγιο από συνδυασμούς λίγων μόλις μονάδων. Από την άλλη, η θεωρία δεν συνοδεύτηκε από κάποιο μηχανισμό που να εξάγει τη δομική ανάλυση των αντικειμένων σε μια εικόνα, ενώ αποτυχάνει να διακρίνει αντικείμενα με παρόμοια δομή σε επίπεδο geons. Οι προσπάθειες δομικής γεωμετρικής αναπαράστασης συνεχίστηκαν, με το [177] να εστιάζει σε ανάλυση κυλίνδρων και το [199] σε ανάλυση επιπέδων.

Η προσπάθεια αναπαράστασης των αντικειμένων έδωσε το έναυσμα για δύο νέες προσεγγίσεις, τα μοντέλα βασισμένα στην εμφάνιση και τις μεθόδους ανάλυσης χαρακτηριστικών. Το [240] είναι μια αρκετά πρωτοπόρα εργασία: δημιουργεί ένα μοντέλο προσώπου βασισμένο σε “ιδιοπρόσωπα”, τα οποία είναι ιδιοδιανύσματα πρωταρχικής ανάλυσης σε ένα σύνολο εικόνων προσώπων. Η πρωτοπορία εδώ έγκειται σε πολλά επίπεδα: Διαχωρίζεται η τρισδιάστατη από τη δισδιάστατη προσέγγιση, όπως και η ανίχνευση από την αναγνώριση. Επιπλέον, εισέρχεται η Μηχανική Μάθηση στο προσκήνιο, με την εξαγωγή μαθηματικών χαρακτηριστικών για τη δημιουργία μοντέλου προσώπου αλλά και για τη διάκριση των προσώπων μεταξύ τους, αντί για μεθόδους Όρασης Υπολογιστών αποκλειστικά. Η προσέγγιση αυτή γενικεύεται



Σχήμα 4.3: Η εργασία των Viola-Jones αποτέλεσε σταθμό στην έρευνα πάνω στην ανίχνευση αντικειμένων χάρη στην απλότητα, ταχύτητα και αποτελεσματικότητά της. Στην εικόνα φαίνονται τα αποτελέσματα της εκτέλεσης του αλγορίθμου ανίχνευσης προσώπου πάνω στις δοθείσες εικόνες. Παρότι η αρχική σχεδίαση έγινε για ανίχνευση προσώπου, η μέθοδος γενικεύεται και δίνει ικανοποιητικά αποτελέσματα για μια ποικιλία αντικειμένων υπό ορισμένες προϋποθέσεις και περιορισμούς στους μετασχηματισμούς που μπορούν να υφίστανται τα ανιχνευόμενα αντικείμενα και απαιτεί μικρό αριθμό θετικών δειγμάτων εκπαίδευσης.

Εικόνα από το [245].

και αποκτά κάποιον φορμαλισμό στο [163], το οποίο απεικονίζει και τις γεωμετρικές πολλαπλότητες που αναπαριστούν ένα αντικείμενο. Η γενική ιδέα είναι η εξαγωγή εμπειρικών μοντέλων με κάποια ικανότητα γενίκευσης. Φορμαλισμό επιχειρεί επίσης να αποδώσει και το [231], από τη σκοπιά της εκτίμησης συνάρτησης. Παράλληλα, μια παρόμοια αλλά από διαφορετική σκοπιά προσέγγιση ακολουθείται στο [232], όπου αξιοποιούνται ιστογράμματα χρώματος. Οι μέθοδοι μοντέλων εμφάνισης είναι οι πρώτες που θίγουν την απόκριση σε πραγματικό χρόνο. Εν τούτοις, υπάρχουν σαφείς αδυναμίες των χαρακτηριστικών που χρησιμοποιούνται με αποτέλεσμα την ευαισθησία σε απλούς γεωμετρικούς μετασχηματισμούς και επικαλύψεις.

Οι μέθοδοι ανάλυσης χαρακτηριστικών είχαν μεγαλύτερη επιτυχία σε επίδοση και ταχύτητα. Ωστόσο η εκπαίδευση γίνεται σε εικόνες που περιέχουν αποκλειστικά το ζητούμενο αντικείμενο και για τον εντοπισμό του αντικειμένου στην εικόνα απαιτείται μια αναζήτηση με κυλιόμενα παράθυρα πάνω στην εικόνα. Γίνονται σημαντικές πρόοδοι σε ανίχνευση αντικειμένων γενικότερα αλλά και πιο ειδικά στην ανίχνευση προσώπου, η οποία ξεκίνησε με το [240] και συνεχίστηκε στο [11]. Σταθμό στην ανίχνευση προσώπων αλλά και αντικειμένων γενικότερα αποτελεί αναμφίβολα το [245]. Η εργασία αυτή εισήγαγε την έννοια των ολοκληρωτικών εικόνων κι επιτάχυνε εξαιρετικά την αναζήτηση πάνω στην εικόνα. Ο ανιχνευτής προσώπου των Viola-Jones είναι γρήγορος και εύρωστος, παρότι απλός, και χρησιμοποιήθηκε για αρκετά χρόνια λόγω αυτών των προτερημάτων του. Ακόμα, εν έτει 2017, το σύστημα αυτό παραμένει ως μια γρήγορη επιλογή για απλά καθήκοντα ανίχνευσης. Τέλος, άξιο αναφοράς αυτή την εποχή είναι και το [209], το οποίο εκμεταλλεύεται την πρόοδο της Μηχανικής Μάθησης και επαναφέρει το ζήτημα της αποσύνθεσης του αντικειμένου σε δομικές μονάδες, με αρκετά διαφορετική φύση πλέον.

Η άνοδος της Μηχανικής Μάθησης και η ταυτόχρονη εξέλιξη του hardware ανοίγουν το δρόμο για μια νέα κατεύθυνση που συνδυάζει εξαγωγή τοπικών χαρακτηριστικών, αμετάβλητες δομές και μοντέλα εμφάνισης. Τα [140], [25], [138] εισάγουν αναλλοίωτους μετασχηματισμούς και σημεία ενδιαφέροντος, δείχνοντας πώς αυτά μπορούν να χρησιμοποιηθούν στην ανίχνευση αντικειμένων. Στα [150], [151] η ανίχνευση γίνεται ορίζοντας μια μετρική απόστασης και λύνοντας ένα πρόβλημα βελτιστοποίησης ως προς αυτή. Το [74] αξιοποιεί τις καμπύλες σαν χαρακτηριστικά των αντικειμένων, το [264] εξάγει τοπικά αμετάβλητα χαρακτηριστικά και εστιάζει στο μετασχηματισμό

τους με μέθοδο πυρήνα για αποτελεσματικότερη ταξινόμηση και το [72] χρησιμοποιεί μια μη επιβλεπόμενη μέθοδο για την εκτίμηση των παραμέτρων των μετασχηματισμών που χρησιμοποιεί. Το [49] αποτέλεσε ένα σημαντικό σταθμό εδώ, χρησιμοποιούμενο για χρόνια στην ανίχνευση ανθρώπων. Η εργασία αυτή αναπαριστούσε τοπικά την εικόνα με χαρακτηριστικά HOG και ταξινομούσε με SVM. Τέλος, το [46] ήταν από τις πρώτες εργασίες που επέκτειναν το μοντέλο των χαρακτηριστικών σε “σάκο” (bag-of-features), μια ιδέα που είχε ήδη εμφανιστεί πριν αρκετά χρόνια στην κατηγοριοποίηση κειμένου [206].

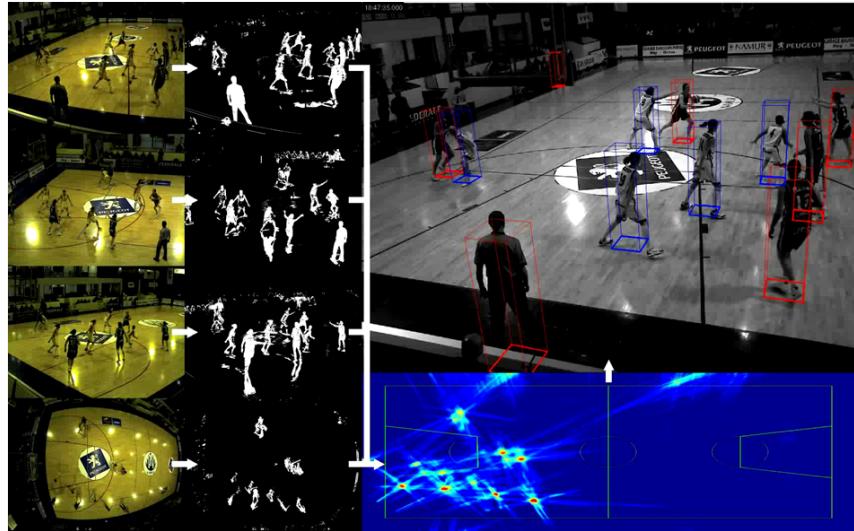
Τα τελευταία χρόνια η έρευνα πάνω στην ανίχνευση αντικειμένων έχει γενικευθεί και δεν υπάρχει μαζικός προσανατολισμός. Η εξέλιξη του πεδίου και του υλικού έχουν επιτρέψει την εστίαση πλέον σε πιο δύσκολα εγχειρήματα, όπως ο εντοπισμός παραμορφώσιμων αντικειμένων. Τα [69], [68] αποτελούν δύο βασικές και πρωτοπόρες εργασίες στην κατεύθυνση αυτή, οι οποίες επαναφέρουν το ζήτημα της διάσπασης ενός σύνθετου αντικειμένου σε μέρη. Το [175] αξιοποιεί δεδομένα βάθους για να πετύχει ανίχνευση μερών του σώματος σε πραγματικό χρόνο, ζήτημα το οποίο πλέον είναι εξέχουσας σημασίας. Προς αυτή την κατεύθυνση γίνονται προσπάθειες όπως του [53] και τελικά τα συνελικτικά νευρωνικά δίκτυα, μετά την τεράστια επιτυχία τους στην ταξινόμηση [123], [89], χρησιμοποιούνται και για ανίχνευση στα [183], [81], με το [96] να πετυχαίνει το πιο πρόσφατο state-of-the-art αποτέλεσμα. Ιδιαίτερη έμφαση έχει δοθεί, λόγω των πολλαπλών εφαρμογών, στην ανίχνευση χεριών τα τελευταία χρόνια [160], [180], [270], [271], [84], με διαφορετικές προσεγγίσεις είτε για απευθείας εφαρμογές [76], [52], είτε για αλληλεπίδραση με αντικείμενα [113], [197]. Τέλος, μια ακόμα αξιόλογη τάση είναι τα μέτρα ύπαρξη αντικειμένου (objectness), τα οποία παράγουν χάρτες που υποδηλώνουν την ύπαρξη ή μη αντικειμένου σε κάθε περιοχή [33] και που μπορούν να αξιοποιηθούν για εξαγωγή προσκηνίου σε εικόνες [198], [269], [225].

Κλείνοντας αυτή την ιστορική αναδρομή, θα σχολιάσουμε δύο πράγματα. Το πρώτο αφορά τον ορισμό της ανίχνευσης αντικειμένων που επιζητά την ελάχιστη σε εμβαδόν περιοχή. Η συνθήκη αυτή όταν τηρηθεί αυστηρά καταλήγει στο συγγενικό πρόβλημα της σημασιολογικής κατάτμησης [237], [134], [266], [137]. Το δεύτερο σχόλιο αφορά μια εκτίμηση του επιπέδου στο οποίο βρίσκεται η ανίχνευση αντικειμένων στη σημερινή εποχή. Η ιστορική πορεία δείχνει τεράστια εξέλιξη, ωστόσο ακόμα δεν υπάρχει ενιαίος ανιχνευτής αντικειμένων για κάθε σύνολο δεδομένων και μάλιστα η έρευνα δεν περιστρέφεται γύρω από κοινά σύνολα δεδομένων ώστε να υπάρξει σύγκριση. Επομένως, ένα σύστημα τεχνητής όρασης που θα μπορούσε να πλησιάσει την ανθρώπινη αντίληψη σε φυσικό περιβάλλον είναι ακόμα ένας μακρινός στόχος.

4.2 Ανίχνευση Προσκηνίου και Παρακολούθηση Αντικειμένων σε Βίντεο

4.2.1 Γενικά

Αν δούμε το βίντεο ως ακολουθία εικόνων με σταθερή συχνότητα, είναι φανερό ότι αποδίδουμε μία επιπλέον διάσταση στη συλλογή εικόνων, τον χρόνο. Ωστόσο, μία απεικόνιση του βίντεο ως συνάρτηση τριών διαστάσεων θα έκρυβε τη συσχέτιση των διαδοχικών frames, η οποία επιβάλλει έναν περιορισμό στις μεταβολές των χωρικών συντεταγμένων σε καθορισμένα χρονικά διαστήματα, δηλαδή, αν θεωρήσουμε



Σχήμα 4.4: Ζήτημα της ανίχνευσης προσκηνίου είναι η εύρεση όχι μόνο των κινούμενων αντικειμένων αλλά και των κινήσεων ενδιαφέροντος. Παρότι ο ορισμός αυτός δεν είναι σαφής, διαισθητικά ζητάμε την κίνηση που αφορά την απεικονιζόμενη δράση στην οποία και εστιάζουμε και όχι πιθανές άλλες δράσεις που μπορεί να συμβαίνουν στο περιβάλλον, κάτι που ονομάζουμε παρασκήνιο. Στο εικονιζόμενο παράδειγμα, προσκήνιο αποτελούν οι αθλήτριες και η μπάλα, αλλά όχι κάποια κίνηση που μπορεί να πραγματοποιείται στην εξέδρα. Εικόνα από το [54].

συνεχή ροή frames, το βίντεο είναι μια περιορισμένη επιφάνεια τριών διαστάσεων όπου οι δύο πρώτες συντεταγμένες είναι συναρτήσεις της τρίτης. Είναι και διαισθητικά αλλά και μαθηματικά φανερό ότι μπορούμε επομένως να μιλάμε για δύο χρονικά μεταβαλλόμενες χωρικές συναρτήσεις, εισάγοντας την έννοια της κίνησης και της ταχύτητας. Η ανίχνευση προσκηνίου και η παρακολούθηση αντικειμένων μόνο τότε αποκτούν υπόσταση. Εύκολα κανείς θα αντιλαμβανόταν ότι η παρακολούθηση αντικειμένων σε βίντεο, εφόσον έχει αίσθηση της έννοιας του αντικειμένου, είναι ο προσδιορισμός της θέσης του αντικειμένου σε κάθε χρονική στιγμή. Η παρακολούθηση είναι δηλαδή η γενίκευση της ανίχνευσης σε ακολουθία συσχετιζόμενων εικόνων. Για το ποιο είναι το προσκήνιο σε μια εικόνα ή βίντεο ωστόσο δεν υπάρχει μαθηματικός ορισμός, πρόκειται για ένα πρόβλημα μη καλώς ορισμένο. Στην περίπτωση των ακίνητων εικόνων είδαμε ότι τα τελευταία χρόνια νέοι ορισμοί δίνονται και προσπάθειες γίνονται για το λεγόμενο μέτρο “objectness”, την ύπαρξη αντικειμένου του προσκηνίου στην κάθε θέση. Στα βίντεο συνήθως ορίζουμε ως προσκήνιο τα κινούμενα αντικείμενα, δηλαδή τις θέσεις μεταβολών μεταξύ διαδοχικών frames. Το δυικό πρόβλημα του προσκηνίου είναι η εξαγωγή μοντέλου παρασκηνίου. Τα δύο προβλήματα είναι πρακτικά ισοδύναμα κι οι όροι χρησιμοποιούνται εναλλακτικά ανάλογα με το σκοπό δράσης.

Είναι φανερό ότι η παρακολούθηση αντικειμένων αλλά κυρίως η ανίχνευση προσκηνίου, που δεν είναι καλώς ορισμένο πρόβλημα, παρουσιάζουν ποικίλες προκλήσεις. Οι αλλαγές στον φωτισμό, η σκιάση, ο θόρυβος και η κίνηση της κάμερας είναι ίσως αυτές που έχουν μοντελοποιηθεί αποτελεσματικότερα. Σημαντικές δυσκολίες επιβάλλει το ίδιο το παρασκήνιο. Συχνά, υπάρχουν επικαλύψεις του αντικειμένου κίνησης με αντικείμενα του παρασκηνίου ή το παρασκήνιο παρέχει “καμουφλάζ” σε αντικείμενα του προσκηνίου. Ακόμα χειρότερα είναι τα πράγματα όταν έχουμε να κάνουμε με δυναμικό υπόβαθρο: αν θυμηθούμε τον ορισμό που επιχειρήσαμε να δώσουμε για το προσκήνιο, βλέπουμε ότι κάθε κίνηση, η οποία θα προκαλεί αλλαγή μεταξύ δύο διαδοχικών frames, πρέπει να ερμηνεύεται ως προσκήνιο. Αυτό όμως δεν ισχύει, καθώς

αντικείμενα του παρασκηνίου μπορεί να μεταβάλλουν την εμφάνισή τους, για παράδειγμα μια ανοιχτή τηλεόραση ή ένα αυτοκίνητο που διέρχεται αρκετά μέτρα πίσω από αυτό που καταγράφει η κάμερα, ή ακόμα και οι καιρικές συνθήκες να δυσχεραίνουν τη διάκριση του αντικειμένου [179]. Από τα τελευταία παραδείγματα γίνεται εμφανές το πόσο εύθραυστος είναι ο ορισμός που δώσαμε για το προσκήνιο. Οι άνθρωποι τείνουν να θεωρούν προσκήνιο αυτή τη δράση οποία εξελίσσεται μπροστά τους κι έχει “ενδιαφέρον” για το δράστη και το σκοπό δράσης. Ακόμα κι αν ο δράστης σταματήσει να κινείται, οπότε αποτελεί παρασκήνιο με τον αρχικό ορισμό, μπορεί να δρα, άρα να αποτελεί προσκήνιο. Τέλος, υπάρχουν και αντικειμενικές δυσκολίες που πρέπει να ληφθούν υπόψιν, όπως η μοντελοποίηση της κίνησης ενός αντικειμένου με άγνωστη συνάρτηση ταχύτητας (συμπεριλαμβανομένου και γωνιακής ταχύτητας) και η ανεπάρκεια δεδομένων για εξαγωγή μοντέλου περιβάλλοντος.

Όσο δύσκολο κι αν είναι το πρόβλημα της ανίχνευσης προσκηνίου και της παρακολούθησης αντικειμένων, τόσο χρήσιμη είναι η εφαρμογή των αποτελεσμάτων του σε πολλές πτυχές της καθημερινότητας. Συστήματα επίβλεψης, ανιχνευτές κίνησης, προγράμματα αλληλεπίδρασης ανθρώπου-μηχανής, εξαγωγή περιεχομένου από βίντεο και αναγνώριση χειρονομιών σε πραγματικό χρόνο είναι μερικές από τις πολυάριθμες εφαρμογές των ανιχνευτών κίνησης και προσκηνίου. Πιο συγκεκριμένα, τα αυτόματα αυτοκίνητα του μέλλοντος πρέπει να διαθέτουν μηχανισμούς εντοπισμού πεζών ή επικίνδυνων αντικειμένων σε πραγματικό χρόνο [59], ενώ συστήματα υποβοήθησης ηλικιωμένων πρέπει να ανιχνεύουν επιτυχώς τις κινήσεις που τα αφορούν [188]. Στη σφαίρα αυτή, πολλές έρευνες έχουν γίνει και γίνονται πάνω στην παρακολούθηση αντικειμένων και στην ανίχνευση προσκηνίου.

4.2.2 Ιστορικά Στοιχεία

Η πρώτη προσέγγιση παρακολούθησης που εφαρμόστηκε ήταν με τη χρήση σημείων. Συγκεκριμένα, το αντικείμενο αναπαρίσταται ως ένα σύνολο σημείων και παρακολουθείται κάθε σημείο ξεχωριστά. Στα μέσα της δεκαετίας του '80 κυριαρχούν οι στατιστικές μέθοδοι παρακολούθησης σημείων και πρώτο το [24] κάνει χρήση του φίλτρου Kalman [1] για την παρακολούθηση σημείων σε θορυβώδεις εικόνες. Στο [13] συνεχίζεται η ίδια προσέγγιση, ενώ το [195] επεκτείνει τη χρήση του φίλτρου για να εκτιμήσει τρισδιάστατες τροχιές από δισδιάστατη κίνηση. Σε παρόμοιο πλαίσιο, τα [28] και [182] χρησιμοποιούν τον πιο ευσταθή αλγόριθμο JPDAF, ενώ τα [229], [44], [27] κάνουν χρήση πιθανοτικών ή υπολογιστικά αποδοτικότερων παραλλαγών του πιο εύρωστου αλγορίθμου MHT [185]. Συγγενική αλλά πιο προχωρημένη, η προσέγγιση του [98] χρησιμοποιεί στατιστικές μεθόδους και φίλτρα σωματιδίων. Παράλληλα με τις στατιστικές αναπτύσσονται και ντετερμινιστικές μέθοδοι παρακολούθησης, με τα [212], [205] και [181] να κάνουν την αρχή και τα [101], [242], [213] να βελτιώνουν τις προσεγγίσεις αυτές. Γενικά, η αναπαράσταση των αντικειμένων ως σύνολα σημείων την εποχή εκείνη είχε περιορισμένη επιτυχία καθώς βασίζονταν πάνω σε αυστηρές υποθέσεις για την κατανομή θορύβου και τη μορφή των αντικειμένων οι οποίες συχνά δε μοντελοποιούν μια γενικευμένη πραγματικότητα.

Μια διαφορετική προσέγγιση του προβλήματος της οπτικής παρακολούθησης αντικειμένων ήταν αυτή της λεγόμενης “παρακολούθησης πυρήνα”, όπου το αντικείμενο αναπαρίσταται από ένα γεωμετρικό σχήμα ή μια συμπαγή περιοχή του χώρου και μελετάται η κίνηση αυτής της περιοχής, συνήθως με προσεγγίσεις κλασσικών γεωμετρικών μετασχηματισμών ή υπολογισμό οπτικής ροής. Δύο σχολές διαμορφώθηκαν εδώ, ανάλογα με την αναπαράσταση του αντικειμένου. Η πρώτη δίνει έμφαση

στον εντοπισμό του αντικειμένου και στην παρακολούθηση κίνησης ανανεώνοντας τον εντοπισμό, ενώ η αναπαράσταση γίνεται με πρότυπα [15] ή μοντέλα μίξης [108], ιστογράμματα χρώματος [75] και άλλες σχετικές μεθόδους [136]. Ιδιαίτερης αναφοράς χρίζουν τα [144], [217] και [40], [42] τα οποία αποτελούν ακόμα και σήμερα ρεαλιστικές και επίκαιρες εναλλακτικές. Ξεχωριστές για την πρώτη σχολή καθώς πετυχαίνουν παρακολούθηση πολλών αντικειμένων ταυτόχρονα είναι οι εργασίες των [102], που μοντελοποιεί παρασκήνιο και προσκήνιο με μοντέλα μίξης γκαουσιανών, και [235] που απεικονίζει τα frames σαν ένα σύνολο επιπέδων (παρασκήνιο και αντικείμενα) ορίζοντας ως παραμέτρους το σχήμα, την κίνηση και την εμφάνιση. Η δεύτερη σχολή, αντί να παρακολουθεί στη διάρκεια του βίντεο ένα αντικείμενο εξάγοντας το μοντέλο του από πιθανοτικές μεθόδους, πρώτα δημιουργεί ένα μοντέλο συνδυάζοντας εικόνες από διαφορετικές όψεις έτσι ώστε να αποκτά ευρωστία σε γεωμετρικούς μετασχηματισμούς. Το πλεονέκτημα αυτών των μεθόδων είναι η γνώση ότι παρακολουθείται το ίδιο αντικείμενο κι όχι όποιο αντικείμενο κινείται. Το [16] χρησιμοποιεί Ανάλυση Πρωταρχικών Συνιστωσών (PCA) για να αναπαραστήσει και να ανακατασκευάσει το αντικείμενο, μεταφέροντας τη σύγκριση προτύπων (εναλλακτική του ταίριασματος) στο χώρο των ιδιοδιανυσμάτων. Σε παρόμοια κατεύθυνση, το [6] χρησιμοποιεί Μηχανές Διανυσματικής Υποστήριξης (SVM) για να εκπαιδεύσει τον παρακολουθητή χρησιμοποιώντας ως αρνητικά δείγματα περιοχές του παρασκηνίου που μοιάζουν με το αντικείμενο.

Οι προσεγγίσεις παρακολούθησης πυρήνα βασίζονται στην υπόθεση της δυνατότητας αναπαράστασης των αντικειμένων με απλά γεωμετρικά σχήματα τα οποία παραμένουν σχετικά αναλλοίωτα στο χρόνο. Η υπόθεση αυτή παρέχει ταχύτητα στο σύστημα αλλά δεν είναι πάντα εύλογη, συν ότι το παρασκήνιο μπορεί να απεικονίζεται με παρόμοια γεωμετρικά σχήματα. Η μεταβλητότητα, η παραμορφωσιμότητα και η συνθετότητα των αντικειμένων ήταν η αιτία αναζήτησης πιο εκλεπτυσμένων μεθόδων αναπαράστασης και παρακολούθησης, όπως η “παρακολούθηση σιλουέτας”. Μια απλή αλλά αδύναμη προσέγγιση είναι το ταίριασμα σχήματος με την υπόθεση ότι μεταξύ διαδοχικών frames μπορεί να πραγματοποιηθεί μόνο μετάθεση του αντικειμένου και με χρήση του μετασχηματισμού Hausdorff [100], [132]. Μια εναλλακτική στο ταίριασμα σχήματος είναι η μοντελοποίηση του αντικειμένου με ιστογράμματα χρώματος και ακμές [114], ενώ το [87] χρησιμοποιεί πληροφορία από τις ακμές εντός της περιοχής της σιλουέτας για να εντοπίσει το αντικείμενο. Τέλος, το [207] χειρίζεται διαφορετικά το πρόβλημα, εφαρμόζοντας τον μετασχηματισμό Hough στο χώρο των ταχυτήων.

Η παρακολούθηση σιλουέτας εκφράστηκε ποιοτικότερα από τις μεθόδους καμπυλών. Οι μέθοδοι αυτές παρακολουθούν την εξέλιξη καμπυλών σε διαδοχικά frames και μπορούν να διακριθούν σε δύο βασικές κατευθύνσεις. Η πρώτη, αναπαριστά το σχήμα και την κίνηση της καμπύλης ως εσωτερικές μεταβλητές στο χώρο καταστάσεων. Η ιδέα ξεκινά ήδη από το 1992 με το [236] και αναπτύσσεται στα [103], [172]. Το [148] επεκτείνει το μοντέλο ώστε να διαχειρίζεται επικαλύψεις και να παρακολουθεί πληθώρα αντικειμένων, ενώ το [32] μοντελοποιεί την εξέλιξη της καμπύλης με Κρυφά Μαρκοβιανά Μοντέλα (HMM). Η δεύτερη κατεύθυνση στράφηκε προς την ελαχιστοποίηση του συναρτησοειδούς της ενέργειας καμπύλης. Η ενέργεια της καμπύλης οριζόταν σε όρους χρονικής πληροφορίας με τη μορφή είτε χρονικών παραγώγων (οπτική ροή) [12], [153], [45], είτε με στατιστικά εμφάνισης παραγόμενα από τις περιοχές αντικειμένων και τις περιοχές παρασκηνίου [261], [194]. Στην πρώτη περίπτωση έχουμε χρήση του Λογισμού Μεταβολών ως βασικό εργαλείο, ενώ στη δεύτερη χρησιμοποιούνται ευριστικές μέθοδοι. Ωστόσο, παρά τη δύναμη και την ευελιξία τους, οι μέθοδοι παρακολούθησης σιλουέτας πάσχουν στην περίπτωση ανάμιξης ή διαχωρισμού των αντικειμένων, για παράδειγμα όταν ένας άνθρωπος αφήνει ή



Frame 46

Frame 57

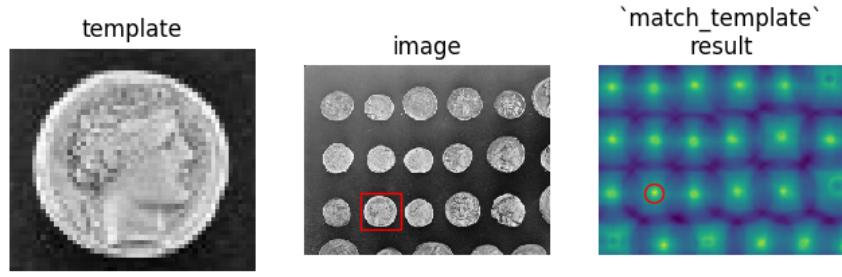
Frame 100

Σχήμα 4.5: Παρακολούθηση κινούμενου αυτοκινήτου σε πραγματικό χρόνο με τη βοήθεια Κάρτας Γραφικών Υπολογιστή (GPU). Παρατηρούμε ότι η κάμερα κινείται και ως εκ τούτου η σχετική θέση κάμερας και περιβάλλοντος μεταβάλλεται διαρκώς. Εν τούτοις το σύστημα κατορθώνει να απομονώνει το αντικείμενο ενδιαφέροντος και να κατατάσσει την υπόλοιπη εικόνα ως παρασκήνιο. Οι κάρτες γραφικών συνέβαλλαν δραστικά στην αύξηση της ταχύτητας υπολογισμών για τις απαιτητικές εφαρμογές της σημεριωής εποχής. Εικόνα από το [159].

λαμβάνει ένα αντικείμενο του περιβάλλοντος, αλλά και (λιγότερο) στην περίπτωση ισχυρών επικαλύψεων των αντικειμένων. Μια σύνοψη και σύγκριση των παραπάνω αλγορίθμων γίνεται στο [262], ενώ μια πιο πρόσφατη προσέγγιση γίνεται στο [127].

Στα τελευταία χρόνια η έρευνα έχει γενικευθεί και πλέον η απαιτητικότητα των εφαρμογών έχει ωθήσει προς νέες κατευθύνσεις την παρακολούθηση αντικειμένων. Μεγάλη έμφαση δίνεται στην απόκριση σε πραγματικό χρόνο [131], [120], απαίτηση που είχε ήδη εμφανιστεί και παλιότερα [41], [118]. Η εξέλιξη αυτή υποβοηθείται από τις σύγχρονες καινοτομίες σε hardware και software [159], [147], [260]. νέες μέθοδοι γεννιούνται [155], [158], ενώ ταυτόχρονα οι παλιότερες συνεχίζουν να αναπτύσσονται [55]. Όπως και στην ανίχνευση αντικειμένων, έτσι και στην παρακολούθηση, έχει δοθεί σήμερα μια ξεχωριστή ροή μελέτης για την περίπτωση των ανθρώπινων χεριών [31], [223], [216], καθώς είναι το κύριο όργανο δράσης των ανθρώπων ενώ παράλληλα παρουσιάζει τεράστια ποικιλομορφία και κατ'επέκταση μοντελοποιείται δύσκολα. Για διαφορετικές εφαρμογές χρησιμοποιούνται και διαφορετικές οπτικές κάμερας, οπότε έχουμε ειδικά συστήματα σχεδιασμένα για κάμερες βάθους [224] και για “εγκεντρικά” βίντεο [133], [271].

Η ανίχνευση προσκηνίου, ως συγγενική της παρακολούθησης εργασία αναπτύχθηκε παράλληλα και συχνά χρησιμοποιήθηκε για την παρακολούθηση [102]. Κύρια ιδέα είναι η χρήση γκαουσιανών μοντέλων μίξης για την περιγραφή του προσκηνίου και του παρασκηνίου με την αρχή να γίνεται το 1999 [227] και την ιδέα να βελτιώνεται λίγο αργότερα [112]. Οι ιδέες αυτές παρέμειναν στο επίκεντρο της έρευνας για τα επόμενα χρόνια [246] και το [21] παρέχει μια σύνοψη μεγάλου αριθμού αυτών. Η χρήση μοντέλων μίξης παραμένει σαν μια δυνατή εναλλακτική μέχρι σήμερα, ωστόσο τα τελευταία χρόνια έχουν αναπτυχθεί και άλλες μέθοδοι. Τα [107], [22] προσεγγίζουν το ζήτημα από τη δυική του μορφή (εξαγωγή παρασκηνίου). Καθώς η ανάγκη για απόκριση πραγματικού χρόνου γίνεται πιο πιεστική, νέες εργασίες ασχολούνται με την ανίχνευση προσκηνίου με χρήση αλγορίθμων αυτο-οργανούμενης εξαγωγής παρασκηνίου (SOBS) [149], προσαρμοστικής κατάτμησης σε επίπεδο πίξελ (PBAS) [94] και εξαγωγής οπτικού παρασκηνίου (ViBe) [9]. Μάλιστα στο [124] γίνεται μια



Σχήμα 4.6: Η λογική της μεθόδου Ταιριάσματος Προτύπων: Δοθεντος ενός προτύπου, αναζητούμε τη βέλτιστη τοποθέτησή του στην εικόνα ώστε να μεγιστοποιείται μια συνάρτηση ομοιότητας μεταξύ προτύπου και τηματος εικόνας. Η συνάρτηση ομοιότητας μπορεί να επιλεχθεί κατάλληλα σύμφωνα με το πρόβλημα και τις απαιτήσεις του σχεδιαστή. Συχνά, χρησιμοποιείται κατώφλι στη συνάρτηση ομοιότητας και τιμές μικρότερες αυτού σημαίνουν μη ταίριασμα και απουσία του προτύπου από την εικόνα.

υλοποίηση του ViBe σε FPGA. Τέλος, τα νευρωνικά δίκτυα χρησιμοποιούνται επίσης στην ανίχνευση προσκηνίου [201] και μάλιστα πλέον έχουν χρησιμοποιηθεί και βαθιά συνελικτικά νευρωνικά δίκτυα [23], [7].

4.3 Θεωρητικό Υπόβαθρο

4.3.1 Ταίριασμα Προτύπων (Template Matching)

Η μέθοδος ταιριάσματος προτύπων (template matching) είναι μία απλή και σχετικά γρήγορη αλλά περιορισμένης αποτελεσματικότητας μέθοδος εντοπισμού αντικειμένων σε εικόνες. Συνοπτικά, έχοντας στη διάθεσή της μια αναπαράσταση του αντικειμένου, μία δηλαδή εικόνα του, αναζητά την περιοχή της εικόνας για την οποία μεγιστοποιείται μια συνάρτηση ταιριάσματος μεταξύ της περιοχής αυτής και του προτύπου για το αντικείμενο. Διαισθητικά, πρόκειται για έναν ανιχνευτή αντικειμένων εκπαιδευμένο με μόνο ένα δείγμα για το αντικείμενο που ζητάμε. Είναι φανερό ότι ένας τέτοιος ανιχνευτής, ή ένα οποιοδήποτε σύστημα με δυνατότητα μάθησης, θα παρουσίαζε υψηλή υπερπροσαρμογή στο πρότυπο: τίποτα δεν αποτελεί δείγμα αυτής της κατηγορίας αντικειμένου εκτός αν ταιριάζει ακριβώς με το πρότυπο. Θα δούμε τα πλεονεκτήματα και μειονεκτήματα ενός τέτοιου ανιχνευτή αντικειμένων αφού εξετάσουμε την υλοποίησή του.

Τρεις απλές υλοποιήσεις ταιριάσματος προτύπων είναι οι μέθοδοι Φίλτρου Μηδενικής Μέσης Τιμής (Zero-Mean Filter), Αθροίσματος Τετραγωνικών Διαφορών (SSD: Sum of Squared Differences) και Κανονικοποιημένης Ετεροσυσχέτισης. Στην πρώτη μέθοδο, φιλτράρουμε την εικόνα με το πρότυπο, έχοντας αφαιρέσει από τη μέση τιμή του και αναζητούμε μέγιστα της εξόδου. Στη δεύτερη, υπολογίζουμε για κάθε περιοχή της εικόνας το άθροισμα των τετραγώνων των διαφορών μεταξύ των pixels της εικόνας και των pixels του προτύπου, αναζητώντας ελάχιστα της εξόδου, ή ισοδύναμα, μέγιστα της αντίθετης συνάρτησης της εξόδου. Τέλος, η τρίτη μέθοδος συνδυάζει τις δύο προηγούμενες εισάγοντας την κανονικοποιημένη ετεροσυσχέτιση

μεταξύ προτύπου και τμήματος εικόνας. Η μέθοδος αυτή είναι η πιο αργή αλλά και η πιο αποτελεσματική, καθώς πετυχαίνει αμεταβλητότητα τοπικά στη μέση ένταση της εικόνας και στην αντίθεση. Αμέσως γρήγορότερη είναι η μέθοδος SSD, η οποία όμως είναι ευαίσθητη στη μεταβολή της έντασης της εικόνας. Τέλος, ακόμα πιο γρήγορη είναι η πρώτη μέθοδος, αλλά ταυτόχρονα είναι και η λιγότερη εύρωστη.

Θα περιγράψουμε με λίγη περισσότερη λεπτομέρεια τώρα τη μέθοδο SSD για ταίριασμα προτύπων. Έστω $f[x, y]$ μια εικόνα και $g[x, y]$ ένα πρότυπο. Υπολογίζουμε το άθροισμα των τετραγωνικών διαφορών σε κάθε περιοχή που προκύπτει από την ολίσθηση του προτύπου πάνω στην εικόνα ως εξής:

$$h[x, y] = \sum_{(k,l)} (g[k, l] - f[k + x, l + y])^2 \quad (4.1)$$

Λαμβάνουμε ως έξοδο την $1 - h[x, y]$, η οποία θα είναι μέγιστη, κοντά στην τιμή 1, όταν το άθροισμα στην έκφραση της h λαμβάνει χαμηλές τιμές, οπότε πρότυπο και εικόνα παρουσιάζουν υψηλή ταύτιση. Εφαρμόζουμε ένα κατώφλι στην έξοδο της $1 - h$ έτσι ώστε να κρατήσουμε μόνο τις περιοχές υψηλής ταύτισης ως πιθανές περιοχές του αντικειμένου. Είναι φανερό από τον ορισμό της h , ότι απόλυτη ταύτιση συμβαίνει μόνο όταν $h = 0$, οπότε πρότυπο και εικόνα ταιριάζουν σχηματικά και χρωματικά. Ο ορισμός της h , αφήνει τη μέθοδο ευάλωτη σε διακυμάνσεις έντασης και φωτισμού. Πράγματι, ένα οποιοδήποτε scaling της εικόνας χρωματικά ($f \rightarrow \alpha * f$, α πραγματικός αριθμός) θα μας δώσει σαν έξοδο την ενέργεια του τμήματος της εικόνας που εξετάζουμε επί έναν πολλαπλασιαστικό παράγοντα $(1 - \alpha)^2$. Δηλαδή, όσο μεγαλύτερη η μεταβολή της φωτεινότητας (υποθέτοντας ομοιόμορφη μεταβολή τοπικά) και άρα όσο πιο μεγάλη η απόσταση του α από τη μονάδα, τόσο μεγαλύτερη η αποτυχία ταύτισης. Θα δούμε τώρα πώς η μέθοδος αυτή μπορεί να υλοποιηθεί αποδοτικά, έτσι ώστε να αξιοποιηθεί η ταχύτητά της.

Ξεκινάμε από τη μορφή της h που δίνεται στην εξίσωση (4.1). Εύκολα διαπιστώνουμε ότι

$$h[x, y] = \sum_{(k,l)} f^2[k + x, l + y] + \sum_{(k,l)} g^2[k, l] - s \sum_{(k,l)} (g[k, l]f[k + x, l + y]) \quad (4.2)$$

Τώρα είναι φανερό ότι ο πρώτος όρος είναι σταθερός ως προς x και y και αρκεί να υπολογιστεί μία φορά. Για τον δεύτερο όρο, αντί να αθροίζουμε κάθε φορά όλα τα pixels, μπορούμε να χρησιμοποιήσουμε τη μέθοδο των ολοκληρωτικών εικόνων [245], δηλαδή να υπολογίσουμε σε κάθε θέση το συσσωρευτικό άθροισμα όλων των τιμών των pixels με τιμές x και y μικρότερες είτε ίσες με αυτές της τρέχουσας θέσης. Από εκεί και πέρα, το πρόβλημα υπολογισμού ενός αθροίσματος pixels σε μια συγκεκριμένη περιοχή εικόνας είναι της κλάσης $O(1)$. Τέλος, ο τρίτος όρος του αθροίσματος περιέχει ένα άθροισμα το οποίο αποτελεί τη συσχέτιση f και g . Για τον αποδοτικό υπολογισμό της συσχέτισης, αφού εφαρμόσουμε κατάλληλο zero-padding στις f και g , περιστρέφουμε την g κατά 90 μοίρες και υπολογίζουμε το εσωτερικό γινόμενο των μετασχηματισμών Fourier της περιστραμμένης g και της f . Η συσχέτιση είναι ίση με το πραγματικό μέρος του αντίστροφου μετασχηματισμού Fourier αυτού του εσωτερικού γινομένου. Οπότε συνολικά, η ποσότητα SSD μπορεί να υπολογιστεί γρήγορα και αποδοτικά.

Έχοντας δει και λεπτομέρειες υλοποίησης, είμαστε σε θέση να διακρίνουμε πλέον τα υπέρ και τα κατά της χρήσης μιας τέτοιας μεθόδου για ανίχνευση αντικειμένων. Από άποψη ταχύτητας, είδαμε ότι με κατάλληλη υλοποίηση ένας αλγόριθμος ταιριάσματος προτύπων μπορεί να γίνει εξαιρετικά γρήγορος. Ας σχολιάσουμε κυρίως όμως το φλέγον ζήτημα της αποτελεσματικότητας αλλά και της δυνατότητας γενίκευσης. Είδαμε ότι για τον έλεγχο ταιριάσματος εφαρμόζεται κατώφλι το οποίο εύκολα απορρίπτει ένα τμήμα εικόνας μόνο και μόνο λόγω διαφορετικού φωτισμού. Ωστόσο για μικρές διακυμάνσεις φωτισμού είναι δυνατό να επιτευχθεί ταίριασμα. Το πρότυπο αποτελεί τη μοναδική εικόνα εκπαίδευσης και άρα τη μοναδική αναπαράσταση του αντικειμένου. Από την μία αυτό αφαιρεί τη διαδικασία εκπαίδευσης, από την άλλη όμως στερεί οποιαδήποτε δυνατότητα γενίκευσης. Ακόμα χειρότερα, η εικόνα προτύπου πιθανότατα περιλαμβάνει και μικρό τμήμα παρασκήνου, καθιστώντας την ανίχνευση ευάλωτη ακόμα και σε μεταθέσεις του ίδιου αντικειμένου μπροστά από διαφορετικό παρασκήνιο. Παρόλα αυτά, η μέθοδος παραμένει χρήσιμη στην ανίχνευση αντικειμένων με χαμηλή οπτική μεταβλητότητα και κίνηση σε βίντεο. Υπάρχουν περιπτώσεις που σε ένα βίντεο μας αρκεί να εντοπίζουμε σταθερά πρότυπα αντικειμένων τα οποία μένουν πρακτικά αναλλοίωτα για μεγάλο χρονικό διάστημα. Σε αυτές τις περιπτώσεις είναι που το ταίριασμα προτύπων μπορεί να δώσει εξαιρετικά και γρήγορα αποτελέσματα.

4.3.2 Εξαγωγή Προσκηνίου με Μοντέλα Μίξης Γκαουσιανών (GMM)

Μια δημοφιλής μέθοδος εξαγωγής προσκηνίου είναι η χρήση Μοντέλων Μίξης Γκαουσιανών. Η μέθοδος εισήχθη στο [227] και βελτιώθηκε στο [112] ώστε να μοντελοποιεί τη μεταβολή στη σκίαση και να μην τη συνυπολογίζει στην κίνηση. Η βασική ιδέα είναι η μοντελοποίηση κάθε πίξελ ως ένα άθροισμα γκαουσιανών με βάρη (μίξη). Επομένως, η πιθανότητα ένα συγκεκριμένο πίξελ να έχει τιμή x_N τη χρονική στιγμή N εκφράζεται από τη σχέση:

$$p(x_N) = \sum_{k=1}^K w_k \eta(x_N; \theta_k) \quad (4.3)$$

με την παράμετρο w_k να εκφράζει το βάρος της k -οστής Γκαουσιανής $\eta(x, \theta_k)$, με τη συνάρτηση αυτή να δίνεται από τη σχέση

$$\eta(x_N; \theta_k) = \eta(x_N; \mu_k, \Sigma_k) = \frac{1}{2\pi^{\frac{D}{2}} |\Sigma_k|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k (x-\mu_k)} \quad (4.4)$$

όπου μ_k είναι η μέση τιμή και Σ_k η συνδιακύμανση της k -οστής συνιστώσας.

Η μέθοδος εκκινεί με την αρχική εκτίμηση του μοντέλου μίξης μέσω συναρτήσεων ανανέωσης προσδοκώμενων επαρκών στατιστικών. Μετά τα L πρώτα δείγματα χρησιμοποιείται η παραλλαγή των L -πρόσφατων δειγμάτων. Η αρχική εκτίμηση βελτιώνει την ακρίβεια της εκτίμησης προσκηνίου επιτρέποντας ταχεία σύγκλιση σε ένα ευσταθές μοντέλο. Στη συνέχεια λαμβάνονται υπόψιν μόνο τα L πρόσφατα παράθυρα ώστε να δοθεί προτεραιότητα στα πιο πρόσφατα δεδομένα κι έτσι το σύστημα προσαρμόζεται ευκολότερα στις αλλαγές που παρακολουθεί. Για τα L πρώτα frames, ο επαναληπτικός αλγόριθμος περιγράφεται από τις παρακάτω εξισώσεις:

$$\hat{w}_k^{N+1} = \hat{w}_k^N + \frac{1}{N+1} (\hat{p}(\omega_k | x_{N+1}) - \hat{w}_k^N) \quad (4.5)$$

$$\hat{\mu}_k^{N+1} = \hat{\mu}_k^N + \frac{\hat{p}(\omega_k | x_{N+1})}{\sum_{i=1}^{N+1} \hat{p}(\omega_k | x_i)} (x_{N+1} - \hat{\mu}_k^N) \quad (4.6)$$

$$\hat{\Sigma}_k^{N+1} = \hat{\Sigma}_k^N + \frac{\hat{p}(\omega_k | x_{N+1})}{\sum_{i=1}^{N+1} \hat{p}(\omega_k | x_i)} ((x_{N+1} - \hat{\mu}_k^N)(x_{N+1} - \hat{\mu}_k^N)^T - \hat{\Sigma}_k^N) \quad (4.7)$$

Μετά τα πρώτα L frames οι εξισώσεις λαμβάνουν τη μορφή:

$$\hat{w}_k^{N+1} = \hat{w}_k^N + \frac{1}{L} (\hat{p}(\omega_k | x_{N+1}) - \hat{w}_k^N) \quad (4.8)$$

$$\hat{\mu}_k^{N+1} = \hat{\mu}_k^N + \frac{1}{L} \left(\frac{\hat{p}(\omega_k | x_{N+1}) x_{N+1}}{\hat{w}_k^{N+1}} - \hat{\mu}_k^N \right) \quad (4.9)$$

$$\hat{\Sigma}_k^{N+1} = \hat{\Sigma}_k^N + \frac{1}{L} \left(\frac{\hat{p}(\omega_k | x_{N+1})(x_{N+1} - \hat{\mu}_k^N)(x_{N+1} - \hat{\mu}_k^N)^T}{\hat{w}_k^{N+1}} - \hat{\Sigma}_k^N \right) \quad (4.10)$$

Η μοντελοποίηση της σκίασης γίνεται με τη βοήθεια χρώματος και φωτεινότητας. Συγκεκριμένα, αν E ένα διόνυσμα της μέσης RGB τιμής των πίξελ υποβάθρου, $\|E\|$ η προσδοκώμενη γραμμή χρωματικότητας (chromaticity line), d η χρωματική παραμόρφωση (chromatic distortion) και τ ένα κατώφλι φωτεινότητας, τότε για δεδομένη ένταση φωτεινότητας I , η παραμόρφωση φωτεινότητας α και η παραμόρφωση χρώματος (color distortion) c , μπορούν να υπολογιστούν από τις σχέσεις

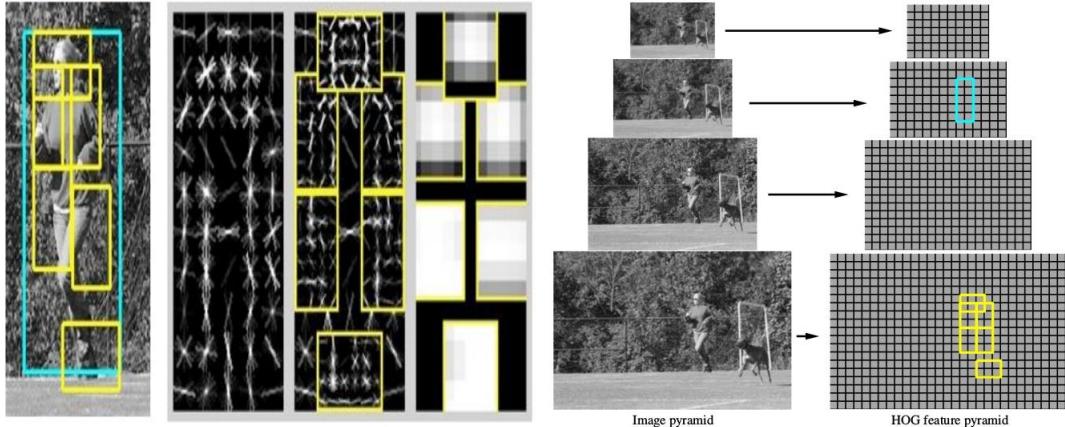
$$\alpha = \arg \min_z (1 - zE)^2 \quad (4.11)$$

$$c = \|I - \alpha E\| \quad (4.12)$$

Υπό την υπόθεση σφαιρικής Γκαουσιανής κατανομής για κάθε συνιστώσα μίξης, η τυπική απόκλιση μπορεί να τεθεί ίση με d . Ένα παρατηρούμενο πίξελ μπορεί να θεωρηθεί κομμάτι κινούμενης σκιάς αν η τιμή του α είναι μεταξύ 2.5 τυπικών αποκλίσεων και $\tau < c < 1$.

4.3.3 Ανίχνευση Ανθρώπων με Χρήση Μοντέλων Παραμορφώσιμων Τμημάτων

Η μέθοδος των Μοντέλων Παραμορφώσιμων Τμημάτων (DPM: Deformable Part Models) χρησιμοποιείται ευρέως στην ανίχνευση αντικειμένων λόγω της εύκολης εκπαίδευσης και της γρήγορης και αποτελεσματικής ανίχνευσης. Τα DPM βασίζονται στην



Σχήμα 4.7: Ενδιάμεσα στάδια κατά την εφαρμογή της μεθόδου των Μοντέλων Παραμορφώσιμων Τμημάτων. **Αριστερά:** Το μοντέλο για την ανίχνευση ανθρώπων. Στο σχήμα φαίνονται τα σταθερά μέρη του ανθρώπου (άκρα και σώμα) τα οποία μπορούν να κινούνται σχετικά και να παραμένουν ενωμένα με χαλαρούς συνδέσμους. Ακόμα, βλέπουμε την απεικόνιση του μοντέλου στο πεδίο των περιγραφητών HOG και τα μέρη του όταν αναλυθούν. **Δεξιά:** Η κάθε εικόνα αναλύεται σε πολλαπλές κλίμακες έτσι ώστε να ενσωματώσει την πληροφορία του μεγέθους. Βλέπουμε πώς η πυραμίδα στο επίπεδο των κλιμάκων εικόνων μεταφέρεται στο χώρο χαρακτηριστικών HOG. Εικόνα από το [69].

υπόθεση της ικανότητας επαρκούς αναπαράστασης των αντικειμένων με συμπαγή δομικά στοιχεία τα οποία συνδέονται μεταξύ τους με μη συμπαγείς συνδέσμους. Παρότι η μέθοδος χρησιμοποιείται γενικά στην ανίχνευση αντικειμένων, για τη χρήση της σε αυτή την εργασία παρουσιάζουμε συνοπτικά τη λειτουργία της στην ανίχνευση ανθρώπων κατά τη φάση ανίχνευσης και όχι εκπαίδευσης. Παραπέμπουμε τον ενδιαφερόμενο στο [68] για περισσότερες λεπτομέρειες.

Η μέθοδος των DPM εκκινεί από την μοντελοποίηση κάθε αντικειμένου με μία «ρίζα», η οποία είναι το μοντέλο όλου του αντικειμένου και τα μέρη ή τμήματα, τα οποία αναλύουν το αντικείμενο σε δομικά στοιχεία. Τα συμπαγή μέρη έχουν τη δυνατότητα να κινούνται, οπότε ένα αντικείμενο έχει τη δυνατότητα να παραμορφώνεται όταν συμβαίνει η κίνηση αυτή, η οποία προκαλεί μεταβολές στις σχετικές θέσεις των μερών ως προς τη ρίζα. Με αυτό τον τρόπο επιτυγχάνεται κάποια αμεταβλητότητα στο ακριβές σχήμα του αντικειμένου. Ταυτόχρονα, η ανάλυση χαρακτηριστικών γίνεται σε πολλαπλές κλίμακες, οπότε εξασφαλίζεται αμεταβλητότητα ως προς το μέγεθος του αντικειμένου. Η εκπαίδευση βασίζεται σε ασθενώς κατηγοριοποιημένα δεδομένα, καθώς αρκούν μόνο τα ορθογώνια που περιέχουν το αντικείμενο και τα μέρη του για την εξαγωγή των φίλτρων ρίζας και τμημάτων. Τελικά εξάγονται μοντέλα ρίζας και τμημάτων τα οποία είναι m-συνιστώσων, δηλαδή, φαινομενικά, χρησιμοποιούνται m διαφορετικά μοντέλα για την αναπαράσταση των μερών και της ρίζας ενός αντικειμένου. Η σχεδίαση αυτή αποδίδει στο σύστημα αναισθησία ως προς τις διαφορετικές όψεις των αντικειμένων.

Η μέθοδος εκκινεί με την εξαγωγή τοπικών χαρακτηριστικών τύπου HOG και μειώνει τη διάσταση με Ανάλυση Πρωταρχικών Συνιστώσων (PCA). Στη συνέχεια, χρησιμοποιεί τη μέθοδο «κυλιόμενου παραθύρου» για τις θέσεις του φίλτρου ρίζας σε κάθε κλίμακα. Τα φίλτρα τμημάτων εφαρμόζονται σε δύο κλίμακες υψηλότερης ανάλυσης και σε συντεταγμένες σχετικές ως προς τη θέση του φίλτρου ρίζας. Κάθε συνδυασμός θέσεων φίλτρων και κλίμακας αποκτά ένα σκορ, το οποίο υποθέτωντας η φίλτρα τμημάτων δίνεται από την εξίσωση

$$score(p_0, p_1, \dots, p_P) = \sum_{i=0}^n F_i * \phi(H, p_i) - \sum_{i=1}^n d_i * \phi_d(dx_i, dy_i) + b \quad (4.13)$$

Στην παραπάνω εξίσωση, p_i είναι ο συμβολισμός του μέρους i , με το p_0 να δηλώνει τη ρίζα. Ο πρώτος όρος αποτελεί το άθροισμα των σκορ της εφαρμογής κάθε φίλτρου F_i του τμήματος p_i πάνω στη συνάρτηση απόδοσης σκορ για την κλίμακα H , που έχει σχηματιστεί κατά την εκπαίδευση. Όσο υψηλότερη τιμή λαμβάνει το άθροισμα αυτό, τόσο πιο βέβαιοι είμαστε για μια έγκυρη ανίχνευση. Ο δεύτερος όρος, ο οποίος αφαιρείται από τα σκορ ανίχνευσης, είναι τα κόστη παραμόρφωσης. Αν δηλαδή d_i είναι η σχετική μετατόπιση του p_i ως προς τη θέση που κατέχει το p_i στο μοντέλο ρίζας που έχει τοποθετηθεί στη θέση που εξετάζουμε, ϕ_d είναι η συνάρτηση κόστους για μια τέτοια μετατόπιση, τότε το γινόμενο είναι το κόστος παραμόρφωσης και τείνει να επιβαρύνει υψηλές παραμορφώσεις από το μοντέλο που έχει προκύψει από την εκπαίδευση. Στην αυθεντική έκδοση της εργασίας, οι συναρτήσεις ϕ υλοποιούνται με Μηχανές Διανυσμάτων Υποστήριξης (SVM). Τέλος ο όρος b , είναι απλά ένα bias. Κρατάμε το μέγιστο της συνάρτησης αυτής πάνω στα p_i για όλες τις m -συνιστώσες. Υψηλό σκορ ρίζας για μια περιοχή μέσω της συνάρτησης αυτής ενδεικνύει το αντικείμενο. Μια μετεπέξεργασία η οποία περιλαμβάνει καταπίεση μη μεγίστων επιστρέφει το βέλτιστο ορθογώνιο ανίχνευσης.

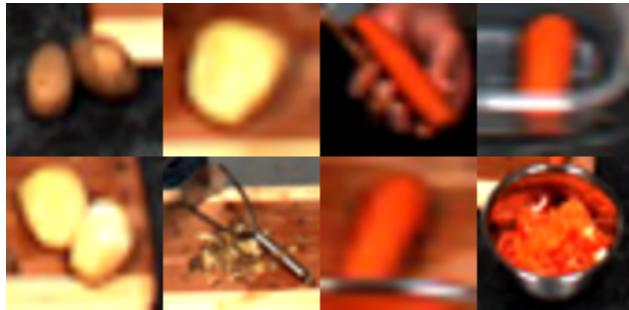
Συμπερασματικά, η μέθοδος των DPM μπορεί να μοντελοποιήσει πολύπλευρες προκλήσεις που καθιστούν την ανίχνευση αντικειμένων ένα δύσκολο πρόβλημα. Άλλο βασικό πλεονέκτημα της μεθόδου είναι η υλοποίησή της με χρήση δυναμικού προγραμματισμού και γενικευμένων μετασχηματισμών απόστασης, στοιχείο που την καθιστά αρκετά γρήγορη. Φυσικά είναι πιο αργή από το ταίριασμα προτύπων αλλά ταυτόχρονα γενικεύει υπερβολικά καλύτερα. Σε επόμενη εργασία [70], η ταχύτητα της μεθόδου αυξήθηκε δραματικά, δίνοντας περισσότερο μέγεθος στη δημοτικότητά της.

4.4 Προτεινόμενο Σύστημα

4.4.1 Το Σύνολο Δεδομένων για Ανίχνευση Αντικειμένων

Το σύνολο δεδομένων που χρησιμοποιούμε παρέχει (στην νεότερή του μορφή) επισημειώσεις για σχέσεις αντικειμένων-ρημάτων ανά τμήμα που έχει κατηγοριοποιηθεί σε κάποια από τις κατηγορίες ρημάτων. Ωστόσο δεν παρέχει καμία επισημείωση για την όψη ή τη δομή των αντικειμένων αυτών. Οπότε για το πρόβλημα της ανίχνευσης αντικειμένων προβήκαμε σε δικές επισημειώσεις. Κατά τη διαδικασία αυτή διαπιστώσαμε προκλήσεις που υποθάλπουν στο σύνολο δεδομένων αυτό σχετικά με την ανίχνευση αντικειμένων και τις οποίες αναφέρουμε εδώ:

- Συμβαίνουν συχνές, μερικές ή ολικές επικαλύψεις του αντικειμένου με άλλα αντικείμενα αλλά κυρίως με μέρη του ανθρώπινου σώματος, ιδιαίτερα χέρια.
- Η απόσταση από τη σταθερή κάμερα είναι μεγάλη για αναγνώριση αντικειμένων τέτοιου μεγέθους (όπως κομμένα λαχανικά ή μικρά μαγειρικά εργαλεία) με ακρίβεια.



Σχήμα 4.8: Παρά την αλλαγή στη δομή και στη φύση τους, τα αντικείμενα συχνά εκπροσωπούνται από την ίδια ετικέτα κλάσης. Για παράδειγμα, μια πατάτα μπορεί να κοπεί σε κομμάτια, τα οποία διατηρούν την ιδιότητα της πατάτας. Ομοίως και μια σαλάτα καρότων, εντάσσεται στην κατηγορία καρότων. Τέτοιες διενέξεις στην εμφάνιση και την ετικέτα συμβαίνουν στο σύνολο δεδομένων που χρησιμοποιείται σε αυτή την εργασία, γεγονός που δυσχεραίνει σημαντικά τον εντοπισμό και την παρακολούθηση των αντικειμένων στη διάρκεια του βίντεο.

- Υπάρχει μικρή μεταβλητότητα μεταξύ των διαφόρων κλάσεων αντικειμένων, όπως διάκριση μεταξύ παρόμοιων εμφανισιακά φρούτων σε μορφή κύβων.
- Ταυτόχρονα, υπάρχει μεγάλη μεταβλητότητα μεταξύ αντικειμένων της ίδιας κλάσης, όπως τα διαφορετικά είδη δοχείων.
- Υπάρχουν αντικείμενα τα οποία σχετίζονται με περιοχές του χώρου και δεν μπορούν να μοντελοποιηθούν επαρκώς, όπως ψυγείο, ντουλάπια και συρτάρια.
- Ακόμα πιο δύσκολα, υπάρχουν αντικείμενα τα οποία αλλάζουν σύσταση ή δομή κατά τη διάρκεια του βίντεο, αλλά συνεχίζουν να κατηγοριοποιούνται με την ίδια ετικέτα. Για παράδειγμα, φρούτα τα οποία από ακέραια μορφή κόβονται και στη συνέχεια πολτοποιούνται.
- Η έλλειψη πρότερης γνώσης για τα αντικείμενα που θα εμφανιστούν. Παρότι το λεξιλόγιο που χρησιμοποιούμε είναι περιορισμένο, συχνά εμφανίζονται συσκευασίες αγνώστου περιεχομένου, όπως μπουκάλια ή κουτιά.
- Η ανεπάρκεια γενικού μοντέλου για τα αντικείμενα της ίδιας κλάσης σε διαφορετικά βίντεο. Παρότι η πρόκληση αυτή είναι άμεσα συνδεδεμένη με την υψηλή ενδομεταβλητότητα, εν τούτοις αναφέρεται ξεχωριστά για να τονίσει τις διαφορές στις συνθήκες λήψης, όπως ο διαφορετικός φωτισμός, αλλά και τις δομικές διαφοροποιήσεις λόγω διαφορετικής επεξεργασίας των αντικειμένων σε διαφορετικά βίντεο.

Ιδιαίτερη αναφορά αξίζει να γίνει για έναν παράγοντα που δυσκολεύει σημασιολογικά την έγκυρη γενική ανίχνευση αντικειμένων στα frames, τη χρήση των αντικειμένων. Ας σκεφτούμε για παράδειγμα ένα χώρο μαγειρικής όπου διάφορα υλικά και εργαλεία βρίσκονται πάνω σε ένα πάγκο. Ο δράστης στο βίντεο αλληλεπιδρά με ορισμένα από αυτά, τα οποία πρέπει να εντοπιστούν. Ωστόσο εδώ εμφανίζονται τα εξής προβλήματα. Αρχικά, ο πάγκος μπορεί να περιέχει πληθώρα αντικειμένων τα οποία επίσης θα ανιχνευτούν αλλά δε θα πρέπει να ληφθούν υπόψιν στη σημασιολογία καθώς δε χρησιμοποιούνται. Τώρα από τα αντικείμενα που χρησιμοποιούνται, τα εργαλεία (όπως το μαχαίρι) υπερκαλύπτονται από τα χέρια του δράστη ή τα άλλα χρησιμοποιούμενα αντικείμενα στα οποία εφαρμόζονται. Αυτό αποκλείει τη δυνατότητα εντοπισμού τους από ένα σύστημα που βασίζεται αποκλειστικά στην όραση.

Από την άλλη, τα σκεύη που χρησιμοποιούνται μπορεί να παραμένουν ακίνητα (όπως ένα ταψί στο οποίο προσθέτουμε υλικά) και να μην φαίνεται η άμεση χρήση τους ή να χρησιμοποιούνται μόνο σε περιπτώσεις υψηλής κίνησης ή και αλληλεπίδρασης με τα χέρια και άρα υψηλής οπτικής μεταβλητότητας (όπως ένας τρίφτης). Όπως ήδη αναφέραμε εξάλλου, αντικείμενα όπως οι πρώτες ύλες αλλάζουν σύσταση και δομή κατά τη διάρκεια της δράσης. Τέλος, στις επισημειώσεις και στη σημασιολογία λαμβάνονται υπόψιν αντικείμενα τα οποία δεν μπορούμε να παρατηρήσουμε. Για παράδειγμα στη δράση “εξαγωγή από ψυγείο”, γνωρίζουμε ποια αντικείμενα συνδέονται σημασιολογικά, αλλά δεν μπορούμε να δούμε ποιο είναι το αντικείμενο που εξάγεται στη συγκεκριμένη περίπτωση παρά μόνο αφού εξαχθεί, οπότε έχει τελειώσει και η δράση. Θα δούμε στη συνέχεια πώς μοντελοποιούμε το πρόβλημα της ανίχνευσης αντικειμένων υπό αυτές τις δυσκολίες.

4.4.2 Η Εισαγωγή της Περιοχής Ενδιαφέροντος

Τονίσαμε νωρίτερα τις δυσκολίες εντοπισμού των αντικειμένων σε εικόνες όπως τα frames του βίντεο. Καταλήξαμε στο ότι πέρα από τις προκλήσεις που ενέχει η ανίχνευση αυτή καθαυτή, πρέπει σημασιολογικά να καταλήξουμε στο αν ένα αντικείμενο χρησιμοποιείται στη διάρκεια μιας δράσης. Επιπλέον, πρέπει να αποφανθούμε για τον εντοπισμό αντικειμένων τα οποία δεν είναι οπτικά ευδιάκριτα. Στο σημείο αυτό θα προχωρήσουμε σε μια παραδοχή η οποία, παρότι ασθενής, μπορεί να συρρικνώσει αρκετά τις περιοχές μελέτης. Η παραδοχή αυτή αφορά το σχηματισμό περιοχών ενδιαφέροντος οι οποίες περιλαμβάνουν άνθρωπο ή αποτελούν προσκήνιο. Ο διαχωρισμός σε περιοχές είναι σύμφωνος με την διαίσθησή μας: πρώτα, η δράση εκτελείται από άνθρωπο, άρα είναι βέβαιο ότι κοντά στα αντικείμενα δράσης θα βρίσκεται ο δράστης. Έπειτα, η δράση συνήθως σχετίζεται με κίνηση, οπότε το να ορίσουμε το προσκήνιο ως χώρο δράσης είναι μια κίνηση προς τη σωστή κατεύθυνση.

Σε περισσότερη λεπτομέρεια, χρησιμοποιούμε τον ανιχνευτή του [68] για ανίχνευση του ανθρώπου, προεκπαιδευμένο στο PASCAL VOC 2007 γνωρίζοντας εκ των προτέρων ότι μόνο ένας άνθρωπος θα εμφανίζεται στο βίντεο. Τρέχουμε τον ανιχνευτή ανθρώπων ανά 10 frames (περίπου 0.35 δευτερόλεπτα) υποθέτοντας ότι σε εργασίες μαγειρικής οι κινήσεις του δράστη θα είναι περιορισμένες στο χώρο και με μικρή σχετικά ταχύτητα, οπότε η θέση του δε θα αλλάζει σημαντικά στα 10 αυτά frames. Έτσι, θεωρούμε για τα 10 αυτά frames ταυτόσημα αποτελέσματα ανίχνευσης. Σε frames όπου ο ανιχνευτής αποτυγχάνει να εντοπίσει περιοχή ανθρώπου, εξισώνουμε την περιοχή ανθρώπου με αυτή της τελευταίας ανίχνευσης. Στην ειδική περίπτωση που ο ανιχνευτής δώσει κενή έξοδο στο πρώτο frame, θεωρούμε ότι ακόμα δεν έχει εισέλθει ο άνθρωπος κι έτσι αφήνουμε κενό το αποτέλεσμα. Σε frames όπου ο ανιχνευτής δώσει περισσότερες από μία περιοχές ανθρώπου, λαμβάνουμε το ελάχιστο ορθογώνιο που περικλείει την ένωση των περιοχών αυτών.

Για ανίχνευση προσκηνίου χρησιμοποιούμε μοντέλα GMM. Το αποτέλεσμα είναι συνήθως ένα σύνολο περιοχών τις οποίες συνενώνουμε και πάλι όπως και στην περίπτωση των περιοχών ανθρώπου. Ο αλγόριθμος ανίχνευσης προσκηνίου τρέχει όλα τα frames του βίντεο και αρχικοποιεί ένα μοντέλο παρασκηνίου από τα N πρώτα frames του βίντεο, ενώ δεν γίνεται ειδική διαχείριση των frames στα οποία έχουμε κενό αποτέλεσμα ανίχνευσης. Η μεταβλητή N είναι ρυθμιζόμενη παράμετρος και η τιμή της επιλέχθηκε ξεχωριστά για κάθε βίντεο (από 30 ως 95), ως συνάρτηση του συνολικού μήκους του βίντεο αλλά και του μήκους του αρχικού τμήματος βίντεο πριν την είσοδο



Σχήμα 4.9: Ένα παράδειγμα εξαγωγής περιοχής ενδιαφέροντος. Σε πρώτη φάση τρέχουμε ανιχνευτή ανθρώπων βασισμένο σε DPM. Σε δεύτερη φάση, τρέχουμε ανιχνευση προσκηνίου με GMM. Συνενώνουμε το πλήθος των περιοχών που προκύπτουν σε όλη τους την έκταση. Επεκτείνουμε την περιοχή ενδιαφέροντος κατά ένα μικρό αριθμό πίξελ δεξιά και αριστερά, ενώ την αφήνουμε να λάβει όλο το ύψος της εικόνας. Έτσι μπορούμε να συλλάβουμε τμήματα αντικειμένων χώρου τα οποία μας βοηθούν στην ανιχνευσή τους όταν χρειάζεται, χωρίς όμως να κείνται εντός της περιοχής προσκηνίου ή εύρεσης ανθρώπου.

του ανθρώπου και την εκκίνηση των δράσεων. Επιπλέον, καθορίστηκε η τιμή της παραμέτρου ρυθμού μάθησης στην τιμή 0.003, ώστε να δίνουμε χρονικό περιθώριο στην εξέλιξη δράσεων στις οποίες ο δράστης διατηρεί μεγάλο μέρος του σώματός του σταθερό. Τέτοιου είδους δράσεις εμφανίζονται συχνά στο σύνολο δεδομένων μας καθώς συχνά έχουμε εργασίες λεπτομέρειας όπου μόνο τα άκρα των χεριών κουνιούνται, όπως ο τεμαχισμός ενός φρούτου σε φέτες. Τέλος, επιλέγουμε ως ελάχιστο εμβαδόν μιας υποψήφιας περιοχής προσκηνίου τα 250 τετραγωνικά pixels.

Για την εξαγωγή των περιοχών ενδιαφέροντος, έχοντας το πολύ μία περιοχή ανθρώπου και το πολύ μία περιοχή προσκηνίου, προχωράμε αρχικά στη συνένωση αυτών των περιοχών. Αν και οι δύο περιοχές είναι κενές, τότε δεν υπάρχει περιοχή ενδιαφέροντος αφού μάλλον δεν υπάρχει άνθρωπος να δράσει. Σε διαφορετική περίπτωση, λαμβάνουμε και πάλι το ελάχιστο ορθογώνιο που περικλείει την ένωση των δύο περιοχών. Στη συνέχεια, επεκτείνουμε την περιοχή ενδιαφέροντος ως εξής: α) στον άξονα x, διευρύνουμε κατά 15 pixels το μέγιστο αριστερά και δεξιά, όπου αυτό είναι δυνατό και β) στον άξονα y, αφήνουμε την περιοχή ενδιαφέροντος να επεκταθεί σε όλο το ύψος της εικόνας ώστε να μπορεί συμπεριλάβει αντικείμενα σε όλο το ύψος, από τον πάγκο και τα αντικείμενα επί αυτού μέχρι τα ψηλότερα ντουλάπια.

Κατά τη διαδικασία κατασκευής της περιοχής ενδιαφέροντος είναι φανερά τα πλεονεκτήματα που εισάγει αυτή η σχεδίαση. Συσχετίζει άνθρωπο, περιοχές και αντικείμενα ενδιαφέροντος, εκτείνεται και εξελίσσεται δυναμικά ώστε να μην αποκλείει αντικείμενα ενδιαφέροντος και υπολογίζεται σχετικά εύκολα υπολογιστικά. Ωστόσο,

τόσο ο ορισμός της όσο και οι παραδοχές που τον συνοδεύουν αφήνουν χώρο για άλλες προκλήσεις. Ξεκινώντας από τη μεταχείριση της εξόδου του ανιχνευτή ανθρώπου και λαμβάνοντας υπόψιν την αραιή χρονικά ανίχνευση και τη μέθοδο παρεμβολής που χρησιμοποιούμε βλέπουμε ότι το κόστος μιας αποτυχίας ανίχνευσης μεταφέρει την παρελθοντική θέση στο μέλλον, κάτι το οποίο μπορεί να προκαλέσει αρκετά λανθασμένη εκτίμηση της θέσης του ανθρώπου. Δεδομένου ότι το μοντέλο ανιχνευτή που εφαρμόζεται είναι προεκπαίδευμένο, υπάρχει χώρος για ανεπαρκείς εκτιμήσεις. Ακόμα χειρότερα, το σφάλμα μιας λανθασμένης ανίχνευσης μπορεί να συσσωρεύεται καθώς η αποτυχία επαναλαμβάνεται σε διαδοχικά frames. Στην περίπτωση εύρεσης παραπάνω της μιας περιοχής ανθρώπου, η συνένωση των περιοχών μπορεί να περικλείει περιοχές εντελώς άσχετες με την πραγματική περιοχή και άρα λανθασμένη επέκταση της περιοχής ενδιαφέροντος. Το ίδιο αποτέλεσμα μπορεί να προκαλέσει η παρόμοια μεταχείριση των πολλαπλών περιοχών προσκηνίου, οι οποίες μπορεί να περιέχουν και σφάλματα λόγω μεταβολών σκίασης και φωτισμού. Ο χαμηλός ρυθμός μάθησης καθυστερεί την εκμάθηση νέου μοντέλου υποβάθρου και συνεισφέρει στο να μην απορρίπτει προσκήνιο το οποίο παραμένει σταθερό για μικρό χρονικό διάστημα. Αν από την άλλη μέρος του προσκηνίου χαρακτηριστεί ως παρασκήνιο, ο μικρός ρυθμός μάθησης δυσχεραίνει την επαναφορά του στο προσκήνιο. Τέλος, η συνένωση περιοχών ανθρώπου και προσκηνίου και η επέκταση του αποτελέσματος είναι κάπως αυθαίρετη.

Σίγουρα πολλά από τα παραπάνω ζητήματα μπορούν να μοντελοποιηθούν εναλλακτικά, με διαφορετικά πλεονεκτήματα και μειονεκτήματα. Σε αυτό το σημείο θα σταθύμε στη δύναμη της μοντελοποίησης που τελικά χρησιμοποιήσαμε μπροστά στα παραπάνω ζητήματα, τονίζοντας όμως ότι αυτά συνεχίζουν να υπάρχουν ως κίνδυνοι. Το μοντέλο ανιχνευτή έχει προεκπαίδευτεί σε ένα μεγάλο και γενικό σύνολο δεδομένων κι έτσι αναμένουμε να γενικεύει με ευρωστία. Οι πολλαπλές περιοχές ανίχνευσης συνήθως παρουσιάζουν πολύ υψηλό ποσοστό επικάλυψης, όπως διαπιστώνεται πειραματικά, οπότε η συνένωσή τους είναι ασφαλής. Η αποκοπή μέρους του προσκηνίου δεν επηρεάζει σημαντικά το αποτέλεσμα καθώς η ένωση με την αρκετά ευρύτερη περιοχή ανθρώπου θα διατηρήσει την περιοχή ενδιαφέροντος πρακτικά αναλλοίωτη. Εξάλλου, αποκοπή προσκηνίου μπορεί να συμβεί σε περίπτωση μακράς σε χρονικό μήκος ακινησίας του δράστη, άρα το αποτέλεσμα της ανίχνευσης ανθρώπου θα είναι πρακτικά σταθερό. Τέλος, η επέκταση των περιοχών ενδιαφέροντος είναι η μόνη λύση για να εντοπίσουμε και αντικείμενα με τις δυσκολίες ανίχνευσης που έχουν αναφερθεί νωρίτερα. Αν και δεν είναι μαθηματικά φορμαλισμένη, η χρήση αυτής της επέκτασης μπορεί να αποδίδει πειραματικά, να περιλαμβάνει γειτονικά και χρήσιμα στη δράση αντικείμενα αλλά και να αποκόπτει ταυτόχρονα παρευρισκόμενα αλλά μη χρησιμοποιούμενα αντικείμενα.

4.4.3 Η Ανίχνευση Αντικειμένων

Όπως σε κάθε πρόβλημα ανίχνευσης αντικειμένων, έτσι και σε αυτή την εργασία, προκειμένου να επιλέξουμε τον κατάλληλο ανιχνευτή πρέπει να ορίσουμε τις επιθυμητές προδιαγραφές, να εξετάσουμε ενδελεχώς το πώς οι δυσκολίες του συνόλου δεδομένων πρέπει να μοντελοποιηθούν αποτελεσματικά και να προβούμε στους απαραίτητους συμβιβασμούς. Νωρίτερα παρουσιάσαμε τις ιδιαιτερότητες του συνόλου δεδομένων που χρησιμοποιήσαμε σχετικά με την ανίχνευση αντικειμένων. Θα δείξουμε τώρα τη δική μας προσέγγιση και στη συνέχεια το πώς οι επιλογές που κάναμε στη σχεδίαση ανταποκρίνονται στις προκλήσεις του συνόλου δεδομένων.



Σχήμα 4.10: Η οπτική ανίχνευση αντικειμένων μπορεί εύκολα να λειτουργήσει αποτελεσματικά, με τη μέθοδο Ταιριάσματος Προτύπων ή άλλη, σε εικόνες όπου γνωστά αντικείμενα κείνται ακίνητα και πρακτικά χωρίς επικαλύψεις. Εν τούτοις, μόνο η οπτική πληροφορία εισάγει με μεγάλη πιθανότητα σφάλματα μη χρήσης. Στην εικόνα, μπορούμε εύκολα να διακρίνουμε αντικείμενα όπως φρούτα και μαχαίρι πάνω στον πάγκο, ωστόσο ο δράστης δεν τα χρησιμοποιεί. Η πληροφορία ανίχνευσης αυτών των αντικειμένων είναι θόρυβος που επιβαρύνει τον τελικό ταξινομητή σημασιολογίας. Το πρόβλημα λύνεται με την εισαγωγή της περιοχής ενδιαφέροντος, χάρη στην οποία διαπιστώνουμε ότι τα εντοπιζόμενα αντικείμενα δεν βρίσκονται στην περιοχή προσκηνίου και άρα δε λαμβάνονται υπόψιν.

Αρχικά, προβήκαμε σε ανίχνευση αντικειμένων ξεχωριστά στα frames και επιλέξαμε ανιχνευτή Ταιριάσματος Προτύπων με χρήση Αθροίσματος Τετραγωνικών Διαφορών (SSD) και τρέξαμε δειγματοληπτώντας το βίντεο ανά 10 frames, ως ένα συνδυασμό ταχύτητας και κάποιας ευρωστίας. Για κάθε βίντεο, εξαγάγαμε πρότυπα για ένα υποσύνολο των αντικειμένων τα οποία εμφανίζονται στο βίντεο. Τα αντικείμενα τα οποία επιλέχθηκαν για αναπαράσταση μέσω προτύπων ήταν αυτά τα οποία εμφανίζονταν σε αρκετά διαδοχικά frames με σταθερή μορφή και δυνατότητα αναπαράστασης. Για παράδειγμα, ένα φρούτο που κείται για κάποια δευτερόλεπτα στον πάγκο ακίνητο, μπορεί να εντοπισθεί αποτελεσματικά. Επιπλέον, αντικείμενα όπως το ψυγείο, οι ντουλάπες και τα συρτάρια, αναπαρίστανται από το μέρος του επίπλου που εμφανίζεται μόλις αυτά χρησιμοποιηθούν (ανοίξουν). Από την άλλη, αντικείμενα δυσδιάκριτα, όπως ένα μαχαίρι ή συνεχώς κινούμενα αντικείμενα, όπως μια κουτάλα, δεν αναπαρίστανται οπτικά καθώς δεν αρκεί μια εικόνα για να εντοπισθούν σε μεγάλη διάρκεια βίντεο. Περαιτέρω, σε κάθε βίντεο, μπορεί να χρησιμοποιούνται περισσότερα του ενός πρότυπα για την ίδια κατηγορία αντικειμένων. Μάλιστα για κάθε βίντεο γίνεται διαφορετική αναπαράσταση για την ίδια κατηγορία αντικειμένων, δηλαδή δεν χρησιμοποιούμε το ίδιο πρότυπο για πολλά βίντεο. Τέλος, αποφεύχθηκε η αναπαράσταση σταθερών χωρικών αντικειμένων, όπως εστία μαγειρέματος ή θήκη δοχείων μπαχαρικών, καθώς ο συνεχής τους εντοπισμός μας είναι άχρηστος αφού εμφανίζονται στα περισσότερα frames όμως συχνά δεν επηρεάζουν τη δράση λόγω της εμφάνισής τους και μόνο.

Θα σταθούμε εδώ για να αφουγκραστούμε τα πλεονεκτήματα που έχει μια τέτοια επιλογή δεδομένων των ιδιαιτεροτήτων του συνόλου δεδομένων. Η απόσταση από την κάμερα, παράγοντας που δυσχεραίνει την ανίχνευση των αντικειμένων, πλέον δεν επηρεάζει: έχουμε να ανιχνεύσουμε ένα σταθερό πρότυπο κι όχι να επισημάνουμε ακριβώς μια περιοχή του χώρου χωρίς πρότερη πληροφορία. Η αυστηρή ανίχνευση προτύπων μπορεί να αντισταθμίσει την χαμηλή μεταβλητότητα μεταξύ των διαφορετικών κλάσεων. Προς αυτή την κατεύθυνση, χρησιμοποιούμε αυστηρό κατώφλι στην έξοδο του SSD συστήματος, ίσο με 0.99. Το κατώφλι αυτό απορρίπτει ακόμα και παρόμοια αντικείμενα. Ωστόσο υπάρχει ο κίνδυνος να απορρίψει και διαφορετικές εμφανίσεις του ίδιου αντικειμένου. Επίσης η στρατηγική αυτή προστατεύει από την υψηλή μεταβλητότητα μεταξύ αντικειμένων της ίδιας κατηγορίας ή ακόμα και την αλλαγή σύστασης των αντικειμένων: μια ακέραια ντομάτα και μια σάλτσα ντομάτας απεικονίζονται δικαίως ως διαφορετικά αντικείμενα αλλά αναφέρονται στην ίδια κλάση. Η επιλογή αναπαράστασης των αντικειμένων περιοχών (όπως το ψυγείο) με το μεταβλητό τους μέρος επιλύει και το ζήτημα μοντελοποίησής τους. Τέλος, με την χρήση διαφορετικών προτύπων ανά βίντεο, επιλύουμε και το ζήτημα μεταβλητότητας των αντικειμένων μεταξύ των βίντεο. Δηλαδή η επιλογή αυτού του μοντέλου αντικρούει 6 από τους 8 αναφερθέντες παράγοντες δυσκολίας ανίχνευσης.

Στη συνέχεια, εξετάζουμε το πρόβλημα της χρήσης. Είναι φανερό ότι σε κάθε δράση μας ενδιαφέρουν μόνο τα αντικείμενα που χρησιμοποιούνται στη διάρκειά της. Εν τούτοις, η χρήση των αντικειμένων, συχνότερα με τα χέρια, οδηγεί σε σημαντικές επικαλύψεις των αντικειμένων και τεράστια οπτική μεταβλητότητα. Η πρώτη απάντησή μας σε αυτό είναι η παρακολούθηση αντικειμένων τα οποία μπορούν να παραμένουν αναλλοίωτα ή να αναπαρίστανται από κάποιο σταθερό μέρος τους, π.χ. τα κομμάτια ενός φρούτου που κόβεται τα οποία έχουν ήδη κοπεί, παραμένουν σταθερά καθώς αυτό συνεχίζει να κόβεται. Η παρατήρηση οπτικά αναλλοίωτων αντικειμένων ωστόσο ενέχει άλλους κινδύνους: τα αντικείμενα αυτά μπορούν να εντοπίζονται σε ποικίλες χρονικές στιγμές χωρίς να αφορούν πραγματικά τη δράση. Αν, για παράδειγμα, ένα μαγειρικό σκεύος περιέχει φαγητό το οποίο βράζει πάνω στην εστία, ενώ ταυτόχρονα ο δράστης ασχολείται με μια άλλη εργασία, όπως προετοιμασία των υλικών που θα βράσουν, τότε το μαγειρικό σκεύος συνεχίζει να εντοπίζεται παρότι δεν αφορά τη δράση. Με άλλα λόγια, απαιτούμε από τον ανιχνευτή όχι απλά να εντοπίζει την εμφάνιση ή μη των αντικειμένων αλλά να λαμβάνει μια απόφαση σχετικά με το αν η εμφάνιση αυτή παρουσιάζει ενδιαφέρον για τη δράση. Για το λόγο αυτό, εισαγάγαμε την περιοχή ενδιαφέροντος, η κατασκευή της οποίας περιγράφηκε νωρίτερα. Κρατάμε ως ενδιαφέροντα τα αντικείμενα των οποίων το ορθογώνιο ανίχνευσης εμφανίζει μη κενή τομή με την περιοχή ενδιαφέροντος. Έτσι, περιμένουμε ότι στο προηγούμενο παράδειγμα, το μαγειρικό σκεύος δε θα περιέχεται στην περιοχή ενδιαφέροντος, στην οποία θα περιέχονται τα εργαλεία και τα υλικά που όντως αφορούν τη δράση. Φυσικά, υπάρχουν περιπτώσεις που αυτή η υπόθεση δεν επαληθεύεται, ειδικά αν λάβουμε υπόψιν την δειγματοληψία του βίντεο τόσο στην ανίχνευση αντικειμένων όσο και στην ανίχνευση ανθρώπων. Προς αυτή την κατεύθυνση, επιβάλλαμε περιορισμό στον συνολικό αριθμό των αναπαραστάσεων για κάθε αντικείμενο έτσι ώστε να ελαττώσουμε τις μη ενδιαφέρουσες ανιχνεύσεις.

Υπό τις παραπάνω παραδοχές και σχεδιαστικές επιλογές, τα αποτελέσματα ανίχνευσης είναι αρκετά εύρωστα και συνεπή. Μοντελοποιούνται ικανοποιητικά οι περισσότερες από τις προκλήσεις που εμφανίζει το σύνολο δεδομένων, ενώ αποκτούμε πληροφορία για τη χρησιμότητα της εμφάνισης στη δράση, μέσω της περιοχής ενδιαφέροντος. Ωστόσο, παρά τα προηγούμενα, εμφανίζονται ταυτόχρονα άλλα προβλήματα.

Περιορίσαμε την ανίχνευση σε σχετικά σταθερά αντικείμενα με όριο στον αριθμό αναπαραστάσεων ανά αντικείμενο, αφήνοντας πολλαπλές εμφανίσεις αντικειμένων χωρίς αναπαράσταση. Ταυτόχρονα, το αυστηρό κατώφλι ταιριάσματος αποκλείει την ανίχνευση αντικειμένων χωρίς αναπαράσταση. Αναλογιζόμενοι αυτές τις δυσκολίες αξιοποιούμε επιπλέον αίσθηση για να ανιχνεύσουμε μεγαλύτερο εύρος αντικειμένων, μέσω της ομιλίας στο βίντεο.

4.4.4 Η Αξιοποίηση της Ομιλίας στην Ανίχνευση Αντικειμένων

Με την εισαγωγή της περιοχής ενδιαφέροντος θίξαμε τα ζητήματα ταυτοποίησης της χρήσης ή μη και του συσχετισμού των παρατηρούμενων αντικειμένων με τον δράστη. Μια άλλη τεχνική είναι να αξιοποιήσουμε τη συμπληρωματική φύση του λόγου και της οπτικής αντίληψης για να λάβουμε πληροφορία για τα χρησιμοποιούμενα αντικείμενα. Για το λόγο αυτό προβήκαμε στη δημιουργία υποτίτλων για τα βίντεο του test συνόλου δεδομένων. Η ιδέα είναι ότι η χρήση των υποτίτλων περιέχει πλούσια πληροφορία για τα αντικείμενα (εργαλεία και υλικά) που σχετίζονται άμεσα με τη δράση αλλά και για αντικείμενα τα οποία εμφανίζονται πρώτη φορά. Από τη γωνία του εντοπισμού αντικειμένων, ήδη γίνεται εμφανές το ότι τα δύο προβλήματα ανίχνευσης (με χρήση όρασης και με χρήση λόγου) είναι δυικά και συμπληρωματικά από τη φύση τους. Το πρώτο εστιάζει σε οπτικά αναλλοίωτα χαρακτηριστικά και εντοπίζει αντικείμενα που σχετίζονται με τη δράση χωρικά αλλά όχι άμεσα με το δράστη (πχ τα σκεύη, αλλά όχι τα εργαλεία), οπότε ένα τέτοιο σύστημα είναι εύρωστο σε αντικείμενα που παραμένουν ακίνητα αλλά χρήσιμα. Από την άλλη, ένα σύστημα εντοπισμού βασισμένο στο λόγο εστιάζει σε γλωσσικά αναλλοίωτα χαρακτηριστικά και τείνει να ανιχνεύει αντικείμενα που αναφέρονται συχνά και έχουν άμεση σχέση με το χρήστη, όπως τα εργαλεία και τα υλικά, τα οποία συνήθως έχουν άμεση αλληλεπίδραση με τα χέρια και παραμορφώνονται ή μεταβάλλονται δομικά.

Η προσέγγιση εξαγωγής των εμφανίσεων των αντικειμένων γλωσσικά είναι αντίστοιχη με τη διαδικασία που εφαρμόστηκε στην περίπτωση της οπτικής ανίχνευσης. Αντί για την οπτική αμεταβλητότητα των σκευών ή αντικειμένων χώρου, εκμεταλλεύμαστε τη λεκτική αμεταβλητότητα των άμεσα χειριζόμενων αντικειμένων. Έτσι, υλοποιούμε ένα γλωσσικό template matching, αναζητώντας λέξεις κλειδιά μέσα στους υποτίτλους. Οι λέξεις αυτές ταυτίζονται με τις κλάσεις αντικειμένων του λεξιλογίου που χρησιμοποιούμε και η εμφάνιση μιας τέτοιας λέξης κατά μήκος ενός υποτίτλου αυτομάτως δηλώνει την ανίχνευση του αντικειμένου αυτού σε όλη τη διάρκεια του υποτίτλου. Επιπλέον, θεωρούμε συχνή τη χρήση κάποιου αντικειμένου και αφού έχει αναφερθεί, οπότε επεκτείνουμε το διάστημα μεταφράζοντας το παραπάνω εγχειρήμα με χρήση ασαφούς λογικής: το αντικείμενο θα εμφανίζεται “τη στιγμή που αναφέρεται και λίγο μετά”. Τη φράση αυτή αποδίδει μια συνάρτησης μετοχής (η οποία μπορεί διαισθητικά να ληφθεί ως η εκ των υστέρων κατανομή της πιθανότητας εμφάνισης του αντικειμένου στο βίντεο ως προς το χρόνο, με δεδομένη την παρατήρηση του ονόματος κλάσης στον τρέχοντα υπότιτλο. Η συνάρτηση αυτή είναι μοναδιαία στο χρονικό διάστημα που ορίζει ο υπότιτλος, ενώ φθίνει γραμμικά μετά το χρονικό πέρας του υποτίτλου, μέχρι μηδενισμού, για χρονικό διάστημα ίσο με το 10% της διάρκειας του υποτίτλου.

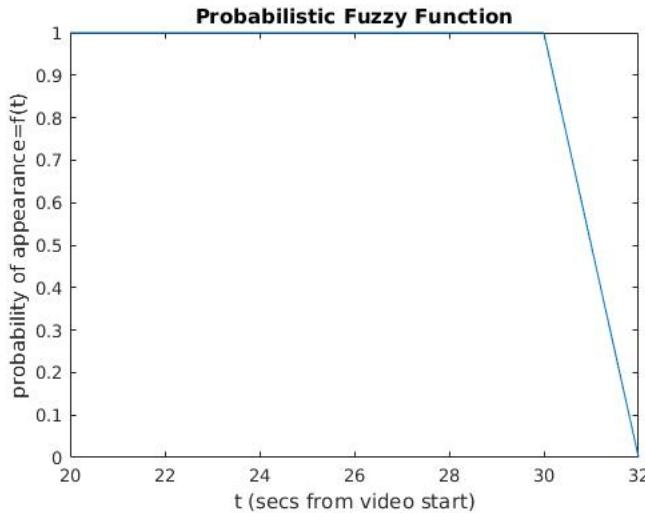


Σχήμα 4.11: Πολλές φορές η ανίχνευση αντικειμένων σε μια εικόνα είναι δυσεπίλυτο πρόβλημα ακόμα και για τον άνθρωπο. Δεν αρκεί μια εικόνα, κατά πάσα πιθανότητα, έτσι ώστε να συμπεράνουμε με ασφάλεια τι κρατάει στα χέρια του ο άνθρωπος της εικόνας. Σε αυτό μπορεί να συμβάλλει η εισαγωγή πρότερης γνώσης ή η ομιλία. Σε τέτοια βίντεο είναι αρκετά συχνή η συμπερίληψη των χρησιμοποιούμενων αντικειμένων στον λόγο. Συνδυάζοντας κατάλληλα λόγο και εικόνα λαμβάνουμε αρκετά πιο ισχυρές ενδείξεις για την εμφάνιση ενός αντικειμένου.

4.4.5 Συνδυασμός Εικόνας και Γλώσσας

Στα παραπάνω, τονίστηκε η συμπληρωματικότητα του λόγου και της οπτικής αντίληψης. Είναι φανερό ότι η μέγιστη ευρωστία αποκτάται συνενώνοντας τις δύο αυτές μεθόδους. Συνδυάζουμε τα αποτελέσματα ως εξής: Πρώτα, μετατρέπουμε τα αποτελέσματα οπτικής ανίχνευσης των αντικειμένων σε δυαδικά ιστογράμματα. Αυτό γίνεται ελέγχοντας τα frames στα οποία υπάρχει ορθογώνιο ανίχνευσης το οποίο περιέχεται έστω και μερικώς στην περιοχή ενδιαφέροντος και σημειώνοντας τιμή 1 για τα frames αυτά. Για αντικείμενα τα οποία αναπαρίστανται με περισσότερα από ένα πρότυπα, συνενώνουμε τα ιστογράμματα που προκύπτουν από την ανίχνευση των διαφορετικών προτύπων έτσι ώστε να έχουμε τιμή 1 αν έστω ένα από τα πρότυπα του αντικειμένου εντοπίσθηκε. Στη συνέχεια εξάγουμε τα ιστογράμματα από τους υποτίτλους, λαμβάνοντας υπόψιν την ασαφή συνάρτηση που πρέπει να εισάγουμε, οπότε εδώ δεν έχουμε δυαδικά ιστογράμματα. Αθροίζουμε τα παραπάνω ιστογράμματα (οπτικής ή γλωσσικής αντίληψης) για κάθε βίντεο και κλάση αντικειμένου. Είναι φανερό ότι τώρα θα έχουμε τιμές μεγαλύτερες της μονάδας. Ψαλιδίζουμε στο 1.1 κάθε τιμή μεγαλύτερη του 1.1. Με αυτόν τον τρόπο δίνουμε επιπλέον δύναμη στα frames όπου γλώσσα και όραση συμφωνούν αλλά, συγχρόνως, δεν τα αφήνουμε να έχουν τόση δύναμη ώστε να εκσφενδονίσουν οποιοδήποτε διάστημα στο οποίο συμμετέχουν σε μια τιμή που θα υποδηλώνει μεγάλη βεβαιότητα για την εμφάνιση του αντικειμένου σε όλο το διάστημά του.

Τώρα για την αναπαράσταση ενός τμήματος βίντεο, λαμβάνουμε τα αθροιστικά ιστογράμματα πάνω στα frames που το αποτελούν. Προκύπτει ένα διάνυσμα όπου κάθε



Σχήμα 4.12: Θεωρούμε αυξημένη την πιθανότητα ένα αντικείμενο να εμφανίζεται ”όσο διαρκεί ο υπότιτλος και λίγο μετά”. Η φράση αυτή ερμηνεύεται μαθηματικά με μια ασαφή συνάρτηση μετοχής, η οποία έχει τιμή 1 στα χρονικά όρια του υποτίτλου και φθίνει γραμμικά μέχρι μηδενισμού από το πέρας του υποτίτλου μέχρι και 10% της διάρκειάς του έπειτα από αυτόν. Για παράδειγμα, η συνάρτηση μετοχής για ένα αντικείμενο που αναφέρεται στον υπότιτλο στο χρονικό παράθυρο μεταξύ 20 και 30 δευτερολέπτων έχει τη μορφή της εικόνας.

Θέση περιέχει ένα αριθμό, που όσο μεγαλύτερος είναι τόσο πιο σίγουροι είμαστε για την εμφάνιση του αντικειμένου στο frame. Διαιρούμε το διάνυσμα με το μήκος του διαστήματος σε frames, λαμβάνουμε δηλαδή το μέσο ιστόγραμμα κατά μήκος του τμήματος βίντεο. Μετατρέπουμε το ιστόγραμμα σε δυαδικό επιλέγοντας κατώφλι 0.8. Το κατώφλι αυτό απορρίπτει λανθασμένες ευρέσεις αντικειμένων μέσα στην περιοχή ενδιαφέροντος, όπου ο όρος «λανθασμένες» εδώ, αφορά την αδιαφορία της εμφάνισης ως προς τη δράση, ενώ, ταυτόχρονα, ενισχύει τμήματα μεγάλης βεβαιότητας χωρίς αυτά να επισκιάζουν τμήματα μικρότερης βεβαιότητας. Τελικά λαμβάνουμε ένα δυαδικό διάνυσμα μήκους ίσο με το πλήθος των αντικειμένων του λεξιλογίου.

Πριν κλείσουμε το κεφάλαιο αυτό πρέπει να σχολιάσουμε δύο πράγματα. Αρχικά, σε θεωρητικό επίπεδο, θα δούμε πώς ο συνδυασμός γλώσσας και όρασης, μαζί με τις σχεδιαστικές μας επιλογές, οδήγησαν σε μια δυνατή μοντελοποίηση των δυσκολιών του συνόλου δεδομένων. Οι 8 απαριθμημένες προκλήσεις που αφορούν την οπτική αντίληψη δεν επηρεάζουν τη γλωσσική αντίληψη, αλλά είδαμε πώς, ακόμα και μέσω της οπτικής αντίληψης, αντιμετωπίσθηκαν δυσκολίες όπως η μεταβλητότητα, η απόσταση από την κάμερα, η αναπαράσταση χώρων και η αλλαγή δομής των αντικειμένων στο χρόνο. Παράλληλα, η χρησιμότητα των εμφανίσεων προσεγγίστηκε με την χρήση της περιοχής ενδιαφέροντος, ενώ το άθροισμα των γλωσσικών και οπτικών ιστογραμμάτων μπορεί να αντισταθμίζει λάθη του ενός από τα δύο συστήματα. Το άλλο θέμα στο οποίο πρέπει να σταθούμε είναι, σε πειραματικό επίπεδο, να αξιολογήσουμε την απόδοση της ανίχνευσης αντικειμένων ανά τμήμα ενδιαφέροντος. Το πρόβλημα αυτό είναι any-of, δηλαδή κάθε τμήμα μπορεί να ανήκει σε πολλές κατηγορίες ταυτόχρονα, με άλλα λόγια στο πρόβλημά μας, κάθε τμήμα να περιέχει πολλά από τα αντικείμενα. Οπότε θα χρησιμοποιήσουμε μετρικές multilabel ταξινομητών. Ως ground truth λαμβάνουμε τις επισημειώσεις του MPII Cooking Dataset 2 για τα αντικείμενα που χρησιμοποιούνται σε κάθε δράση. Τα αποτελέσματα φαίνονται στον πίνακα 4.1.

Metric	Value
Accuracy	77.95%
Recall	0.91
Precision	0.87
F1-score	0.88

Πίνακας 4.1: Αποτελέσματα 4 μετρικών για την ανίχνευση αντικειμένων πάνω σε όλα τα βίντεο του συνόλου δεδομένων test που χρησιμοποιήσαμε. Παρατηρούμε ότι με σύζευξη λόγου και εικόνας μπορούμε να πετύχουμε ικανοποιητικά αποτελέσματα, πολύ πιο εύρωστα από ό,τι με χρήση μόνο ενός από τους δύο τρόπους.

Τα παραπάνω αποτελέσματα και οι δυσκολίες που αντιμετωπίσθηκαν μας υποδεικνύουν κάποια συμπεράσματα για τη συνθετότητα της ανθρώπινης αντίληψης όταν μιλάμε για ανίχνευση αντικειμένων. Παρατηρώντας τη δυνατότητα του ανθρώπου να εντοπίζει τόσο εύκολα το αντικείμενο της επιλογής του σε μια εικόνα που δεν έχει ξαναδεί, πράγματι κανείς εντυπωσιάζεται αν αναλογισθεί πόσο δύσκολο υπολογιστικά πρόβλημα είναι. Εμείς στεκόμαστε στη δυνατότητα του ανθρώπου να αντιληφθεί τα αντικείμενα αυτού του συνόλου δεδομένων για να βγάλουμε κάποια πορίσματα και όχι σε ψυχοφυσικές έρευνες. Αρχικά, ο άνθρωπος δε χρησιμοποιεί μία αντίληψη μόνο για να εξάγει ένα συμπέρασμα. Υπάρχει πάντα πληροφορία από τα συμφραζόμενα αλλά και από παραμορφώσιμα μοντέλα για τη δομή του αντικειμένου. Η σημασιολογία είναι επίσης κάτι πρωτίστης σημασίας στη λήψη αποφάσεων. Στη δική μας περίπτωση, δυσδιάκριτα μαγειρικά μπορούν να διακριθούν από τη χρήση τους, το χειρισμό τους, αλλά και τη σύνδεση με άλλα υλικά. Μάλιστα η ανθρώπινη εμπειρία μπορεί να ανακατασκευάσει τμήματα του αντικειμένου σε εικόνες που αυτό επικαλύπτεται ακόμα και πλήρως, μόνο από τη σημασιολογία της δράσης. Ακόμα και το σενάριο μπορεί να εξαχθεί και να δώσει διαφορετική πρότερη πιθανότητα σε αντικείμενα που μπορούν να εμφανιστούν, μοντελοποιώντας έτσι την έλλειψη πρότερης γνώσης για τα αντικείμενα που εμφανίζονται πρώτη φορά. Τέλος, ο άνθρωπος αντιλαμβάνεται τη χρονική αλληλουχία και τα αποτελέσματα που αυτή επιφέρει. Με αυτόν τον τρόπο, μπορεί να γνωρίζει την ύπαρξη ενός αντικειμένου στο χώρο χωρίς να την αντιλαμβάνεται οπτικά. Ταυτόχρονα, αισθάνεται τις μεταβολές στη φύση των αντικειμένων και συχνά είναι σε θέση να τις προβλέψει από την αλληλεπίδρασή τους με τα άλλα αντικείμενα. Συνολικά, μια σχεδιαστική μέθοδος ταιριάσματος, όπως αυτή που επιλέξαμε, καθόλου δεν συνδέεται με τη μοντελοποίηση και τη συνθετότητα της ανθρώπινης αντίληψης, παρότι αποκρίνεται στις προκλήσεις του συνόλου δεδομένων και δίνει καλά αποτελέσματα.

Κεφάλαιο 5

Σημασιολογικό Υποσύστημα: Ο Τύπος Λαβής

Ο τύπος λαβής (grasping type) είναι πρακτικά ο τρόπος με τον οποίο ένας άνθρωπος πιάνει ένα αντικείμενο με τα χέρια του προκειμένου να προβεί σε κάποια ενέργεια. Στην αναγνώριση δράσεων, μπορούμε να εξάγουμε χρήσιμες πληροφορίες από τον τύπο λαβής, τόσο για τον τύπο της δράσης, όσο και για τα όρια μιας δράσης. Από την πλευρά του συστήματος που θα δράσει μετά την αναγνώριση, ανθρώπου ή ρομπότ, ο τύπος λαβής είναι απαραίτητη πληροφορία για την αλληλεπίδραση με τα αντικείμενα.

Στη βιβλιογραφία εμφανίζονται ποικίλες μορφές τύπων λαβής, οι οποίες κατατάσσονται σε κατηγορίες ανάλογα με τα εκάστοτε κριτήρια [67]. Από αυτές, τις ταξινομήσεις, η πιο χρήσιμη στην αναγνώριση δράσεων είναι η κατηγοριοποίηση με βάση τη λειτουργία, σε λαβές ακριβείας και δύναμης [106].

Σε αυτή την εργασία, η προσέγγισή μας στην αναγνώριση τύπου λαβής εκκινεί με την ανίχνευση των χεριών του ανθρώπου-δράστη στο βίντεο. Συνεχίζουμε με την εκπαίδευση πάνω στις εικόνες χεριών για την εξαγωγή του τύπου λαβής με χρήση συνελικτικών χαρακτηριστικών και τελικά εξάγουμε τα ιστογράμματα τύπων λαβής για τα βίντεο των δεδομένων ελέγχου (test). Το κεφάλαιο οργανώνεται ως εξής: Πρώτα



Σχήμα 5.1: Ο σχηματισμός των χεριών στο χειρισμό των αντικειμένων εμπεριέχει σημασιολογική πληροφορία η οποία μπορεί να υποδεικνύει την πρόθεση ή τη δράση αυτή καθαυτή. Στην εικόνα αριστερά για παράδειγμα, ο τρόπος λαβής του μαχαίριού είναι προσανατολισμένος στη δύναμη και μπορούμε να υποθέσουμε ότι το μαχαίρι πρόκειται να χρησιμοποιηθεί σε κάποια εργασία. Αντίθετα στη μεσαία εικόνα, ο τύπος λαβής είναι προσανατολισμένος στην ακρίβεια. Η φυσική απόκριση στη λαβή αυτή είναι να δεχτούμε το μαχαίρι που μας αποδίδεται, όπως στην εικόνα δεξιά. Εικόνα από [255].

γίνεται μια αναφορά σε νευρωνικά δίκτυα και συνελικτικά χαρακτηριστικά. Στη συνέχεια παρουσιάζουμε τη μέθοδο ανίχνευσης χεριών και τα χαρακτηριστικά BING. Τελικά δείχνουμε τη σύζευξη όλων αυτών των μεθόδων με χρήση μη επιβλεπόμενης μάθησης για την εξαγωγή των ιστογραμμάτων τύπων λαβής.

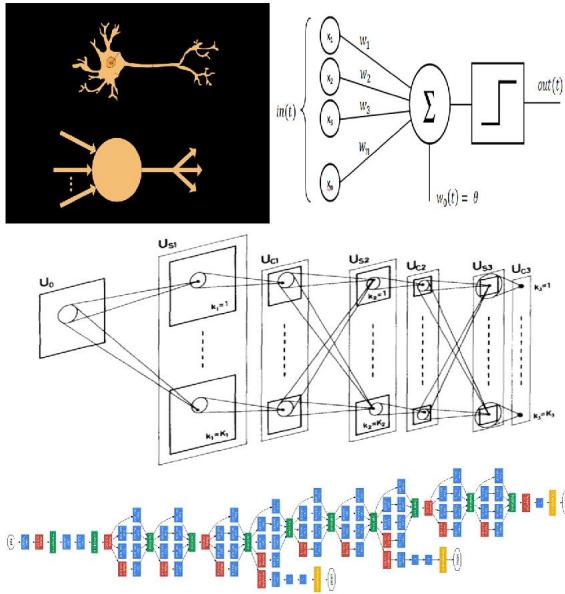
5.1 Θεωρητικό Υπόβαθρο

5.1.1 Νευρωνικά Δίκτυα

Η ιδέα του νευρωνικού δικτύου εισάγεται για πρώτη φορά στα τέλη της δεκαετίας του '50 με τον περίφημο αλγόριθμο Perceptron [196]. Ο αλγόριθμος στοχεύει στο να προσαρμόσει στα δεδομένα εκπαίδευσης έναν γραμμικό ταξινομητή έτσι ώστε να ελαχιστοποιήσει κάποιο κριτήριο κόστους που σχετίζεται με την ακρίβεια ταξινόμησης, "δείχνοντας" επαναλαμβανόμενα τα δείγματα εκπαίδευσης στον ταξινομητή. Αποδεικνύεται ότι ο αλγόριθμος Perceptron συγκλίνει και η δομική μονάδα αυτή, δηλαδή ο προκύπτων ταξινομητής, είναι ένας νευρώνας, δομικός λίθος ενός νευρωνικού δικτύου. Ωστόσο, η εμφάνιση ικανοποιητικών αποτελεσμάτων μετά τη σύγκλιση προϋποθέτει γραμμική διαχωρισμότητα. Για επιπλέον διαχωρισμό του χώρου δεδομένων σε περιοχές, μπορούμε να συνδυάσουμε δομικές μονάδες Perceptron σε πολλαπλά επίπεδα. Αποδεικνύεται ότι με ένα νευρωνικό δίκτυο τριών επιπέδων μπορούμε να προσεγγίσουμε οποιαδήποτε συνάρτηση.

Το 1980, εισάγονται για πρώτη φορά οι συνάψεις με τοπικότητα μέσω αυτοοργανούμενων χαρτών και τα δίκτυα γίνονται βαθύτερα. Το Neocognitron [79], αποτελεί την εξέλιξη του προκατόχου του [78] και αποτελεί το πρώτο Συνελικτικό Νευρωνικό δίκτυο, επηρεασμένο δομικά από τον οπτικό φλοιό της γάτας [97]. Καθώς η τεχνολογία εξελίσσεται, προτείνονται διαφοροποιήσεις [130] και απλοποιήσεις των Συνελικτικών Δικτύων [10], [220] και δομικές αλλαγές [38] με την τεχνολογία των καρτών γραφικών υπολογιστή (GPU) να συμβάλλει στην αποτελεσματική εκπαίδευση βαθιών δικτύων [39]. Η επανάσταση των Συνελικτικών Νευρωνικών Δικτύων γίνεται το 2012 [123] με το μοντέλο AlexNet να κερδίζει τον ετήσιο διαγωνισμό αναγνώρισης αντικειμένων ImageNet Challenge σημειώνοντας μεγάλη αύξηση του ποσοστού επιτυχίας και ανοίγοντας το δρόμο για νέα εκτεταμένη έρευνα στο πεδίο των Νευρωνικών Δικτύων. Τα τελευταία χρόνια, τα Συνελικτικά Νευρωνικά Δίκτυα έχουν εξελιχθεί σε δομή και βάθος, με ικανότητα να πετυχαίνουν πολύ υψηλές επιδόσεις αναγνώρισης και ταυτόχρονα να είναι εξαιρετικά γρήγορα [263], [233], [89], [60]. Επιπλέον, συναντούν ευρεία εφαρμογή πέραν της ταξινόμησης όπως η συντακτική ανάλυση [30], η περιγραφή εικόνων [258], [243], η ανίχνευση αντικειμένων [186] και η κατάτμηση εικόνων [137], ενώ έχει αρχίσει και η χρήση τους απευθείας σε βίντεο [116].

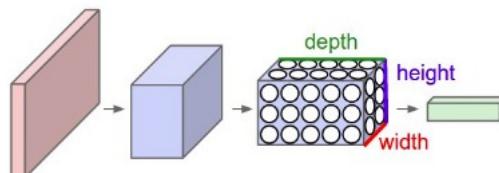
Θα δούμε τώρα σε μεγαλύτερο βάθος τη δομή και τη λειτουργία των Συνελικτικών Νευρωνικών Δικτύων. Τα κλασσικά νευρωνικά δίκτυα είναι ένας χάρτης νευρώνων που ενώνονται με πολλαπλασιαστικές συνάψεις και οι παράμετροί τους (βάρη) προσαρμόζονται στη διαδικασία της εκπαίδευσης. Με αυτόν τον τρόπο, κάθε νευρώνας εκτελεί έναν γραμμικό μετασχηματισμό ακολουθούμενο από μια μη γραμμικότητα. Το συνολικό δίκτυο εκφράζει έναν διαφορικό μετασχηματισμό της εισόδου σε σκορ κλάσεων. Τα συνελικτικά δίκτυα διατηρούν τις παραπάνω ιδιότητες αλλά στηρίζονται στην υπόθεση ότι η είσοδος θα είναι εικόνα, οπότε ποικίλες χρήσιμες ιδιότητες



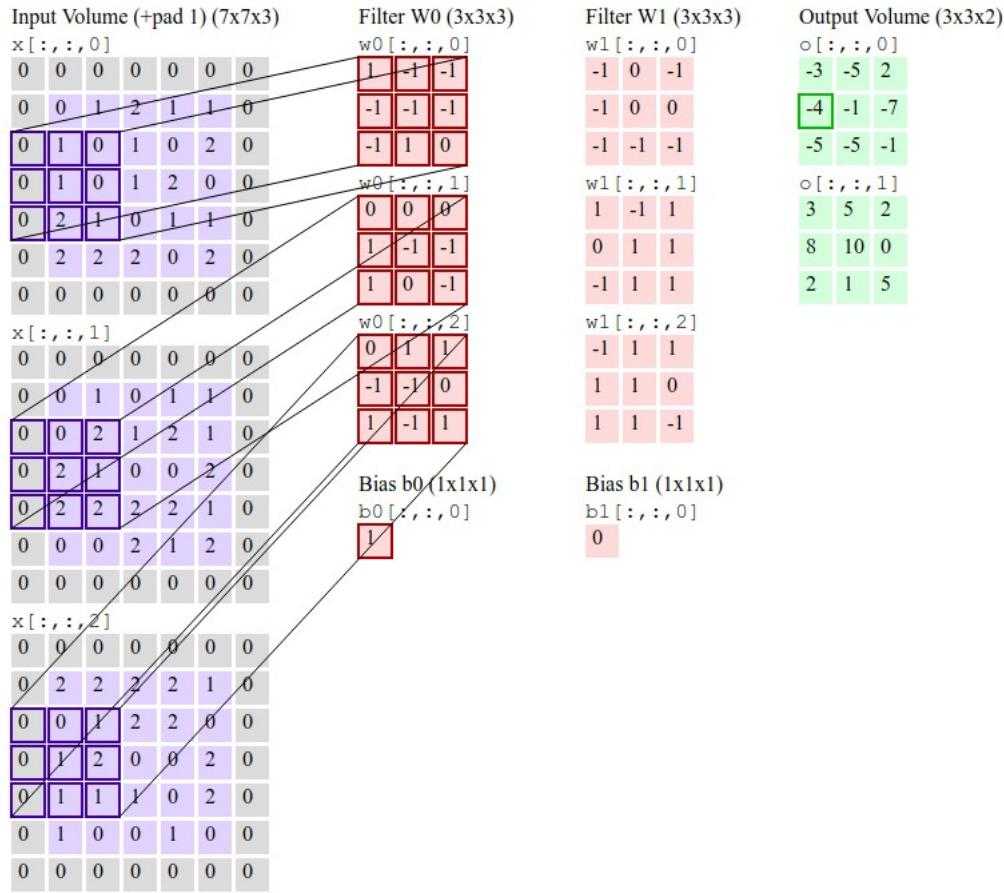
Σχήμα 5.2: Η εξέλιξη των νευρωνικών στην πορεία τους. Εκκινώντας εμπνεόμενα από τα βιολογικά νευρικά κύτταρα, τα πρώτα νευρωνικά δίκτυα αποτελούνταν από μια γραμμική συνδεσμολογία απλών νευρώνων σε επίπεδα. Στο Neocognitron [79] εμφανίζεται η οργάνωση σε επίπεδα με διαφορετικό συναπτικό πεδίο που απεικονίζεται σε διαφορετικά τμήματα εισόδου. Οι νευρώνες και κατά συνέπεια τα εκφραζόμενα χαρακτηριστικά αποκτούν μια τοπικότητα η φαίνεται να χρησιμεύει ιδιαίτερα στην ανάλυση εικόνων και βίντεο. Στα τελευταία χρόνια, η χρήση ταχύτερου υλικού (hardware) και η ανάπτυξη προγραμματιστικών τεχνικών (παραλληλία) επέτρεψαν τη σχεδίαση και εφαρμογή βαθιών συνελικτικών δικτύων, όπως το εικονιζόμενο GoogLeNet [233], τα οποία έγιναν ασυναγώνιστα σε πολλά πεδία της Μηχανικής Μάθησης.

μπορούν επιπλέον να ενσωματωθούν στη δομή τους. Αυτό οδηγεί σε αλλαγή της χωρικής διάταξης με αραιές συνδέσεις, οι οποίες απαιτούν λιγότερα βάρη. Έτσι τα συνελικτικά δίκτυα είναι πιο γρήγορα, πιο ακριβή και πιο εύρωστα.

Η πρώτη κύρια διαφορά των συνελικτικών από τα κλασσικά νευρωνικά δίκτυα είναι ότι κάθε επίπεδο είναι τώρα τρισδιάστατο και έχει πλάτος, μήκος και βάθος. Προφανώς το τελευταίο στάδιο είναι διάνυσμα με διάσταση ίση με τον αριθμό των κλάσεων. Η ακόλουθη εικόνα δείχνει ένα συνελικτικό δίκτυο και επισημειώνει τις διαστάσεις ενός συνελικτικού σταδίου:



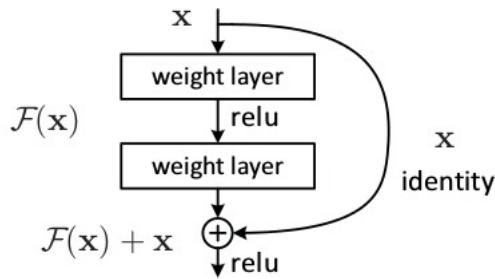
Σχήμα 5.3: Ένα συνελικτικό νευρωνικό δίκτυο οργανώνει τους νευρώνες του σε τρεις διαστάσεις, πλάτος, ύψος και βάθος, όπως οπτικοποιούνται εδώ σε ένα από τα επίπεδά του. Κάθε επίπεδο ενός συνελικτικού νευρωνικού δικτύου μετασχηματίζει έναν τρισδιάστατο όγκο εισόδου σε έναν τρισδιάστατο χώρο ενεργοποιήσεων νευρώνων.



Σχήμα 5.4: Η δράση της συνέλιξης. Μια σειρά από φίλτρα κυλίονται επί της εισόδου και το αποτέλεσμα του τοπικού εσωτερικού γινομένου απεικονίζεται απευθείας στην έξοδο. Η έξοδος θα έχει διάσταση βάθους όσο και το πλήθος των φίλτρων. Οι χωρικές διαστάσεις προκύπτουν από το πλήθος των σημείων όπου υπολογίζονται τοπικά τα εσωτερικά γινόμενα καθώς το φίλτρο κυλίεται. Συνήθως, βήμα αλίσθησης μοναδιαίο σε συνδυασμό με κατάλληλη αρχική προσαύξηση με μηδενικά αφήνουν αναλλοίωτες τις χωρικές διαστάσεις εξόδου ως προς την είσοδο, ενώ η μεταβλητή είναι η διάσταση βάθους.

Τα δίκτυα αυτά, αντίθετα με τα συμβατικά νευρωνικά, περιέχουν διάφορους τύπους σταδίων. Ο πιο βασικός είναι σίγουρα τα συνελικτικά στάδια, τα οποία είναι ο κύριος πυρήνας της εξαγωγής χαρακτηριστικών και υπολογίζουν την έξοδο νευρώνων που συνδέονται σε τοπικές περιοχές της εισόδου. Ακολουθούν τα pooling layers, τα οποία εκτελούν υποδειγματοληψία προκειμένου βαθμιαία να μειώνουν το μέγεθος των σταδίων και των παραμέτρων και να προστατεύουν από υπερπροσαρμογή (overfitting) και κόστος υπολογισμών. Συνήθως κοντά στο στάδιο εξόδου, τα συνελικτικά νευρωνικά δίκτυα περιέχουν κλασσικά, πλήρως συνδεδεμένα στάδια, όπως αυτά των κλασσικών νευρωνικών δικτύων, τα οποία παράγουν τα σκορ κλάσεων. Τέλος, μια κατηγορία σταδίων είναι τα ReLU (γραμμικοί περιοριστές), τα οποία είναι στην ουσία συναρτήσεις ενεργοποίησης που ακολουθούν τον υπολογισμό των συνελικτικών σταδίων, αντικαθιστώντας τις μη γραμμικές συναρτήσεις που χρησιμοποιούν τα συμβατικά νευρωνικά.

Θα αναλύσουμε σε λίγο μεγαλύτερο βάθος τη δομή ενός συνελικτικού σταδίου. Πρακτικά σε κάθε τέτοιο στάδιο η είσοδος είναι ένας όγκος $W \times H \times D$ και η έξοδος ένας



Σχήμα 5.5: Η μορφή των residual συναρτήσεων. Ένας εμπροσθόδρομος βρόχος παρακάμπτει ενδιάμεσα στάδια και αθροίζεται στην έξοδό τους. Το σχήμα αυτό κρύβει την πεμπτουσία των δικτύων ResNet, τα οποία βελτιστοποιούνται και πετυχαίνουν κορυφαίες αποδόσεις, ακόμα και συγκρινόμενα με άλλα συστήματα. Εικόνα από [89]

νέος όγκος $W' \times H' \times D'$. Επομένως χρειάζεται να καθορίσουμε μια συνάρτηση που θα κάνει αυτόν τον μετασχηματισμό. Το στάδιο πρέπει να εφαρμόσει N φίλτρα στην είσοδο. Αυτή είναι η τρίτη διάσταση της εξόδου, $D' = N$. Τώρα, κάθε νευρώνας ενός φίλτρου συνδέεται σε ένα μικρό κλάσμα της εισόδου, το λεγόμενο συναπτικό πεδίο, που καθορίζεται από τη θέση του νευρώνα στο φίλτρο. Καθώς θέλουμε να εφαρμόσουμε το φίλτρο σε όλη την εικόνα, έχουμε πολλούς νευρώνες ανά φίλτρο. Στην πράξη, κυλάμε το φίλτρο πάνω στην αρχική εικόνα και υπολογίζουμε το εσωτερικό γινόμενο, δηλαδή συνελίσσουμε το φίλτρο με την εικόνα, επιτρέποντας πιθανόν μετακινήσεις με μεγαλύτερο βήμα από 1. Συνοψίζοντας, νευρώνες με ίδιο μήκος και πλάτος έχουν το ίδιο συναπτικό πεδίο, ενώ νευρώνες με ίδιο βάθος είναι αντίτυπα του ίδιου φίλτρου. Κάθε φίλτρο εφαρμόζεται στο συναπτικό του πεδίο και η έξοδος αθροίζεται σε όλο το βάθος εισόδου αποδίδοντας μια τιμή εξόδου. Η σύμπτυξη αυτών των εξόδων δημιουργεί τον $W' \times H' \times D'$ χώρο εξόδου.

5.1.2 Το Δίκτυο ResNet

Το 2015, στο [89] εισήχθη μια νέα διάταξη των επιπέδων εντός του συνελικτικού δικτύου. Παρατηρήθηκε ότι αναδιατάσσοντας τις συνδέσεις μεταξύ των επιπέδων, τα στάδια αθρούνται στο να μαθαίνουν συναρτήσεις “υπολοίπου” (residual functions) με αναφορά στην είσοδο, της μορφής $F(x) + x$ και ότι τα δίκτυα αυτής της μορφής εκπαιδεύονται ευκολότερα, μπορούν να είναι πολύ βαθύτερα και πετυχαίνουν πολύ μεγαλύτερη ακρίβεια. Το προκύπτον δίκτυο ονομάστηκε ResNet από τη μορφή των συναρτήσεών του, που είχαν την residual μορφή.

Η ιδέα του ResNet εκκινεί από την παρατήρηση ότι παρότι το βάθος των συνελικτικών δικτύων φαίνεται να συνεισφέρει σημαντικά στην ακρίβεια που επιτυγχάνουν, τελικά η απόδοσή τους φτάνει σε κορεσμό και περαιτέρω αύξηση του βάθους συνεισφέρει αρνητικά στην ακρίβεια καθώς δυσκολεύει πολύ η βελτιστοποίηση (degradation problem). Οι residual τοπολογίες φαίνεται να λύνουν αυτό το πρόβλημα: η τοπολογία βελτιστοποιείται εύκολα με μεθόδους Βαθμωτής Κατάβασης (Stochastic Gradient Descend) και βάθος, ακόμα και 10 φορές μεγαλύτερο είναι εφικτό, ενώ η απόδοσή τους είναι ολοένα και μεγαλύτερη. Η εξήγηση για αυτό, είναι η αρχικοποίηση προς το ταυτοτικό ταίριασμα (identity mapping). Δηλαδή, προσεγγίζοντας την $H(x) - x$, είναι ευκολότερο να επιτευχθεί σύγκλιση αν αυτό είναι βέλτιστο. Η επιτυχία του ResNet δικαιώνει πειραματικά μια τέτοια σχεδίαση, χωρίς ωστόσο απόδειξη βελτιστότητας.

5.1.3 Συνελικτικά Χαρακτηριστικά και Μεταβατική Εκμάθηση (Transfer Learning)

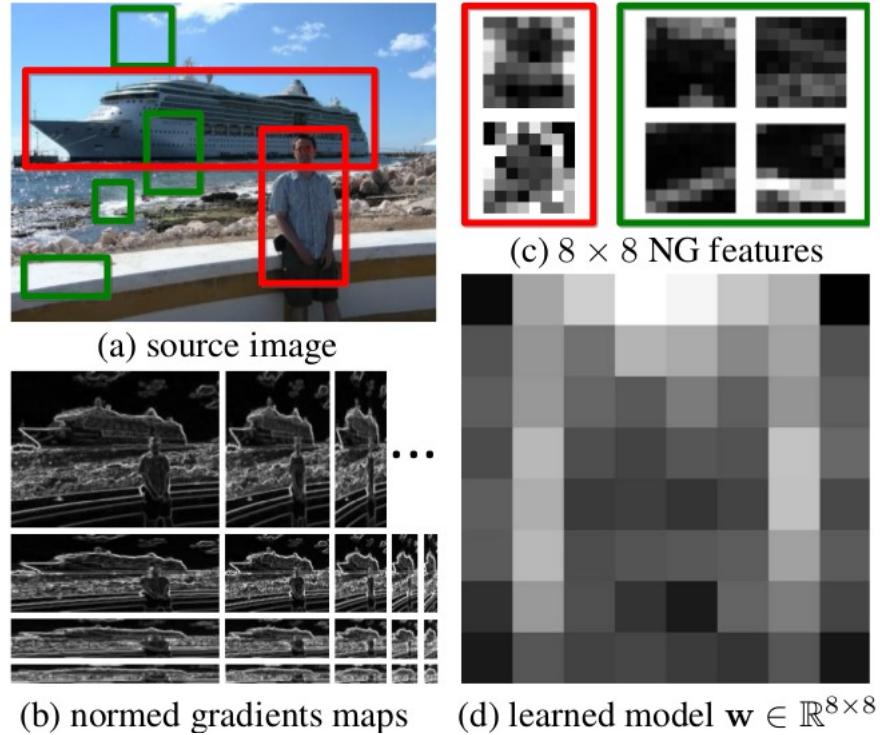
Η επιτυχία των νευρωνικών δικτύων έχει ωθήσει στην ευρεία χρήση τους. Εν τούτοις η εκπαίδευσή τους είναι δύσκολη και χρονοβόρα. Αυτό που είναι ενθαρρυντικό είναι η εύκολη προσαρμογή τους σε νέα καθήκοντα και σύνολα δεδομένων. Ένα βαθύ νευρωνικό δίκτυο, προεκπαιδευμένο σε κάποιο σύνολο δεδομένων, μπορεί να χρησιμοποιηθεί ως αρχικοποίηση για ένα νέο δίκτυο σε ένα νεο σύνολο δεδομένων. Ακόμα πιο ενθαρρυντικό: τα πρώτα στάδια των δικτύων σε παρόμοια καθήκοντα κωδικοποιούν πληροφορία χαμηλού επιπέδου και ελάχιστα μεταβάλλονται ανά σύνολο δεδομένων. Επομένως, αρχικοποιώντας ένα βαθύ δίκτυο με τα βάρη ενός προεκπαιδευμένου, μπορούμε να επανεκπαιδεύσουμε μόνο το τελευταίο στάδιο και να το προσαρμόσουμε στα νέα δεδομένα με επιτυχία. Μάλιστα πειράματα δείχνουν ότι τα προεκπαιδευμένα δίκτυα επιτυγχάνουν υψηλότερες επιδόσεις από δίκτυα που εκπαιδεύονται απευθείας στα νέα δεδομένα. Η τεχνική αυτή εφαρμόζεται ευρύτατα και ονομάζεται transfer learning.

Μπορούμε επομένως να δούμε ένα συνελικτικό νευρωνικό δίκτυο σαν μια υπολογιστική μονάδα εξαγωγής χαρακτηριστικών μεγάλης διακριτότητας και αναπαραστικής δυνατότητας. Τα χαρακτηριστικά αυτά εισάγονται σε έναν γραμμικό ταξινομητή, το τελευταίο στάδιο του δικτύου και λαμβάνουμε τις πιθανότητες κλάσεων. Γίνεται επομένως φανερό ότι μπορούμε να αντικαταστήσουμε το στάδιο εξόδου με κάποιον άλλο ταξινομητή. Μια άλλη προσέγγιση παρουσιάζεται στο [234], όπου μελετάται η χρήση SVM loss σαν συνάρτηση κόστους του δικτύου.

5.1.4 BING: Binarized Normed Gradients for Objectness Estimation

Η έννοια του μέτρου objectness αφορά στην ύπαρξη ή μη αντικειμένου σε ένα τμήμα εικόνας. Ουσιαστικά ένα σύστημα objectness προτείνει περιοχές της εικόνας όπου υπάρχουν αντικείμενα. Η ιδιότητα αυτή είναι χρήσιμη σε ανιχνευτές παραθύρων με μέθοδο κυλιόμενου παραθύρου καθώς συρρικνώνει δραστικά την περιοχή αναζήτησης.

Μια γρήγορη υλοποίηση objectness προτείνεται στο [33], όπου εισάγεται η έννοια των BING. Τα αρχικά BING μεταφράζονται σε δυαδικά κωδικοποιημένες κανονικοποιημένες παραγώγους και στηρίζονται στην παρατήρηση ότι εικόνες αντικειμένων, ανεξαρτήτως κλάσης, με ικανοποιητική ακρίβεια, μπορούν να διακριθούν από εικόνες μη αντικειμένων παρατηρώντας τις παραγώγους τους αφού προηγηθεί μια αλλαγή μεγέθους σε μια καθορισμένη τιμή. Συγκεκριμένα η αυθεντική εργασία πρότεινε το μετασχηματισμό μεγέθους σε 8×8 και τη χρήση της νόρμας των παραγώγων ως διακριτικό διάνυσμα χαρακτηριστικών 64 διαστάσεων. Εν συνεχεία, κωδικοποιούν δυαδικά τα χαρακτηριστικά αυτά ώστε να πετύχουν υψηλότερη ταχύτητα υπολογισμών με πράξεις δυαδικών τελεστών.



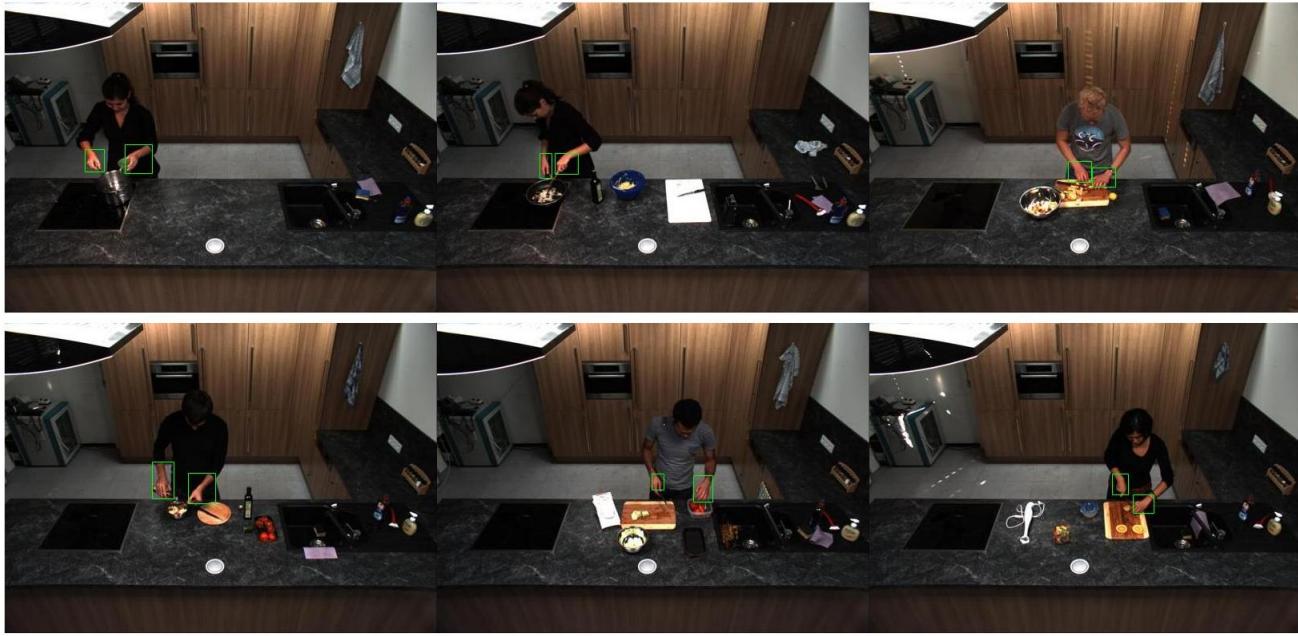
Σχήμα 5.6: Παρόλο που ο χώρος των παραθύρων αντικειμένων (κόκκινο) και του παρασκηνίου (πράσινο) παρουσιάζει τεράστιες διακυμάνσεις στο χώρο των εικόνων, σε κατάλληλες κλίμακες, τα κανονικοποιημένα διανύσματα παραγώγων εμφανίζουν μεγάλη συσχέτιση. Με ένα απλό μοντέλο 64 διαστάσεων, μπορούμε να λάβουμε αξιόπιστες προτάσεις για την ύπαρξη ή μη αντικειμένου. Εικόνα από [33].

5.2 Προτεινόμενο Σύστημα

5.2.1 Ανίχνευση Χεριών

Για την ανίχνευση των χεριών εκπαιδεύσαμε έναν ταξινομητή κυλιόμενου παραθύρου συνδυάζοντας χαρακτηριστικά τύπου HOF, ιστογράμματα χρώματος και BING. Πιο συγκεκριμένα, χρησιμοποιήσαμε σταθερό μέγεθος παραθύρου ίσο με 31×31 , οπότε τα διανύσματα HOF είχαν σταθερό μέγεθος. Για τα χαρακτηριστικά BING διατηρήσαμε τις αυθεντικές παραμέτρους και μετασχηματίζαμε κάθε παράθυρο σε εικόνα μεγέθους 8×8 . Στη συνέχεια υπολογίζαμε τις παραγώγους της μετασχηματισμένης εικόνας και λαμβάνουμε τη νόρμα-1 αυτών. Τέλος, για τα ιστογράμματα χρώματος συνδυάσαμε συνιστώσες των χώρων HSV και YcbCr, επηρεασμένοι από το [214], κρατώντας το κανάλι H από τον πρώτο χώρο και τα κανάλια Cb, Cr από τον δεύτερο χώρο. Τα ιστογράμματα των καναλιών συνενώνονται με παράθεση (concatenation). Με ίδιο τρόπο συνενώνονται τα διαφορετικά είδη χαρακτηριστικών (HOF, BING, ιστογράμματα χρώματος). Τα χαρακτηριστικά είναι συμπληρωματικά: τα χαρακτηριστικά HOF κωδικοποιούν τη συνολική τοπική αναπαράσταση, τα ιστογράμματα χρώματος το χρώμα και τα χαρακτηριστικά BING το σχήμα.

Με τα χαρακτηριστικά αυτά εκπαιδεύσαμε έναν ταξινομητή Τυχαίου Δάσους (Random Forest) [91]. Ο ταξινομητής αυτός δεν είναι παρά μια συλλογή δέντρων αποφάσεων όπου η τελική απόφαση προκύπτει με στάθμιση των αποφάσεων των δέντρων. Τελικά ο προκύπτων ταξινομητής είναι αποτελεσματικός αλλά και γρήγορος, καθώς

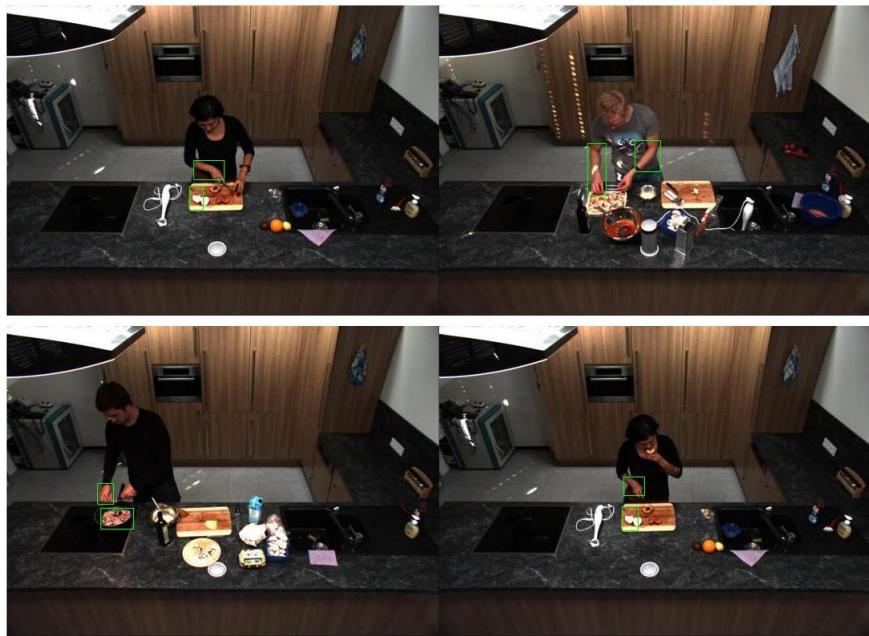


Σχήμα 5.7: Παραδείγματα επιτυχούς ανίχνευσης χεριών. Τα προκύπτοντα παράθυρα μπορούν να εισέλθουν στο επόμενο στάδιο εξαγωγής χαρακτηριστικών και τελικά να ταξινομηθούν ως τύποι λαβής.

οι μόνες πράξεις που έχει να εκτελέσει είναι η σύγκριση και η στάθμιση. Εκπαιδεύουμε λοιπόν έναν τέτοιο ταξινομητή σε δεδομένα τύπου “χέρι” εναντίον “όχι χέρι”. Καθώς δεν διαθέταμε σύνολο δεδομένων απόλυτα σχετικό, λάβαμε επισημειώσεις για το χέρι από ένα σύνολο [191], [4] που προοριζόταν για ταξινόμηση πόζας και στο οποίο η επισημείωση του χεριού ήταν σημειακή και όχι σαν περιοχή. Επιπλέον, στο σύνολο υπήρχαν εικόνες όπου το χέρι επικαλυπτόταν από άλλα αντικείμενα, εν τούτοις επισημειωνόταν. Καθαρίσαμε όσο το δυνατόν αυτές τις εικόνες. Παρείχαμε ως δεδομένα εκπαίδευσης από κάθε εικόνα τις εικόνες του χεριού, δύο εικόνες υποβάθρου και, αραιά, εικόνες κεφαλιού. Με αυτόν τον τρόπο δηλώνουμε ρητά ότι δεν αρκούν εικόνες δέρματος ως θετικά δείγματα αλλά εικόνες χεριών. Το κριτήριο βελτιστοποίησης του ταξινομητή επιλέχθηκε να είναι η μετρική Precision, καθώς θέλουμε τον λιγότερο δυνατό αριθμό εσφαλμένων θετικών εκτιμήσεων. Με cross-validation μετρήσαμε 94% precision στη φάση εκπαίδευσης.

Στη φάση ελέγχου, κυλάμε παράθυρο 31×31 πάνω στην εικόνα με βήμα 10 σε κάθε διάσταση, οπότε λαμβάνουμε επικαλυπτόμενα παράθυρα. Το βήμα επιλέχθηκε στην τιμή αυτή ως συμβιβασμός μεταξύ ακρίβειας και ταχύτητας. Για κάθε παράθυρο εξάγουμε τα χαρακτηριστικά και αφήνουμε τον ταξινομητή να αποφασίσει αν η εικόνα είναι εικόνα χεριού ή όχι. Παρότι η εκπαίδευση παρουσιάζει πολύ καλή μετρική precision, είναι φανερό ότι λόγω του μεγάλου αριθμού παραθύρων ανά εικόνα, δεν μπορούμε να αποφύγουμε την εμφάνιση εσφαλμένων θετικών εκτιμήσεων. Αντιμετωπίζουμε αυτό το φαινόμενο με επιπλέον επεξεργασία των αποτελεσμάτων σε 3 στάδια.

Στο πρώτο στάδιο γίνεται η εκτίμηση των superpixels [225]. Πρακτικά αντιστοιχίζουμε στην εικόνα έναν χάρτη πιθανοτήτων ύπαρξης χεριού. Η δημιουργία αυτού του χάρτη γίνεται ως εξής: Αρχικοποιούμε το χάρτη στο 0. Για κάθε παράθυρο που έχει ταξινομηθεί ως παράθυρο χεριού προσθέτουμε 1 στην αντίστοιχη περιοχή στο

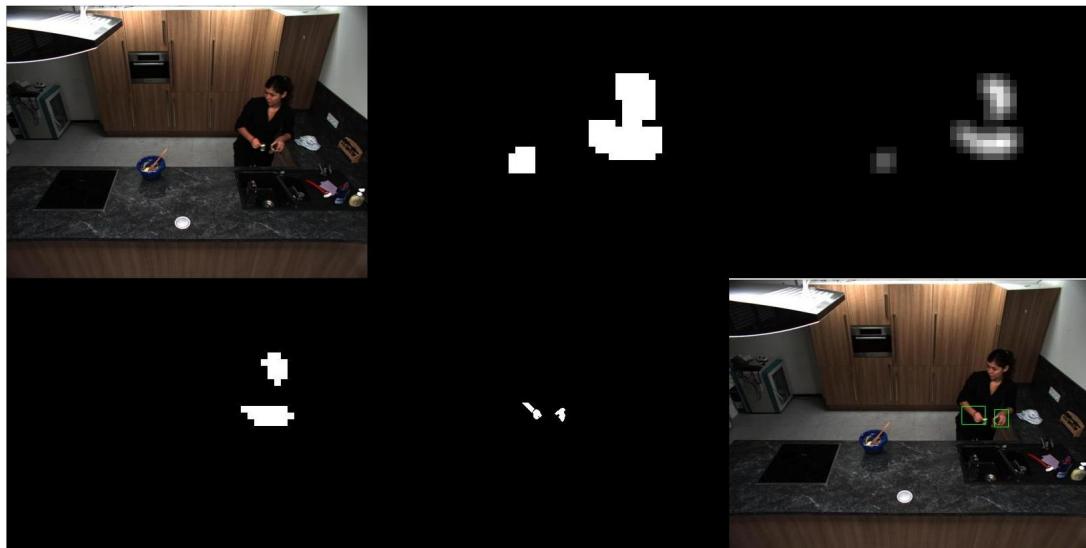


Σχήμα 5.8: Περιπτώσεις όπου η ανίχνευση χεριών αποτυγχάνει. Είναι πιθανό, περιοχές κοντά σε χέρια να ταξινομηθούν ως χέρια λόγω του χρωματικού κατωφλίου και των περιορισμών έκτασης που εφαρμόζονται στο τικό στάδιο του αλγορίθμου ανίχνευσης. Εν τούτοις, προτιμήθηκε η ύπαρξη τέτοιων σφαλμάτων μπροστά στα οφέλη μιας τέτοιας σχεδίασης. Κατά την εξαγωγή του τύπου λαβής, τα σφάλματα εισάγουν θόρυβο που αντισταθμίζουμε με την εισαγωγή επιπλέον κατηγοριών στο στάδιο μη επιβλεπόμενου διαχωρισμού.

χάρτη. Τελικά κανονικοποιούμε τον προκύπτοντα χάρτη, λαμβάνοντας μια γκρίζα εικόνα όπου η τιμή κάθε pixel δηλώνει την πιθανότητα το pixel αυτό να είναι pixel χεριού. Για να λάβουμε τα superpixels χρειάζεται να μετατρέψουμε το χάρτη πιθανοτήτων σε δυαδική εικόνα. Αυτό γίνεται με τη χρηση κατωφλίου, το οποίο στην περίπτωσή μας επιλέχθηκε ίσο με 0.5. Τα superpixels είναι επομένως συμπαγείς περιοχές pixels με ενιαία ιδιότητα, που εδώ η ταξινόμησή τους ως pixels χεριού. Η χρήση superpixels φιλτράρει επιπλέον τις λανθασμένες θετικές ταξινομήσεις στην περίπτωση που δεν έχουν ισχυρή γειτονιά, δηλαδή γειτονικές περιοχές χεριού.

Στο δεύτερο στάδιο υπολογίζουμε τα κουτιά των διακριτών superpixels και λαμβάνουμε μια εκτίμηση του εμβαδού τους. Κρατάμε το πολύ 3 κουτιά, αυτά με το μεγαλύτερο εμβαδόν, με την περίπτωση των τριών κουτιών να συμβαίνει μόνο όταν υπάρχει ισοβαθμία εμβαδών. Υποθέτουμε ότι τα χέρια θα είναι σε παρόμοιο ύψος, οπότε θα υπάρχει μια οριζόντια ευθεία η θα διέρχεται μέσα από όλα τα κουτιά χεριών. Αν δεν υπάρχει τέτοια ευθεία, κρατάμε μόνο τα κουτιά για τα οποία μπορεί να βρεθεί οριζόντια ευθεία η οποία να διέρχεται από μέσα από τα ίδια και το χαμηλότερο σε ύψος κουτί. Η επιλογή αυτή γίνεται ώστε να κόψει εσφαλμένα θετικά κουτιά στην περιοχή του κεφαλιού. Τέλος, απορρίπτονται τα κουτιά με εμβαδόν μικρότερο των 300 τετραγωνικών pixels, εκτός αν αυτό αποτελεί το μέγιστο εμβαδόν περιοχής χεριού από αυτές που ανιχνεύθηκαν.

Στο τρίτο στάδιο, διαχωρίζονται οι περιοχές που έχουν μείνει σε υποπεριοχές μεγαλύτερης ακριβείας. Το κίνητρο πίσω από αυτό είναι ότι συχνά τα δύο χέρια βρίσκονται



Σχήμα 5.9: Βήματα του αλγορίθμου ανίχνευσης χεριών. Η αρίθμηση εφαρμόζεται από πάνω αριστερά προς τα κάτω δεξιά. **Εικόνα 1:** Η αρχική εικόνα. **Εικόνα 2:** Η περιοχή που καλύπτουν τα παράθυρα με μηδενική πιθανότητα συμπεριληψης χεριού. Τα παράθυρα αυτά προκύπτουν από την κύλιση ενός παραθύρου πάνω στην εικόνα, την εξαγωγή τοπικών χαρακτηριστικών και την ταξινόμηση αυτών. **Εικόνα 3:** Ο χάρτης πιθανοτήτων ύπαρξης χεριού. Ο χάρτης αυτός προκύπτει μετά από κανονικοποίηση της εικόνας 2. **Εικόνα 4:** Τα superpixels. Ο χάρτης των superpixels προκύπτει από την κατωφλίωση του χάρτη πιθανοτήτων. Βλέπουμε ότι υπάρχουν δύο συμπαγείς περιοχές πλέον. Καθώς στο άξονα γ δεν υπάρχει επικάλυψη, η πάνω περιοχή αποκόπτεται. Η κάτω περιοχή χρειάζεται επιπλέον ανάλυση για διαχωρισμό των χεριών. **Εικόνα 5:** Με την εφαρμογή χρωματικών κατωφλίων και μορφολογικών τελεστών απομονώνουμε τις περιοχές των χεριών. **Εικόνα 6:** Η τελική εικόνα ανίχνευσης.

κοντά μεταξύ τους οπότε οι περιοχές τους επικαλύπτονται στα superpixels και προκύπτει μία περιοχή μόνο και για τα δύο χέρια. Ο επιπλέον διαχωρισμός γίνεται με κατανομή χρώματος. Ο αλγόριθμος βασίζεται στο [121] και αξιοποιεί την χρωματική πληροφορία των χώρων RGB, HSV και YCbCr. Σύμφωνα με την εργασία αυτή, ικανοποιητική χρωματική συνθήκη για δέρμα είναι η

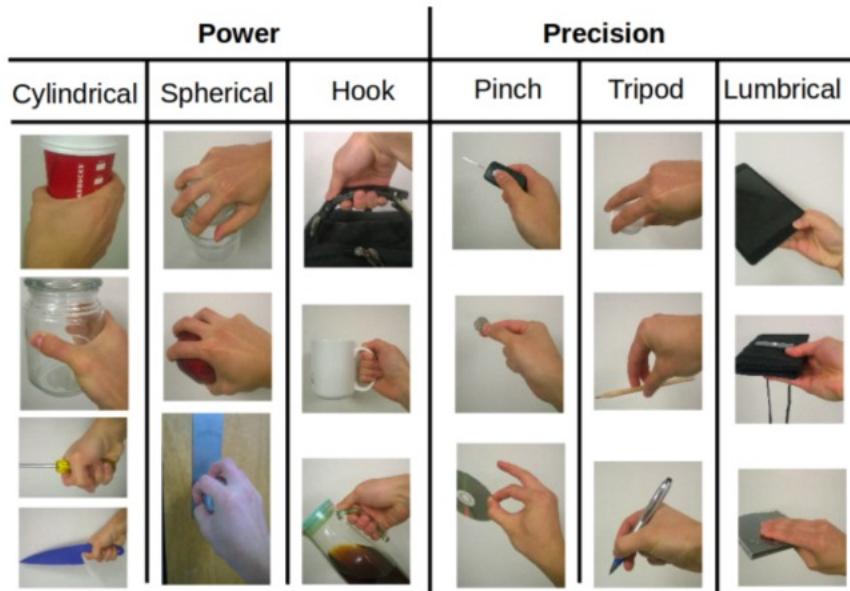
$$\begin{aligned}
& ((R > 95) \cap (G > 40) \cap (B > 20) \cap (R > G) \cap (R > B) \cap (|R - G| > 15) \cap (A > 15)) \\
& \quad \cap \\
& \quad \left(((0 \leq H \leq 50) \cap (0.23 \leq S \leq 0.68)) \right. \\
& \quad \cup \\
& \quad ((Cr > 135) \cap (Cb > 85) \cap (Y > 80) \\
& \quad \cap (Cr \leq (1.5862Cb) + 20) \\
& \quad \cap (Cr \geq 0.3448Cb + 76.2069) \\
& \quad \cap (Cr \leq -4.5652Cb + 234.5652) \\
& \quad \cap (Cr \leq -1.15Cb + 301.75) \\
& \quad \left. \cap (Cr \leq -2.2857Cb + 432.85) \right) \\
& \tag{5.1}
\end{aligned}$$

Εφαρμόζουμε αυτή τη συνθήκη για να παράγουμε δυαδικές εικόνες από κάθε superpixel χεριού. Για να κόψουμε το θόρυβο εφαρμόζουμε μορφολογικό opeining με κύκλο ακτίνας 2. Αν αυτό κόβει όλη την εικόνα, χρησιμοποιούμε κύκλο ακτίνας 1. Αν και αυτό κόβει όλη την εικόνα, εφαρμόζουμε closing με κύκλο ακτίνας 2. Η τελευταία περίπτωση λειτουργεί στην εμφάνιση χεριών υπό γωνία τέτοια ώστε να φαίνεται μικρό μέρος τους. Υπολογίζουμε τα νέα superpixels και κρατάμε και πάλι το πολύ 3 περιοχές, τις μεγαλύτερες σε εμβαδόν. Κόβουμε όλες τις περιοχές με εμβαδόν μικρότερο του 99. Με αυτόν τον τρόπο πετάμε αρκετό θόρυβο και διαχωρίζουμε τα χέρια τα οποία αρχικά βρίσκονταν εντός του ίδιου superpixel. Ακολουθεί ένα τελικό στάδιο στο οποίο κρατάμε τα δύο (αν υπάρχουν) μεγαλύτερα σε εμβαδόν superpixels από την ένωση όλων των περιοχών και επιστρέφουμε αυτά ως τελικές περιοχές χεριών.

5.2.2 Εξαγωγή Τύπου Λαβής

Ο τύπος λαβής μπορεί να περιέχει πληροφορία για τη χρήση των σχετιζόμενων με τη δράση αντικειμένων και τελικά με την ίδια τη δράση. Πειράματα στο [265] δείχνουν την αξία του τύπου λαβής στην αναγνώριση του σκοπού του χειριστή αντικειμένων. Επιπλέον, αποτελούν και χαρακτηριστικά χρήσιμα στην κατάτμηση του βίντεο σε δράσεις. Στο [255] παρουσιάζεται μια χρήσιμη ταξινόμηση τύπων λαβής για πλούσια πληροφορία σχετικά με την αναγνώριση δράσης και σκοπού. Η ταξινόμηση αυτή εστιάζει στη λειτουργικότητα και χωρίζει τις λαβές σε τρεις μεγάλες κατηγορίες: προσανατολισμένες σε δύναμη, σε ικανότητα και καθημερινές. Η ταξινόμηση αυτή είναι παρόμοια με την πιο γενική, η οποία διαχωρίζει σε λαβές δύναμης, ακριβείας και έκταση/ξεκούραση. Εδώ διαχωρίζουμε επιπλέον τις λαβές δύναμης σε κυλινδρικές, σφαιρικές και σε μορφή γάντζου, ενώ επίσης διαχωρίζουμε τις λαβές ακριβείας σε τσιμπήματος, τριποδικές και παράδοσης (lumbrical). Οπότε μαζί με τις περιπτώσεις έκτασης και ξεκούρασης, όπου δεν πραγματοποιείται κάποια λαβή από τις παραπάνω, έχουμε 8 κατηγορίες λαβών.

Για την αναγνώριση του τύπου λαβών δεν διαθέτουμε κάποιο σύνολο δεδομένων που να μπορεί να προσαρμοστεί στα δικά μας δεδομένα, οπότε αναγκαστήκαμε να κινηθούμε με μη επιβλεπόμενη μάθηση. Συγκεκριμένα, σε πρώτη φάση υπολογίζουμε



Σχήμα 5.10: Οι τύποι λαβής στους οποίους στηριζόμαστε ως χρήσιμη πληροφορία για την αναγνώριση δράσεων. Λαβές που δεν ταξινομούνται σε αυτές τις κατηγορίες θα είναι λαβές έκτασης και ξεκούρασης. Εικόνα από [255].

τα συνελικτικά χαρακτηριστικά για τις εικόνες χεριών που εντοπίζονται αυτόματα στα βίντεο εκπαίδευσης, χρησιμοποιώντας το ResNet. Στη συνέχεια χρησιμοποιούμε τη μέθοδο των k-μέσων για να εξάγουμε 10 κέντρα κατηγοριών. Επιλέξαμε 10 κέντρα αντί για 8 ώστε να κόψουμε περιπτώσεις θορύβου από την ανίχνευση χεριών, καθώς δεν υπάρχει Ground Truth επισημείωση. Λαμβάνοντας τα 10 κέντρα, εκπαιδεύουμε και εφαρμόζουμε ταξινόμηση με χρήση του αλγορίθμου Στοχαστικής Βαθμωτής Κατάβασης (Stochastic Gradient Descend, SGD) στα δείγματα εκπαίδευσης και δημιουργούμε τα ιστογράμματα τύπων λαβής για τα δεδομένα εκπαίδευσης. Να σημειωθεί ότι σε κάθε frame που εξετάζουμε, θα εμφανίζονται το πολύ δύο χέρια, ωστόσο δεν είναι εκ των προτέρων γνωστό το ποιο από τα δύο θα εμφανίζεται ή ποιο από τα δύο χέρια αντιστοιχεί στο αριστερό και ποιο στο δεξί. Οπότε αθροίζουμε τα ιστογράμματα για τα δύο χέρια που εμφανίζονται, εκμεταλλεύμενοι την προσεταιριστική ιδιότητα. Αν εμφανίζεται μόνο ένα χέρι στο frame τότε το δεύτερο διάνυσμα είναι το μηδενικό διάνυσμα 10 διαστάσεων. Εφαρμόζουμε την ίδια διαδικασία στα δείγματα εκπαίδευσης για να λάβουμε τα ιστογράμματα εκπαίδευσης. Στο κεφάλαιο 6 θα δείξουμε πώς συνενώνουμε αυτά τα χαρακτηριστικά και τη συνεισφορά τους στην ταξινόμηση.

Κεφάλαιο 6

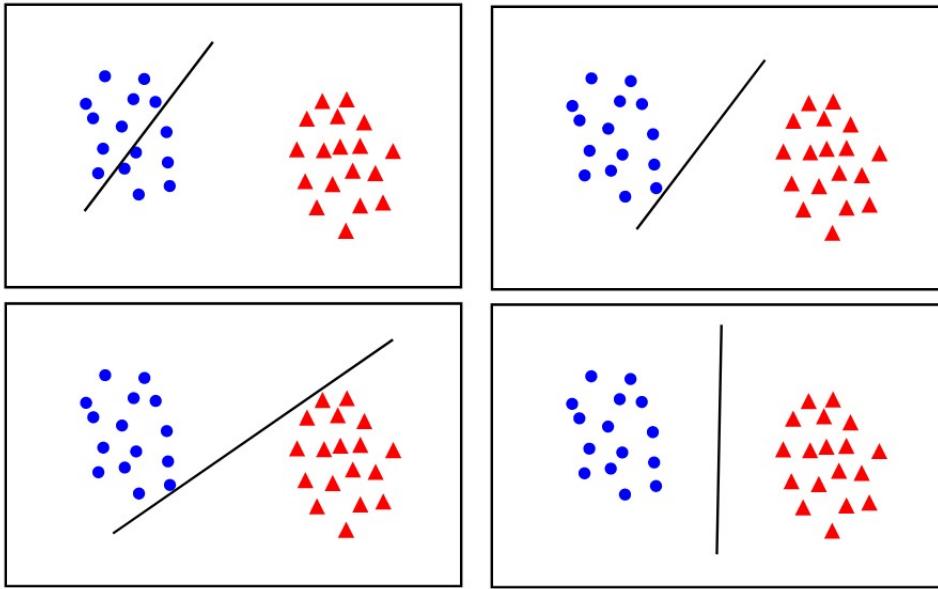
Πειρόματα και Αποτελέσματα Αναγνώρισης Δράσεων

Έχοντας αναλύσει την εξαγωγή των χαρακτηριστικών από τα διαφορετικά κανάλια πληροφορίας, φτάνουμε στο σημείο της σύμμειξης των καναλιών προκειμένου να προκύψει μια ενιαία αναπαράσταση, κατάλληλη για ταξινόμηση. Αναζητούμε ένα συνδυασμό απόδοσης και επίδοσης, γνωρίζοντας ότι πιθανόν οι δύο αυτοί παράγοντες να είναι αντιμαχόμενοι. Στο κεφάλαιο αυτό θα εξετάσουμε θα εξετάσουμε την ταξινόμηση των εξαγόμενων χαρακτηριστικών των τημάτων βίντεο με αρκετές διαφορετικές μεθόδους, συγκρίνοντας τα αποτελέσματά τους και τελικά ανάγοντας τη σύγκριση με άλλες μεθόδους της παγκόσμιας βιβλιογραφίας. Το πρόβλημα της ταξινόμησης γενικά, στην επιβλεπόμενη μάθηση, ορίζεται ως η εύρεση συνάρτησης βέλτιστης απεικόνισης νέων δεδομένων σε κάποια ή κάποιες από τις γνωστές κατηγορίες, έχοντας προηγηθεί εκπαίδευση του συστήματος με δεδομένα των οποίων οι ετικέτες είναι γνωστές. Αρχικά θα παρουσιάσουμε συνοπτικά τη φιλοσοφία της μεθόδου ταξινόμησης με χρήση Μηχανών Διανυσμάτων Υποστήριξης (SVM) και στη συνέχεια θα δούμε τις διάφορες μεθόδους που χρησιμοποιήσαμε κατά την ταξινόμηση, πρώτα θεωρητικά κι έπειτα πειραματικά.

6.1 Θεωρητικό Υπόβαθρο

6.1.1 Μηχανές Διανυσμάτων Υποστήριξης (SVM)

Οι Μηχανές Διανυσμάτων Υποστήριξης (στο εξής SVM, από τον αγγλικό όρο Support Vector Machine) είναι γενικευμένα, μη πιθανοτικά, γραμμικά μοντέλα ταξινομητών επιβλεπόμενης μάθησης. Ο αλγόριθμος SVM εισήχθη στο [43] ως Δίκτυο Διανυσμάτων Υποστήριξης. Ο στόχος του αλγορίθμου SVM είναι ο διαχωρισμός, με βάση τα δείγματα εκπαίδευσης, του χώρου χαρακτηριστικών σε υποχώρους με τρόπο τέτοιο ώστε να μεγιστοποιείται η απόσταση μεταξύ των υποχώρων διαφορετικών κλάσεων. Η ταξινόμηση γίνεται σύμφωνα με τον χώρο στον οποίο εμπίπτουν τα νέα δείγματα, οπότε ο αλγόριθμος SVM εκτελείται σε γραμμικό χρόνο, αφού αρκεί να υπολογισθεί

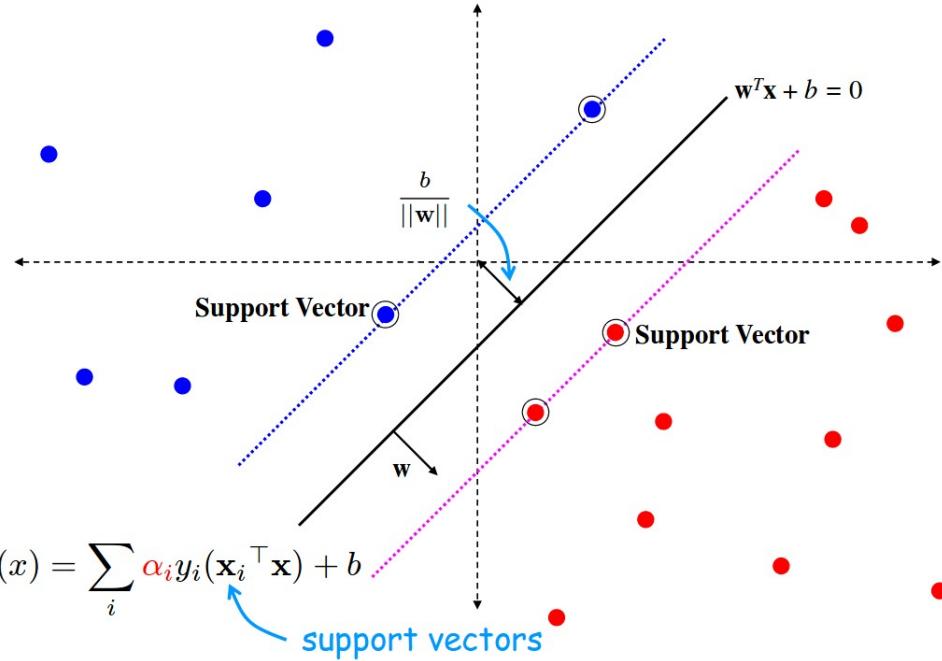


Σχήμα 6.1: Είναι φανερό ότι σε ένα πρόβλημα ταξινόμησης δύο γραμμικά διαχωρίσιμων κλάσεων υπάρχουν άπειρες λύσεις. Ποια από αυτές είναι η προτιμότερη; Αν το κριτήριο είναι η ικανότητα γενίκευσης, τότε βέλτιστη λύση, για τα υπάρχοντα δεδομένα, είναι η γραμμή που αφήνει μέγιστο περιθώριο και για τις δύο κλάσεις, έτσι ώστε να διατηρείται ισορροπία μεταξύ των χώρων σφαλμάτων των κλάσεων. Ο αλγόριθμος SVM αναζητά τη βέλτιστη αυτή γραμμή, μέσα από ένα πρόβλημα ελαχιστοποίησης κόστους. Η αναζήτηση βέλτιστης λύσης και όχι απλά λύσης διαχωρίζει τον ταξινομητή SVM από τους υπόλοιπους γραμμικούς ταξινομητές. Εικόνα από [171]

ένα εσωτερικό γινόμενο των βαρών των χαρακτηριστικών, τα οποία βάρη προβάλλουν το διάνυσμα προς ανίχνευση στον κατάλληλο υποχώρο, με το ίδιο το διάνυσμα προς ταξινόμηση. Έπομένως, η εκπαίδευση ανάγεται σε ένα πρόβλημα βελτιστοποίησης, το οποίο γεωμετρικά ισοδυναμεί με τον βελτιστού διαχωρισμό του χώρου χαρακτηριστικών με χρήση υπερεπιπέδων. Το αντικείμενο είναι πλούσιο σε χρήση μαθηματικών εργαλείων, η αναφορά των οποίων ξεφεύγει από τον σκοπό αυτής της εργασίας. Παραπέμπουμε τον ενδιαφερόμενο στην παγκόσμια βιβλιογραφία, εκκινώντας από την αυθεντική εργασία [43]. Θα εξετάσουμε όμως εδώ τις διάφορες παραλλαγές που θα χρησιμεύσουν στην πορεία της εργασίας μας.

6.1.1.1 Μη Γραμμική SVM: Το Τέχνασμα Πυρήνα

Από τον ορισμό του, ο αλγόριθμος SVM είναι γραμμικός, οπότε δε λειτουργεί ικανοποιητικά για μη γραμμικά διαχωρίσιμες κλάσεις. Ωστόσο, συχνά χρησιμοποιείται μια παραλλαγή του αλγορίθμου, καθιστώντας τον επαρκή και για μη γραμμικά προβλήματα. Στην πράξη, μετασχηματίζουμε τον χώρο χαρακτηριστικών σε έναν άλλο χώρο, συνήθως μεγαλύτερης διάστασης, στον οποίο τα χαρακτηριστικά αποκτούν ιδιότητες γραμμικής διαχωρισμότητας. Η ταξινόμηση γίνεται στον νέο αυτό χώρο, οπότε κατά τη φάση εκπαίδευσης, κάθε δείγματα υφίσταται μια μη γραμμική απεικόνιση και το πρόβλημα βελτιστοποίησης ανάγεται σε εύρεση των υπερεπιπέδων που διαχωρίζουν βέλτιστα τις περιοχές αυτού του νέου χώρου. Αν ωστόσο μεταβούμε πίσω στον αρχικό χώρο χαρακτηριστικών, θα δούμε ότι τα υπερεπίπεδα έχουν λάβει μορφή υπερεπιφανειών, δίνοντας την εντύπωση του μη γραμμικού ταξινομητή. Η συνάρτηση μεταφοράς στον νέο υπερχώρο ονομάζεται συνάρτηση πυρήνα (kernel) και



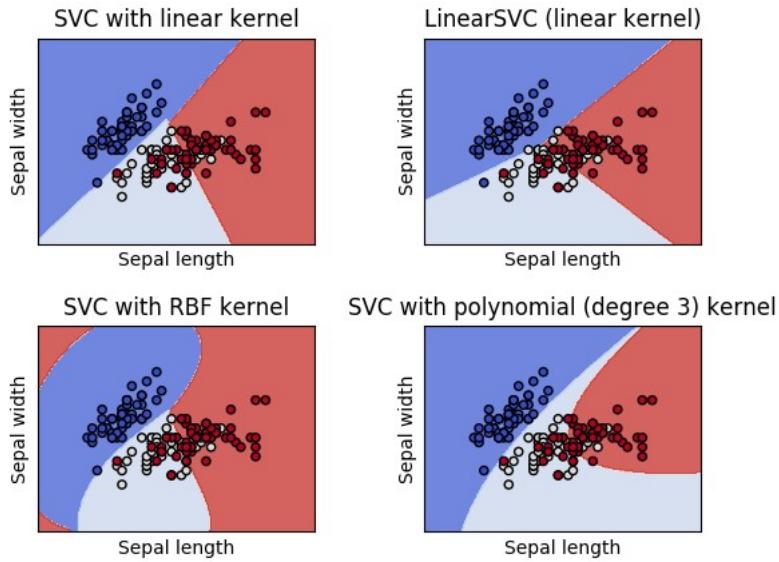
Σχήμα 6.2: Η μέθοδος ταξινόμησης SVM αναζητά τους βέλτιστους εκπροσώπους των ορίων των κλάσεων, τα λεγόμενα Διανύσματα Υποστήριξης (Support Vectors). Σύμφωνα με αυτά τα διανύσματα, οριοθετεί γραμμικά τους χώρους των κλάσεων και λαμβάνει τη μεσοπαράλληλο των γραμμικών συνόρων ως τη βέλτιστη διαχωριστική γραμμή μεταξύ των κλάσεων. Σε παραπάνω διαστάσεις, η διαχωριστική επιφάνεια είναι ένα υπερεπίπεδο. Με μετασχηματισμό πυρήνα, το υπερεπίπεδο αυτό μπορεί να αναπαριστά μια υπερεπιφάνεια η οποία προκύπτει από μια μη γραμμική απεικόνιση του χώρου χαρακτηριστικών σε έναν άλλο, διαφορετικής πιθανόν διάστασης. Εικόνα από [171]

ως εκ τούτου η όλη διαδικασία αναφέρεται συχνά ως τέχνασμ πυρήνα (kernel trick) λόγω της ψευδαίσθησης της μη γραμμικότητας.

Στη βιβλιογραφία έχουν αναφερθεί αρκετές συναρτήσεις πυρήνα. Κλασσική περίπτωση είναι αυτή του τετραγωνικού ή του πολυωνυμικού πυρήνα, που μεταφέρει τα δείγματα σε έναν χώρο όπου ένα πολυώνυμο θα απεικονίζεται ως ευθεία (υπερεπίπεδο). Στην παρούσα εργασία χρησιμοποιούμε τον συνδυασμό χαρακτηριστικών με χ^2 πυρήνα, οποίος βρέθηκε ότι συνδυάζεται γόνιμα με την προσέγγιση του Σάκου Λέξεων στην αναγνώριση τόσο εικόνων (αντικειμένων) όσο και βίντεο (δράσεων) [128], [248]. Η προσέγγιση η οποία ακολουθούμε σε αυτή τη συνεργασία είναι ίδια με του [248] και στοχεύει στην από κοινού αναπαράσταση των διαφορετικών χαρακτηριστικών. Αν έχουμε N διαφορετικά κανάλια χαρακτηριστικών, τότε η συνάρτηση πυρήνα περιγράφεται από τη σχέση:

$$K(x_i, x_j) = \exp\left(-\sum_{c=1}^N \frac{1}{A^c} D(x_i^c, x_j^c)\right) \quad (6.1)$$

όπου $D(x_i^c, x_j^c)$ είναι η χ^2 του βίντεο x_i από το βίντεο x_j στα χαρακτηριστικά του καναλιού c . A^c είναι η μέση τιμή χ^2 αποστάσεων μεταξύ των δειγμάτων εκπαίδευσης στο κανάλι c . Η χ^2 απόσταση δύο διανυσμάτων x και y δίνεται από την εξίσωση:



Σχήμα 6.3: Ο ίδιος χώρος δεδομένων μπορεί να διαχωριστεί αρκετά διαφορετικά αν μεσολαβήσει μη γραμμική απεικόνιση σε έναν άλλο χώρο όπου μια μη γραμμική διαχωριστική επιφάνεια μπορεί να γίνει διαχωριστική. Για παράδειγμα, σε έναν χώρο που προκύπτει από έναν πολυωνυμικό μετασχηματισμού βαθμού n , ένα πολυώνυμο βαθμού n μπορεί να αναπαριστά μια ευθεία. Μετασχηματισμοί αυτού του είδους χρησιμοποιούνται για να διευκολύνουν τη γραμμική διαχωρισμότητα και την αποδοτική χρήση της μεθόδου SVM. Εικόνα από [171]

$$D(x, y) = \exp\left(-\gamma \sum_i \frac{(x[i] - y[i])^2}{x[i] + y[i]}\right) \quad (6.2)$$

Η προσέγγιση αυτή απεικονίζει κάθε δείγμα με ένα διάνυσμα πραγματικών αριθμών μήκους ίσο με τον αριθμό των δειγμάτων εκπαίδευσης. Η μέθοδος έχει αποδειχθεί εύρωστη στο συνδυασμό πολλαπλών καναλιών διαφορετικής πληροφορίας.

6.1.1.2 Πιθανοτική SVM: Η Μέθοδος Platt

Οι μη πιθανοτικοί ταξινομητές επιστρέφουν την κατηγορία στην οποία τοποθετούν ένα δείγμα σύμφωνα με κάποια μετρική. Ωστόσο η μετρική αυτή δεν έχει φυσική ερμηνεία και μεταξύ των τιμών που λαμβάνει δε μπορεί να γίνει απευθείας σύγκριση. Από την άλλη, δε μας παρέχεται κάποιο μέτρο βεβαιότητας της ταξινόμησης έτσι ώστε να μπορεί να χρησιμοποιηθεί σε επόμενο στάδιο για κάποιο συνδυασμό με άλλους πιθανοτικούς ταξινομητές. Η λύση που προτάθηκε σε αυτό ήταν η πιθανοτική αντιστάθμιση (probability calibration). Πιο σχετική με την περίπτωση των SVM είναι η μέθοδος Κλιμάκωσης Platt (Platt Scaling) η οποία προτάθηκε στο [176]. Η μέθοδος επιχειρεί να προσεγγίσει με μια εκθετική κατανομή πιθανοτήτων την πραγματική κατανομή πιθανοτήτων εξόδου, υπολογίζοντας παραμέτρους σύμφωνα με τα δείγματα εκπαίδευσης. Αν για παράδειγμα ο ταξινομητής κατατάσσει σε κατηγορίες το δείγμα x βασιζόμενος σε μια υπολογισθείσα συνάρτηση $f(x)$, τότε η μέθοδος Platt εκτιμά μια κατανομή

$$P(y|x) = \frac{1}{1 + \exp(Af(x) + b)} \quad (6.3)$$

με τα A και b να αποτελούν δυο βαθμωτές παραμέτρους που μαθαίνονται κατά τη διάρκεια εκπαίδευσης.

6.1.2 Το Σχήμα Tf-Idf

Στην θεωρία πληροφορίας, το σχήμα Tf-Idf (Term frequency - Inverse document frequency) είναι μια αριθμητική στατιστική μετρική που αναδεικνύει τη σημασία μιας λέξης σε ένα κείμενο από μια συλλογή εγγράφων. Η τιμή της μετρικής αυτής αυξάνει ανάλογα με τον αριθμό των φορών που η λέξη εμφανίζεται μέσα στο κείμενο, αλλά η εμφάνιση αντισταθμίζεται από τη συχνότητα της λέξης στο σύνολο των κειμένων της συλλογής. Έτσι λαμβάνεται υπόψιν το γεγονός ότι κάποια λέξη μπορεί να είναι γενικού σκοπού και να εμφανίζεται σε πολλά κείμενα πολλές φορές, χωρίς αυτό να αποδίδει κατι ριστό στην εμφάνισή της. Το σχήμα Tf-Idf συνδυάζει τους δύο όρους Term frequency και Inverse document frequency. Ο πρώτος όρος εισήχθη στο [145] και εκφράζει τον αριθμό των φορών που ένας όρος εμφανίζεται σε ένα κείμενο. Ο δεύτερος όρος εμφανίστηκε στο [111] και πρόκειται για το αντίστροφο της συχνότητας εμφάνισης της λέξης στο σύνολο των κειμένων. Έστω λοιπόν d ένα κείμενο από μια συλλογή κειμένων D και t μια λέξη. Τότε η μετρική Tf-Idf για την εμφάνιση της λέξης t στο κείμενο d έχει την έκφραση

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D) \quad (6.4)$$

Η αναλυτική έκφραση του σχήματος ποικίλει ανάλογα με τη χρήση ή τα δεδομένα. Πιο συνήθεις εκφράσεις οι $1 + \log f_{t,d}$ και $\log(1 + \frac{N}{n_t})$ οι οποίες στηρίζονται στο κομμάτι Tf και στο Idf αντίστοιχα.

6.2 Η Δική Μας Προσέγγιση και Τα Πειραματικά αποτελέσματα

6.2.1 Οι Ρυθμίσεις του Συνόλου Δεδομένων

Σε αυτό το σημείο παρουσιάζουμε τις ρυθμίσεις εκπαίδευσης και ελέγχου πάνω στο σύνολο δεδομένων MPII Cooking Activities. Το σύνολο περιέχει 44 βίντεο. Από αυτά, χρησιμοποιήσαμε τα 24 για εκπαίδευση, με τρόπο που να εξασφαλίζει την εμφάνιση όλων των κατηγοριών τόσο στα δεδομένα εκπαίδευσης όσο και στα δεδομένα ελέγχου. Παρότι το πλήθος των δράσεων είναι 65, επιλέγουμε τις 61 από αυτές, συσχετίζοντάς τες με τις εναπομείνασες 4 με δράσεις από το λεξιλόγιο μας. Οι επισημειώσεις αντικειμένων υπάρχουν στη νεότερη έκδοση του συνόλου δεδομένων. Από τις κατηγορίες αντικειμένων αυτές, κρατάμε μόνο τις 92, αφού μόνο αυτές εμφανίζονται στα βίντεο εκπαίδευσης και στα βίντεο ελέγχου. Λεπτομέρειες για τις παρεμβάσεις στο σύνολο δεδομένων αναφέρονται στο Παράρτημα A. Παρά την ελαφρά τροποποίηση όσον αφορά το πλήθος των δράσεων προς αναγνώριση, τα πειράματα δείχνουν να

συμφωνούν ως προς τα αποτελέσματα, με την αυθεντική υλοποίηση των [191], την οποία και υλοποιήσαμε ξανά.

Για όλα τα πειράματα ταξινόμησης έγινε μελέτη πάνω σε ποικιλία ταξινομητών και αλγορίθμων. Τελικά, επιλέχθηκε ο ταξινομητής SVM καθώς συνδυάζει υψηλή απόδοση και επίδοση. Οι παράμετροι του ταξινομητή επιλέχθηκαν με cross-validation ξεχωριστά για κάθε πείραμα. Η μετρική σε όλα τα πειράματα είναι η σταθμισμένη Mean Average Precision (mAP), η οποία χρησιμοποιήθηκε από την πρώτη εργασία πάνω στο συγκεκριμένο σύνολο δεδομένων [191] και έκτοτε παραμένει ως μετρική για ενιαία σύγκριση αποτελεσμάτων στο σύνολο αυτό. Η μετρική Precision ορίζεται από την εξίσωση:

$$\text{Precision} = \frac{|\{\text{TruePositives}\}|}{|\{\text{PredictedPositives}\}|} \quad (6.5)$$

Διαισθητικά, η μετρική Precision εξετάζει το κατά πόσο ένα σύστημα είναι ικανό να απορρίπτει τα αρνητικά δείγματα. Είναι φανερό ότι υψηλή τιμή Precision δεν εξασφαλίζει ακριβή ταξινόμηση. Επίσης είναι φανερό ότι η μετρική είναι ορισμένη για το πρόβλημα δύο κλάσεων. Θα δούμε τη γενίκευση στην περίπτωση των πολλών κλάσεων αφού ορίσουμε την ποσότητα Average Precision (AP) ως εξής:

$$\text{AveragePrecision} = \int_0^1 p(r)dr \quad (6.6)$$

Ν Τώρα μπορούμε να λάβουμε τον σταθμισμένο μέσο των τιμών AP για κάθε κλάση (θεωρώντας N δυαδικά προβλήματα, με N το πλήθος των κλάσεων) για να εκφράσουμε μια μετρική κατάλληλη για το πρόβλημα πολλών κλάσεων ως εξής:

$$\text{weighted mAP} = \frac{\sum_{i=1}^N w_i AP[i]}{\sum_{i=1}^N w_i} \quad (6.7)$$

6.2.2 Αξιοποίηση Πληροφορίας Όρασης Χαμηλού Επιπέδου

Σε πρώτη φάση επιχειρούμε να δούμε τη δύναμη των οπτικών χαρακτηριστικών. Η πληροφορία αυτή κωδικοποιεί την κίνηση και μπορεί να εξασφαλίσει ότι μια κίνηση συμβαίνει, σε αντίθεση με τη σημασιολογία που δεν μπορεί να εγγυηθεί κάτι τέτοιο. Ελέγχουμε διαφορετικούς τρόπους συνδυασμού των χαρακτηριστικών. Ο πρώτος είναι και ο προτεινόμενος από την αυθεντική εργασία των Πυκνών Τροχιών και του [191] και κάνει χρήση του πυρήνα χ^2 ώστε να συνδυάσει τα κανονικοποιημένα (ώστε να αθροίζουν στη μονάδα) ιστογράμματα για τους έξι τύπους οπτικών χαρακτηριστικών. Τα δείγματα εκπαίδευσης είναι 2888, οπότε και η διάσταση του διανύσματος που προκύπτει από αυτό τον μετασχηματισμό είναι 2888. Ένα τέτοιο διάνυσμα περιέχει την οπτική πληροφορία ενός τμήματος βίντεο. Μια άλλη μέθοδος συνδυασμού είναι η κωδικοποίηση των διανυσμάτων με το σχήμα Tf-Idf και η παράθεσή τους (concatenation). Το προκύπτον διάνυσμα έχει μήκος $4 \times 4000 + 3336 + 1536 = 20872$, τιμή η οποία επιβραδύνει αλλά δεν δυσχεραίνει τη μέθοδο SVM στην απόδοση, καθώς ο αλγόριθμος μπορεί να λειτουργεί ικανοποιητικά σε διαστάσεις πολύ μεγαλύτερες

από το πλήθος των δειγμάτων εκπαίδευσης. Τέλος, συνδυάζουμε τις παραπάνω μεθόδους, πρώτα μεταχηματίζοντας κατά Tf-Idf και μετά εφαρμόζοντας τον μη γραμμικό πυρήνα. Τα αποτελέσματα φαίνονται εδώ:

Method	mAP
Original method with χ^2 kernels ([191])	57.9
Original method with χ^2 kernels (ours)	58.4
Tf-Idf features stacked method	51.88
Tf-Idf combined with Chi-Squared kernels	58.27

Πίνακας 6.1: Αποτελέσματα των μεθόδων που αξιοποιούν μόνο πληροφορία χαμηλού επιπέδου Όρασης Υπολογιστών. Η προτεινόμενη μέθοδος φαίνεται ότι υπερισχύει των υπολοίπων και ότι η χρήση χ^2 πυρήνα είναι απαραίτητη.

Καταρχάς, στον πίνακα παρουσιάσαμε, πέραν των τριών δικών μας αποτελεσμάτων, το αποτέλεσμα του [191]. Ο λόγος είναι η σύγκριση της δικής μας υλοποίησης με τη δική τους, του ίδιου αλγορίθμου, αλλά με 61 κατηγορίες αντί για 65. Βλέπουμε ότι αφαιρώντας 4 κατηγορίες αποκτούμε μικρό προβάδισμα, όμως γενικά το αποτέλεσμα δεν αλλοιώνεται σημαντικά και έτσι τα αποτελέσματα της εργασίας μας παραμένουν συγκρίσιμα με αυτά της παγκόσμιας βιβλιογραφίας. Τώρα συγκρίνοντας τις τρεις δικές μας υλοποιήσεις, βλέπουμε ότι η κωδικοποίηση Tf-Idf είναι πολύ πιο ασθενής από την μετατροπή των χαρακτηριστικών με χ^2 πυρήνες. Ο συνδυασμός τους είναι ελάχιστα λιγότερο ισχυρός από τη μέθοδο χ^2 πυρήνων μόνη της, δείχνοντας ότι η χρήση αυτού του μετασχηματισμού εμπεριέχει την πιο επαρκή αναπαραστική δύναμη. Θα δούμε στη συνέχεια ωστόσο ότι αυτό ανατρέπεται όταν εισάγονται επιπλέον χαρακτηριστικά.

6.2.3 Αξιοποίηση Σημασιολογικής Πληροφορίας: Αντικείμενα

Στη συνέχεια εισάγουμε την σημασιολογική πληροφορία των αντικειμένων. Η πληροφορία αυτή εμπεριέχει τη σχέση δράσεων και αντικειμένων, εξασφαλίζοντας ότι μια σχέση πραγματοποιείται. Εξετάζουμε 6 τρόπους συνδυασμού χαρακτηριστικών. Οι πρώτοι τρεις χρησιμοποιούν τον συνδυασμό Tf-Idf και χ^2 πυρήνων για την αναπαράσταση των οπτικών χαρακτηριστικών, οπότε η διάσταση αυτού του τμήματος πληροφορίας είναι 2888. Για τα ιστογράμματα αντικειμένων δοκιμάζουμε απλή συνένωση, μετασχηματισμό Tf-Idf και συνένωση και μετασχηματισμό Tf-Idf και συμπερίληψη στον πυρήνα χ^2 ως έβδομο κανάλι πληροφορίας. Ως αποτέλεσμα, η διάσταση του διανύσματος ενός τμήματος βίντεο είναι 2980 (ως 2888+92) για τις δύο πρώτες περιπτώσεις και 2888 για την τρίτη. Πέραν αυτών των συνδυασμών, δοκιμάζουμε μετατροπή χ^2 πυρήνα για τα οπτικά χαρακτηριστικά (χωρίς Tf-Idf) και απλή συνένωση (τελική διάσταση 2980) ή συμπερίληψη στον πυρήνα ως έβδομο κανάλι (τελική διάσταση 2888) για τα ιστογράμματα αντικειμένων. Τέλος, μια ακόμα μεθόδος συνδυασμού είναι η συνένωση των οπτικών χαρακτηριστικών αφού προηγηθεί κωδικοποίηση με το σχήμα Tf-Idf και έπειτα συνένωση με τα ιστογράμματα αντικειμένων, οδηγώντας σε μια διάσταση 20964. Σε όλες τις μεθόδους χρησιμοποιήθηκαν οι ίδιες επισημειώσεις αντικειμένων, οι οποίες λαμβάνονται από το συγγενικό σύνολο δεδομένων MPII Cooking 2. Για να μοντελοποιήσουμε τον θόρυβο από την ανίχνευση αντικειμένων, εισάγουμε τεχνητά θόρυβο στα δείγματα εκπαίδευσης για τις περιπτώσεις όπου για ένα τμήμα οι επισημειώσεις δείχνουν ότι δεν υπάρχουν αντικείμενα. Ο θόρυβος αφορά 1, 2, 3 ή 4 τυχαία αντικείμενα τα οποία εισάγονται με πιθανότητα

0.15 για την κάθε περίπτωση, ενώ με πιθανότητα 0.4 αφήνουμε το διάνυσμα χωρίς θόρυβο. Στον πίνακα συγκρίνουμε τα αποτελέσματα αυτών των τεχνικών:

Method	mAP
χ^2 LLV features combined with objects	65.1
χ^2 LLV features and χ^2 Tf-Idf objects	58.33
χ^2 Tf-Idf LLV features, combined with objects	64.8
χ^2 Tf-Idf LLV features and Tf-Idf objects	62.86
χ^2 Tf-Idf LLV features and χ^2 Tf-Idf objects	54.76
Stacked Tf-Idf LLV features combined with objects	65.74

Πίνακας 6.2: Αποτελέσματα των μεθόδων που αξιοποιούν και πληροφορία αντικειμένων σε σύγκριση. Βλέπουμε ότι τα ποσοστά ανεβήκαν και ότι η ανάγκη για χ^2 μετασχηματισμό υποχωρεί.

Πρώτη παρατήρηση είναι σίγουρα η ενίσχυση των αποτελεσμάτων κατά πολύ μεγάλο περιθώριο κέρδους. Η συμπληρωματικότητα των καναλιών οπτικής και σημασιολογικής πληροφορίας γίνεται φανερή και πειραματικά. Μάλιστα, έχει ενδιαφέρον η παρατήρηση ότι ο συνδυασμός των καναλιών εμπλουτίζει τόσο τη διακριτική ικανότητα που δεν είναι απαραίτητος ο μετασχηματισμός με χ^2 πυρήνες για τα οπτικά χαρακτηριστικά. Για το τμήμα αντικειμένων, η πιο αξιέπαινη αναπαραστατική δύναμη εμφανίζεται με την απλή συνένωση, χωρίς επιπλέον μετασχηματισμούς. Τέλος, βλέπουμε ότι εισάγοντας την πληροφορία αντικειμένων, το σχήμα Tf-Idf μπορεί να ξεπεράσει την απόδοση του χ^2 μετασχηματισμού για τα οπτικά χαρακτηριστικά. Αυτή είναι μια ιδιαίτερα σημαντική παρατήρηση και από άποψη επίδοσης, καθώς ο χ^2 μετασχηματισμός είναι αρκετά δαπανηρός σε σχέση με το σχήμα Tf-Idf.

6.2.4 Αξιοποίηση Σημασιολογικής Πληροφορίας: Τύποι Λαβής

Η τελευταία μορφή πληροφορίας που εισάγουμε είναι αυτή των τύπων λαβής (grasping types). Το κανάλι αυτό εμπεριέχει το χειρισμό των αντικειμένων και το σχηματισμό των χεριών. Δοκιμάζουμε την απευθείας συνένωση των ιστογραμμάτων τύπων λαβής και αντικειμένων με τα κατά χ^2 μετασχηματισμένα, τα κατά Tf-Idf μετασχηματισμένα και σωρρευμένα και τα απλώς σωρρευμένα οπτικά χαρακτηριστικά, οδηγώντας σε διαστάσεις 2990, 20974 και 20974 αντίστοιχα. Τα αποτελέσματα φαίνονται συγκριτικά στον πίνακα:

Method	mAP
χ^2 LLV features combined with objects and grasping	65.15
Stacked Tf-Idf LLV features combined with objects and grasping	66.45
Stacked LLV features combined with objects and grasping	66.25

Πίνακας 6.3: Αποτελέσματα των μεθόδων που αξιοποιούν και πληροφορία τύπων λαβής. Υπάρχει μικρή βελτίωση στην απόδοση, ενώ η ανάγκη για χ^2 μετασχηματισμό εξαλείφεται.

Από τον παραπάνω πίνακα μπορούμε να εξάγουμε πολλά χρήσιμα συμπεράσματα. Καταρχάς, η εισαγωγή πληροφορίας για τον τύπο λαβής υποβοηθά την ταξινόμηση, έστω κατά ένα μικρό περιθώριο κέρδους, το οποίο είναι αναμενόμενο λόγω της απουσίας επισημειώσεων για χέρια και τύπους λαβής στο σύνολο δεδομένων. Η αύξηση

της απόδοσης είναι μεγαλύτερη στην περίπτωση του συνδυασμού με κατά Tf-Idf μετασχηματισμένα οπτικά χαρακτηριστικά, δείχνοντας ότι ο χ^2 μετασχηματισμός μπορεί να οδηγήσει σε κόρο την απόδοση, χωρίς όμως να έχουμε επαρκείς αποδείξεις για αυτό. Η κωδικοποίηση με το σχήμα Tf-Idf μπορεί επομένως να αντικαταστήσει τον χ^2 μετασχηματισμό σε αυτούς τους συνδυασμούς. Από την άλλη, μικρή διαφορά υπάρχει και από την απλή συνένωση, η οποία είναι φανερά πιο αποδοτική χρονικά.

6.2.5 Επανεκτίμηση Πιθανοτήτων Εξόδου

Στο κεφάλαιο 2 προβήκαμε στην υπόθεση ότι αντικείμενα και χαρακτηριστικά χαμηλού-επιπέδου όρασης εμφανίζονται ανεξάρτητα, άρα

$$P(\alpha | vocabulary, features) = P(\alpha | vocabulary) \times P(\alpha | features) \quad (6.8)$$

Ήρθε επομένως η ώρα να ελέγξουμε την καρποφορία αυτής της επιλογής. Εκπαιδεύουμε τρεις ταξινομητές SVM με εκτίμηση πιθανοτήτων μέσω Κλιμάκωσης Platt. Για τον πρώτο ταξινομητή συνενώνουμε ιστογράμματα αντικειμένων με κατά Tf-Idf μετασχηματισμένα ιστογράμματα οπτικών χαρακτηριστικών. Προσθέτουμε στα παραπάνω και τα ιστογράμματα τύπων λαβής για να εκπαιδεύσουμε τον δεύτερο ταξινομητή. Ο τελευταίος ταξινομητής εκπαιδεύεται πάνω στη σώρρευση μη μετασχηματισμένων ιστογραμμάτων οπτικών χαρακτηριστικών, αντικειμένων και τύπων λαβής. Από τους ταξινομητές αυτούς λαμβάνουμε την κατανομή $P(\alpha | features)$. Στηριζόμαστε στα ιστογράμματα αντικειμένων που ανιχνεύθησαν για να εξάγουμε την κατανομή $P(\alpha | vocabulary)$. Στην προκειμένη, ο όρος *vocabulary* αφορά μόνο τα αντικείμενα. Η διαφοροποίηση έγκειται στο ότι η κατανομή αυτή εξάγεται από πληροφορία κειμένου, οπότε στην πράξη πρόκειται για έναν συνδυασμό διαφορετικών καναλιών πληροφορίας.

Το σύνολο δεδομένων MPII Cooking 2, που αποτελεί συνέχεια, όπως έχουμε αναφέρει, του MPII Cooking Activities, περιέχει σύνολο περιγραφών για τα βίντεο και τις δράσεις τους. Το σύνολο περιγραφών έχει γραφεί σε φυσική γλώσσα και χρησιμοποιεί τις λέξεις του λεξιλογίου μας, οπότε μπορεί να χρησιμεύσει για την εξαγωγή των σχέσεων δράσεων-αντικειμένων. Από τα κείμενα αυτά εξάγονται οι επισημειώσεις ιστογραμμάτων αντικειμένων. Ουσιαστικά, η χρήση των λεξιών του λεξιλογίου στην ίδια κύρια πρόταση αυξάνει την πιθανότητα συσχέτισής τους. Για κάθε δράση μετράμε τη σχετική συχνότητα συνδυασμού της με κάθε λέξη του λεξιλογίου μας (ως προς τα αντικείμενα) και ορίζουμε αυτή ως πιθανότητα της δράσης δεδομένης της εμφάνισης της λέξης. Επειδή υπάρχουν δράσεις που δεν έχουν συνδυαστεί με όλα τα αντικείμενα, εφαρμόζουμε ομαλοποίηση Laplace στην φάση εξαγωγής των σχετικών συχνοτήτων, οι οποίες τώρα θα δίνονται από τη σχέση

$$P(\alpha | object_i) = \frac{T_{\alpha,i} + 1}{\sum_{i \in Objects} (T_{\alpha,i} + 1)} \quad (6.9)$$

δηλαδή προσθέτουμε 1 στις εμφανίσεις κάθε αντικειμένου μαζί με τη δράση α . Έτσι αποφεύγουμε να γενικεύσουμε τη μηδενική πιθανότητα εμφάνισης λόγω πιθανώς περιορισμένου συνόλου θετικών δειγμάτων. Κατά τη φάση ελέγχου, εφαρμόζουμε τη σχέση 6.8 θεωρώντας ως *vocabulary* τα αντικείμενα τα οποία εμφανίστηκαν. Δεν

συμπεριλαμβάνουμε τη μη εμφάνιση στον υπολογισμό των πιθανοτήτων καθώς πειραματικά διαπιστώσαμε ότι δεν προσφέρει πολλά. Τρέξαμε τα πειράματα για τους 3 ταξινομητές χρησιμοποιώντας τα υπολογισμένα αλλά και τα πραγματικά αντικείμενα (επισημειώσεις) ώστε να εξαντλήσουμε τα οφέλη αυτών των τεχνικών. Τα αποτελέσματα φαίνονται στον πίνακα:

Method	mAP
Stacked Tf-Idf LLV features combined with objects	66.43
Stacked Tf-Idf LLV features combined with GT objects	73.11
Stacked Tf-Idf LLV features combined with objects and grasping	67.1
Stacked Tf-Idf LLV features combined with GT objects and grasping	73.17
Stacked LLV features combined with objects and grasping	67.25
Stacked LLV features combined with GT objects and grasping	73.4

Πίνακας 6.4: Αποτελέσματα των μεθόδων που προσαρμόζουν τις πιθανότητες στην έξοδο. Η υπόθεση στατιστικής ανεξαρτησίας των δύο καναλιών, οπτικής και γλωσσικής πληροφορίας ενσωματώνει σημασιολογία υψηλότερου επιπέδου η οποία συνεισφέρει σημαντικά στην αποτελεσματική ταξινόμηση.

Όπως γίνεται φανερό, η απλή σώρρευση δίνει το καλύτερο αποτέλεσμα από όλες τις μεθόδους. Υπάρχει, επιπλέον, σαφής διαφοροποίηση μεταξύ των αποτελεσμάτων των αληθινών και των υπολογισμένων αντικειμένων, παρά την ακρίβεια υπολογισμού των αντικειμένων, όπως δείχαμε στο κεφάλαιο 4. Δηλαδή ο αλγόριθμος εξαρτάται ισχυρά από την εύρωστη ανίχνευση αντικειμένων. Από την άλλη, αξιοποιώντας την εξέλιξη των μεθόδων ανίχνευσης αντικειμένων, η σχεδιαστική επιλογή δείχνει ότι ο συνδυασμός αυτών των καναλιών πληροφορίας με τις παραπάνω υποθέσεις μπορεί να ωθήσει σε υψηλής ακριβείας συστήματα αναγνώρισης δράσεων. Μάλιστα, το γεγονός ότι η απλή σώρρευση είναι η ταχύτερη και ακριβέστερη μέθοδος, δείχνει την αποσυμπλοκή της μη γραμμικότητας και τη μη αναγκαιότητα πυρήνων.

6.2.6 Συνολική Παρουσίαση Αποτελεσμάτων και Σύγκριση με Παγκόσμια Βιβλιογραφία

Συνολικά τα αποτελέσματα όλων των μεθόδων που εφαρμόσαμε παρουσιάζονται στον πίνακα:

Method	mAP
Original method with χ^2 kernels ([191])	57.9
Original method with χ^2 kernels (ours)	58.4
Tf-Idf features stacked method	51.88
Tf-Idf combined with Chi-Squared kernels	58.27
χ^2 LLV features combined with objects	65.1
χ^2 LLV features and χ^2 Tf-Idf objects	58.33
χ^2 Tf-Idf LLV features, combined with objects	64.8
χ^2 Tf-Idf LLV features and Tf-Idf objects	62.86
χ^2 Tf-Idf LLV features and χ^2 Tf-Idf objects	54.76
Stacked Tf-Idf LLV features combined with objects	65.74
χ^2 LLV features combined with objects and grasping	65.15
Stacked Tf-Idf LLV features combined with objects and grasping	66.45
Stacked LLV features combined with objects and grasping	66.25
Stacked Tf-Idf LLV features combined with objects	66.43
Stacked Tf-Idf LLV features combined with GT objects	73.11
Stacked Tf-Idf LLV features combined with objects and grasping	67.1
Stacked Tf-Idf LLV features combined with GT objects and grasping	73.17
Stacked LLV features combined with objects and grasping	67.25
Stacked LLV features combined with GT objects and grasping	73.4

Πίνακας 6.5: Συνολικά αποτελέσματα όλων των μεθόδων που δοκιμάσαμε. Βλέπουμε ότι το εύρος είναι 15%, μια αξιόλογη βελτίωση από τη χρήση μόνο οπτικής πληροφορίας. Επιπλέον, η απλή σώρρευση χαρακτηριστικών είναι αρκετά ταχύτερη του υπολογισμού χ^2 πυρήνα.

Θα συγκρίνουμε τώρα τα αποτελέσματά μας με αυτά της παγκόσμιας βιβλιογραφίας.

Method	mAP
P-CNN + IDT-FV [36]	71.4
Interaction Part Mining [268]	72.4
Holistic + Pose [191]	57.9
Video Darwin [73]	72.0
Hierarchical Mid-Level Actions [230]	66.8
Higher-order Pooling [34]	73.1
GRP + IDT-FV [35]	75.5
Stacked LLV features combined with objects and grasping (ours)	67.25
Stacked LLV features combined with GT objects and grasping (ours)	73.4

Πίνακας 6.6: Σύγκριση αποτελεσμάτων της παγκόσμιας βιβλιογραφίας για το σύνολο δεδομένων MPII Cooking Activities. Επισημειώνονται με έντονα γράμματα η καλύτερη επίδοση [35] και η δική μας καλύτερη επίδοση, η οποία καταλαμβάνει τη δεύτερη θέση.

Τα παραπάνω αποτελέσματα δείχνουν τη δύναμη της σχεδίασής μας με τον συνδυασμό πολλών καναλιών πληροφορίας. Ταυτόχρονα, διαφαίνεται η σημασία της αυτονομίας των μονάδων σχεδιαστικά. Μπορούμε επομένως να αντικαταστήσουμε κάποια υπομονάδα, όπως το υποσύστημα εξαγωγής χαρακτηριστικών όρασης χαμηλού επιπέδου, με μια πιο ανανεωμένη έκδοση. Συζητούμε διάφορες ερευνητικές προτάσεις στο Κεφάλαιο 8.

Κεφάλαιο 7

Κατάτμηση Δράσεων

Η κατάτμηση ενός βίντεο δράσεων είναι απλά ο διαχωρισμός του σε τμήματα στα οποία συμβαίνει μόνο μία δράση. Οι βασικές προσεγγίσεις κατάτμησης στηρίζονται είτε σε επιβλεπόμενες είτε σε μη επιβλεπόμενες είτε σε ημιεπιβλεπόμενες μεθόδους μάθησης. Η πρώτη κατηγορία αξιοποιεί μοντέλα δράσεων και σύμφωνα με αυτά επιχειρεί να λύσει το πρόβλημα της κατάτμησης ως ένα πρόβλημα βελτιστοποίησης πάνω στα σκορ βεβαιότητας για κάθε θραύσμα. Οι μη επιβλεπόμενες μέθοδοι συνήθως στηρίζονται στις ομοιότητες μιας οικογένειας βίντεο προκειμένου να αποφασίσουν για τα όρια μιας δράσης που επαναλαμβάνεται σε όλα τα βίντεο. Τέλος, οι ημιεπιβλεπόμενες μέθοδοι επιχειρούν να σπάσουν το βίντεο σε συμπαγή κομμάτια έχοντας σαν οδηγό ένα αρχείο κειμένου. Οι δύο τελευταίες περιπτώσεις μπορούν να οδηγήσουν και στην αυτόματη εξαγωγή αναπαράστασης για τη δράση. Σε αυτή την εργασία εφαρμόζουμε τεχνικές επιβλεπόμενης μάθησης, οπότε θα εστιάσουμε σε αναφορές σε αυτές. Παρόλα αυτά, θα δοθούν και κάποια ενδιαφέροντα παραδείγματα που αξιοποιούν οποιαδήποτε από τις άλλες δύο μεθόδους επίσης.

7.1 Ιδέες και Τεχνικές Κατάτμησης Βίντεο

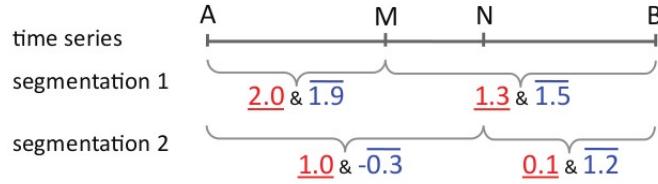
Πολλές εργασίες επιχείρησαν να προσεγγίσουν μαθηματικά το πρόβλημα καταλήγοντας σε ένα πρόβλημα βελτιστοποίησης. Το [146] προσεγγίζει το ζήτημα με κρυφά μαρκοβιανά μοντέλα που ενσωματώνουν τη δυναμική των δράσεων και χρησιμοποιεί τον ταξινομητή AdaBoost για την τελική απόφαση. Το [218] εφαρμόζει ημιμαρκοβιανά μοντέλα λαμβάνοντας υπόψιν χαρακτηριστικά εντός των τμημάτων αλλά και στα σύνορα αυτών με τα γειτονικά τμήματα, ενώ η αναπαράσταση των γειτονικών τμημάτων επηρεάζει το τρέχον τμήμα. Το πρόβλημα ελαχιστοποίησης επιλύεται με έναν αλγόριθμο τύπου Viterbi. Στο [8] συνδυάζονται κρυφά μαρκοβιανά μοντέλα με ιεραρχικές διαδικασίες Dirichlet επιλύοντας πρόβλημα κατάτμησης με άγνωστο αριθμό κλάσεων. Το [122] λαμβάνει υπόψιν τον μετασχηματισμό Hough για την δημιουργία προτάσεων ορίων θραυσμάτων. Ένας ταξινομητής SVM χρησιμοποιείται για να αποδώσει σκορ στα θραύσματα και η βέλτιστη κατάτμηση επιτυγχάνεται με χρήση δυναμικού προγραμματισμού. Η μέθοδος αυτή είναι λειτουργική ακόμα και σε περιπτώσεις όπου υπάρχουν άγνωστες κλάσεις.

Άλλες εργασίες δίνουν έμφαση στο είδος των χαρακτηριστικών που διακρίνουν τα τμήματα μεταξύ τους. Το [109] χρησιμοποιεί χαρακτηριστικά HOG και οπτικής ροής για να εξάγει μοντέλα για κάθε frame. Αφού ταξινομήσει κάθε frame ξεχωριστά, αθροίζει τις πιθανότητες των frames για να εξάγει την δράση. Η ταξινόμηση των frames γίνεται με χρήση δενδρικής δομής όπου οι κόμβοι αναπαριστούν χαρακτηριστικά κίνησης τα οποία κωδικοποιούν την αλληλουχία frames για κάθε δράση. Το [51] κάνει χρήση οπτικών χαρακτηριστικών χαμηλού επιπέδου με προσέγγιση σάκου λέξεων. Ο διαχωρισμός γίνεται με ένα στοχαστικό σχήμα που περιλαμβάνει διαδικασίες Dirichlet. Το [26] συγκεντρώνει χαρακτηριστικά τοπικών παραγώγων και οπτικής ροής και τα κωδικοποιεί με Fisher Vectors. Σε ένα κινούμενο παράθυρο σταθερού μήκους προβαίνει σε ταξινόμηση ανά θέση και αθροίζει τις πιθανότητες για κάθε ανά pixel σύμφωνα με τα αποτελέσματα ταξινόμησης σε κάθε παράθυρο που το περιλαμβάνει. Η μέγιστη πιθανότητα δράσης αντιστοιχεί στη δράση στην οποία ταξινομούμε το κάθε pixel. Σε όχι πολύ διαφορετική τάση, το [215] αξιοποιεί χαρακτηριστικά χρώματος, σχήματος και κίνησης για να διακρίνει κινήσεις σε ένα κλειστό περιβάλλον.

Συγγενικές εργασίες είναι αυτές της συνόψισης (summarization) βίντεο. Το [169] ασχολείται με την πολυτροπική συνόψιση βασισμένη σε ένα σύνολο ευριστικών ανάλογα των τομέα. Το [2] συγκεντρώνει πολυτροπικά χαρακτηριστικά και προσπαθεί να τα συνδυάσει με χρήση σημασιολογίας. Σε πιο πρόσφατες έρευνες, το [178] συνοψίζει βίντεο με εκ των προτέρων γνωστή θεματολογία. Χρησιμοποιεί σημασιολογικές ευριστικές για το σπάσιμο σε τμήματα τα οποία ταξινομεί στη συνέχεια. Το [189] στοχεύει σε κάτι πιο σύνθετο: στην εξαγωγή λεκτικής περιγραφής για βίντεο, συγκεκριμένα ταινίες. Μια σχετική εργασία [190] πετυχαίνει γλωσσική περιγραφή για ένα βίντεο σε πολλαπλά επίπεδα βάθους, δηλαδή από απλή λεζάντα μέχρι περιγραφική λεπτομέρεια.

Σχετικές με τις μαθηματικές μεθόδους των αλγορίθμων κατάτμησης είναι και ορισμένες εργασίες μη επιβλεπόμενης κατάτμησης. Το [82] κινείται προς αυτή την κατεύθυνση αξιοποιώντας το σχήμα Kernelized Temporal Cut, το οποίο ενσωματώνει χώρους Hilbert για να διαχειριστεί μη παραμετρικά προβλήματα υψηλής διαστατικότητας. Το [222] προχωράει σε μη επιβλεπόμενη κατάτμηση με Γκαουσιανά Μοντέλα Μίξης (Gaussian Mixture Models, GMM) και ταξινομεί τα θραύσματα με επιβλεπόμενη μάθηση. Στο [85] παρουσιάζεται μια διαφορετική άποψη για το ζήτημα κατάτμησης, αυτό της παράλληλης κατάτμησης με δύο βίντεο. Ο στόχος είναι η αναπαράσταση των βίντεο με χαρακτηριστικά τα οποία ευθυγραμμίζονται και στη συνέχεια αξιολογούνται με μαρκοβιανά τυχαία πεδία (Markov Random Fields, MRF) ώστε να γίνει το σπάσιμο του βίντεο. Μια άλλη προσέγγιση στο [252] αναπαριστά τις δράσεις με λέξεις, δομικά στοιχεία, οι οποίες επιτρέπουν τη μοντελοποίηση των σχέσεων μεταξύ των δράσεων στο εύρος του βίντεο.

Φυσικά υπάρχουν πολλές ακόμα ξεχωριστές κατευθύνσεις. Για παράδειγμα το [37] εφαρμόζει γενετικούς αλγορίθμους στην κατάτμηση του βίντεο. Το [143] εξάγει τα σημαντικά frames που μπορούν να αντιπροσωπεύσουν τις δράσεις στο βίντεο και βρίσκει μόνο με αυτά τα κατάλληλα τμήματα. Το [61] χρησιμοποιεί ασθενώς επιβλεπόμενη μάθηση για να εξάγει όχι μόνο τις δράσεις αλλά και τα πρόσωπα από ένα βίντεο. Εν έτει 2017, τα συνελικτικά νευρωνικά δίκτυα έχουν εισέλθει και στην κατάτμηση βίντεο. Συγκεκριμένα, το [129] χρησιμοποίησε χωροχρονικά (spatiotemporal) δίκτυα για ταυτόχρονη κατάτμηση και αναγνώριση δράσεων. Το [267] συνδυάζει μια δομή δικτύου με μονάδες LSTM (Long Short-Term Memory). Τέλος, το [57] εισήγαγε μία αρχιτεκτονική που αναμιγνύει τα χρονικά με τα επανερχόμενα (recurrent) νευρωνικά δίκτυα και πετυχαίνει state-of-the-art αποτελέσματα.



Σχήμα 7.1: Κατά το σπάσιμο ενός τμήματος AB έχουμε πολλαπλές επιλογές ως το πλήθος και το μήκος των επιμέρους τμημάτων. Ποια είναι όμως η βέλτιστη; Σύμφωνα με τους Hoai et al. [92], αυτό προκύπτει αναζητώντας το μέγιστο περιθώριο κέρδους μεταξύ της πιο πιθανής κλάσης και της αμέσως λιγότερο πιθανής. Οπότε στην εικόνα του σχήματος, όπου έχουμε δύο κλάσεις με σκορ που φαίνονται στο σχήμα για κάθε τμήμα, προτιμότερο ειναι το σπάσιμο στο σημείο N. Εικόνα από [92]

7.2 Ο αλγόριθμος των Hoai et al.

Η πιο σχετική εργασία με τη δική μας, όσον αφορά το κομμάτι της κατάτμησης δράσεων, είναι αυτή των [92]. Η μέθοδός τους εκκινεί με την εκμάθηση ενός μοντέλου ταξινομητή για αναγνώριση δράσεων. Όπως κι εμείς στην εργασία μας, έτσι και εκείνοι, χρησιμοποιούν ταξινομητές SVM. Για το βέλτιστο σπάσιμο του βίντεο σε τμήματα, ορίζουν το πρόβλημα έτσι ώστε να αναχθεί σε ένα πρόβλημα ελαχιστοποίησης, το οποίο και λύνουν με έναν αλγόριθμο δυναμικού προγραμματισμού. Αναλύουμε τα επιμέρους βήματα ξεχωριστά.

7.2.1 Εκπαίδευση SVM

Έστω n το πλήθος βίντεο X^i με $i = 1 \dots n$. Για τα βίντεο αυτά είναι γνωστά τα όρια κάθε δράσης, έστω πλήθος k_i , στα $0 = s_1^i < s_2^i < \dots < s_{k_i+1}^i = \text{len}(X^i)$ και αντιστοιχούν στις ετικέτες κλάσεων $y_1^i, \dots, y_{k_i}^i, i = 1, \dots, m$. Υπό αυτές τις συνθήκες, στόχος της εκπαίδευσης είναι η ελαχιστοποίηση ως προς $w_i, \xi_t^i \geq 0$ της ποσότητας:

$$\frac{1}{2m} \sum_{j=1}^m \|w_i\|^2 + C \sum_{i=1}^n \sum_{t=1}^{k_i} \xi_t^i \quad (7.1)$$

έτσι ώστε να ισχύει

$$(w_{y_t^i} - w_y)^T \phi(X_{(s_t^i, s_{t+1}^i)}^i) \geq 1 - \xi_t^i \quad \forall i, t, y \neq y_t^i \quad (7.2)$$

όπου $X_{(s_t^i, s_{t+1}^i)}^i$ το τμήμα του βίντεο X^i μεταξύ $[s_t + 1, s_{t+1}]$, $\phi(\cdot)$ η συνάρτηση χαρακτηριστικών και $w_y^T \phi(X_{(s_t^i, s_{t+1}^i)}^i)$ τα SVM scores για την ανάθεση του τμήματος αυτού στην κλάση y . Ακόμα, στις παραπάνω εξισώσεις χρησιμοποιούνται η παράμετρος C (regularization parameter) που δημιουργεί τις συνθήκες ισορροπίας μεταξύ μεγάλων περιθωρίων κέρδους και ελαφρύτερης παραβίασης των περιορισμών και οι μεταβλητές ξ_t^i (slack variables), που επιτρέπουν μικρή παραβίαση των περιορισμών. Σύμφωνα λοιπόν με αυτή τη σύμβαση, η σχέση 7.2 απαιτεί το τμήμα να αντιστοιχίζεται στην

κλάση y με μεγάλη βεβαιότητα (σχετικά μεγαλύτερο σκορ από κάθε άλλη κλάση για κάποιο περιθώριο κέρδους) και θέλουμε ταυτόχρονα ελαχιστοποίηση των ξ_t^i και των w_i , προβλήματα αντιμαχόμενα, οπότε παρέχουμε αντιστάθμιση με την παράμετρο C .

7.2.2 Κατάτμηση με Καταπίεση μη-Μεγίστων

Μέσω της εκπαίδευσης έχουμε εξάγει τα βάρη w_i . Για την κατάτμηση ζητάμε, για ένα άγνωστο βίντεο X , την εύρεση k και ακολουθίας s_1, \dots, s_{k+1} έτσι ώστε να ελαχιστοποιείται ως προς k, s_t, y_t, ξ_t η ποσότητα

$$\sum_{t=1}^k \xi_t \quad (7.3)$$

έτσι ώστε να ισχύει

$$l_{min} \leq s_{t+1} - s_t \leq l_{max} \quad \forall t \text{ και } (w_{y_t} - w_y)^T \phi(X_{(s_t, s_{t+1})}) \geq 1 - \xi_t \quad \forall t, y \neq y_t \quad (7.4)$$

με $s_1 = 0$ και $s_{k+1} = len(X)$, όπου τα l_{min}, l_{max} είναι το ελάχιστο και το μέγιστο μήκος τμήματος αντίστοιχα και υπολογίζονται κατά τη φάση εκπαίδευσης. Η κύρια ιδέα είναι να σπάσουμε το βίντεο έτσι ώστε τα τμήματα να αντιστοιχίζονται σε μία κλάση με μεγάλη βεβαιότητα. Η εξίσωση 7.3 απαιτεί να καταπιεστούν οι λιγότερο (μη μέγιστα) πιθανές κλάσεις, αντί να μεγιστοποιεί απευθείας SVM scores.

7.2.3 Δυναμικός Προγραμματισμός

Θεωρούμε τη βέλτιστη κατάτμηση για το τμήμα από 0 ως u ,

$$f(u) = \min_{k, s_t, y_t, \xi_t \geq 0} \sum_{t=1}^k \xi_t \quad (7.5)$$

υπό τους περιορισμούς της εξίσωσης 7.4 μόνο που τώρα $s_{k+1} = u$. Για κάθε συνδυασμό (u, l) με $u \in (0, len(X))$, $l \in [l_{min}, l_{max}]$ ορίζουμε

$$\xi(u, l) = \max(0, 1 - (w_{\hat{y}} - w_{\tilde{y}})^T \phi(X_{(u-l, u]})) \quad (7.6)$$

με \hat{y}, \tilde{y} τις θέσεις πρώτου και δεύτερου μεγίστου της ποσότητας $w_y^T \phi(X_{(u-l, u]})$. Σκοπός είναι ο υπολογισμός του ελαχίστου $f(len(X))$ με την εξίσωση

$$f(u) = \arg \min_l (\xi(u, l) + f(u - l)) \quad (7.7)$$

η οποία λύνεται με χρήση δυναμικού προγραμματισμού. Η πολυπλοκότητα αυτού του αλγορίθμου είναι $O(m(l_{max} - l_{min} + 1)len(X))$, δηλαδή $O(n^2)$.

7.3 Προτεινόμενος Αλγόριθμος

Επιχειρηματολογούμε ότι ο αλγόριθμος των [92] έχει ένα σαφές μειονέκτημα σε σύνολα δεδομένων τα οποία ενσωματώνουν την κατηγορία δράσεων υποβάθρου στις κλάσεις τους: επιβραβεύει την κατάτμηση σε μεγάλα τμήματα τα οποία κατατάσσονται στην κλάση δράσεων υποβάθρου. Πράγματι, ο αλγόριθμος επιχειρεί να ελαχιστοποιήσει ένα άθροισμα, άρα μια μη φραγμένη ποσότητα. Δεδομένου ότι οι όροι που αθροίζονται είναι όλοι μη αρνητικοί, το άθροισμα επιβαρύνεται με τη σώρρευση πολλών μικρών τμημάτων. Αντίθετα, ένα τμήμα που περιλαμβάνει πολλές μικρές ενέργειες, μπορεί να ταξινομηθεί συνολικά ως δράση υποβάθρου. Αυτό βέβαια εξαρτάται και από τη δύναμη του ταξινομητή, αλλά μόνο ιδανικά ο ταξινομητής θα επιβραβεύει με θριαμβευτικά μεγάλο κέρδος την πιο πιθανή κλάση. Επιπλέον, το κόστος του αλγορίθμου είναι αρκετό, ειδικά όταν η εξαγωγή των χαρακτηριστικών τμήματων είναι βαριά.

Προτείνουμε λοιπόν τον εξής αλγόριθμο, ο οποίος βελτιώνει αυτόν των [92]. Σε πρώτη φάση, θεωρώντας την ακολουθία αντικειμένων (ή γενικότερα λεξιλογίου) ως ισχυρό χαρακτηριστικό αναγνώρισης δράσεων, τεμαχίζουμε το βίντεο σε τμήματα στα οποία το διάνυσμα αντικειμένων παραμένει σταθερό. Δηλαδή, ορίζουμε τα τμήματα των δράσεων ως τα σημεία διαφοροποίησης των αντικειμένων που παρατηρούμε. Στη συνέχεια εφαρμόζουμε έναν αλγόριθμο δυναμικού προγραμματισμού, ο οποίος επιβραβεύει τη μέγιστη βεβαιότητα των τμημάτων, αντί για τη μέγιστη συνολική βεβαιότητα θραύσης. Συγκεκριμένα, σε κάθε τμήμα με σταθερό διάνυσμα αντικειμένων, εξετάζουμε όλα τα πιθανά συμπληρωματικά υποτμήματα, 2 σε πλήθος, που αν ενωθούν σε συνθετούν το βίντεο. Δηλαδή, σπάμε το υποτμήμα σε δύο μικρότερα με όλους τους δυνατούς τρόπους. Έστω μια θέση u μες στο υποτμήμα. Αριθμούμε για ευκολία τις θέσεις αυτές από 0 ως $len(X_i)$, για το υποτμήμα X_i . Τότε ζητάμε το u που ελαχιστοποιεί την ποσότητα

$$\begin{aligned} f(u) &= \min((\xi(u, u), \xi(len(X_i), len(X_i) - u))) \\ &= \min(\max(0, 1 - (w_{\hat{y}} - w_{\tilde{y}})^T \phi(X_{(0, u]})), \max(0, 1 - (w_{\hat{y}} - w_{\tilde{y}})^T \phi(X_{(u, len(X_i)]}))) \end{aligned} \quad (7.8)$$

Προσεγγίζουμε την ποσότητα $(w_{\hat{y}} - w_{\tilde{y}})^T \phi(X_{(u-l, u]})$ με τις αντίστοιχες πιθανότητες για τις δύο πιο πιθανές κλάσεις, οπότε η ποσότητα αυτή γράφεται ισοδύναμα $\Delta P_{(u-l, u]}$, που πρόκειται για μια μη αρνητική ποσότητα μικρότερη ή ίση του 1. Άρα η παραπάνω σχέση 7.8 γράφεται τώρα

$$f(u) = \min(1 - \Delta P_{(0, u]}, 1 - \Delta P_{(u, len(X_i)]}) \quad (7.9)$$

Είναι φανερό ότι από τα δύο τμήματα που τελικά θα προκύψουν, το ένα δεν θα δέχεται επιπλέον σπάσιμο, αφού έχει επιτευχθεί το ελάχιστο σκορ. Οπότε τρέχουμε τον ίδιο αλγόριθμο στο δεύτερο τμήμα. Σταματάμε τις επαναλήψεις όταν ένα υποτμήμα ταξινομηθεί ολόκληρο με βέλτιστο σκορ. Στο σημείο αυτό αναφέρουμε ότι στα άκρα

πρέπει να ισχύει μια διαφορετική συνθήκη. Όταν εξετάζουμε ένα τμήμα ως σύνολο τότε αποδίδουμε για τη θέση αυτή σκορ ίσο με το σκορ του τμήματος. Δηλαδή αποδίδουμε άπειρο σκορ σε τμήμα μηδενικού μήκους. Τέλος, συγχωνεύουμε διαδοχικά τμήματα που έχουν ταξινομηθεί στην ίδια κλάση.

Από άποψη πολυπλοκότητας, ο αλγόριθμος εκκινεί σπάζοντας το βίντεο σε κομμάτια, τάξης $O(1)$. Σε κάθε τμήμα μήκους l , διατρέχουμε l υπολογισμούς για να βρούμε το ελάχιστο σκορ και στη συνέχεια σπάμε το τμήμα και διατρέχουμε μόνο το ένα θραύσμα. Γίνεται φανερό ότι ο αλγόριθμος εντός της ταξινομίας είναι στην καλύτερη περίπτωση $O(n)$ (το τμήμα ταξινομείται ενιαία βέλτιστα), στη μέση περίπτωση $O(n \log n)$ (σπάμε σε $O(\log n)$ κομμάτια για τα οποία εκτελούμε $O(n)$ υπολογισμούς) και στη χειρότερη περίπτωση $O(n^2)$ (τα βέλτιστα τμήματα είναι πολυ μικρού μήκους και πρακτικά περνάμε ολόκληρο το τμήμα πολλές φορές). Δεδομένου ότι τα θραύσματα είναι της τάξης $O(1)$, ο συνολικός αλγόριθμος έχει μέση πολυπλοκότητα $O(n \log n)$. Το γεγονός αυτό καθιστά τον αλγόριθμο αρκετά πιο αποδοτικό από τον προτεινόμενο των [92].

Παρουσιάζουμε εδώ τα δομικά μέρη της παραπάνω λογικής διαδικασίας. Ο αλγόριθμός μας εκκινεί από την κατάτμηση του βίντεο με βάση τις αλλαγές στα εμφανιζόμενα αντικείμενα. Με κάποια απλότητα, δείχνουμε εδώ μια τέτοια συνάρτηση:

Algorithm 1 Πρώτη φάση του προτεινόμενου αλγορίθμου: κατάτμηση του βίντεο σύμφωνα με τα σημεία μεταβολής των εμφανίσεων αντικειμένων

```

function OBJECTSEGMENTATION(video)
    objectSegments=[]
    for f = 1 : length(frames) do
        if objectFeatures(f,:) ≠ objectFeatures(f - 1,:) then
            concatenate(objectSegments,f)
        end if
    end for
    return objectSegments
end function

```

Έχοντας λάβει τα αρχικά τμήματα, αναλύουμε το κάθε ένα από αυτά ξεχωριστά εφαρμόζοντας μια μέθοδο δυναμικού προγραμματισμού. Συνενώνουμε τα επιμέρους τμήματα που προκύπτουν και έχουμε το σύνολο των θραυσμάτων. Ένα βήμα που δε φαίνεται στον παρακάτω αλγόριθμο είναι η συνένωση διαδοχικών τμημάτων με ίδια ετικέτα κλάσης. Όπου στους παρακάτω αλγορίθμους φαίνεται ως όρισμα συνάρτησης ένα διάστημα της μορφής $[A, B]$, αυτό αφορά όλους τους ακεραίους από το A (συμπεριλαμβανομένου του A), μέχρι το B (μη συμπεριλαμβάνοντας το B).

Algorithm 2 Κύριος κορμός του προτεινόμενου αλγορίθμου: κλήση της συνάρτησης δυναμικού προγραμματισμού για κάθε τμήμα βίντεο που έχει προκύψει από το σπάσιμο σύμφωνα με τις μεταβολές αντικειμένων και σώρρευση των αποτελεσμάτων. Τα αποτελέσματα λαμβάνονται από τις λίστες FinalClassification και FinalSegmentation.

```

function CORESEGMENTATION(video)
    FinalSegmentation = []
        ▷ start frames of computed segments
    FinalClassification = []
        ▷ classes of computed segments
    objectSegments=OBJECTSEGMENTATION(video)
    for s = 2 : length(objectSegments) do
        finalSegments, finalClasses=DYNAMICSEGMENTATION([objectSegments(s-1), objectSegments(s)])
        concatenate(FinalSegmentation, finalSegments)
        concatenate(FinalClassification, finalClasses)
    end for
end function

```

Ο πυρήνας της συνάρτησης δυναμικού προγραμματισμού φαίνεται αμέσως παρακάτω. Αφού εξαχθούν τα χαρακτηριστικά και υπολογιστεί η SVM Loss με χρήση των πιθανοτήτων των δύο πιθανότερων κλάσεων, αντιπροσωπεύουμε έναν συνδυασμό τμημάτων, δηλαδή μια τομή, με την ελάχιστη των δύο πιθανοτήτων για τα τμήματα. Κρατάμε την τομή με το ελάχιστο κόστος και ορίζουμε ως πιθανό εσωτερικό σημείο θραύσης την τομή αυτή. Το τμήμα με την ελάχιστη πιθανότητα έχει πλέον ταξινομηθεί βέλτιστα και τρέχουμε την ίδια συνάρτηση για το δεύτερο τμήμα.

Algorithm 3 Συνάρτηση Δυναμικού Προγραμματισμού για την επαναληπτική κατάτμηση ενός δοθέντος τμήματος με σταθερές επισημειώσεις αντικειμένων.

```

function DYNAMICSEGMENTATION(segment)
    [Loss, Classes, Segments] = PROBABILITYCOMPUTATION(segment)
    segmentStart = argmin(Loss)
    segmentClass = Classes(segmentStart)
    if then segmentStart == 1 finalSegments = segment(segmentStart)
    finalClasses = segmentClass
    else if Segments(segmentStart) == 1 then
        newFinalSegments, newFinalClasses = DYNAMICSEGMENTATION([segment(segmentStart), segment(end)])
        finalSegments = concatenate(segment(segmentStart), newFinalSegments)
        finalClasses = concatenate(segmentClass, newFinalClasses)
    else
        newFinalSegments, newFinalClasses = DYNAMICSEGMENTATION([segment(segmentStart), segment(end)])
        finalSegments = concatenate(newFinalSegments, segment(segmentStart))
        finalClasses = concatenate(newFinalClasses, segmentClass)
    end if return finalSegments, finalClasses
end function

```

Ο υπολογισμός των πιθανοτήτων απαιτεί τον υπολογισμό τους και την διαμόρφωσή τους από την πληροφορία κειμένου. Σε πρώτη φάση εξάγονται τα χαρακτηριστικά για το κάθε τμήμα βίντεο. Γίνεται ταξινόμηση για κάθε ένα από αυτά και υπολογίζουμε την SVM Loss με χρήση πιθανοτήτων, όπως περιγράφηκε νωρίτερα. Επιστρέφουμε την τιμή της SVM Loss για κάθε τμήμα, την προκύπτουσα κλάση στην οποία

ταξινομείται και το υποτμήμα του (πρώτο ή δεύτερο) στο οποίο αντιστοιχεί η τιμή SVM Loss αυτή.

Algorithm 4 Αλγόριθμος εξαγωγής SVM Loss και κλάσεων μέσω πιθανοτήτων. Χωρίζουμε κάθε πιθανό τμήμα σε Left και Right, τα οποία δείχνουν το πώς το αρχικό τμήμα έχει σπάσει σε δύο.

```

function PROBABILITYCOMPUTATION(segment)
    for s = 1 : length(segment) do
        featuresLeft = EXTRACTFEATURES([segment(1), segment(s)])
        featuresRight = EXTRACTFEATURES([segment(s), segments(end)])
        probabilitiesLeft = CLASSIFY(featuresLeft)
        probabilitiesRight = CLASSIFY(featuresRight)
        sortedPLeft = sort(probabilitiesLeft)
        sortedPRight = sort(probabilitiesRight)
        lossLeft = 1 - (sortedPLeft(1) - sortedPLeft(2))
        lossRight = 1 - (sortedPRight(1) - sortedPRight(2))
        classLeft = argmin(probabilitiesLeft)
        classRight = argmin(probabilitiesRight)
        Loss(s) = min(lossLeft, lossRight)
        Segments(s) = argmin(lossLeft, lossRight)
        classes = concatenate(classLeft, classRight)
        Classes(s) = classes(Segments(s))
    end for
    return Loss, Classes, Segments
end function

```

Τρέξαμε τον αλγόριθμο κατάτμησης σε 18 από τα 20 test βίντεο, λόγω περιορισμών υπολογιστικών πόρων σε θέματα μνήμης. Για την αξιολόγηση των αποτελεσμάτων χρησιμοποιούμε τη μετρική Mean Average Precision που περιγράψαμε στο προηγούμενο κεφάλαιο, ελέγχοντας την μετρική ανά frame βίντεο. Σε 323262 test frames, ο αλγόριθμός μας πιάνει απόδοση 46.4% mAP. Καθώς, από όσο γνωρίζουμε δεν έχουν δημοσιευτεί αποτελέσματα κατάτμησης δράσεων για το συγκεκριμένο σύνολο δεδομένων, δεν μπορούμε να συγκρίνουμε αυτό το αποτέλεσμα με άλλα της βιβλιογραφίας.

Το πλήθος των δράσεων που αναγνωρίζουμε (61) και το μεγάλο μήκος των βίντεο σε frames (18000 frames κατά μέσο όρο) αποτρέπουν μια λεπτομερή οπτικοποίηση των αποτελεσμάτων στο χρονικό ορίζοντα. Αντί αυτού, παραθέτουμε μια ένδειξη της απόδοσης του αλγορίθμου στο πρόβλημα της κατάτμησης σε δράσεις ή υπόβαθρο. Πρόκειται για ένα δυαδικό πρόβλημα το οποίο είναι ενδιαφέρον λόγω της ανισορροπίας του πλήθους των δειγμάτων των κλάσεων. Πιο αναλυτικά, η κλάση υποβάθρου καταλαμβάνει το 35% περίπου των δειγμάτων εκπαίδευσης, ποσοστό υπερβολικά μεγαλύτερο από το αντίστοιχο όλων των υπολοίπων κλάσεων. Επομένως αναμένουμε η κλάση αυτή να εμφανίζεται αρκετά συχνά και στα test videos, οπότε η ικανότητα διαχωρισμού των δράσεων υποβάθρου από τις υπόλοιπες είναι σημαντικό ζήτημα. Ως μεμονωμένο πείραμα, μετρήσαμε την μετρική mAP για το δυαδικό πρόβλημα υπόβαθρο ή μη υπόβαθρο και βρήκαμε 80.5% mAP. Το σχήμα 7.2 δείχνει τα αποτελέσματα αυτής της διάκρισης σε ένα βίντεο ελέγχου.



Σχήμα 7.2: Αποτελέσματα της κατάτμησης του βίντεο s19-d01 για το δυαδικό πρόβλημα δράσεων υποβάθρου εναντίον δράσεων μη υποβάθρου. Ποιοτικά (εικόνα) αλλά και ποσοτικά (80.5% mAP) τα αποτελέσματα είναι ικανοποιητικά. Η παρούσα εικόνα δείχνει με μαύρο τις περιοχές δράσεων υποβάθρου και με λευκό τις δράσεις προσκηνίου. Όπως φαίνεται, το σχήμα περικλείει δύο γραμμες αποτελεσμάτων. Στην πάνω γραμμή φαίνεται η πραγματική (ground truth) κατάτμηση και στην κάτω η υπολογισμένη κατάτμηση με χρήση του αλγορίθμου που προτείνουμε.

Κεφάλαιο 8

Συμπεράσματα-Επίλογος

8.1 Συμβολή της Διπλωματικής Εργασίας

Στην εργασία αυτή αντιμετωπίσαμε τα ζητήματα αναγνώρισης και κατάτμησης δράσεων λεπτομέρειας σε βίντεο με χρήση πολλών καναλιών πληροφορίας. Δώσαμε βάρος στη σημαντικότητα της σημασιολογίας στην αναγνώριση δράσεων και δείξαμε ότι η σύζευξη αυτής με ισχυρά χαρακτηριστικά χαμηλού επιπέδου Όρασης Υπολογιστών μπορεί να βελτιώσει αισθητά το αποτέλεσμα ταξινόμησης σε σχέση με τη χρήση αποκλειστικά οπτικών χαρακτηριστικών.

Η εργασία μας εκκίνησε με την παρουσίαση ενός γενικευμένου, ενιαίου συστήματος κατάτμησης και αναγνώρισης δράσεων το οποίο συνδυάζει πολλαπλά κανάλια πληροφορίας όπως Όραση, Κείμενο και Ομιλία (μέσω υποτίτλων). Ο συνδυασμός των διαφορετικών καναλιών φαίνεται να συνεισφέρει τόσο στην αναπαράσταση των δεδομένων, όσο και στην αποτελεσματικότερη αναγνώριση και τελικά κατάτμηση. Μαζί με τη γενικευμένη σχεδίαση, παρουσιάσαμε και συγκεκριμένη υλοποίηση του προτεινόμενου συστήματος, εξηγώντας τις παραμέτρους που επιλέξαμε και εφαρμόζοντάς το σε ένα απαιτητικό σύνολο δεδομένων βίντεο [191].

Καθώς το σύνολο δεδομένων δεν παρείχε το σύνολο των επισημειώσεων που θα θέλαμε να χρησιμοποιήσουμε, προβήκαμε στη δημιουργία δικών μας επισημειώσεων. Πιο αναλυτικά, με τη βοήθεια επισημειωτών, συλλέχθηκαν αρχεία υποτίτλων για τα βίντεο. Ακόμη, προβήκαμε σε επισημειώσεις αντικειμένων ανά 100 frames βίντεο, κοιτώντας για αντικείμενα που εξασφαλίζουν την ελάχιστη οπτική διασπορά στη διάρκεια του βίντεο. Επιπλέον, χρησιμοποιήσαμε μια νεότερη μορφή του συνόλου δεδομένων [192] για να συλλέξουμε τις σχέσεις αντικειμένων-δράσεων μέσω κειμενικών δεδομένων που περιέχονται στις επισημειώσεις αυτού. Τέλος, τροποποιώντας τις επισημειώσεις ενός συγγενικού συνόλου δεδομένων [4] για αναγνώριση πόζας, εξαγάγαμε επισημειώσεις για τις θέσεις των χεριών, οπότε μπορέσαμε να χτίσουμε μια βάση εκπαίδευσης ανιχνευτή χεριών.

Συνεχίσαμε με την υλοποίηση του προτεινόμενου συστήματος παρουσιάζοντας την κατασκευή κάθε υπομονάδας ξεχωριστά, αφού όπως τονίζουμε, ο σχεδιασμός επιτρέπει την αυτόνομη ενημέρωση ενός υποσυστήματος. Χρησιμοποιήσαμε τη μέθοδο των Πυκνών Τροχιών, συνδυαζόμενη με χαρακτηριστικά στάσης σώματος, για εξαγωγή

χαρακτηριστικών Όρασης χαμηλού επιπέδου (low-level features). Ταυτόχρονα, συνδυάσαμε πληροφορία από την οπτική ανίχνευση αντικειμένων με αντίστοιχη από τη μελέτη υποτίτλων για να εξάγουμε τα αντικείμενα που εμφανίζονται και χρησιμοποιούνται στη διάρκεια του βίντεο και χρησιμοποιήσαμε την εμφάνιση των αντικειμένων ως χαρακτηριστικό αναπαράστασης των δράσεων. Κατά την ανίχνευση των αντικειμένων εισαγάγαμε την περιοχή ενδιαφέροντος, η οποία εκτείνεται γύρω από τον άνθρωπο και τις περιοχές κίνησης προσκηνίου. Θεωρήσαμε έγκυρες τις εμφανίσεις των αντικειμένων μόνο μέσα σε αυτή την περιοχή, ελαχιστοποιώντας έτσι την συμπερίληψη εμφάνισης αντικειμένων άσχετων με τη δράση στα χαρακτηριστικά. Δείξαμε ότι ο συνδυασμός υποτίτλων και οπτικής αντίληψης μπορεί να επιτύχει υψηλές επιδόσεις ανίχνευσης.

Στη συνέχεια εξαγάγαμε πληροφορία από τον τύπο λαβής (grasping type) και την χρησιμοποιήσαμε στην αναπαράσταση των δράσεων. Για το σκοπό αυτό υλοποιήσαμε έναν ανιχνευτή χεριών και τροφοδοτήσαμε την έξοδό του σε ένα συνελικτικό νευρωνικό δίκτυο (ResNet) στο οποίο είχε αφαιρεθεί το τελικό στάδιο. Έτσι, είδαμε το δίκτυο σαν μια μονάδα εξαγωγής συνελικτικών χαρακτηριστικών, τα οποία χρησιμοποιήσαμε αφενός για να ομαδοποιήσουμε τους τύπους λαβής σε 10 κατηγορίες και αφετέρου για να ταξινομήσουμε τις προκύπτουσες εικόνες χεριών σε κλάσεις. Δείξαμε ότι η εισαγωγή των ταξινομημένων τύπων λαβής ως χαρακτηριστικά στην περιγραφή των δράσεων βοηθάει ελαφρώς στην ταξινόμηση και δικαιολογήσαμε τη μικρή συνεισφορά με την απουσία ετικετών και επισημειώσεων για τύπους λαβής.

Για την ταξινόμηση, πειραματιστήκαμε με ποικίλες μεθόδους συνδυασμού χαρακτηριστικών και ταξινομητές. Μετά από σύγκριση διαφόρων αλγορίθμων, στο κομμάτι του ταξινομητή καταλήξαμε στις γραμμικές Μηχανές Διανυσμάτων Υποστήριξης (SVM) λόγω της ταχύτητας και της απόδοσης που πετύχαιναν. Όσον αφορά τον συνδυασμό, δοκιμάσαμε πρώτα τη δύναμη των χαμηλού επιπέδου χαρακτηριστικών, στη συνέχεια τα συνδυάσαμε με χαρακτηριστικά εμφάνισης αντικειμένων και τελικά με χαρακτηριστικά τύπων λαβής. Ταυτόχρονα, πειραματιστήκαμε με τους τρόπους συνδυασμού των διαφορετικών χαρακτηριστικών. Δείξαμε ότι όταν χρησιμοποιηθούν μόνα τους τα χαρακτηριστικά χαμηλού επιπέδου Όρασης, είναι ανάγκη να υποστούν ένα μη γραμμικό μετασχηματισμό, όπως ο χ^2 πυρήνας, ενώ η συνένωση αποτυγχάνει. Με την εισαγωγή των χαρακτηριστικών εμφανίσεων αντικειμένων αιρείται αυτή η ανάγκη και το σχήμα Tf-Idf μπορεί να χρησιμοποιηθεί αντί όποιου άλλου μετασχηματισμού, οδηγώντας ταυτόχρονα σε καλύτερη απόδοση και επίδοση. Παρόμοια αποτελέσματα δίνει και η εισαγωγή των χαρακτηριστικών τύπων λαβής. Τέλος, εκμεταλλευτήκαμε τα δεδομένα κειμένου για να εξάγουμε σχέσεις αντικειμένων-δράσεων και να προσαρμόσουμε τις πιθανοτήτες εξόδου. Η πράξη αυτή αφενός ωθεί σε πολύ βελτιωμένες επιδόσεις και αφετέρου αίρει και την ανάγκη για κωδικοποίηση των διαφορετικών χαρακτηριστικών, αφού η απλή συνένωσή τους δίνει το καλύτερο αποτέλεσμα (73.1 mAP). Πειράματα θεωρώντας γνωστές τις εμφανίσεις των αντικειμένων έδειξαν ότι η μέθοδος αυτή μπορεί να συγκριθεί με τα state-of-the-art αποτελέσματα.

Εφαρμόσαμε τον ταξινόμητη που έδωσε το καλύτερο αποτέλεσμα αναγνώρισης που επιτύχαμε για να προβούμε σε κατάτμηση δράσεων. Προτείναμε έναν νέο αλγόριθμο κατάτμησης, ο οποίος εκκινεί από την αρχική τμήση με βάση τις μεταβολές των εμφανιζόμενων αντικειμένων και κάνει χρήση δυναμικού προγραμματισμού για τον περαιτέρω βέλτιστο διαχωρισμό των επιμέρους τμημάτων με βάση την ελαχιστοποίηση της τροποποιημένης SVM loss $1 - (P_{max1} - P_{max2})$, με P_{max1} , P_{max2} τις πιθανότητες της πιο πιθανής και της αμέσως πιο πιθανής κλάσης αντίστοιχα. Τα διαδοχικά τμήματα που προκύπτουν με ίδια κλάση συνενώνονται. Αξιολογούμε την επιτυχία του αλγορίθμου μας χρησιμοποιώντας ως μετρική την ανά frame mAP. Συγκεκριμένα,

πετυχαίνουμε επίδοση 46.4 % mAP ανά frame. Καθώς δεν υπάρχουν αποτελέσματα στη βιβλιογραφία για αυτό το πρόβλημα, δεν προβαίνουμε σε κάποια σύγκριση.

Πέραν των παραπάνω, στη διάρκεια της εργασίας συχνά σταθήκαμε σε ιστορικά στοιχεία σχετικά με την αναγνώριση και την κατάτμηση δράσεων με χρήση διαφορετικών καναλιών πληροφορίας, την ανίχνευση αντικειμένων σε εικόνες και την εξαγωγή προσκηνίου. Ταυτόχρονα, είδαμε θεωρητικά πολλές μεθόδους και αλγορίθμους, όπως η μέθοδος των Πυκνών Τροχιών και τα χαρακτηριστικά πόζας, η ανίχνευση αντικειμένων με χρήση Μοντέλων Παραμορφώσιμων Τμημάτων, η εξαγωγή προσκηνίου με χρήση Γκαουσιανών Μοντέλων Μίξης, τα Συνελικτικά Νευρωνικά Δίκτυα και πιο αναλυτικά το δίκτυο ResNet, τα χαρακτηριστικά BING και ο αλγόριθμος κατάτμησης των [92]. Μάλιστα στον τελευταίο, αναλύσαμε τα βήματά του και τη λειτουργία του και αναδείξαμε τις αδυναμίες του, οπότε στηριχθήκαμε σε αυτόν για να προτείνουμε έναν νέο αλγόριθμο κατάτμησης βίντεο. Τέλος, μελετήσαμε και παραθέσαμε μεθόδους και αποτελέσματα σχετικών εργασιών με τη δική μας, είτε για την επιρροή τους στη δική μας είτε για σύγκριση.

Συνοψίζοντας, οι κυριότερες συνεισφορές της εργασίας αυτής είναι οι εξής:

- Η πρόταση ενός γενικευμένου ενιαίου συστήματος αναγνώρισης και κατάτμησης δράσεων σε αφαιρετική μορφή, σχεδιασμένο έτσι ώστε να αξιοποιεί πληροφορία πολλαπλών καναλιών και κάθε υπομονάδα του να είναι ανεξάρτητη από τις υπόλοιπες.
- Η υλοποίηση ενός τέτοιου συστήματος, η οποία καταλήγει σε απόδοση ταξινόμησης δράσεων που συγκρίνεται με τις κορυφαίες επιδόσεις της βιβλιογραφίας.
- Ο πειραματισμός με χαρακτηριστικά διαφορετικών καναλιών και η πειραματική απόδειξη της δύναμης του συνδυασμού πληροφορίας χαμηλού επιπέδου με σημασιολογική πληροφορία υψηλού επιπέδου.
- Ο πειραματισμός με διαφορετικές μεθόδους συνδυασμού χαρακτηριστικών και η πειραματική απόδειξη ότι με ισχυρά χαρακτηριστικά, όπως συνδυασμένα οπτικά και σημασιολογικά χαρακτηριστικά, αιρέται η ανάγκη για μη γραμμικό μετασχηματισμό και ότι αρκεί η συνένωση των χαρακτηριστικών σε ένα διάνυσμα που ταξινομείται με γραμμική SVM.
- Η ενίσχυση των πιθανοτήτων εξόδου του αλγορίθμου αναγνώρισης, η οποία φαίνεται να προσφέρει επιπλέον κέρδος στην απόδοση του ταξινομητή δράσεων.
- Η πρόταση και υλοποίηση ενός νέου αλγορίθμου κατάτμησης βίντεο δράσεων, ο οποίος κάνει χρήση πιθανοτήτων και δυναμικού προγραμματισμού.
- Η πρόταση και υλοποίηση ενός αλγορίθμου ανίχνευσης χεριών σε εικόνες, ο οποίος συνδυάζει χαρακτηριστικά BING, HOG και ιστογράμματα χρώματος και είναι αρκετά γρήγορος και εύρωστος.
- Η πρόταση και παροχή νέων επισημειώσεων στο σύνολο δεδομένων [191]. Ακόμα και αν οι επισημειώσεις αυτές δεν είναι άμεσα χρήσιμες, παραμένουν ως ένδειξη πιθανής μελλοντικής εργασίας πάνω σε ένα σύνολο δεδομένων που θα παρέχει, πέραν των επισημειωμένων βίντεο δράσεων, πληροφορία και για ανίχνευση αντικειμένων, χεριών και υποτίτλους ή ήχο.

8.2 Προτάσεις Για Μελλοντική Έρευνα

Καθώς η εργασία αυτή, όπως και κάθε εργασία, είναι πεπερασμένη, υπάρχουν ζητήματα τα οποία θα επιθυμήσουμε να εξετάσουμε σε μελλοντικές ερευνές μας. Πιστεύουμε ότι ορισμένες πρακτικές από αυτές είναι πολλά υποσχόμενες και μπορούν να αποδόσουν εξαιρετικά. Όπως εξάλλου τονίσαμε στη διάρκεια αυτής της εργασίας, η ανεξάρτητη σχεδίαση των υποσυστημάτων του προτεινόμενου συστήματος διευκολύνει την αυτόνομη ενημέρωση κάθε υπομονάδας ξεχωριστά, οπότε τα πειράματα που προτείνουμε μπορούν να γίνουν επηρεάζοντας μόνο το αντίστοιχο κομμάτι του συνολικού συστήματος.

Σε πρώτη φάση, εξετάζουμε πιθανές βελτιώσεις του υποσυστήματος οπτικής πληροφορίας χαμηλού επιπέδου. Η μέθοδος που χρησιμοποιήθηκε σε αυτή την εργασία είναι οι Πυκνές Τροχιές [248], συνδυαζόμενες με χαρακτηριστικά πόζας [191]. Ως φυσική συνέχεια βλέπουμε τη χρήση των Βελτιωμένων Πυκνών Τροχιών [247]. Εξάλλου, δύο από τις εργασίες [35], [36] που συγκρίνουμε με τη δική μας στο κεφάλαιο 6, πάνω στο ίδιο σύνολο δεδομένων, χρησιμοποιούν αυτή τη μέθοδο και κωδικοποιούν τα χαρακτηριστικά με Fisher Vectors, με την πρώτη εργασία μάλιστα να πετυχαίνει το τρέχον state-of-the-art αποτέλεσμα. Επιπλέον, οι δύο εργασίες αυτές χρησιμοποιούν Συνελικτικά Νευρωνικά Δίκτυα, για διαφορετικούς σκοπούς η κάθεμία. Μια ιδέα είναι να συνδυάσουμε συνελικτικά χαρακτηριστικά με χαρακτηριστικά Πυκνών Τροχιών, όπως στο [35]. Τέλος, ως βελτίωση πάνω στα χαρακτηριστικά πόζας, οι [36] χρησιμοποιούν νευρωνικά δίκτυα, κατεύθυνση που επίσης είναι ενδιαφέροντα. Οι διαφορετικές μορφές χαρακτηριστικών μπορούν να συνδυαστούν με διαφορετικές μεθόδους σύμμειξης καναλιών πριν τον ταξινομητή.

Στην ανίχνευση αντικειμένων, θα ήταν χρήσιμο να δοκιμάσουμε διαφορετικούς αλγορίθμους ανίχνευσης, πιο εύρωστους και γενικούς από το Ταίριασμα Προτύπων. Επιπλέον, πιο σύνθετη λογική μπορεί να χρησιμοποιηθεί και στην ανάλυση των υποτίτλων, αφού κι εκεί εφαρμόζουμε γλωσσικό Ταίριασμα Προτύπων. Για παράδειγμα, μια ιδέα όσον αφορά την οπτική ανίχνευση είναι η εφαρμογή των Μοντέλων Παραμορφώσιμων Τμημάτων [69]. Μια άλλη ιδέα είναι ο συνδυασμός χαρακτηριστικών, όπως στην περίπτωση του ανιχνευτή χεριών. Το [268] εξάγει προτάσεις αντικειμένων με χαρακτηριστικά BING και παρακολουθεί τις περιοχές αυτές πυκνά στη διάρκεια του βίντεο. Στο [152] εκπαιδεύεται ένας ανιχνευτής αντικειμένων (τροφών) με Συνελικτικά Νευρωνικά Δίκτυα. Στην ίδια εργασία, η ανάλυση υποτίτλων γίνεται αφού προηγηθεί συντακτική ανάλυση η οποία εξάγει τα μέρη του λόγου. Στη συνέχεια, οι συγγραφείς κρατάνε τα ουσιαστικά που εμφανίζονται γύρω από ένα ρήμα που επιλέγεται από μια γνωστή λίστα. Τέλος, το [3], χρησιμοποιεί συντακτικές σχέσεις ρήματος-αντικειμένου για να εξάγει κατηγορίες δράσεων με μη επιβλεπόμενο τρόπο, οι οποίες σχετίζονται άμεσα με μια κλάση αντικειμένων (π.χ. αλλαγή λάστιχου).

Σίγουρα μια σημαντική εργασία θα ήταν η εφαρμογή του συστήματός μας πάνω σε ένα άλλο σύνολο δεδομένων με διαφορετικές επισημειώσεις. Ιδανικά, θα θέλαμε επιπλέον, επισημειώσεις για ανίχνευση αντικειμένων και χεριών/τύπων λαβής, καθώς και υπότιτλους ή έστω ηχητικό κανάλι για να τους εξάγουμε. Στο κομμάτι της εξαγωγής τύπων λαβής ειδικά, θα μπορούσαμε σε αυτή την περίπτωση να διαπιστώσουμε πειραματικά τη συνεισφορά της πληροφορίας αυτής στην αναγνώριση των δράσεων, αφού όπως είδαμε, στην περίπτωσή μας είναι αρκετά μικρή, κάτι που δικαιολογήσαμε μέσω της εκπαίδευσης με τα αποτελέσματα ομαδοποίησης, χωρίς δηλαδή ετικέτες. Στο κομμάτι των υποτίτλων, θα είχε ενδιαφέρον η χρήση πραγματικών υποτίτλων, χωρίς τους περιορισμούς που εισαγάγαμε στη δημιουργία τους. Τέλος, η

εφαρμογή του αλγορίθμου κατάτμησης που προτείνουμε θα είχε ενδιαφέρον σε ένα σύνολο δεδομένων όπου υπάρχουν και άλλα αποτελέσματα από σχετικές εργασίες στην παγκόσμια βιβλιογραφία, έτσι ώστε να μπορούμε να συγκρίνουμε τα αποτελέσματά μας.

Τέλος, στην περίπτωση της κατάτμησης, θεωρήσαμε σιωπηρά ότι τα χαρακτηριστικά εμφανίσεων αντικειμένων μένουν πρακτικά σταθερά σε μεγάλα τμήματα βίντεο. Η υπόθεση αυτή δε μας ενόχλησε, αφού χειριστήκαμε την κατάτμηση με τις ground truth επισημειώσεις εμφάνισης αντικειμένων. Εν τούτοις, αν επιχειρούσμε να κάνουμε το ίδιο με τα υπολογισμένα ιστογράμματα εμφανίσεων των αντικειμένων μέσω ανίχνευσης, τα αποτελέσματα θα ήταν διαφορετικά, καθώς θα υπάρχει αρκετός θόρυβος. Αν για παράδειγμα σε ένα frame εντοπισθεί λανθασμένα ένα αντικείμενο, τότε θα έχουμε σπάσιμο του βίντεο εκεί. Το πρόβλημα αυτο λύνεται μερικώς με την τελική συνένωση των διαδοχικών περιοχών που ταξινομήθηκαν στην ίδια κλάση. Από την άλλη, είναι ανάγκη να μελετηθούν μέθοδοι αποθορυβοποίησης των ιστογραμμάτων αντικειμένων. Η εισαγωγή Κρυφών Μαρκοβιανών Μοντέλων (HMM) θα μπορούσε να συμβάλλει προς την κατεύθυνση αυτή.

Κλείνοντας, θα θέλαμε να επισημάνουμε ότι, καθώς η τεχνολογία εξελίσσεται ραγδαία, η παρούσα εργασία αναπόφευκτα θα ξεπεραστεί. Ευχόμαστε καλοπροαίρετα να γίνει αυτό, όμως ταυτόχρονα ελπίζουμε ότι η εργασία μας θα συνεισφέρει στην εξέλιξη κι ότι πάνω της μπορούν να βασιστούν νέες ερευνητικές πορείες που θα αποδειχθούν ισχυρότερες.

Παραρτήματα

Παράρτημα A

Το Σύνολο Δεδομένων Και Οι Τροποποιήσεις Στις Οποίες Προβήκαμε

A.1 MPII Cooking Activities Dataset [191]

Το σύνολο δεδομένων MPII Cooking Activities Dataset περιλαμβάνει 44 βίντεο μαγειρικής στα οποία εμφανίζονται 65 κατηγορίες δράσεων. Από αυτές, οι 64 αφορούν δράσεις σχετικές με τη μαγειρική άμεσα (π.χ. πλύσιμο και κοπή φρούτων) ή έμμεσα (π.χ. εξαγωγή υλικού από το ψυγείο). Μία επιπλέον κατηγορία είναι η κλάση δράσεων υποβάθρου (background activity), που περιλαμβάνει τα τμήματα του βίντεο στα οποία δε συμβαίνει κάποια δράση ενδιαφέροντος από τις υπόλοιπες 64. Ο πίνακας A.1 περιέχει το σύνολο των δράσεων οι οποίες εμφανίζονται στο σύνολο MPII Cooking Activities Dataset. Από αυτές τις κατηγορίες, κρίνουμε ότι οι δράσεις put on cutting-board, read και smell ελάχιστα ενδιαφέρουν τη ροή της δράσης, οπότε κατατάσσονται στις δράσεις υποβάθρου. Ταυτόχρονα, η δράση wash hands δεν αφορά τη ροή και δεν διακρίνεται αρκετά από την δράση wash objects. Οπότε συνενώνουμε τις δύο κατηγορίες σε μία, με ενιαίο τίτλο wash objects. Επομένως, χρησιμοποιούμε τις 61 από τις 65 δράσεις.

Τα 44 βίντεο ακολουθούν μια σύμβαση ονομασίας τύπου sSS-dDD, με τα SS και DD να αποτελούν διψήφιους αριθμούς που υποδεικνύουν το υποκείμενο, δράστη του βίντεο που εκτελεί τη συνταγή, (s από το subject) και το πιάτο-συνταγή (d από το dish) αντίστοιχα. Για παράδειγμα, έγκυρη ονομασία είναι η s19-d07, που υποδηλώνει ότι ο δράστης 19 εκτελεί τη συνταγή 7. Στο σύνολο των 44 βίντεο, εμφανίζονται διάφοροι συνδυασμοί υποκειμένων από το 8 ως το 20 και πιάτων από το 1 ως το 14. Είναι φανερό ότι το σύνολο δεδομένων περιέχει ένα υποσύνολο όλων των πιθανών συνδυασμών δραστών και συνταγών. Τα βίντεο που περιέχονται φαίνονται στον πίνακα A.2. Από αυτά, επιλέχθηκαν ως βίντεο εκπαίδευσης όσα αφορούν τους δράστες 11 ως 15. Επειδή κάποιες κατηγορίες δεν περιέχονταν στα βίντεο αυτά, προσθέσαμε στα βίντεο εκπαίδευσης επιπλέον τα s08-d02 και s10-d10. Όλα τα βίντεο είναι πυκνά δειγματοληπτημένα στα 29.4 frames ανά δευτερόλεπτο. Η συνολική διάρκεια των βίντεο,

Actions in MPII Cooking Activities Dataset

Background activity, change temperature, cut apart, cut dice, cut in, cut off ends, cut out inside, cut slices, cut stripes, dry, fill water from tap, grate, lid: put on, lid: remove, mix, move from X to Y, open egg, open tin, open/close cupboard, open/close drawer, open/close fridge, open/close oven, package X, peel, plug in/out, pour, pull out, puree, put in bowl, put in pan/pot, put on bread/dough, put on cutting-board, put on plate, read, remove from package, rip open, scratch off, screw close, screw open, shake, smell, spice, spread, squeeze, stamp, stir, strew, take & put in cupboard, take & put in drawer, take & put in fridge, take & put in oven, take & put in spice holder, take ingredient apart, take out from cupboard, take out from drawer, take out from fridge, take out from oven, take out from spice holder, taste, throw in garbage, unroll dough, wash hands, wash objects, whisk, wipe clean.

Πίνακας A.1: Οι 65 δράσεις του συνόλου δεδομένων MPII Cooking Activities Dataset. Στα πειράματά μας, οι δράσεις put on cutting-board, read και smell ενσωματώνονται στην κατηγορία background activity και η δράση wash hands συνενώνεται με τη δράση wash hands.

Videos in MPII Cooking Activities Dataset

s08-d02, s08-d04, s08-d11, s08-d14, s10-d02, s10-d10, s10-d11, s11-d01, s11-d06, s11-d11, s11-d12, s11-d13, s11-d14, s12-d05, s12-d07, s12-d09, s12-d10, s12-d14, s13-d08, s13-d09, s13-d11, s13-d12, s13-d13, s14-d08, s14-d09, s14-d11, s15-d03, s15-d07, s15-d14, s16-d01, s16-d06, s16-d09, s16-d11, s17-d02, s17-d05, s17-d13, s18-d11, s19-d01, s19-d06, s19-d07, s19-d09, s19-d10, s19-d12, s20-d07.

Πίνακας A.2: Τα 44 βίντεο του συνόλου δεδομένων MPII Cooking Activities Dataset. Από αυτά, χρησιμοποιούμε για εκπαίδευση όσα αφορούν τους δράστες 11 ως 15 και τα s08-d02 και s10-d10.

εκπαίδευσης και ελέγχου είναι περίπου 8 ώρες, επομένως ο αριθμός των συνολικών frames ξεπερνά τις 800000.

Στις επισημειώσεις των βίντεο παρέχονται τα όρια (πρώτο και τελευταίο frame) κάθε δράσης, ο κωδικός της δράσης (ένας ακέραιος από 1 ως 65) και το όνομα της δράσης. Επομένως τα βίντεο έχουν επισημειώσεις ιδανικές για αξιολόγηση αναγνώρισης και κατάτμησης δράσεων. Καθώς η αυθεντική εργασία [191] που εισήγαγε το σύνολο δεδομένων αξιοποιεί τη μέθοδο Πυκνών Τροχιών [248] για την αναπαράσταση των βίντεο, συνδυαζόμενη με χαρακτηριστικά πόζας [5], οι ερευνητές παραθέτουν τα υπολογισμένα χαρακτηριστικά (HOG, HOF, MBH, Περιγραφητής Τροχιάς, Μοντέλα Σώματος, FFT) για το σύνολο των βίντεο. Η σημασία και η φύση των χαρακτηριστικών αυτών είναι θέμα του κεφαλαίου 3. Τα χαρακτηριστικά παρέχονται σε κωδικοποιημένη μορφή με τη μέθοδο k-μέσων και k=4000, ενώ για ευκολία χειρισμού παρέχονται τα ολοκληρωτικά ιστογράμματα χαρακτηριστικών. Ο ενδιαφερόμενος μπορεί να βρει περισσότερες πληροφορίες και υλικό για το σύνολο MPII Cooking Activities Dataset στην τοποθεσία .

A.2 MPII Cooking 2 Dataset [192]

Το σύνολο MPII Cooking 2 Dataset αποτελεί επέκταση του MPII Cooking Activities Dataset και περιλαμβάνει 273 βίντεο μαγειρικής με περισσότερες δράσεις και επισημειώσεις. Η σύμβαση για την ονομασία των βίντεο παραμένει. Δεν θα αναφερθούμε

Objects in MPII Cooking 2 Dataset
apple, baking tray, blender, bottle, bowl, box grater, bread, bread knife, butter, carrot, cheese, chefs knife, chocolate, colander, corn, cucumber, cup, cupboard, dough, drawer, egg, eggshell, electricity column, electricity plug, flat grater, fridge, front peeler, frying pan, garbage, garlic clove, glass, ham, hot chocolate powder bag, jar, kiwi, knife, kohlrabi, ladle, lemon, lid, masher, measuring pitcher, milk, mushroom, oil, onion, orange, oven, paper, paper box, peach, pear, peel, pepper, pineapple, plastic bag, platic bottle, plastic box, plastic paper bag, plate, plum, pot, potato, puree, salad, salami, salt, seed, side peeler, sink, soup, spatula, spice, spice holder, spice shaker, sponge, sponge cloth, spoon, squeezer, stone, stove, tap, teaspoon, tin, tin opener, tomato, towel, water, wire whisk, wrapping paper, yolk, zucchini.

Πίνακας A.3: Οι 92 κατηγορίες αντικειμένων που εμφανίζονται στα βίντεο του συνόλου MPII Cooking Activities Dataset και που χρησιμοποιούμε στα πειράματά μας.

με λεπτομέρεια στα βίντεο και τις δράσεις του συνόλου καθώς δεν χρησιμοποιούνται αυτά καθαυτά τα δεδομένα στη δική μας εργασία. Θα εστιάσουμε στις επιπλέον επισημειώσεις τις οποίες αξιοποιούμε, τονίζοντας πρώτα ότι οι δράσεις και τα βίντεο του MPII Cooking Activities Dataset είναι υποσύνολα των δράσεων και των βίντεο του MPII Cooking 2 Dataset.

Το σύνολο MPII Cooking 2 Dataset περιέχει επισημειώσεις αρκετά πλουσιότερες αυτών του MPII Cooking Activities Dataset. Αρχικά, παρέχονται επισημειώσεις δράσεων για κάθε βίντεο στη μορφή: έναρξη (σε χρόνο και frames) και λήξη δράσης, είδος δράσης, δράστης και σύνολο αντικειμένων που πλαισιώνουν τη δράση. Τα αντικείμενα αυτά χωρίζονται επιπλέον σε εργαλεία και υλικά, αλλά η διάκριση αυτή δεν αφορά την εργασία μας. Με βάση τις επισημειώσεις αυτές, αποκτούμε έναν χρονικό χάρτη των αντικειμένων που εμφανίζονται στο βίντεο. Βέβαια υπογραμμίζουμε το ότι οι επισημειώσεις αυτές αφορούν τα αντικείμενα που χρησιμοποιούνται στην κάθε δράση και όχι αυτά που εμφανίζονται οπτικά στο βίντεο σε κάθε χρονική στιγμή. Τα αντικείμενα που εμφανίζονται κατατάσσονται σε 155 κατηγορίες, από τις οποίες μας ενδιαφέρουν μόνο οι 92, καθώς αυτές εμφανίζονται στα βίντεο του MPII Cooking Activities Dataset. Επιπλέον, στις 92 κατηγορίες δεν περιλαμβάνονται συγκεκριμένα ουσιαστικά όπως χέρι και πάγκος, καθώς δεν δίνουν ιδιαίτερη πληροφορία για τη δράση. Το σύνολο των ουσιαστικών που χρησιμοποιούνται στα πειράματά μας φαίνεται στον πίνακα A.3.

Σε παλιότερη εργασίαν των ερευνητών [184], είχε σχεδιαστεί σύστημα αυτόματης εξαγωγής λεκτικών περιγραφών για τα βίντεο της πρώτης έκδοσης του MPII Cooking 2 Dataset. Για την εγκυρότητα αυτών των περιγραφών, είχε συλλεχθεί από τους ερευνητές ένα σύνολο ανθρώπινων περιγραφών για τις δράσεις και τα αντικείμενα των βίντεο. Οι επισημειώσεις του MPII Cooking 2 Dataset ενσωματώνουν τις περιγραφές αυτές, προβαίνοντας σε εξαγωγή από κοινού εμφανίσεων δράσεων και αντικειμένων. Μια από κοινού εμφάνιση πραγματοποιείται όταν το ρήμα της δράσης και τα ουσιαστικά των αντικειμένων συνυπάρχουν στην ίδια πρόταση. Με βάση αυτές τις περιγραφές μπορούμε να εξάγουμε την πρότερη πιθανότητα εμφάνισης της δράσης α δεδομένης της εμφάνισης του αντικειμένου *object*, $P(\alpha|object)$. Η διαδικασία αυτή αναλύεται στο κεφάλαιο 6. Περισσότερες λεπτομέρειες και υλικό για το σύνολο MPII Cooking 2 Dataset μπορούν να βρεθούν στην τοποθεσία .

Start Time	End Time	Subtitle
714.7200	722.7210	Pour enough oil to fill the bottom of your frying-pan.
723.0950	726.9380	Spread it uniformly.
728.4170	730.6840	Ok, turn on the stove in this point.
730.7240	734.8660	Choose a high temperature for start.

Πίνακας A.4: Παράδειγμα υποτίτλων για το βίντεο s08-d04. Όλοι οι χρόνοι αναφέρονται σε δευτερόλεπτα από την αρχή του βίντεο.

A.3 Επιπλέον Επισημειώσεις Στις Οποίες Προβήκαμε

A.3.1 Υπότιτλοι

Για την αξιοποίηση της ομιλίας ως τρόπο έκφρασης επιπλέον πληροφορίας, προβαίνουμε σε ανάλυση υποτίτλων, ώστε να αποφύγουμε τα σφάλματα που θα εισήγαγε ένα σύστημα αυτόματης αναγνώρισης λόγου, καθώς αυτό ξεφεύγει από τον σκοπό αυτής της εργασίας. Επειδή δεν υπάρχουν υπότιτλοι για το MPII Cooking Activities Dataset, δημιουργήσαμε εμείς. Σε ένα σύνολο επισημειωτών δόθηκε το βίντεο και οι επισημειώσεις δράσεων και αντικειμένων από το MPII Cooking 2 Dataset. Επιβάλλαμε τον περιορισμό της χρήσης των ονομάτων των αντικειμένων απευθείας και όχι συνωνύμων τους, καθώς στη χρήση των υποτίτλων αναζητούμε μόνο τα ονόματα των αντικειμένων όπως αυτά παρέχονται από το MPII Cooking 2 Dataset. Όσον αφορά τους χρόνους ενός επιλόγου, επιβάλλαμε να εκτείνονται εντός της διάρκειας της δράσης και το χρονικό τους μήκος να περιορίζεται κάτω από τα 15 δευτερόλεπτα, ανάλογα πάντα με το μέγεθος του διαλόγου σε λέξεις. Προσπαθήσαμε να διατηρήσουμε τη φυσικότητα των υποτίτλων και ταυτόχρονα να διευκολύνουμε την ανίχνευση αντικειμένων μέσω αυτών. Στον πίνακα A.4 φαίνεται ένα παράδειγμα τμήματος των υποτίτλων για το βίντεο s08-d04.

A.3.2 Επισημειώσεις Αντικειμένων

Το σύνολο MPII Cooking 2 Dataset περιέχει επισημειώσεις για τα αντικείμενα που χρησιμοποιούνται ανά δράση. Ωστόσο, οπτική επισημείωση, για τη θέση δηλαδή του χρησιμοποιούμενου αντικειμένου στα frames, δεν υπάρχει. Προκειμένου να εντοπίσουμε οπτικά τα αντικείμενα στα frames του βίντεο, είναι φανερό ότι πρέπει να έχουμε δεδομένα για την εμφάνιση τους. Καθώς η μεταβλητότητα των αντικειμένων είναι εξαιρετικά μεγάλη (όπως αναλύουμε στο κεφάλαιο 4), δημιουργούμε μοντέλα αντικειμένων για κάθε βίντεο ξεχωριστά. Η δημιουργία επισημειώσεων έγινε σε δύο στάδια, τα οποία εξυπηρετούν διαφορετικούς σκοπούς.

Σε πρώτη φάση, δειγματοληπτούμε το βίντεο με περίοδο 100 frames και σε κάθε ένα σημείωνουμε τα αντικείμενα που εμφανίζονται. Η διαδικασία αυτή γίνεται με τη λειτουργία `trainingImageLabeler` του λογισμικού MATLAB. Με το λογισμικό αυτό, είναι εύκολο να δημιουργηθεί ένα ορθογώνιο που περιλαμβάνει την περιοχή ενδιαφέροντος. Ο χρήστης εισάγει τα ορθογώνια και τις ετικέτες και το πρόγραμμα εξάγει τις χωρικές συντεταγμένες πάνω στο frame. Ένα παράδειγμα φαίνεται στην παρακάτω εικόνα:



Σχήμα A.1: Παράδειγμα δημιουργίας επισημειώσεων με το λογισμικό trainingImageLabeler του MATLAB. Εισάγουμε τα ορθογώνια και την ετικέτα και το λογισμικό εξάγει τις συντεταγμένες. Η χρήση του πακέτου αυτου διευκολύνει τη μαζική δημιουργία θετικών δειγμάτων εκπαίδευσης για ανιχνευτές αντικειμένων.

Προσπαθούμε να σημειώνουμε τα αντικείμενα τα οποία μένουν σχετικά σταθερά εμφανισιακά σε αρκετά frames κι έτσι μπορούν να εντοπιστούν από έναν ανιχνευτή όπως ο [245]. Έτσι, επιλέγουμε να δημιουργούμε αντιπροσώπους κλάσεων με τμήματα αντικειμένων που παραμένουν σταθερά και εντός της περιοχής ενδιαφέροντος. Για παράδειγμα στην παραπάνω εικόνα, το πορτοκάλι που κείται δίπλα στον στίφτη μπορεί να εντοπιστεί. Εν τούτοις δεν είναι αυτό το πορτοκάλι που βρίσκεται εν χρήση αλλά ένα τμήμα του που μένει ακίνητο. Παρόμοια στρατηγική ακολουθείται για τις υπόλοιπες κατηγορίες αντικειμένων. Ως ένα ακόμα παράδειγμα, τα έπιπλα αναπαρίστανται με την εμφάνιση ή μη του τμήματός τους που φαίνεται όταν αυτά χρησιμοποιούνται (ανοίγουν). Στην παρακάτω εικόνα φαίνεται ένα τέτοιο παράδειγμα για την περίπτωση του ψυγείου.



Σχήμα Α.2: Παράδειγμα αναπαράστασης σταθερού αντικειμένου με το τμήμα του το οποίο εμφανίζεται μόνο όταν το αντικείμενο χρησιμοποιείται. Έτσι, εξασφαλίζεται ότι το αντικείμενο θα εντοπίζεται μόνο σε περιπτώσεις ενδιαφέροντος και όχι σε όλη τη διάρκεια του βίντεο.

Οι επισημειώσεις αυτές λειτουργούν ως θετικά δείγματα για εκπαίδευση ενός ανιχνευτή αντικειμένων. Λόγω της υψηλής μεταβλητότητας, επιλέξαμε τη χρήση Ταιριάσματος Προτύπων (Template Matching) ως ανιχνευτή αντικειμένων κι έτσι απαιτούνταν να συλλέξουμε πρότυπα. Έχοντας λοιπόν τις πλήρεις επισημειώσεις μέσω του trainingImageLabeler, επιλέξαμε αυτές που εμφανίζοταν συχνά και για μεγάλη χρονική διάρκεια ως αντιτροσώπους κλάσεων αντικειμένων. Πρακτικά, συρρικνώνουμε τα θετικά δείγματα έτσι ώστε να ελαχιστοποιείται η απώλεια που αυτό συνεπάγεται στην μεταβλητότητα και στον ικανοποιητικό εντοπισμό. Στο κεφάλαιο 4 αναλύουμε περισσότερα για τη μέθοδο εντοπισμού των αντικειμένων.

A.3.3 Επισημειώσεις Χεριών

Στο κεφάλαιο 5 δείχνουμε τον εντοπισμό χεριών και την εξαγωγή του τύπου λαβής (grasping type). Στο κεφάλαιο 6, τα αποτελέσματα φαίνονται να υποβοηθούνται από αυτή την πληροφορία. Εν τούτοις, σύνολο δεδομένων δεν υπήρχε ούτε σχετικά με τον εντοπισμό χεριών, ούτε σχετικά με την εξαγωγή τύπου λαβής, που να σχετίζονται άμεσα με το MPII Cooking Activities Dataset. Επομένως, έπρεπε να δημιουργήσουμε, με βάση τις διαθέσιμες πληροφορίες, μια βάση για τον εντοπισμό χεριών αρχικά.

Παρότι υπάρχουν διάφοροι επιτυχείς αλγόριθμοι και σύνολα δεδομένων [160] για ανιχνευση χεριών, ακολουθούμε την πρόταση του [192] και επανεκπαιδεύουμε έναν ανιχνευτή χεριών με εικόνες που επιλέγονται αραιά από το σύνολο δεδομένων MPII Cooking Activities Dataset. Για τις εικόνες εκπαίδευσης, συνενώνουμε τα σύνολα δεδομένων πόζας [4] και [191]. Οι επισημειώσεις των συνόλων αυτών είναι σύνολα σημείων που αναπαριστούν τις θέσεις διαφόρων συνδέσμων του άνω τμήματος του σώματος. Δύο από αυτές τις επισημειώσεις αφορούν το αριστερό και το δεξί χέρι. Ωστόσο, οι επισημειώσεις δεν εξασφαλίζουν ότι τα χέρια φαίνονται στην εικόνα,



Σχήμα A.3: Παραδείγματα επισημειώσεων πόζας. Με μπλε απεικονίζονται τα σημεία που χαρακτηρίζονται ως χέρια μέσω των επισημειώσεων πόζας. Δείχνουμε τις περιοχές ως τετράγωνα 10×10 γύρω από τα σημεία επισημείωσης για να είναι εμφανείς οπτικά στον αναγνώστη. Βλέπουμε ότι ενώ στη δεξιά εικόνα οι επισημειώσεις ταυτίζονται με τη θέση των χεριών, τα οποία είναι εμφανή οπτικά, στην εικόνα αριστερά οι επισημειώσεις δείχνουν τη θέση που βρίσκεται το χέρι πίσω από το σώμα του ανθρώπου. Είναι φανερό ότι εικόνες σαν την αριστερή δεν μπορούν να χρησιμεύσουν σαν θετικά δείγματα για έναν ανιχνευτή χεριών και έτσι αποκλείονται από τα θετικά δεδομένα εκπαίδευσης και χρησιμοποιούνται μόνο για την εξαγωγή αρνητικών παραδειγμάτων.

παρά μόνο δείχνουν το πού αυτά βρίσκονται. Δύο παραδείγματα φαίνονται στην εικόνα A.3, όπου στην δεύτερη περίπτωση βλέπουμε μια φυσιολογική απεικόνιση των χεριών, ενώ στην πρώτη, βλέπουμε σημεία να απεικονίζουν τη θέση των χεριών πίσω από το υπόλοιπο σώμα του ανθρώπου. Οπότε σε πρώτη φάση, σκανάραμε το σύνολο αυτών των εικόνων και κρατήσαμε μόνο αυτές για τις οποίες φαίνονταν τα χέρια. Για εικόνες που φαινόταν μόνο το ένα χέρι, κρατήσαμε ετικέτες ώστε στη φάση εξαγωγής θετικών δειγμάτων να λάβουμε την εικόνα μόνο του χεριού που εμφανίζεται. Σε δεύτερη φάση, εξαγάγαμε ορθογώνια που περιέκλειαν τα χέρια. Πειραματικά διαπιστώσαμε ότι κουτιά σταθερού μεγέθους 31×31 είναι επαρκή για τη συμπερίληψη του χεριού ανεξαρτήτως της απόστασης από την κάμερα. Τέλος, λάβαμε αρνητικά δείγματα κυλώντας ένα παράθυρο με βήμα 5 σε οριζόντιο και κάθετο άξονα ανά εικόνα και λαμβάνοντας έτσι δύο τυχαία δείγματα μη χεριών από κάθε εικόνα. Αν 6 εικόνες, χρησιμοποήσαμε το άνω τμήμα του κεφαλιού ως επιπλέον αρνητικό δείγμα, ενώ επίσης ανά 6, αλλά με άλλη αρχή μέτρησης, χρησιμοποήσαμε το κάτω μέρος του κεφαλιού για αρνητικό δείγμα. Η διαδικασία που ακολουθεί περιγράφεται στο κεφάλαιο 5.

Η έλλειψη επισημειώσεων για τον τύπο λαβής μας ανάγκασε να χρησιμοποιήσουμε μη επιβλεπόμενη μάθησης, όπως περιγράφεται στο κεφάλαιο 5. Επομένως, δημιουργούμε ένα σύνολο δεδομένων τύπων λαβής με αυτόματο τρόπο, λύνοντας το πρόβλημα απουσίας επισημειώσεων και δεν προβαίνουμε σε επιπλέον διαδικασίες επισημειώσεων.

Βιβλιογραφία

- [1] “A New Approach to Linear Filtering and Prediction Problems”. Στο: *Control Theory* (2009). doi: [10.1109/9780470544334.ch9](https://doi.org/10.1109/9780470544334.ch9).
- [2] Marc Al-Hames κ.ά. “Multimodal Integration for Meeting Group Action Segmentation and Recognition”. Στο: *Machine Learning for Multimodal Interaction, Lecture Notes in Computer Science* (2006), σσ. 52–63. doi: [10.1007/11677482_5](https://doi.org/10.1007/11677482_5).
- [3] Jean-Baptiste Alayrac κ.ά. “Unsupervised Learning from Narrated Instruction Videos”. Στο: *Proceedings of the 2016 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2016* (2016). doi: [10.1109/cvpr.2016.495](https://doi.org/10.1109/cvpr.2016.495).
- [4] Sikandar Amin κ.ά. “Multi-view Pictorial Structures for 3D Human Pose Estimation”. Στο: *Proceedings of the 2013 British Machine Vision Conference* (Ιαν. 2013), σσ. 45.1–45.11. doi: [10.5244/C.27.45](https://doi.org/10.5244/C.27.45).
- [5] M. Andriluka, S. Roth και B. Schiele. “Pictorial structures revisited: People detection and articulated pose estimation”. Στο: *Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2009* (2009). doi: [10.1109/cvprw.2009.5206754](https://doi.org/10.1109/cvprw.2009.5206754).
- [6] S. Avidan. “Support Vector Tracking”. Στο: *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001* (2001). doi: [10.1109/cvpr.2001.990474](https://doi.org/10.1109/cvpr.2001.990474).
- [7] Mohammadreza Babaee, Duc Tung Dinh και Gerhard Rigoll. “A Deep Convolutional Neural Network for Video Sequence Background Subtraction”. Στο: *Pattern Recognition* (2017). doi: [10.1016/j.patcog.2017.09.040](https://doi.org/10.1016/j.patcog.2017.09.040).
- [8] Ava Bargi, Richard Yi Da Xu και Massimo Piccardi. “An online HDP-HMM for joint action segmentation and classification in motion capture data”. Στο: *Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. CVPR 2012* (2012). doi: [10.1109/cvprw.2012.6239230](https://doi.org/10.1109/cvprw.2012.6239230).
- [9] Olivier Barnich και Marc Van Droogenbroeck. “ViBE: A powerful random technique to estimate the background in video sequences”. Στο: *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP 2009* (2009). doi: [10.1109/icassp.2009.4959741](https://doi.org/10.1109/icassp.2009.4959741).
- [10] Sven Behnke. “Hierarchical Neural Networks for Image Interpretation”. Στο: *Lecture Notes in Computer Science* (2003). doi: [10.1007/b11963](https://doi.org/10.1007/b11963).

- [11] P.n. Belhumeur, J.p. Hespanha και D.j. Kriegman. “Eigenfaces vs. Fisherfaces: recognition using class specific linear projection”. Στο: *Proceedings of the 1997 IEEE Transactions on Pattern Analysis and Machine Intelligence* 19.7 (1997), σσ. 711–720. doi: [10.1109/34.598228](https://doi.org/10.1109/34.598228).
- [12] Marcelo Bertalmio, Guillermo Sapiro και Gregory Randall. “Morphing Active Contours”. Στο: *Scale-Space Theories in Computer Vision, Lecture Notes in Computer Science* (1999), 46–57. doi: [10.1007/3-540-48236-9_5](https://doi.org/10.1007/3-540-48236-9_5).
- [13] David Beymer και Kurt Konolige. “Real-Time Tracking of Multiple People Using Continuous Detection”. Στο: *Proceedings of the 1999 IEEE International Workshop on Intelligent Environments*. 1999.
- [14] Irving Biederman. “Recognition-by-components: A theory of human image understanding.” Στο: *Psychological Review* 94.2 (1987), σσ. 115–117. doi: [10.1037/0033-295x.94.2.115](https://doi.org/10.1037/0033-295x.94.2.115).
- [15] S. Birchfield. “Elliptical head tracking using intensity gradients and color histograms”. Στο: *Proceedings of the 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 1998* (1998). doi: [10.1109/cvpr.1998.698614](https://doi.org/10.1109/cvpr.1998.698614).
- [16] Michael J. Black και Allan D. Jepson. “EigenTracking: Robust matching and tracking of articulated objects using a view-based representation”. Στο: *Computer Vision – ECCV 1996, Lecture Notes in Computer Science* (1996), 329–342. doi: [10.1007/bfb0015548](https://doi.org/10.1007/bfb0015548).
- [17] Aaron F. Bobick και James W. Davis. “The recognition of human movement using temporal templates”. Στο: *Proceedings of the 2001 IEEE Transactions on Pattern Analysis and Machine Intelligence* 23.3 (2001), σσ. 257–267.
- [18] A.f. Bobick και Y.a. Ivanov. “Action recognition using probabilistic parsing”. Στο: *Proceedings of the 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 1998* (1998). doi: [10.1109/cvpr.1998.698609](https://doi.org/10.1109/cvpr.1998.698609).
- [19] P. Bojanowski κ.ά. “Weakly-Supervised Alignment of Video with Text”. Στο: *Proceedings of the 2015 IEEE International Conference on Computer Vision. ICCV 2015* (2015). doi: [10.1109/iccv.2015.507](https://doi.org/10.1109/iccv.2015.507).
- [20] Piotr Bojanowski κ.ά. “Weakly Supervised Action Labeling in Videos under Ordering Constraints”. Στο: *Computer Vision – ECCV 2014, Lecture Notes in Computer Science* (2014), 628–643. doi: [10.1007/978-3-319-10602-1_41](https://doi.org/10.1007/978-3-319-10602-1_41).
- [21] Thierry Bouwmans, Fida El Baf και Bertrand Vachon. “Background Modeling using Mixture of Gaussians for Foreground Detection - A Survey”. Στο: *Recent Patents on Computer Science* 1.3 (2010), 219–237. doi: [10.2174/1874479610801030219](https://doi.org/10.2174/1874479610801030219).
- [22] Thierry Bouwmans και El Hadi Zahzah. “Robust PCA via Principal Component Pursuit: A review for a comparative evaluation in video surveillance”. Στο: *Computer Vision and Image Understanding* 122 (2014), 22–34. doi: [10.1016/j.cviu.2013.11.009](https://doi.org/10.1016/j.cviu.2013.11.009).
- [23] Marc Braham και Marc Van Droogenbroeck. “Deep background subtraction with scene-specific convolutional neural networks”. Στο: *Proceedings of the 2016 International Conference on Systems, Signals and Image Processing. IWSSIP 2016* (2016). doi: [10.1109/iwssip.2016.7502717](https://doi.org/10.1109/iwssip.2016.7502717).

- [24] Ted J. Broida και Rama Chellappa. “Estimation of Object Motion Parameters from Noisy Images”. Στο: *Proceedings of the 1986 IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-8.1* (1986), 90–99. doi: [10.1109/tpami.1986.4767755](https://doi.org/10.1109/tpami.1986.4767755).
- [25] M. Brown και D. Lowe. “Invariant Features from Interest Point Groups”. Στο: *Proceedings of the 2002 British Machine Vision Conference* (2002). doi: [10.5244/c.16.23](https://doi.org/10.5244/c.16.23).
- [26] Johanna Carvajal κ.ά. “Joint Recognition and Segmentation of Actions via Probabilistic Integration of Spatio-Temporal Fisher Vectors”. Στο: *Trends and Applications in Knowledge Discovery and Data Mining, Lecture Notes in Computer Science* (2016), σσ. 115–127. doi: [10.1007/978-3-319-42996-0_10](https://doi.org/10.1007/978-3-319-42996-0_10).
- [27] Tat-Jen Cham και J.m. Rehg. “A multiple hypothesis approach to figure tracking”. Στο: *Proceedings of the 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 1999* (1999). doi: [10.1109/cvpr.1999.784636](https://doi.org/10.1109/cvpr.1999.784636).
- [28] Y.-L. Chang και J.k. Aggarwal. “3D structure reconstruction from an ego motion sequence using statistical estimation and detection theory”. Στο: *Proceedings of the 1991 IEEE Workshop on Visual Motion* (1991). doi: [10.1109/wvm.1991.212797](https://doi.org/10.1109/wvm.1991.212797).
- [29] Chen Chen, Kui Liu και Nasser Kehtarnavaz. “Real-time human action recognition based on depth motion maps”. Στο: *Journal of Real-Time Image Processing* 12.1 (2013), σσ. 155–163. doi: [10.1007/s11554-013-0370-1](https://doi.org/10.1007/s11554-013-0370-1).
- [30] Danqi Chen και Christopher Manning. “A Fast and Accurate Dependency Parser using Neural Networks”. Στο: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. EMNLP 2014* (2014). doi: [10.3115/v1/d14-1082](https://doi.org/10.3115/v1/d14-1082).
- [31] Mingmin Chen Chenand Zhang, Kaijia Qiu και Zhigeng Pan. “Real-Time Robust Hand Tracking Based on Camshift and Motion Velocity”. Στο: *Proceedings of the 2014 International Conference on Digital Home* (2014). doi: [10.1109/icdh.2014.11](https://doi.org/10.1109/icdh.2014.11).
- [32] Yunqiang Chen, Yong Rui και T.s. Huang. “JPDAF based HMM for real-time contour tracking”. Στο: *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001* (2001). doi: [10.1109/cvpr.2001.990521](https://doi.org/10.1109/cvpr.2001.990521).
- [33] Ming-Ming Cheng κ.ά. “BING: Binarized Normed Gradients for Objectness Estimation at 300fps”. Στο: *Proceedings of the 2014 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2014* (2014). doi: [10.1109/cvpr.2014.414](https://doi.org/10.1109/cvpr.2014.414).
- [34] Anoop Cherian, Piotr Koniusz και Stephen Gould. “Higher-Order Pooling of CNN Features via Kernel Linearization for Action Recognition”. Στο: *Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision. WACV 2017* (2017). doi: [10.1109/wacv.2017.22](https://doi.org/10.1109/wacv.2017.22).
- [35] Anoop Cherian κ.ά. “Generalized Rank Pooling for Activity Recognition”. Στο: *CoRR* abs/1704.02112 (2017).
- [36] Guilhem Cheron, Ivan Laptev και Cordelia Schmid. “P-CNN: Pose-Based CNN Features for Action Recognition”. Στο: *Proceedings of the 2015 IEEE International Conference on Computer Vision. ICCV 2015* (2015). doi: [10.1109/iccv.2015.368](https://doi.org/10.1109/iccv.2015.368).

- [37] P. Chiu κ.ά. “A genetic algorithm for video segmentation and summarization”. Στο: *Proceedings of the 2000 IEEE International Conference on Multimedia and Expo. ICME 2000* (2000). doi: [10.1109/icme.2000.871011](https://doi.org/10.1109/icme.2000.871011).
- [38] Dan C. Cireşan κ.ά. “Flexible, High Performance Convolutional Neural Networks for Image Classification”. Στο: *Proceedings of the 2011 International Joint Conference on Artificial Intelligence - Volume Two*. IJCAI'11. AAAI Press, 2011, σσ. 1237–1242.
- [39] D. Ciresan, U. Meier και J. Schmidhuber. “Multi-column deep neural networks for image classification”. Στο: *Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2012* (2012). doi: [10.1109/cvpr.2012.6248110](https://doi.org/10.1109/cvpr.2012.6248110).
- [40] D. Comaniciu και P. Meer. “Mean shift: a robust approach toward feature space analysis”. Στο: *Proceedings of the 2002 IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.5 (2002), 603–619. doi: [10.1109/tpami.2002.1000236](https://doi.org/10.1109/tpami.2002.1000236).
- [41] D. Comaniciu, V. Ramesh και P. Meer. “Real-time tracking of non-rigid objects using mean shift”. Στο: *Proceedings of the 2000 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2000* (2000). doi: [10.1109/cvpr.2000.854761](https://doi.org/10.1109/cvpr.2000.854761).
- [42] Dorin Comaniciu, Visvanathan Ramesh και Peter Meer. “Kernel-Based Object Tracking”. Στο: *Proceedings of the 2003 IEEE Transactions on Pattern Analysis and Machine Intelligence* 25.5 (Μάι. 2003), σσ. 564–575.
- [43] Corinna Cortes και Vladimir Vapnik. “Support-vector networks”. Στο: *Machine Learning* 20.3 (1995), 273–297. doi: [10.1007/bf00994018](https://doi.org/10.1007/bf00994018).
- [44] I.j. Cox και S.l. Hingorani. “An efficient implementation and evaluation of Reids multiple hypothesis tracking algorithm for visual tracking”. Στο: *Proceedings of the 1996 International Conference on Pattern Recognition. ICPR 1996* (1996). doi: [10.1109/icpr.1994.576318](https://doi.org/10.1109/icpr.1994.576318).
- [45] Daniel Cremers και Christoph Schnörr. “Statistical shape knowledge in variational motion segmentation”. Στο: *Image and Vision Computing* 21.1 (2003), 77–86. doi: [10.1016/s0262-8856\(02\)00128-2](https://doi.org/10.1016/s0262-8856(02)00128-2).
- [46] G. Csurka κ.ά. “Visual categorization with bags of keypoints”. Στο: *Proceedings of the 2004 European Conference on Computer Vision Workshop on Statistical Learning in Computer Vision* (2004), σσ. 1–22.
- [47] *CVPR 2011 Tutorial on Human Activity Recognition*. 2011.
- [48] *CVPR 2014 Tutorial on Emerging Topics in Human Activity Recognition*. 2014.
- [49] Navneet Dalal και Bill Triggs. “Histograms of Oriented Gradients for Human Detection”. Στο: *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2005*. 2005.
- [50] Navneet Dalal, Bill Triggs και Cordelia Schmid. “Human Detection Using Oriented Histograms of Flow and Appearance”. Στο: *Proceedings of the 2006 European Conference on Computer Vision - Volume Part II*. Berlin, Heidelberg: Springer-Verlag, 2006, σσ. 428–441.
- [51] Buchsbaum Daphna, Canini Kevin και Griffiths Thomas. “Segmenting and Recognizing Human Action using Low-level Video Features”. Στο: *Proceedings of the Cognitive Science Society 2011*. Τόμ. 33. 2011.

- [52] Ahmad Yahya Dawod, Md Jan Nordin και Junaidi Abdullah. "Static Hand Gestures: Fingertips Detection Based on Segmented Images". Στο: *Journal of Computer Science* 11.12 (2015), σσ. 1090–1098. doi: [10.3844/jcssp.2015.1090.1098](https://doi.org/10.3844/jcssp.2015.1090.1098).
- [53] Thomas Dean κ.ά. "Fast, Accurate Detection of 100,000 Object Classes on a Single Machine". Στο: *Proceedings of the 2013 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2013* (2013). doi: [10.1109/cvpr.2013.237](https://doi.org/10.1109/cvpr.2013.237).
- [54] Damien Delannay, Nicolas Danhier και Christophe De Vleeschouwer. "Detection and recognition of sports(wo)men from multiple views". Στο: *Proceedings of the 2009 ACM/IEEE International Conference on Distributed Smart Cameras. ICDSC 2009* (2009). doi: [10.1109/icdsc.2009.5289407](https://doi.org/10.1109/icdsc.2009.5289407).
- [55] Barga Deori και Dalton Meitei Thounaojam. "A Survey on Moving Object Tracking in Video". Στο: *International Journal on Information Theory* 3.3 (2014), 31–46. doi: [10.5121/ijit.2014.3304](https://doi.org/10.5121/ijit.2014.3304).
- [56] *Dictionary.com*.
- [57] Li Ding και Chenliang Xu. "TricorNet: A Hybrid Temporal Convolutional and Recurrent Network for Video Action Segmentation". Στο: *CoRR* abs/1705.07818 (2017).
- [58] P. Dollar κ.ά. "Behavior Recognition via Sparse Spatio-Temporal Features". Στο: *Proceedings of the 2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance* (2005). doi: [10.1109/vspets.2005.1570899](https://doi.org/10.1109/vspets.2005.1570899).
- [59] Piotr Dollar, Serge Belongie και Pietro Perona. "The Fastest Pedestrian Detector in the West". Στο: *Proceedings of the 2010 British Machine Vision Conference* (2010). doi: [10.5244/c.24.68](https://doi.org/10.5244/c.24.68).
- [60] Jeff Donahue κ.ά. "Long-term recurrent convolutional networks for visual recognition and description". Στο: *Proceedings of the 2015 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2015* (2015). doi: [10.1109/cvpr.2015.7298878](https://doi.org/10.1109/cvpr.2015.7298878).
- [61] Olivier Duchenne κ.ά. "Automatic annotation of human actions in video". Στο: *Proceedings of the 2009 IEEE International Conference on Computer Vision. ICCV 2009* (2009). doi: [10.1109/iccv.2009.5459279](https://doi.org/10.1109/iccv.2009.5459279).
- [62] Richard O. Duda και Peter E. Hart. "Use of the Hough transformation to detect lines and curves in pictures". Στο: *Communications of the ACM* 15.1 (1972), σσ. 11–15. doi: [10.1145/361237.361242](https://doi.org/10.1145/361237.361242).
- [63] Efros κ.ά. "Recognizing action at a distance". Στο: *Proceedings of the 2003 IEEE International Conference on Computer Vision. ICCV 2003* (2003). doi: [10.1109/iccv.2003.1238420](https://doi.org/10.1109/iccv.2003.1238420).
- [64] Michael W. Eysenck και Mark T. Keane. *Cognitive psychology: a students handbook*. Psychology Press, 2010.
- [65] Gunnar Farnebäck. "Two-frame Motion Estimation Based on Polynomial Expansion". Στο: *Proceedings of the 2003 Scandinavian Conference on Image Analysis. SCIA'03*. Berlin, Heidelberg: Springer-Verlag, 2003, σσ. 363–370.
- [66] O.d. Faugeras και M. Hebert. "The Representation, Recognition, and Locating of 3-D Objects". Στο: *The International Journal of Robotics Research* 5.3 (1986), σσ. 27–52. doi: [10.1177/027836498600500302](https://doi.org/10.1177/027836498600500302).

- [67] Thomas Feix κ.ά. “A Metric for Comparing the Anthropomorphic Motion Capability of Artificial Hands”. Στο: *Proceedings of the 2013 IEEE Transactions on Robotics* 29.1 (2013), 82–93. doi: [10.1109/tro.2012.2217675](https://doi.org/10.1109/tro.2012.2217675).
- [68] P F Felzenszwalb κ.ά. “Object Detection with Discriminatively Trained Part-Based Models”. Στο: *Proceedings of the 2010 IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.9 (2010), σσ. 1627–1645. doi: [10.1109/tpami.2009.167](https://doi.org/10.1109/tpami.2009.167).
- [69] Pedro Felzenszwalb, David Mcallester και Deva Ramanan. “A discriminatively trained, multiscale, deformable part model”. Στο: *Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2008* (2008). doi: [10.1109/cvpr.2008.4587597](https://doi.org/10.1109/cvpr.2008.4587597).
- [70] Pedro F. Felzenszwalb, Ross B. Girshick και David Mcallester. “Cascade object detection with deformable part models”. Στο: *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2010* (2010). doi: [10.1109/cvpr.2010.5539906](https://doi.org/10.1109/cvpr.2010.5539906).
- [71] Pedro F. Felzenszwalb και Daniel P. Huttenlocher. “Pictorial Structures for Object Recognition”. Στο: *International Journal of Computer Vision* 61.1 (2005), 55–79. doi: [10.1023/b:visi.0000042934.15159.49](https://doi.org/10.1023/b:visi.0000042934.15159.49).
- [72] R. Fergus, P. Perona και A. Zisserman. “Object class recognition by unsupervised scale-invariant learning”. Στο: *Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2003* (2003). doi: [10.1109/cvpr.2003.1211479](https://doi.org/10.1109/cvpr.2003.1211479).
- [73] Basura Fernando κ.ά. “Modeling video evolution for action recognition”. Στο: *Proceedings of the 2015 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2015* (2015). doi: [10.1109/cvpr.2015.7299176](https://doi.org/10.1109/cvpr.2015.7299176).
- [74] Vittorio Ferrari, Tinne Tuytelaars και Luc Van Gool. “Object Detection by Contour Segment Networks”. Στο: *Computer Vision - ECCV 2006, Lecture Notes in Computer Science* (2006), σσ. 14–28. doi: [10.1007/11744078_2](https://doi.org/10.1007/11744078_2).
- [75] P. Fieguth και D. Terzopoulos. “Color-based tracking of heads and other mobile objects at video frame rates”. Στο: *Proceedings of the 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 1997* (1997). doi: [10.1109/cvpr.1997.609292](https://doi.org/10.1109/cvpr.1997.609292).
- [76] Ivan Fratric και Slobodan Ribaric. “Real-Time Model-Based Hand Localization for Unsupervised Palmar Image Acquisition”. Στο: *Advances in Biometrics, Lecture Notes in Computer Science* (2009), σσ. 1280–1289. doi: [10.1007/978-3-642-01793-3_129](https://doi.org/10.1007/978-3-642-01793-3_129).
- [77] Yoav Freund και Robert E Schapire. “A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting”. Στο: *Journal of Computer and System Sciences* 55.1 (1997), 119–139. doi: [10.1006/jcss.1997.1504](https://doi.org/10.1006/jcss.1997.1504).
- [78] Kunihiko Fukushima. “Cognitron: A self-organizing multilayered neural network”. Στο: *Biological Cybernetics* 20.3-4 (1975), 121–136. doi: [10.1007/bf00342633](https://doi.org/10.1007/bf00342633).
- [79] Kunihiko Fukushima. “Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position”. Στο: 36 (Φεβ. 1980), σσ. 193–202.
- [80] D.m. Gavrila και L.s. Davis. “3-D model-based tracking of humans in action: a multi-view approach”. Στο: *Proceedings of the 1996 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 1996* (1996). doi: [10.1109/cvpr.1996.517056](https://doi.org/10.1109/cvpr.1996.517056).

- [81] Ross Girshick. "Fast R-CNN". Στο: *Proceedings of the 2015 IEEE International Conference on Computer Vision. ICCV 2015* (2015). doi: [10.1109/iccv.2015.169](https://doi.org/10.1109/iccv.2015.169).
- [82] Dian Gong κ.ά. "Kernelized Temporal Cut for Online Temporal Segmentation and Recognition". Στο: *Computer Vision - ECCV 2012, Lecture Notes in Computer Science* (2012), σσ. 229–243. doi: [10.1007/978-3-642-33712-3_17](https://doi.org/10.1007/978-3-642-33712-3_17).
- [83] W. Eric L. Grimson και Tomas Lozano-Perez. "Localizing Overlapping Parts by Searching the Interpretation Tree". Στο: *Proceedings of the 1987 IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-9.4* (1987), σσ. 469–482. doi: [10.1109/tpami.1987.4767935](https://doi.org/10.1109/tpami.1987.4767935).
- [84] Tomasz Grzejszczak, Michal Kawulok και Adam Galuszka. "Hand landmarks detection and localization in color images". Στο: *Multimedia Tools and Applications* 75.23 (2015), σσ. 16363–16387. doi: [10.1007/s11042-015-2934-5](https://doi.org/10.1007/s11042-015-2934-5).
- [85] Jiaming Guo κ.ά. "Video Co-segmentation for Meaningful Action Extraction". Στο: *Proceedings of the 2013 IEEE International Conference on Computer Vision. ICCV 2013* (2013). doi: [10.1109/iccv.2013.278](https://doi.org/10.1109/iccv.2013.278).
- [86] Robert A Haaf κ.ά. "Object recognition and attention to object components by preschool children and 4-month-old infants". Στο: *Journal of Experimental Child Psychology* 86.2 (2003), σσ. 108 –123.
- [87] I. Haritaoglu, D. Harwood και L.s. Davis. "W4: Real-time surveillance of people and their activities". Στο: *Proceedings of the 2000 IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.8 (2000), 809–830. doi: [10.1109/34.868683](https://doi.org/10.1109/34.868683).
- [88] Chris Harris και Mike Stephens. "A combined corner and edge detector". Στο: *Proceedings of the 1988 Alvey Vision Conference*. 1988, σσ. 147–151.
- [89] Kaiming He κ.ά. "Deep Residual Learning for Image Recognition". Στο: *Proceedings of the 2016 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2016* (2016). doi: [10.1109/cvpr.2016.90](https://doi.org/10.1109/cvpr.2016.90).
- [90] Yutaka Hirano κ.ά. "Industry and Object Recognition: Applications, Applied Research and Challenges". Στο: *Toward Category-Level Object Recognition, Lecture Notes in Computer Science* (2006), σσ. 49–64. doi: [10.1007/11957959_3](https://doi.org/10.1007/11957959_3).
- [91] Tin Kam Ho. "Random decision forests". Στο: *Proceedings of the 1995 International Conference on Document Analysis and Recognition* (1995). doi: [10.1109/icdar.1995.598994](https://doi.org/10.1109/icdar.1995.598994).
- [92] Minh Hoai, Zhen-Zhong Lan και Fernando De La Torre. "Joint segmentation and classification of human actions in video". Στο: *Proceedings of the 2011 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2011* (2011). doi: [10.1109/cvpr.2011.5995470](https://doi.org/10.1109/cvpr.2011.5995470).
- [93] Minh Hoai και Andrew Zisserman. "Improving Human Action Recognition Using Score Distribution and Ranking". Στο: *Computer Vision - ACCV 2014, Lecture Notes in Computer Science* (2014), 3–20. doi: [10.1007/978-3-319-16814-2_1](https://doi.org/10.1007/978-3-319-16814-2_1).
- [94] Martin Hofmann, Philipp Tiefenbacher και Gerhard Rigoll. "Background segmentation with feedback: The Pixel-Based Adaptive Segmenter". Στο: *Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. CVPR 2012* (2012). doi: [10.1109/cvprw.2012.6238925](https://doi.org/10.1109/cvprw.2012.6238925).

- [95] De-An Huang, Li Fei-Fei και Juan Carlos Niebles. “Connectionist Temporal Modeling for Weakly Supervised Action Labeling”. Στο: *Computer Vision – ECCV 2016, Lecture Notes in Computer Science* (2016), 137–153. doi: [10.1007/978-3-319-46493-0_9](https://doi.org/10.1007/978-3-319-46493-0_9).
- [96] Jonathan Huang κ.ά. “Speed/accuracy trade-offs for modern convolutional object detectors”. Στο: *CoRR* abs/1611.10012 (2016).
- [97] D. H. Hubel και T. N. Wiesel. “Receptive fields, binocular interaction and functional architecture in the cat's visual cortex”. Στο: *The Journal of Physiology* 160.1 (1962), 106–154. doi: [10.1113/jphysiol.1962.sp006837](https://doi.org/10.1113/jphysiol.1962.sp006837).
- [98] C. Hue, J.-P. Le Cadre και P. Perez. “Sequential Monte Carlo methods for multiple target tracking and data fusion”. Στο: *Proceedings of the 2002 IEEE Transactions on Signal Processing* 50.2 (2002), 309–325. doi: [10.1109/78.978386](https://doi.org/10.1109/78.978386).
- [99] Daniel P. Huttenlocher και Shimon Ullman. “Recognizing solid objects by alignment with an image”. Στο: *International Journal of Computer Vision* 5.2 (1990), σσ. 195–212.
- [100] D.p. Huttenlocher, J.j. Noh και W.j. Rucklidge. “Tracking non-rigid objects in complex scenes”. Στο: *Proceedings of the 1993 IEEE International Conference on Computer Vision. ICCV 1993* (1993). doi: [10.1109/iccv.1993.378231](https://doi.org/10.1109/iccv.1993.378231).
- [101] S.s. Intille, J.w. Davis και A.f. Bobick. “Real-time closed-world tracking”. Στο: *Proceedings of the 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 1997* (1997). doi: [10.1109/cvpr.1997.609402](https://doi.org/10.1109/cvpr.1997.609402).
- [102] M. Isard και J. MacCormick. “BraMBLe: a Bayesian multiple-blob tracker”. Στο: *Proceedings of the 2001 IEEE International Conference on Computer Vision. ICCV 2001* (2001). doi: [10.1109/iccv.2001.937594](https://doi.org/10.1109/iccv.2001.937594).
- [103] Michael Isard και Andrew Blake. “CONDENSATION: Conditional Density Propagation for Visual Tracking”. Στο: *International Journal of Computer Vision* 29.1 (Αύγ. 1998), σσ. 5–28.
- [104] Arpit Jain κ.ά. “Representing Videos Using Mid-level Discriminative Patches”. Στο: *Proceedings of the 2013 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2013* (2013). doi: [10.1109/cvpr.2013.332](https://doi.org/10.1109/cvpr.2013.332).
- [105] Mihir Jain, Jan C. Van Gemert και Cees G. M. Snoek. “What do 15,000 object categories tell us about classifying and localizing actions?” Στο: *Proceedings of the 2015 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2015* (2015). doi: [10.1109/cvpr.2015.7298599](https://doi.org/10.1109/cvpr.2015.7298599).
- [106] M. Jeannerod. “The Timing of Natural Prehension Movements”. Στο: *Journal of Motor Behavior* 16.3 (1984), 235–254. doi: [10.1080/00222895.1984.10735319](https://doi.org/10.1080/00222895.1984.10735319).
- [107] R. Amali Therese Jenifa, C. Akila και V. Kavitha. “Rapid background subtraction from video sequences”. Στο: *Proceedings of the 2012 International Conference on Computing, Electronics and Electrical Technologies. ICCEET 2012* (2012). doi: [10.1109/icceet.2012.6203780](https://doi.org/10.1109/icceet.2012.6203780).
- [108] A.d.Jepson, D.j. Fleet και T.r. El-Maraghi. “Robust online appearance models for visual tracking”. Στο: *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001* (2001). doi: [10.1109/cvpr.2001.990505](https://doi.org/10.1109/cvpr.2001.990505).

- [109] Zhuolin Jiang, Zhe Lin και Larry S. Davis. “A Tree-Based Approach to Integrated Action Localization, Recognition and Segmentation”. Στο: *Trends and Topics in Computer Vision, Lecture Notes in Computer Science* (2012), σσ. 114–127. doi: [10.1007/978-3-642-35749-7_9](https://doi.org/10.1007/978-3-642-35749-7_9).
- [110] Gunnar Johansson. “Visual perception of biological motion and a model for its analysis”. Στο: *Perception and Psychophysics* 14.2 (1973), σσ. 201–211. doi: [10.3758/bf03212378](https://doi.org/10.3758/bf03212378).
- [111] Karen Spärck Jones. “A statistical interpretation of term specificity and its application in retrieval”. Στο: *Journal of Documentation* 60.5 (1972), 493–502. doi: [10.1108/00220410410560573](https://doi.org/10.1108/00220410410560573).
- [112] P. Kaewtrakulpong και R. Bowden. “An Improved Adaptive Background Mixture Model for Real-time Tracking with Shadow Detection”. Στο: *Video-Based Surveillance Systems* (2002), 135–144. doi: [10.1007/978-1-4615-0913-4_11](https://doi.org/10.1007/978-1-4615-0913-4_11).
- [113] Byeongkeun Kang κ.ά. “Hand Segmentation for Hand-Object Interaction from Depth map”. Στο: *CoRR* abs/1603.02345 (2016).
- [114] Jinman Kang, I. Cohen και G. Medioni. “Object reacquisition using invariant appearance model”. Στο: *Proceedings of the 2004 International Conference on Pattern Recognition. ICPR 2004.* (2004). doi: [10.1109/icpr.2004.1333883](https://doi.org/10.1109/icpr.2004.1333883).
- [115] Vadim Kantorov και Ivan Laptev. “Efficient Feature Extraction, Encoding, and Classification for Action Recognition”. Στο: *Proceedings of the 2014 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2009* (2014). doi: [10.1109/cvpr.2014.332](https://doi.org/10.1109/cvpr.2014.332).
- [116] Andrej Karpathy κ.ά. “Large-Scale Video Classification with Convolutional Neural Networks”. Στο: *Proceedings of the 2014 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2014* (2014). doi: [10.1109/cvpr.2014.223](https://doi.org/10.1109/cvpr.2014.223).
- [117] Yan Ke, Rahul Sukthankar και Martial Hebert. “Spatio-temporal Shape and Flow Correlation for Action Recognition”. Στο: *Proceedings of the 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2007* (2007). doi: [10.1109/cvpr.2007.383512](https://doi.org/10.1109/cvpr.2007.383512).
- [118] Zuwhan Kim. “Real time object tracking based on dynamic feature grouping with background subtraction”. Στο: *Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2008* (2008). doi: [10.1109/cvpr.2008.4587551](https://doi.org/10.1109/cvpr.2008.4587551).
- [119] Kris Kitani, Yoichi Sato και Akihiro Sugimoto. “Recovering the Basic Structure of Human Activities from a Video-Based Symbol String”. Στο: *Proceedings of the 2007 IEEE Workshop on Motion and Video Computing. WMVC 2007* (2007). doi: [10.1109/wmvc.2007.34](https://doi.org/10.1109/wmvc.2007.34).
- [120] Amedome Min-Dianey Kodjo και Yang Jinhua. “Real-time moving object tracking in video”. Στο: *Proceedings of the 2012 International Conference on Optoelectronics and Microelectronics* (2012). doi: [10.1109/icoom.2012.6316342](https://doi.org/10.1109/icoom.2012.6316342).
- [121] S. Kolkur κ.ά. “Human Skin Detection Using RGB, HSV and YCbCr Color Models”. Στο: *Proceedings of the 2016 International Conference on Communication and Signal Processing. ICCASP 2016* (2016). doi: [10.2991/iccasp-16.2017.51](https://doi.org/10.2991/iccasp-16.2017.51).

- [122] Dimitrios Kosmopoulos, Konstantinos Papoutsakis και Antonis Argyros. “Segmentation and classification of modeled actions in the context of unmodeled ones”. Στο: *Proceedings of the 2014 British Machine Vision Conference* (2014). doi: [10.5244/c.28.95](https://doi.org/10.5244/c.28.95).
- [123] Alex Krizhevsky, Ilya Sutskever και Geoffrey E Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. Στο: *Advances in Neural Information Processing Systems* 25. Επιμέλεια υπό F. Pereira κ.ά. Curran Associates, Inc., 2012, σσ. 1097–1105.
- [124] Tomasz Kryjak, Mateusz Komorkiewicz και Marek Gorgon. “Real-time implementation of foreground object detection from a moving camera using the ViBe algorithm”. Στο: *Computer Science and Information Systems* 11.4 (2014), 1617–1637. doi: [10.2298/csis131218055k](https://doi.org/10.2298/csis131218055k).
- [125] Hilde Kuehne, Ali Arslan και Thomas Serre. “The Language of Actions: Recovering the Syntax and Semantics of Goal-Directed Human Activities”. Στο: *Proceedings of the 2014 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2014* (2014). doi: [10.1109/cvpr.2014.105](https://doi.org/10.1109/cvpr.2014.105).
- [126] Hilde Kuehne, Alexander Richard και Juergen Gall. “Weakly supervised learning of actions from transcripts”. Στο: *Computer Vision and Image Understanding* (2017). doi: [10.1016/j.cviu.2017.06.004](https://doi.org/10.1016/j.cviu.2017.06.004).
- [127] Gondi Lakshmeeswari και K Karthik. “Survey on Algorithms for Object Tracking in Video”. Στο: 141 (Μάι. 2016), σσ. 17–22.
- [128] Ivan Laptev κ.ά. “Learning realistic human actions from movies”. Στο: *Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2008* (2008). doi: [10.1109/cvpr.2008.4587756](https://doi.org/10.1109/cvpr.2008.4587756).
- [129] Colin Lea κ.ά. “Segmental Spatiotemporal CNNs for Fine-Grained Action Segmentation”. Στο: *Computer Vision - ECCV 2016, Lecture Notes in Computer Science* (2016), σσ. 36–52. doi: [10.1007/978-3-319-46487-9_3](https://doi.org/10.1007/978-3-319-46487-9_3).
- [130] Y. Lecun κ.ά. “Gradient-based learning applied to document recognition”. Στο: *Proceedings of the IEEE 1998* 86.11 (1998), 2278–2324. doi: [10.1109/5.726791](https://doi.org/10.1109/5.726791).
- [131] Juan Lei κ.ά. “Real-time object tracking on mobile phones”. Στο: *Proceedings of the First Asian Conference on Pattern Recognition* (2011). doi: [10.1109/acpr.2011.6166663](https://doi.org/10.1109/acpr.2011.6166663).
- [132] Baoxin Li κ.ά. “Model-based temporal object verification using video”. Στο: *Proceedings of the 2001 IEEE Transactions on Image Processing* 10.6 (2001), 897–908. doi: [10.1109/83.923286](https://doi.org/10.1109/83.923286).
- [133] Cheng Li και Kris M. Kitani. “Pixel-Level Hand Detection in Ego-centric Videos”. Στο: *Proceedings of the 2013 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2013* (2013). doi: [10.1109/cvpr.2013.458](https://doi.org/10.1109/cvpr.2013.458).
- [134] Hui Li κ.ά. “A benchmark for semantic image segmentation”. Στο: *Proceedings of the 2013 IEEE International Conference on Multimedia and Expo. ICME 2013* (2013). doi: [10.1109/icme.2013.6607512](https://doi.org/10.1109/icme.2013.6607512).
- [135] Yinxiao Li κ.ά. “Multi-Sensor Surface Analysis for Robotic Ironing”. Στο: *CoRR* abs/1602.04918 (2016).
- [136] Qingshan Liu, Songde Ma και Hanqing Lu. “Head tracking using shapes and adaptive color histograms”. Στο: *Journal of Computer Science and Technology* 17.6 (2002), 859–864. doi: [10.1007/bf02960777](https://doi.org/10.1007/bf02960777).

- [137] Jonathan Long, Evan Shelhamer και Trevor Darrell. “Fully convolutional networks for semantic segmentation”. Στο: *Proceedings of the 2015 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2015* (2015). doi: [10.1109/cvpr.2015.7298965](https://doi.org/10.1109/cvpr.2015.7298965).
- [138] David G. Lowe. “Distinctive Image Features from Scale-Invariant Keypoints”. Στο: *International Journal of Computer Vision* 60.2 (2004), σσ. 91–110. doi: [10.1023/b:visi.0000029664.99615.94](https://doi.org/10.1023/b:visi.0000029664.99615.94).
- [139] David G. Lowe. “Three-dimensional object recognition from single two-dimensional images”. Στο: *Artificial Intelligence* 31.3 (1987), σσ. 355–395. doi: [10.1016/0004-3702\(87\)90070-1](https://doi.org/10.1016/0004-3702(87)90070-1).
- [140] D.g. Lowe. “Object recognition from local scale-invariant features”. Στο: *Proceedings of the 1999 IEEE International Conference on Computer Vision. ICCV 1999* (1999). doi: [10.1109/iccv.1999.790410](https://doi.org/10.1109/iccv.1999.790410).
- [141] T. Lozano-Perez και W.E.L.Grimson. “Off-Line Planning for On-Line Object Localization”. Στο: *Proceedings of the 1986 ACM/IEEE Computer Society Joint Computer Conference*. 1986, σσ. 127–132.
- [142] Cewu Lu κ.ά. “Visual Relationship Detection with Language Priors”. Στο: *Computer Vision – ECCV 2016, Lecture Notes in Computer Science* (2016), 852–869. doi: [10.1007/978-3-319-46448-0_51](https://doi.org/10.1007/978-3-319-46448-0_51).
- [143] Guoliang Lu, Mineichi Kudo και Jun Toyama. “Temporal segmentation and assignment of successive actions in a long-term video”. Στο: *Pattern Recognition Letters* 34.15 (2013), σσ. 1936–1944. doi: [10.1016/j.patrec.2012.10.023](https://doi.org/10.1016/j.patrec.2012.10.023).
- [144] Bruce D. Lucas και Takeo Kanade. “An Iterative Image Registration Technique with an Application to Stereo Vision”. Στο: *Proceedings of the 1981 International Joint Conference on Artificial Intelligence - Volume 2. IJCAI'81*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1981, σσ. 674–679.
- [145] H. P. Luhn. “A Statistical Approach to Mechanized Encoding and Searching of Literary Information”. Στο: *IBM Journal of Research and Development* 1.4 (1957), 309–317. doi: [10.1147/rd.14.0309](https://doi.org/10.1147/rd.14.0309).
- [146] Fengjun Lv και Ramakant Nevatia. “Recognition and Segmentation of 3-D Human Action Using HMM and Multi-class AdaBoost”. Στο: *Computer Vision - ECCV 2006, Lecture Notes in Computer Science* (2006), σσ. 359–372. doi: [10.1007/11744085_28](https://doi.org/10.1007/11744085_28).
- [147] Congyi Lyu κ.ά. “Real-time object tracking system based on field-programmable gate array and convolution neural network”. Στο: *International Journal of Advanced Robotic Systems* 14.1 (2016). doi: [10.1177/1729881416682705](https://doi.org/10.1177/1729881416682705).
- [148] John MacCormick. “A probabilistic exclusion principle for multiple objects”. Στο: *Stochastic Algorithms for Visual Tracking* (2002), 112–123. doi: [10.1007/978-1-4471-0679-1_6](https://doi.org/10.1007/978-1-4471-0679-1_6).
- [149] Lucia Maddalena και Alfredo Petrosino. “The SOBS algorithm: What are the limits?” Στο: *Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. CVPR 2012* (2012). doi: [10.1109/cvprw.2012.6238922](https://doi.org/10.1109/cvprw.2012.6238922).
- [150] Mahamud και Hebert. “Minimum risk distance measure for object recognition”. Στο: *Proceedings of the 2003 IEEE International Conference on Computer Vision. ICCV 2003* (2003). doi: [10.1109/iccv.2003.1238349](https://doi.org/10.1109/iccv.2003.1238349).

- [151] S. Mahamud και M. Hebert. "The optimal distance measure for object detection". Στο: *Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2003* (2003). doi: [10.1109/cvpr.2003.1211361](https://doi.org/10.1109/cvpr.2003.1211361).
- [152] Jonathan Malmaud κ.ά. "What's Cookin'? Interpreting Cooking Videos using Text, Speech and Vision". Στο: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2015). doi: [10.3115/v1/n15-1015](https://doi.org/10.3115/v1/n15-1015).
- [153] A.-R. Mansouri. "Region tracking via level set PDEs without motion computation". Στο: *Proceedings of the 2002 IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.7 (2002), 947–961. doi: [10.1109/tpami.2002.1017621](https://doi.org/10.1109/tpami.2002.1017621).
- [154] D. Marr και H. K. Nishihara. "Representation and Recognition of the Spatial Organization of Three-Dimensional Shapes". Στο: *Proceedings of the Royal Society B: Biological Sciences* 200.1140 (1978), σσ. 269–294. doi: [10.1098/rspb.1978.0020](https://doi.org/10.1098/rspb.1978.0020).
- [155] Shraddha Mehta και Vaishali Kalariya. "Real Time Object Tracking Based on Inter-frame Coding: A Review". Στο: *CoRR* abs/1405.4390 (2014).
- [156] Krystian Mikolajczyk, Bastian Leibe και Bernt Schiele. "Multiple Object Class Detection with a Generative Model". Στο: *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2006*. 2006.
- [157] Krystian Mikolajczyk και Cordelia Schmid. "A Performance Evaluation of Local Descriptors". Στο: *Proceedings of the 2005 IEEE Transactions on Pattern Analysis and Machine Intelligence* 27.10 (Οκτ. 2005), σσ. 1615–1630.
- [158] Ishan Misra, Abhinav Shrivastava και Martial Hebert. "Watch and learn: Semi-supervised learning of object detectors from videos". Στο: *Proceedings of the 2015 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2015* (2015). doi: [10.1109/cvpr.2015.7298982](https://doi.org/10.1109/cvpr.2015.7298982).
- [159] Kinjal B. Mistree, Ashutosh Dutt και Shraddha V. Kothiya. "Real time object tracking for high performance system using GPGPU". Στο: *Proceedings of the 2015 International Conference on Information Processing. ICIP 2015* (2015). doi: [10.1109/infop.2015.7489441](https://doi.org/10.1109/infop.2015.7489441).
- [160] Arpit Mittal, Andrew Zisserman και Philip Torr. "Hand detection using multiple proposals". Στο: *Proceedings of the 2011 British Machine Vision Conference* (2011). doi: [10.5244/c.25.75](https://doi.org/10.5244/c.25.75).
- [161] Darnell Moore και Irfan Essa. "Recognizing Multitasked Activities from Video Using Stochastic Context-free Grammar". Στο: *Proceedings of the 2002 National Conference on Artificial Intelligence*. Menlo Park, CA, USA: American Association for Artificial Intelligence, 2002, σσ. 770–776.
- [162] J. L. Mundy κ.ά. "MORSE: A 3D Object Recognition System Based on Geometric Invariants". Στο: *Proceedings of the 1994 DARPA Image Understanding Workshop*. ARPA, 1994, σσ. 1393–1402.
- [163] Hiroshi Murase και Shree K. Nayar. "Visual learning and recognition of 3-d objects from appearance". Στο: *International Journal of Computer Vision* 14.1 (1995), σσ. 5–24. doi: [10.1007/bf01421486](https://doi.org/10.1007/bf01421486).
- [164] Pradeep Natarajan και Ramakant Nevatia. "Coupled Hidden Semi Markov Models for Activity Recognition". Στο: *Proceedings of the 2007 IEEE Workshop on Motion and Video Computing. WMVC 2007* (2007). doi: [10.1109/wmvc.2007.12](https://doi.org/10.1109/wmvc.2007.12).

- [165] Ram Nevatia, Tao Zhao και Somboon Hongeng. “Hierarchical Language-based Representation of Events in Video Streams”. Στο: *Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshop. CVPR 2003* (2003). doi: [10.1109/cvprw.2003.10038](https://doi.org/10.1109/cvprw.2003.10038).
- [166] Ramakant Nevatia και Thomas O. Binford. “Description and recognition of curved objects?” Στο: *Artificial Intelligence* 8.1 (1977), σσ. 77–98. doi: [10.1016/0004-3702\(77\)90006-6](https://doi.org/10.1016/0004-3702(77)90006-6).
- [167] N.t. Nguyen κ.ά. “Learning and Detecting Activities from Movement Trajectories Using the Hierarchical Hidden Markov Models”. Στο: *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2005* (2005). doi: [10.1109/cvpr.2005.203](https://doi.org/10.1109/cvpr.2005.203).
- [168] N.m. Oliver, B. Rosario και A.p. Pentland. “A Bayesian computer vision system for modeling human interactions”. Στο: *Proceedings of the 2000 IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.8 (2000), σσ. 831–843. doi: [10.1109/34.868684](https://doi.org/10.1109/34.868684).
- [169] Jia-Yu Pan, Hyungjeong Yang και C. Faloutsos. “MMSS: Multi-Modal Story-Oriented Video Summarization”. Στο: *Proceedings of the 2004 IEEE International Conference on Data Mining* (2004). doi: [10.1109/icdm.2004.10033](https://doi.org/10.1109/icdm.2004.10033).
- [170] Sangho Park και J. K. Aggarwal. “A hierarchical Bayesian network for event recognition of human actions and interactions”. Στο: *Multimedia Systems* 10.2 (2004), σσ. 164–179. doi: [10.1007/s00530-004-0148-1](https://doi.org/10.1007/s00530-004-0148-1).
- [171] F. Pedregosa κ.ά. “Scikit-learn: Machine Learning in Python”. Στο: *Journal of Machine Learning Research* 12 (2011), σσ. 2825–2830.
- [172] Natan Peterfreund. “Robust Tracking of Position and Velocity With Kalman Snakes”. Στο: *Proceedings of the 1999 IEEE Transactions on Pattern Analysis and Machine Intelligence* 21.6 (Ιούν. 1999), σσ. 564–569.
- [173] Nicolas Pinto, David D. Cox και James J. DiCarlo. “Why is Real-World Visual Object Recognition Hard?” Στο: *PLoS Computational Biology* 4.1 (2008). doi: [10.1371/journal.pcbi.0040027](https://doi.org/10.1371/journal.pcbi.0040027).
- [174] Hamed Pirsiavash και Deva Ramanan. “Parsing Videos of Actions with Segmental Grammars”. Στο: *Proceedings of the 2014 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2014* (2014). doi: [10.1109/cvpr.2014.85](https://doi.org/10.1109/cvpr.2014.85).
- [175] Christian Plagemann κ.ά. “Real-time identification and localization of body parts from depth images”. Στο: *Proceedings of the 2010 IEEE International Conference on Robotics and Automation. ICRA 2010* (2010). doi: [10.1109/robot.2010.5509559](https://doi.org/10.1109/robot.2010.5509559).
- [176] John C. Platt. “Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods”. Στο: *ADVANCES IN LARGE MARGIN CLASSIFIERS*. MIT Press, 1999, σσ. 61–74.
- [177] J. Ponce, D. Chelberg και W. B. Mann. “Invariant Properties of Straight Homogeneous Generalized Cylinders and Their Contours”. Στο: *Proceedings of the 1989 IEEE Transactions on Pattern Analysis and Machine Intelligence* 11.9 (Σεπτ. 1989), σσ. 951–966.
- [178] Danila Potapov κ.ά. “Category-Specific Video Summarization”. Στο: *Computer Vision - ECCV 2014, Lecture Notes in Computer Science* (2014), σσ. 540–555. doi: [10.1007/978-3-319-10599-4_35](https://doi.org/10.1007/978-3-319-10599-4_35).

- [179] Dilip K. Prasad κ.ά. “Challenges in video based object detection in maritime scenario using computer vision”. Στο: *CoRR* abs/1608.01079 (2016).
- [180] Jagdish Lal Raheja, Karen Das και Ankit Chaudhary. “Fingertip Detection: A Fast Method with Natural Hand”. Στο: *CoRR* abs/1212.0134 (2012).
- [181] K. Rangarajan και M. Shah. “Establishing motion correspondence”. Στο: *Proceedings of the 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 1991* (1991). doi: [10.1109/cvpr.1991.139669](https://doi.org/10.1109/cvpr.1991.139669).
- [182] C. Rasmussen και G.d. Hager. “Probabilistic data association methods for tracking complex visual objects”. Στο: *Proceedings of the 2001 IEEE Transactions on Pattern Analysis and Machine Intelligence* 23.6 (2001), 560–576. doi: [10.1109/tpami.2001.927458](https://doi.org/10.1109/tpami.2001.927458).
- [183] Joseph Redmon κ.ά. “You Only Look Once: Unified, Real-Time Object Detection”. Στο: *Proceedings of the 2016 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2016* (2016). doi: [10.1109/cvpr.2016.91](https://doi.org/10.1109/cvpr.2016.91).
- [184] Michaela Regneri κ.ά. “Grounding Action Descriptions in Videos”. Στο: *Proceedings of the 2013 Transactions of the Association for Computational Linguistics (TACL)* 1 (2013), σσ. 25–36.
- [185] Donald Reid. “An algorithm for tracking multiple targets”. Στο: *Proceedings of the 1978 IEEE Conference on Decision and Control including the 17th Symposium on Adaptive Processes* (1978). doi: [10.1109/cdc.1978.268125](https://doi.org/10.1109/cdc.1978.268125).
- [186] Shaoqing Ren κ.ά. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. Στο: *Proceedings of the 2017 IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.6 (2017), σσ. 1137–1149. doi: [10.1109/tpami.2016.2577031](https://doi.org/10.1109/tpami.2016.2577031).
- [187] Lawrence G. Roberts. *Machine Perception of Three-Dimensional Solids*. Outstanding Dissertations in the Computer Sciences. Garland Publishing, New York, 1963.
- [188] I. Rodomagoulakis κ.ά. “Multimodal human action recognition in assistive human-robot interaction”. Στο: *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP 2016* (2016). doi: [10.1109/icassp.2016.7472168](https://doi.org/10.1109/icassp.2016.7472168).
- [189] Anna Rohrbach κ.ά. “A dataset for Movie Description”. Στο: *Proceedings of the 2015 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2015* (2015). doi: [10.1109/cvpr.2015.7298940](https://doi.org/10.1109/cvpr.2015.7298940).
- [190] Anna Rohrbach κ.ά. “Coherent Multi-sentence Video Description with Variable Level of Detail”. Στο: *Pattern Recognition, Lecture Notes in Computer Science* (2014), σσ. 184–195. doi: [10.1007/978-3-319-11752-2_15](https://doi.org/10.1007/978-3-319-11752-2_15).
- [191] M. Rohrbach κ.ά. “A database for fine grained activity detection of cooking activities”. Στο: *Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2012* (2012). doi: [10.1109/cvpr.2012.6247801](https://doi.org/10.1109/cvpr.2012.6247801).
- [192] Marcus Rohrbach κ.ά. “Recognizing Fine-Grained and Composite Activities Using Hand-Centric Features and Script Data”. Στο: *International Journal of Computer Vision* 119.3 (2015), 346–373. doi: [10.1007/s11263-015-0851-8](https://doi.org/10.1007/s11263-015-0851-8).
- [193] Marcus Rohrbach κ.ά. “Script Data for Attribute-Based Recognition of Composite Activities”. Στο: *Computer Vision – ECCV 2012, Lecture Notes in Computer Science* (2012), 144–157. doi: [10.1007/978-3-642-33718-5_11](https://doi.org/10.1007/978-3-642-33718-5_11).

- [194] Remi Ronfard. “Region-based strategies for active contour models”. Στο: *International Journal of Computer Vision* 13.2 (1994), 229–251. doi: [10.1007/bf01427153](https://doi.org/10.1007/bf01427153).
- [195] R. Rosales και S. Sclaroff. “3D trajectory recovery for tracking multiple objects and trajectory guided recognition of actions”. Στο: *Proceedings of the 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 1999* (1999). doi: [10.1109/cvpr.1999.784618](https://doi.org/10.1109/cvpr.1999.784618).
- [196] F. Rosenblatt. “The perceptron: A probabilistic model for information storage and organization in the brain.” Στο: *Psychological Review* 65.6 (1958), 386–408. doi: [10.1037/h0042519](https://doi.org/10.1037/h0042519).
- [197] Amir Rosenfeld και Shimon Ullman. “Hand-Object Interaction and Precise Localization in Transitive Action Recognition”. Στο: *Proceedings of the 2016 Conference on Computer and Robot Vision. CRV 2016* (2016). doi: [10.1109/crv.2016.27](https://doi.org/10.1109/crv.2016.27).
- [198] Carsten Rother, Vladimir Kolmogorov και Andrew Blake. “”GrabCut” - Interactive Foreground Extraction using Iterated Graph Cuts”. Στο: *Interactive Foreground Extraction using Iterated Graph Cuts* (2004). doi: [10.1145/1186562.1015720](https://doi.org/10.1145/1186562.1015720).
- [199] C. A. Rothwell κ.ά. “Planar object recognition using projective shape representation”. Στο: *International Journal of Computer Vision* 16.1 (1995), σσ. 57–99. doi: [10.1007/bf01428193](https://doi.org/10.1007/bf01428193).
- [200] Charles A. Rothwell κ.ά. “Canonical frames for planar object recognition”. Στο: *Computer Vision - ECCV 1992, Lecture Notes in Computer Science* (1992), σσ. 757–772. doi: [10.1007/3-540-55426-2_86](https://doi.org/10.1007/3-540-55426-2_86).
- [201] Miti Ruchanurucks, Koichi Ogawara και Katsushi Ikeuchi. “Neural Network Based Foreground Segmentation with an Application to Multi-Sensor 3D Modeling”. Στο: *Proceedings of the 2006 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems* (2006). doi: [10.1109/mfi.2006.265586](https://doi.org/10.1109/mfi.2006.265586).
- [202] M.s. Ryoo και J.k. Aggarwal. “Recognition of Composite Human Activities through Context-Free Grammar Based Representation”. Στο: *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2006* (2006). doi: [10.1109/cvpr.2006.242](https://doi.org/10.1109/cvpr.2006.242).
- [203] M.s. Ryoo και J.k. Aggarwal. “Stochastic representation and recognition of high-level group activities: Describing structural uncertainties in human activities”. Στο: *Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. CVPR 2009* (2009). doi: [10.1109/cvpr.2009.5204329](https://doi.org/10.1109/cvpr.2009.5204329).
- [204] S. Sadanand και J.J. Corso. “Action bank: A high-level representation of activity in video”. Στο: *Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2012* (2012). doi: [10.1109/cvpr.2012.6247806](https://doi.org/10.1109/cvpr.2012.6247806).
- [205] V. Salari και I.k. Sethi. “Feature point correspondence in the presence of occlusion”. Στο: *Proceedings of the 1990 IEEE Transactions on Pattern Analysis and Machine Intelligence* 12.1 (1990), 87–91. doi: [10.1109/34.41387](https://doi.org/10.1109/34.41387).
- [206] Gerard Salton και Christopher Buckley. “Term-weighting approaches in automatic text retrieval”. Στο: *Information Processing and Management* 24.5 (1988), σσ. 513–523. doi: [10.1016/0306-4573\(88\)90021-0](https://doi.org/10.1016/0306-4573(88)90021-0).

- [207] Koichi Sato και J.k. Aggarwal. “Temporal spatio-velocity transform and its application to tracking and interaction”. Στο: *Computer Vision and Image Understanding* 96.2 (2004), 100–128. doi: [10.1016/j.cviu.2004.02.003](https://doi.org/10.1016/j.cviu.2004.02.003).
- [208] Silvio Savarese κ.ά. “Spatial-Temporal correlatons for unsupervised action classification”. Στο: *Proceedings of the 2008 IEEE Workshop on Motion and Video Computing. WMVC 2008* (2008). doi: [10.1109/wmvc.2008.4544068](https://doi.org/10.1109/wmvc.2008.4544068).
- [209] Henry Schneiderman και Takeo Kanade. “Object Detection Using the Statistics of Parts”. Στο: *International Journal of Computer Vision* 56.3 (2004), σσ. 151–177. doi: [10.1023/b:visi.0000011202.85607.00](https://doi.org/10.1023/b:visi.0000011202.85607.00).
- [210] C. Schuldt, I. Laptev και B. Caputo. “Recognizing human actions: a local SVM approach”. Στο: *Proceedings of the 2004 International Conference on Pattern Recognition. ICPR 2004* (2004). doi: [10.1109/icpr.2004.1334462](https://doi.org/10.1109/icpr.2004.1334462).
- [211] Ozan Sener κ.ά. “Unsupervised Semantic Parsing of Video Collections”. Στο: *Proceedings of the 2015 IEEE International Conference on Computer Vision. ICCV 2015* (2015). doi: [10.1109/iccv.2015.509](https://doi.org/10.1109/iccv.2015.509).
- [212] Ishwar K. Sethi και Ramesh Jain. “Finding Trajectories of Feature Points in a Monocular Image Sequence”. Στο: *Proceedings of the 1987 IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-9.1* (1987), 56–73. doi: [10.1109/tpami.1987.4767872](https://doi.org/10.1109/tpami.1987.4767872).
- [213] K. Shafique και M. Shah. “A noniterative greedy algorithm for multiframe point correspondence”. Στο: *Proceedings of the 2003 IEEE Transactions on Pattern Analysis and Machine Intelligence* 27.1 (2005), 51–65. doi: [10.1109/tpami.2005.1](https://doi.org/10.1109/tpami.2005.1).
- [214] Khamar Basha Shaik κ.ά. “Comparative Study of Skin Color Detection and Segmentation in HSV and YCbCr Color Space”. Στο: *Procedia Computer Science* 57 (2015), 41–48. doi: [10.1016/j.procs.2015.07.362](https://doi.org/10.1016/j.procs.2015.07.362).
- [215] Ling Shao κ.ά. “Human action segmentation and recognition via motion and shape analysis”. Στο: *Pattern Recognition Letters* 33.4 (2012), σσ. 438–445. doi: [10.1016/j.patrec.2011.05.015](https://doi.org/10.1016/j.patrec.2011.05.015).
- [216] Toby Sharp κ.ά. “Accurate, Robust, and Flexible Real-time Hand Tracking”. Στο: *Proceedings of the 2015 (33rd) Annual ACM Conference on Human Factors in Computing Systems* (2015). doi: [10.1145/2702123.2702179](https://doi.org/10.1145/2702123.2702179).
- [217] Jianbo Shi και Tomasi. “Good features to track”. Στο: *Proceedings of the 1994 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 1994* (1994). doi: [10.1109/cvpr.1994.323794](https://doi.org/10.1109/cvpr.1994.323794).
- [218] Qinfeng Shi κ.ά. “Discriminative human action segmentation and recognition using semi-Markov model”. Στο: *Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2008* (2008). doi: [10.1109/cvpr.2008.4587557](https://doi.org/10.1109/cvpr.2008.4587557).
- [219] Yifan Shi κ.ά. “Propagation networks for recognition of partially ordered sequential action”. Στο: *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2004.* (2004). doi: [10.1109/cvpr.2004.1315255](https://doi.org/10.1109/cvpr.2004.1315255).
- [220] P.y. Simard, D. Steinkraus και J.c. Platt. “Best practices for convolutional neural networks applied to visual document analysis”. Στο: *Proceedings of the 2003 International Conference on Document Analysis and Recognition* (2003). doi: [10.1109/icdar.2003.1227801](https://doi.org/10.1109/icdar.2003.1227801).

- [221] Young Chol Song κ.ά. “Unsupervised Alignment of Actions in Video with Text Descriptions”. Στο: *Proceedings of the 2016 International Joint Conference on Artificial Intelligence*. IJCAI’16. AAAI Press, 2016, σσ. 2025–2031.
- [222] E.h. Spriggs, F. De La Torre και M. Hebert. “Temporal segmentation and activity classification from first-person sensing”. Στο: *Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. CVPR 2009 (2009). doi: [10.1109/cvpr.2009.5204354](https://doi.org/10.1109/cvpr.2009.5204354).
- [223] Srinath Sridhar κ.ά. “Fast and robust hand tracking using detection-guided optimization”. Στο: *Proceedings of the 2015 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. CVPR 2015 (2015). doi: [10.1109/cvpr.2015.7298941](https://doi.org/10.1109/cvpr.2015.7298941).
- [224] Srinath Sridhar κ.ά. “Real-Time Joint Tracking of a Hand Manipulating an Object from RGB-D Input”. Στο: *Computer Vision – ECCV 2016, Lecture Notes in Computer Science* (2016), 294–310. doi: [10.1007/978-3-319-46475-6_19](https://doi.org/10.1007/978-3-319-46475-6_19).
- [225] R Sai Srivatsa και R. Venkatesh Babu. “Salient object detection via objectness measure”. Στο: *Proceedings of the 2015 IEEE International Conference on Image Processing ICIP 2015* (2015). doi: [10.1109/icip.2015.7351654](https://doi.org/10.1109/icip.2015.7351654).
- [226] T. Starner και A. Pentland. “Real-time American Sign Language recognition from video using hidden Markov models”. Στο: *Proceedings of the 1995 International Symposium on Computer Vision*. ISCV 1995 (1995). doi: [10.1109/iscv.1995.477012](https://doi.org/10.1109/iscv.1995.477012).
- [227] Chris Stauffer και W. Eric L. Grimson. “Adaptive Background Mixture Models for Real-Time Tracking”. Στο: *Proceedings of the 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. CVPR 1999. 1999.
- [228] H. Steinhaus. “Sur la division des corp matériels en parties”. Στο: *Bull. Acad. Polon. Sci* 1 (1956), σσ. 801–804.
- [229] Roy L. Streit και Tod E. Luginbuhl. “Maximum Likelihood Method for Probabilistic Multi-Hypothesis Tracking”. Στο: *Signal and Data Processing of Small Targets 1994* (1994). doi: [10.1117/12.179066](https://doi.org/10.1117/12.179066).
- [230] Bing Su κ.ά. “Hierarchical Dynamic Parsing and Encoding for Action Recognition”. Στο: *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV*. Επιμέλεια υπό Bastian Leibe κ.ά. Cham: Springer International Publishing, 2016, σσ. 202–217.
- [231] Kah-Kay Sung και Partha Niyogi. “An Active Learning Formulation for Instance Selection with Applications to Object Detection”. Στο: *Instance Selection and Construction for Data Mining* (2001), σσ. 357–374. doi: [10.1007/978-1-4757-3359-4_20](https://doi.org/10.1007/978-1-4757-3359-4_20).
- [232] Michael J. Swain και Dana H. Ballard. “Color Indexing”. Στο: *Readings in Multimedia Computing and Networking* (2002), σσ. 265–277. doi: [10.1016/b978-155860651-7/50109-1](https://doi.org/10.1016/b978-155860651-7/50109-1).
- [233] Christian Szegedy κ.ά. “Going deeper with convolutions”. Στο: *Proceedings of the 2015 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. CVPR 2015 (2015). doi: [10.1109/cvpr.2015.7298594](https://doi.org/10.1109/cvpr.2015.7298594).
- [234] Yichuan Tang. “Deep Learning using Support Vector Machines”. Στο: *CoRR* abs/1306.0239 (2013).
- [235] Hai Tao, H.s. Sawhney και R. Kumar. “Object tracking with Bayesian estimation of dynamic layer representations”. Στο: *Proceedings of the 2002 IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.1 (2002), 75–89. doi: [10.1109/34.982885](https://doi.org/10.1109/34.982885).

- [236] Demetri Terzopoulos και Richard Szeliski. “Active Vision”. Στο: επιμέλεια υπό Andrew Blake και Alan Yuille. Cambridge, MA, USA: MIT Press, 1993. Κεφ. Tracking with Kalman Snakes, σσ. 3–20.
- [237] Joseph Tighe και Svetlana Lazebnik. “SuperParsing: Scalable Nonparametric Image Parsing with Superpixels”. Στο: *Computer Vision - ECCV 2010, Lecture Notes in Computer Science* (2010), σσ. 352–365. doi: [10.1007/978-3-642-15555-0_26](https://doi.org/10.1007/978-3-642-15555-0_26).
- [238] Son D. Tran και Larry S. Davis. “Event Modeling and Recognition Using Markov Logic Networks”. Στο: *Computer Vision – ECCV 2008, Lecture Notes in Computer Science* (2008), 610–623. doi: [10.1007/978-3-540-88688-4_45](https://doi.org/10.1007/978-3-540-88688-4_45).
- [239] Wen-Hsiang Tsai και King-Sun Fu. “Attributed Grammar-A Tool for Combining Syntactic and Statistical Approaches to Pattern Recognition”. Στο: *Proceedings of the 1980 IEEE Transactions on Systems, Man, and Cybernetics* 10.12 (1980), 873–885. doi: [10.1109/tsmc.1980.4308414](https://doi.org/10.1109/tsmc.1980.4308414).
- [240] M.a. Turk και A.p. Pentland. “Face recognition using eigenfaces”. Στο: *Proceedings of the 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 1991* (1991). doi: [10.1109/cvpr.1991.139758](https://doi.org/10.1109/cvpr.1991.139758).
- [241] A. Vedaldi και K. Lenc. “MatConvNet – Convolutional Neural Networks for MATLAB”. Στο: *Proceedings of the 2015 ACM International Conference on Multimedia*. 2015.
- [242] C.j. Veenman, M.j.t. Reinders και E. Backer. “Resolving motion correspondence for densely moving points”. Στο: *Proceedings of the 2001 IEEE Transactions on Pattern Analysis and Machine Intelligence* 23.1 (2001), 54–72. doi: [10.1109/34.899946](https://doi.org/10.1109/34.899946).
- [243] Oriol Vinyals κ.ά. “Show and tell: A neural image caption generator”. Στο: *Proceedings of the 2015 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2015* (2015). doi: [10.1109/cvpr.2015.7298935](https://doi.org/10.1109/cvpr.2015.7298935).
- [244] P. Viola και M. Jones. “Rapid object detection using a boosted cascade of simple features”. Στο: *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001* (2001). doi: [10.1109/cvpr.2001.990517](https://doi.org/10.1109/cvpr.2001.990517).
- [245] Paul Viola και Michael Jones. “Robust Real-time Object Detection”. Στο: *International Journal of Computer Vision*. 2001.
- [246] Hanzi Wang και D. Suter. “A Re-evaluation of Mixture-of-Gaussian Background Modeling”. Στο: *Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP 2005* (2005). doi: [10.1109/icassp.2005.1415580](https://doi.org/10.1109/icassp.2005.1415580).
- [247] Heng Wang και Cordelia Schmid. “Action Recognition with Improved Trajectories”. Στο: *Proceedings of the 2013 IEEE International Conference on Computer Vision. ICCV 2013* (2013). doi: [10.1109/iccv.2013.441](https://doi.org/10.1109/iccv.2013.441).
- [248] Heng Wang κ.ά. “Action recognition by dense trajectories”. Στο: *Proceedings of the 2011 IEEE International Conference on Computer Vision. ICCV 2011* (2011). doi: [10.1109/cvpr.2011.5995407](https://doi.org/10.1109/cvpr.2011.5995407).
- [249] Heng Wang κ.ά. “Evaluation of local spatio-temporal features for action recognition”. Στο: *Proceedings of the 2009 British Machine Vision Conference*. doi:10.5244/C.23.124. BMVA Press, 2009, σσ. 124.1–124.11.

- [250] Jon A. Webb και J.k. Aggarwal. "Structure from motion of rigid and jointed objects". Στο: *Artificial Intelligence* 19.1 (1982), σσ. 107–130. doi: [10.1016/0004-3702\(82\)90023-6](https://doi.org/10.1016/0004-3702(82)90023-6).
- [251] Shu-Fai Wong, Tae-Kyun Kim και Roberto Cipolla. "Learning Motion Categories using both Semantic and Structural Information". Στο: *Proceedings of the 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2007* (2007). doi: [10.1109/cvpr.2007.383332](https://doi.org/10.1109/cvpr.2007.383332).
- [252] Chenxia Wu κ.ά. "Watch-n-patch: Unsupervised understanding of actions and relations". Στο: *Proceedings of the 2015 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2015* (2015). doi: [10.1109/cvpr.2015.7299065](https://doi.org/10.1109/cvpr.2015.7299065).
- [253] J. Yamato, J. Ohya και K. Ishii. "Recognizing human action in time-sequential images using hidden Markov model". Στο: *Proceedings of the 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 1992* (1992). doi: [10.1109/cvpr.1992.223161](https://doi.org/10.1109/cvpr.1992.223161).
- [254] Shaohua Yang κ.ά. "Grounded Semantic Role Labeling". Στο: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2016). doi: [10.18653/v1/n16-1019](https://doi.org/10.18653/v1/n16-1019).
- [255] Yezhou Yang κ.ά. "Grasp type revisited: A modern perspective on a classical feature for vision". Στο: *Proceedings of the 2015 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2015* (2015). doi: [10.1109/cvpr.2015.7298637](https://doi.org/10.1109/cvpr.2015.7298637).
- [256] Yezhou Yang κ.ά. "Learning the Semantics of Manipulation Action". Στο: *Proceedings of the 2015 (53rd) Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (2015). doi: [10.3115/v1/p15-1066](https://doi.org/10.3115/v1/p15-1066).
- [257] Yezhou Yang κ.ά. "Manipulation action tree bank: A knowledge resource for humanoids". Στο: *Proceedings of the 2014 IEEE-RAS International Conference on Humanoid Robots* (2014). doi: [10.1109/humanoids.2014.7041483](https://doi.org/10.1109/humanoids.2014.7041483).
- [258] Yezhou Yang κ.ά. "Neural Self Talk: Image Understanding via Continuous Questioning and Answering". Στο: *CoRR* abs/1512.03460 (2015).
- [259] Yezhou Yang κ.ά. "Robot Learning Manipulation Action Plans by "Watching" Unconstrained Videos from the World Wide Web". Στο: *Proceedings of the 2015 AAAI Conference on Artificial Intelligence. AAAI'15*. AAAI Press, 2015, σσ. 3686–3692.
- [260] Shinsuke Yasukawa κ.ά. "Real-time object tracking based on scale-invariant features employing bio-inspired hardware". Στο: *Neural Networks* 81 (2016), 29–38. doi: [10.1016/j.neunet.2016.05.002](https://doi.org/10.1016/j.neunet.2016.05.002).
- [261] A. Yilmaz, Xin Li και M. Shah. "Contour-based object tracking with occlusion handling in video acquired using mobile cameras". Στο: *Proceedings of the 2004 IEEE Transactions on Pattern Analysis and Machine Intelligence* 26.11 (2004), 1531–1536. doi: [10.1109/tpami.2004.96](https://doi.org/10.1109/tpami.2004.96).
- [262] Alper Yilmaz, Omar Javed και Mubarak Shah. "Object tracking". Στο: *ACM Computing Surveys* 38.4 (2006). doi: [10.1145/1177352.1177355](https://doi.org/10.1145/1177352.1177355).
- [263] Matthew D. Zeiler και Rob Fergus. "Visualizing and Understanding Convolutional Networks". Στο: *Computer Vision - ECCV 2014, Lecture Notes in Computer Science* (2014), σσ. 818–833. doi: [10.1007/978-3-319-10590-1_53](https://doi.org/10.1007/978-3-319-10590-1_53).

- [264] Jianguo Zhang κ.ά. “Local Features and Kernels for Classification of Texture and Object Categories: An In-Depth Study”. Στο: *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshop. CVPR 2005* (2005). doi: [10.1109/cvprw.2006.121](https://doi.org/10.1109/cvprw.2006.121).
- [265] Yi Zhang κ.ά. “Does the grasp type reveal action intention?” Στο: *Journal of Vision* 15.12 (2015), σ. 1153. doi: [10.1167/15.12.1153](https://doi.org/10.1167/15.12.1153).
- [266] Shuai Zheng κ.ά. “Conditional Random Fields as Recurrent Neural Networks”. Στο: *Proceedings of the 2015 IEEE International Conference on Computer Vision. ICCV 2015* (2015). doi: [10.1109/iccv.2015.179](https://doi.org/10.1109/iccv.2015.179).
- [267] Luowei Zhou, Chenliang Xu και Jason J. Corso. “ProcNets: Learning to Segment Procedures in Untrimmed and Unconstrained Videos”. Στο: *CoRR* abs/1703.09788 (2017).
- [268] Yang Zhou κ.ά. “Interaction part mining: A mid-level approach for fine-grained action recognition”. Στο: *Proceedings of the 2015 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2015* (2015). doi: [10.1109/cvpr.2015.7298953](https://doi.org/10.1109/cvpr.2015.7298953).
- [269] Wangjiang Zhu κ.ά. “Saliency Optimization from Robust Background Detection”. Στο: *Proceedings of the 2014 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2014* (2014). doi: [10.1109/cvpr.2014.360](https://doi.org/10.1109/cvpr.2014.360).
- [270] Xiaolong Zhu, Xuhui Jia και Kwan-Yee K. Wong. “Pixel-Level Hand Detection with Shape-Aware Structured Forests”. Στο: *Computer Vision - ACCV 2014, Lecture Notes in Computer Science* (2015), σσ. 64–78. doi: [10.1007/978-3-319-16817-3_5](https://doi.org/10.1007/978-3-319-16817-3_5).
- [271] Xiaolong Zhu κ.ά. “A two-stage detector for hand detection in ego-centric videos”. Στο: *Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision. WACV 2016* (2016). doi: [10.1109/wacv.2016.7477665](https://doi.org/10.1109/wacv.2016.7477665).
- [272] Andreas Zinnen, Ulf Blanke και Bernt Schiele. “An Analysis of Sensor-Oriented vs. Model-Based Activity Recognition”. Στο: *Proceedings of the 2009 International Symposium on Wearable Computers* (2009).