ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Αναγνώριση ανθρώπινων ενεργειών σε βίντεο με χρήση νευρωνικών δικτύων και λογισμού γεγονότων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΙΩΑΝΝΗ ΠΡΑΠΑ

Επιβλέποντες:	Ανδρεάς-Γεώργιος Σταφυλοπάτης	Καθηγητής Ε.Μ.Π
	Γεώργιος Παλιούρας	Ερευνητής Ε.Κ.Ε.Φ.Ε. “Δημόκριτος”
	Αλέξανδρος Αρτίκης	Ερευνητής Ε.Κ.Ε.Φ.Ε. “Δημόκριτος”
	Νικόλαος Μπασκιώτης	Ερευνητής LIP6 Παρισιού

Αθήνα, Ιούλιος 2017

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Αναγνώριση ανθρώπινων ενεργειών σε βίντεο με χρήση νευρωνικών δικτύων και λογισμού γεγονότων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΙΩΑΝΝΗ ΠΡΑΠΑ

Επιβλέποντες:	Ανδρέας-Γεώργιος Σταφυλοπάτης	Καθηγητής Ε.Μ.Π.
	Γεώργιος Παλιούρας	Ερευνητής Ε.Κ.Ε.Φ.Ε. “Δημόκριτος”
	Αλέξανδρος Αρτίκης	Ερευνητής Ε.Κ.Ε.Φ.Ε. “Δημόκριτος”
	Νικόλαος Μπασκιώτης	Ερευνητής LIP6 Παρισιού

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 11η Ιουλίου 2017.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Ανδρέας-Γεώργιος Σταφυλοπάτης Καθηγητής Ε.Μ.Π.	Γιώργος Στάμου Αναπληρωτής Καθηγητής Ε.Μ.Π.	Παναγιώτης Τσανάκας Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2017

(Υπογραφή)

.....

ΙΩΑΝΝΗΣ ΠΡΑΠΑΣ

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

Copyright © Ιωάννης Πράπας, 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 μή 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.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

Περίληψη

Ο σκοπός αυτής της διπλωματικής εργασίας ήταν σε πρώτο στάδιο η μελέτη σύγχρονων αριθμητικών μεθόδων για την αναγνώριση ανθρώπινων ενεργειών σε βίντεο και στην συνέχεια ο συνδυασμός τους με λογισμό γεγονότων με στόχο την υψηλού επιπέδου κατανόηση βίντεο.

Σε αυτό το πλαίσιο, εξετάσαμε μεθόδους εξαγωγής χαρακτηριστικών κατάλληλων για ανάλυση βίντεο. Υλοποιήσαμε συνελικτικούς αυτοκωδικοποιητές και εξετάσαμε την χρήση τους για αναπαράσταση των καρέ του βίντεο. Τελικά, χρησιμοποιήσαμε και αξιολογήσαμε τα C3D χαρακτηριστικά ως αναπαράσταση κατάλληλη για αναγνώριση απλών ανθρώπινων ενεργειών.

Σαν επόμενο στάδιο στην διαδικασία, ελέγξαμε κατά πόσο οι αναγνωρισμένες απλές ενέργειες μπορούν να βοηθήσουν τον OLED, ένα σύστημα επαγωγικού λογικού προγραμματισμού, να παράγει θεωρίες Λογισμού Γεγονότων για σύνθετες ενέργειες, οι οποίες αφορούν αλληλεπίδραση ατόμων.

Συμβάλλουμε, λοιπόν, στην γεφύρωση του χάσματος που υπάρχει μεταξύ της αδιάφανης (φύση μαύρου κουτιού) αποτελεσματικότητας της βαθιάς μάθησης να βγάζει νόημα από ανεπεξέργαστα δεδομένα και του διαφανούς συμπερασμού της λογικής που μας επιτρέπει να εισάγουμε ανθρώπινη γνώση για την φύση των προβλημάτων υπό επίλυση. Λόγω του ότι οι παρατηρήσεις από τη φύση τους καταγράφουν μέρος της πραγματικότητας, ενώ η λογική εκφράζει συνήθως γενικότερη γνώση, θέλουμε να ενθαρρύνουμε την συνέχιση της έρευνας μακριά από πλήρεις αρχιτεκτονικές μαύρου κουτιού και πιο κοντά σε συγχωνεύσεις τους με συστήματα λογικού συμπερασμού. Τα πειραματικά αποτελέσματα αυτής της εργασίας ενθαρρύνουν περαιτέρω την έρευνα προς αυτήν την κατεύθυνση.

Λέξεις Κλειδιά: αναγνώριση ανθρώπινων ενεργειών, συνελικτικά νευρωνικά δίκτυα, αυτοκωδικοποιητές, C3D χαρακτηριστικά, OLED, λογισμός γεγονότων

Abstract

The main aim of this thesis was to study modern numerical methods for human action recognition in videos and combine them with event calculus towards a high-level understanding.

In the context of low-level human action recognition, we considered feature extraction methods suitable for video analysis. We implemented a convolutional autoencoders and considered their use for frame-level representation. Finally, we used and evaluated the C3D features as a representation for low-level action recognition.

As the next step in the pipeline, we checked how useful the recognized low-level human actions can be for OLED, an Inductive Logic Programming system, capable of learning Event Calculus theories for complex actions, involving the interaction of more than one individuals.

Our contribution lies mainly in bridging the gap between the opaque (black box nature) effectiveness of deep learning to make sense out of raw data and the transparent reasoning of Event Calculus, which allows us to embed human knowledge in the problem-solving process. Because observations capture only certain aspects of the real world, but logic often represents generic knowledge, we would like to encourage research to move away from end-to-end black box architectures and towards hybrids with computational logic. Our experimental results encourage further research towards this direction.

Keywords: human action recognition, convolutional neural networks, autoencoders, C3D features, Event Calculus

Ευχαριστίες

Μια διπλωματική εργασία απαιτεί χρόνο, κόπο και θυσίες. Γίνεται όμως πολύ ευκολότερη όταν έχεις ανθρώπους να σε στηρίζουν και να νοιάζονται αρκετά ώστε να σου επισημαίνουν τα λάθη σου, να παρέχουν τις συμβουλές τους και να σε βοηθούν να προχωράς. Σε αυτό το κομμάτι, ήμουν πολύ τυχερός να περιβάλλομαι σε όλη την διάρκεια της εργασίας από εξαίρετους επιστήμονες τόσο ερευνητικά, όσο και σε ήθος.

Νιώθω πολύ τυχερός για τους επιβλέποντες που είχα, τον Δρ. Γιώργο Παλιούρα και τον Δρ. Αλέξανδρο Αρτίκη από το Ε.Κ.Ε.Φ.Ε. Δημόκριτος και τον Δρ. Νικόλα Μπασκιώτη από το εργαστήριο LIP6 του Παρισιού. Είναι και οι τρεις εξαίρετοι ερευνητές και εξαιρετικοί άνθρωποι. Συστηματικά και υπομονετικά κάθε εβδομάδα παρευρίσκονταν στις συναντήσεις μας, άκουγαν τις ιδέες μου και ας μην ήταν ώριμες τις περισσότερες φορές, μου χάριζαν τις συμβουλές τους και με καθοδήγησαν μέχρι και το τέλος αυτής της εργασίας. Μου έδειξαν πως γίνεται η έρευνα και μου έμαθαν ότι για να προχωράω πρέπει να υποστηρίζω τις ιδέες μου με στέρεα επιχειρήματα. Με τον Γιώργο μένω εντυπωσιασμένος από την μεθοδικότητα του· πάντα λεπτομερής σε βάθος, με καίριες ερωτήσεις, σχόλια και κατευθύνσεις για την συνέχεια. Ευχαριστώ πολύ τον Αλέξανδρο που με πολύ ουσιώδεις ερωτήσεις μου έκανε φανερό πολλές φορές τι δεν ήξερα αρκετά καλά και έπρεπε να μάθω καλύτερα. Ως υπεύθυνος της ομάδας αναγνώρισης ανθρώπινων ενεργειών στον Δημόκριτο, με προώθησε επίσης στα κατάλληλα άτομα και συγγράμματα που θα μπορούσα να συμβουλευτώ. Ο Νικόλας μου παρείχε σημαντική ώθηση τόσο σε θεωρητικά όσο και σε τεχνικά θέματα όσον αφορά τα νευρωνικά δίκτυα και είχε πολύ κρίσιμη συμμετοχή στην πορεία της εργασίας, ειδικά σε σημεία που πήγαινε να ξεφύγει από τον στόχο της. Είμαι πραγματικά ευγνώμων για το κουράγιο του να μπαίνει στο skype και να δίνει μια ευχάριστη παριζιάνικη νότα στις εβδομαδιαίες συναντήσεις μας. Μου προσέφεραν και οι τρεις ελεύθερα την βοήθεια τους, τους ευχαριστώ βαθύτατα και είμαι σίγουρος ότι θα μου λείψουν οι συναντήσεις μας.

Στο πλαίσιο της εργασίας αυτής, είχα επίσης την τύχη να γνωρίσω τον Δρ. Νίκο Κατζούρη, τον συγγραφέα και εμπνευστή ενός εργαλείου ελεύθερου λογισμικού που χρησιμοποιώ στην εργασία. Παρόλο που όλοι μου έλεγαν ότι σφύζει από δουλειά, εκείνος ήταν πάντα διαθέσιμος να μου εξηγήσει πράγματα, και σε πολλές περιπτώσεις ακόμα και να κάνει αλλαγές στον κώδικα για να διευκολύνει κάποια πειράματα μου. Μεγάλη ήταν επίσης η βοήθεια του Χρήστου Σμαϊλη που στην αρχή της ενασχόλησης μου με την όραση υπολογιστών. προσφέρθηκε να μου κάνει μια εισαγωγή στο θέμα με μια τηλεδιάσκεψη, αφού εκείνη την περίοδο βρισκόταν στο Τέξας στα πλαίσια του διδακτορικού του.

Ευχαριστώ τον καθηγητή μου από το Ε.Μ.Π. Ανδρέα Σταφυλοπάτη για την θετική του διάθεση να αναλάβει αυτήν την διπλωματική και επίσης για το γεγονός ότι μου συνέστησε να την γράψω στα Ελληνικά, επιτρέποντάς μου με αυτόν τον τρόπο να εμπλουτίσω την ελληνική βιβλιογραφία στο κομμάτι της αναγνώρισης ανθρώπινων ενεργειών. Έτσι, ήρθα σε επαφή με όρους σχετικούς με μηχανική μάθηση στην γλώσσα μου.

Σε ένα κείμενο ευχαριστιών θα ήταν απείρως μεγάλη αμέλεια να μην αναφέρω τους γονείς μου. Μου παρείχαν τα εφόδια για να εισέλθω αρχικά στην σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Η/Υ του Ε.Μ.Π. και με στήριξαν σε όλη μου την προσπάθεια, μέχρι και το τέλος των σπουδών μου. Είμαι αρκετά σίγουρος ότι δεν θα σταματήσουν να με στηρίζουν και στην συνέχεια της ζωής μου και το θεωρώ φυσικό να τους αφιερώνω αυτή την εργασία.

Τέλος, θα ήθελα να ευχαριστήσω την Δανάη που ήταν μαζί μου από την αρχή αυτής της εργασίας. Είναι τρελά αυθόρμητη, αισιόδοξη, διετέλεσε ως μούσα μου για αυτήν την εργασία και είμαι πολύ τυχερός να την έχω στην ζωή μου. Έβλεπε την μεγάλη εικόνα και πίστευε ότι θα βρω λύση ό,τι έκανα, ακόμα και όταν εγώ είχα απελπιστεί τελείως. Της χρωστάω πολλά ταξίδια που αμελήσαμε να κάνουμε τον τελευταίο καιρό.

Πίνακας περιεχομένων

Κατάλογος σχημάτων

Κατάλογος Πινάκων

1 Εισαγωγή

1.1 Αναγνώριση Ενεργειών σε Βίντεο

Ζούμε στην εποχή των «μεγάλων δεδομένων», όρος που αναφέρεται στην μεγάλη ποσότητα δεδομένων που παράγονται καθημερινά και με ρυθμό διαρκώς αυξανόμενο. Σύμφωνα με την IBM[1] κάθε μέρα παράγουμε 10^{18} bytes δεδομένων, αριθμός που συνεχίζει ακατάπαυστα να αυξάνεται. Ωστόσο, μεγάλος όγκος αυτών των δεδομένων παραμένει ανεκμετάλλευτος, εξαιτίας της αδόμητης φύσης των. Αυτό συμβαίνει σε μεγάλο βαθμό για τα βίντεο που υπολογίζεται[2] ότι μέχρι το 2020 θα αποτελούν το 82% της συνολικής κίνησης στο διαδίκτυο.

Η κατανόηση βίντεο αποτελεί κλάδο της υπολογιστικής όρασης. Η χρήση βαθιάς μάθησης (deep learning) στην υπολογιστική όραση πήρε μεγάλη ώθηση μετά την επιτυχία του [31] στο πολύ απαιτητικό dataset του Imagenet. Η επιτυχία των συνελικτικών νευρωνικών δικτύων (convolutional neural networks) έγκειται σε μεγάλο βαθμό στην ικανότητα τους να βρίσκουν μόνα τους κατάλληλα χαρακτηριστικά για την κατανόηση εικόνας, εκμεταλλευόμενα τις ιδιότητες των εικόνων για να μειώσουν τις εκπαιδεύσιμες παραμέτρους τους. Έτσι, καταφέρνουν να ξεπεράσουν σημαντικά σε αποδοτικότητα σύγχρονες μεθόδους που βασίζονται σε χειροποίητα χαρακτηριστικά. Στην κατανόηση βίντεο, δεν έχει επιτευχθεί ακόμη σημαντική υπεροχή, αφού οι τεχνικές βαθιάς μάθησης επιτυγχάνουν συγκρίσιμα αποτελέσματα με τεχνικές που βασίζονται σε τεχνικές σακιδίων λέξεων (bag-of-words) [40], που χρησιμοποιούν χειροποίητα χαρακτηριστικά [33, 11, 66, 65, 29, 53]. Ωστόσο φαίνεται να είναι θέμα χρόνου, καθώς η έρευνα έχει στραφεί σε μεγάλο βαθμό προς αυτήν την κατεύθυνση. Σημαντική ένδειξη αυτής της ροπής είναι το γεγονός ότι από τα συγγράμματα που έχουν γίνει

[1] https://www-01.ibm.com/software/in/data/bigdata/
[2] http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.html

δεκτά στον τομέα της μηχανικής μάθησης για υπολογιστική όραση στο CVPR 2017[3] (το δημοφιλέστερο συνέδριο όρασης υπολογιστων), η πλειοψηφία αφορά συνελικτικά νευρωνικά δίκτυα και βαθιά μάθηση.

Όμως, παρά την αποτελεσματικότητά τους στο να βγάζουν νόημα από πολλά και πολυδιάστατα δεδομένα, τα νευρωνικά δίκτυα υποφέρουν από κάποιους εγγενείς περιορισμούς, θεωρητικούς αλλά και πρακτικούς. Θεωρητικά, δεν έχει βρεθεί κάποια απόδειξη των συνθηκών που απαιτούνται για να συγκλίνουν σε καλά ελάχιστα. Πρακτικά επομένως, η έρευνα είναι οδηγούμενη από τα πειράματα και την εμπειρία. Ένας ακόμα βασικός τους περιορισμός έγκειται στην αδυναμία εξαγωγής ερμηνείας του μοντέλου που μαθαίνουν, γεγονός που συχνά τις χαρακτηρίζει ως αρχιτεκτονικές μαύρου κουτιού. Η ικανότητά τους για γενίκευση επαφίεται στην γενικότητα[4] του δεδομένων που έχουν χρησιμοποιηθεί κατά την φάση της εκπαίδευσης τους.

Από μια διαφορετική οπτική, έχει γίνει μεγάλη πρόοδος στην ανάλυση συμβολικών ροών δεδομένων χαμηλού επιπέδου για την εξαγωγή υψηλότερου επιπέδου συμπερασμάτων. Όσον αφορά τις συμβολικές ροές δεδομένων, η ύπαρξη τους σε δεδομένα βίντεο δεν είναι προφανής, αλλά επιτακτική σε εφαρμογές που απαιτούν λήψη αποφάσεων σε πραγματικό χρόνο, όπως είναι η επιτήρηση, η αναγνώριση κατάχρησης σε ζωντανές ροές βίντεο αλλά και η υποβοηθούμενη καθημερινή ζωή για άτομα χρήζοντα βοήθειας.

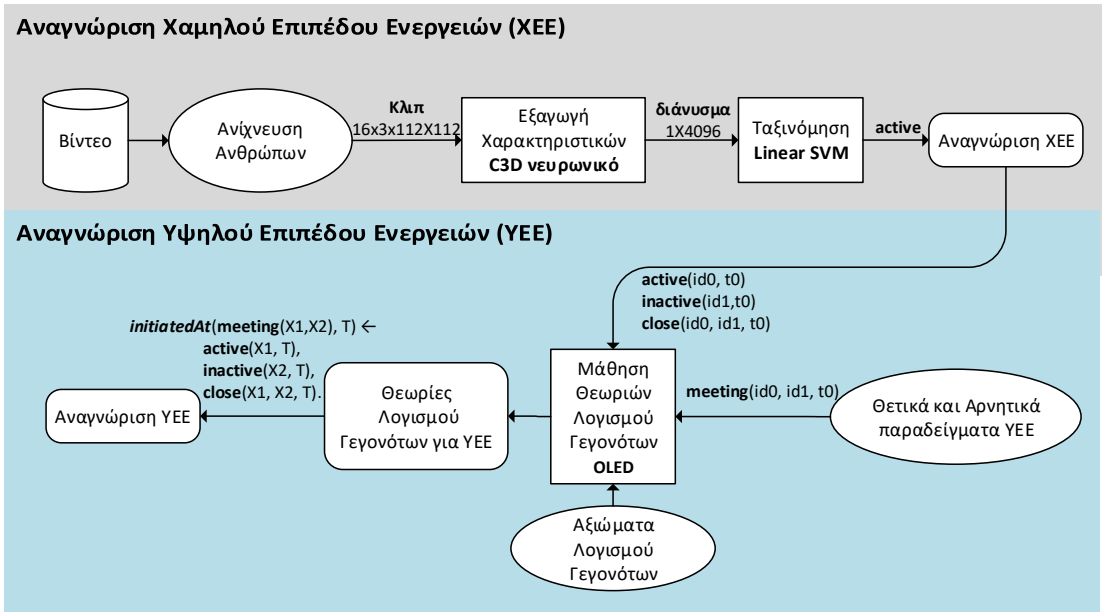
Σε αυτήν την εργασία μελετάμε την χρήση του Λογισμού Γεγονότων [30] ως μέθοδο επεξεργασίας συμβολικών δεδομένων. Τα σύμβολα που περιέχουν οι ροές έχουν προκύψει από κάποιον ταξινομητή που δέχεται στην είσοδο βαθιά μαθημένα χαρακτηριστικά βίντεο. Υποστηρίζουμε, είναι σημαντικό να ενταχθεί η Υπολογιστική Λογική στην διαδικασία μάθησης των βαθιών αρχιτεκτονικών νευρωνικών δικτύων προκειμένου να γίνουν τα αποτελέσματα τους περισσότερο ερμηνεύσιμα, η μάθηση τους περισσότερο διαφανής και η μεροληψία τους περισσότερο διαχειρίσιμη.

1.2 Αντικείμενο διπλωματικής

Στην διπλωματική αυτή μελετήσαμε τρόπους αναγνώρισης απλών ανθρώπινων ενέργειψν και πως αυτές πλέον ως ροές συμβολικών δεδομένων θα μπορούν να χρησιμοποιηθούν από μεθόδους λογικού χρονικού συμπερασμού, , ώστε να αναγνωριστούν σύνθετες ενέργειες μεταξύ περισσότερων ανθρώπων. Για την ανάλυση βίντεο εξετάσαμε την χρήση συνελικτικών αυτοκωδικοποιητών και χρησιμοποιήσαμε χαρακτηριστικά C3D [20] από προεκπαιδευμένο βαθύ νευρωνικό δίκτυο σε συνδυασμό με μηχανές διανυσμάτων υποστήριξης (Support Vector

[3] http://vision.cse.psu.edu/people/chrisF/cvpr_2017/primary_graph_accepted.html
[4] https://blogs.wsj.com/digits/2015/07/01/google-mistakenly-tags-black-people-as-gorillas-showing-limits-of-algorithms/

Machines - SVMs) για την ταξινόμησή τους σε κατηγορίες ενεργειών. Στην συνέχεια, έχοντας τις αναγνωρισμένες απλές ενέργειες, τις χρησιμοποιήσαμε ως είσοδο σε ένα σύστημα μάθησης κανόνων θεωριών Λογισμού Γεγονότων (OLED), ώστε να μπορέσουμε να αναγνωρίσουμε σύνθετες ενέργειες που αφορούν αλληλεπίδραση ατόμων. Αξιολογήσαμε πειραματικά την μέθοδο μας στο σύνολο δεδομένων CAVIAR[5] που ικανοποιεί τις ανάγκες μας για επισημείωση τόσο απλών, όσο και σύνθετων ανθρώπινων ενεργειών. Σε όλα τα πειράματα μας θεωρήσαμε την ανίχνευση ανθρώπων δεδομένη. Η πλήρης πρότασή μας για την διαδικασία αναγνώρισης ανθρώπινων ενεργειών περιγράφεται στην εικόνα 1.1.



Εικόνα 1.1: Προτεινόμενη από την εργασία, διαδικασία, για την αναγνώριση σύνθετων ανθρώπινων ενεργειών σε βίντεο

1.2.1 Συνεισφορά

Η συνεισφορά της διπλωματικής συνοψίζεται ως εξής:

1. Μειώσαμε το χάσμα που υπάρχει ανάμεσα στην βαθιά μάθηση και την λογική, προτείνοντας μια αρχιτεκτονική μάθησης δύο επιπέδων.
2. Μελετήσαμε και αξιολογήσαμε την χρήση C3D χαρακτηριστικών για βραχυπρόθεσμες ενέργειες τόσο στο σύνολο δεδομένων CAVIAR όσο και στο KTH [52].
3. Μελετήσαμε και αξιολογήσαμε την απόδοση του OLED, ενός συστήματος ικανού να μαθαίνει θεωρίες Λογισμού γεγονότων, σε ατελή δεδομένα που έχουν προκύψει αυτόματα από κάποιον ταξινομητή.

[5] http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/

4. Υλοποιήσαμε διάφορες απλές αρχιτεκτονικές συνελικτικών αποκωδικοποιητών και αξιολογήσαμε την χρήση τους για αναπαράσταση καρέ από βίντεο.

1.3 Οργάνωση κειμένου

Η υπόλοιπη διπλωματική οργανώνεται ως εξής:

Στο 2ο Κεφάλαιο παρουσιάζουμε παρόμοιες εργασίες, τόσο στον τομέα της αναγνώρισης ανθρώπινων ενεργειών με αριθμητικές μεθόδους, όσο και στον τομέα της αναγνώρισης σύνθετων ενεργειών από ροές συμβολικών δεδομένων.

Στο 3ο Κεφάλαιο παρουσιάζουμε τα βασικά μέρη των συνελικτικών νευρωνικών δικτύων και των αρχιτεκτονικών που θα χρησιμοποιήσαμε. Επίσης κάνουμε μια επισκόπηση του Λογισμού Γεγονότων (ΕC) και του συστήματος που θα χρησιμοποιήσουμε για την παραγωγή θεωριών ΕC.

Στο 4ο Κεφάλαιο παρουσιάζουμε για τα πειράματα που εκτελέσαμε, την υλοποίηση τους, τα αποτελέσματα που πήραμε και τα συμπεράσματα που εξαγάγαμε.

Στο 5ο Κεφάλαιο ολοκληρώνουμε συνοψίζοντας την εργασία και προτείνοντας κατευθύνσεις για το μέλλον.

2 *Σχετικές εργασίες*

Σκοπός της εργασίας αυτής είναι από ανεπεξέργαστα δεδομένα βίντεο να καταλήξουμε σε συμβολικά δεδομένα που θα μπορέσουν να αποτελέσουν είσοδο σε ένα σύστημα χρονικού λογικού συμπερασμού. Επομένως οι περιοχές ενδιαφέροντος είναι κυρίως δύο:

α) Με είσοδο βίντεο δεδομένα, θέλουμε στην έξοδο να πάρουμε αναγνωρισμένες Χαμηλού Επιπέδου Ενέργειές (ΧΕΕ), όπως ενεργός, ανενεργός, περπάτημα, τρέξιμο, κλπ.

β) Με είσοδο τις αναγνωρισμένες απλές ενέργειες, θέλουμε η τελική έξοδος να μας δώσει Υψηλού Επιπέδου Ενέργειες (ΥΕΕ) που εν γένει αφορούν αλληλεπίδραση ατόμων, όπως συνάντηση, ομαδική κίνηση, πάλη, κλπ.

2.1 Αναγνώριση απλών ανθρώπινων ενεργειών σε βίντεο

Παραφράζοντας τον ορισμό από το [17], ενέργειες είναι «οι πιο βασικές ανθρωποκεντρικές αλληλεπιδράσεις με νόημα». Το πρόβλημα της αναγνώρισης ενεργειών έχει μελετηθεί πολύ και έχει προσεγγιστεί με πολλούς και διάφορους τρόπους. Έχουν χρησιμοποιηθεί υψηλού επιπέδου αναπαραστάσεις που μοντελοποιούν την πόζα [64], το σχήμα [5] ή τα άτομα που εμπλέκονται ως κινούμενους σκελετούς [63]. Στο [50] οι ενέργειες περιγράφονται από ένα μεγάλο σύνολο ανιχνευτών που χρησιμοποιούνται ως βάση για την κατασκευή ενός πολυδιάστατου «χώρου ενεργειών». Ωστόσο, υποστηρίζεται [35] ότι τέτοιες υψηλού επιπέδου αναπαραστάσεις δεν είναι ικανές να συλλάβουν λεπτεπίλεπτες κινήσεις και η έρευνα οδηγείται από σπουδαία αποτελέσματα με χειροποίητους (hand-crafted) τοπικούς περιγραφείς ή αρχιτεκτονικές βαθιάς μάθησης.

Ως χειροποίητα χαρακτηριστικά εννοούμε εκείνα που έχουν προκύψει από την εμπειρία και την γνώση των ειδικών στην όραση υπολογιστών. Δεν σημαίνει ότι χρειάζεται δουλειά με το χέρι για την εξαγωγή τους, αλλά ότι η σύλληψη τους έγινε με βάση το τι διαισθητικά μπορεί να αποτελέσει καλή αναπαράσταση αμετάβλητη σε ασήμαντες σημασιολογικά διαφορές ανάμεσα σε καρέ, όπως αλλαγές κλίμακας, οπτικής, φωτεινότητας, περιστροφής, μετάφρασης, θόρυβος στο παρασκήνιο ή στο προσκήνιο. Σε αυτό το πλαίσιο χρησιμοποιούνται περιγραφείς όψης όπως τα ιστογράμματα προσανατολισμένων κλίσεων (HOG) [11], οι ανεξάρτητοι κλίμακος μετασχηματισμοί (SIFT) [33] που για βίντεο επεκτείνονται από τον χώρο στον χωρόχρονο σε SIFT3D [53] και HOG3D [29]. Για να

συμπεριλάβουν πληροφορία κίνησης συνδυάζονται με περιγραφείς ροής όπως τα ιστογράμματα οπτικής ροής (HOF) [32] και τα ιστογράμματα οριακής κίνησης (MBH) [65] . Στα βίντεο, μεγάλη επιτυχία έχουν δείξει τα χαρακτηριστικά πυκνών τροχιών [65] που ήταν ο προκάτοχος των βελτιωμένων πυκνών τροχιών (iDT) [66], οι οποίες βελτιώνονται λαμβάνοντας επίσης υπόψη και την κίνηση της κάμερας. Γενικά, όλα αυτά τα χαρακτηριστικά συνδυάζονται με τεχνικές σακιδίων λέξεων [40] ή διανυσμάτων Fisher [45, 46] και επιτυγχάνουν αποτελέσματα που βρίσκονται πολύ κοντά στα καλύτερα που αφορούν ανάλυση βίντεο. Ωστόσο, τα τελευταία χρόνια φαίνεται να χάνουν έδαφος από χαρακτηριστικά που προκύπτουν από βαθιά μάθηση.

Σαν βαθιά μάθηση εννοούμε τον κλάδο της μηχανικής μάθησης που περιλαμβάνει στατικά μοντέλα νευρωνικών δικτύων πολλών επιπέδων. Μια σημαντική συνεισφορά στον χώρο της όρασης υπολογιστών ήταν τα Συνελικτικά Νευρωνικά Δίκτυα (ΣΝΔ), σχεδιασμένα αρχικά να αναγνωρίζουν χειρόγραφους χαρακτήρες [38]. Με την κατασκευή μεγαλύτερων συνόλων δεδομένων, αλλά και λόγω τεχνολογικής προόδου φτάσαμε στο [31] που ξεπέρασε κάθε ανταγωνισμό στο πλέον απαιτητικό για την εποχή Imagenet. Τα βαθιά συνελικτικά δίκτυα έχουν χρησιμοποιηθεί για την εξαγωγή χαρακτηριστικών για ανάλυση εικόνας ή βίντεο. Ωστόσο, ένας φανερός περιορισμός τους για την ανάλυση βίντεο είναι ότι δεν λαμβάνουν από μόνα τους υπόψη τον χρόνο. Οι Simonyan και Zisserman [56] το είδαν σαν πρόβλημα που απαιτεί δύο παράλληλα νευρωνικά, ένα χωρικό που δέχεται ως είσοδο καρέ του βίντεο και ένα χρονικό που δέχεται HOF ως είσοδο. Στην δουλειά των Schindler και Van Gool [51] παρουσιάζεται ότι η αναγνώριση ενεργειών μπορεί να γίνει με πληροφορία από 1-7 καρέ. Έτσι, κάποιοι [20, 61] εισήγαγαν χρονική μάθηση επεκτείνοντας σε συνελίξεις 3Δ αντί για συνελίξεις 2Δ. Αρχικά έγινε έκδηλη η αποτελεσματικότητα 3Δ συνελικτικών δικτύων για αναγνώριση ανθρώπινων ενεργειών [20] από μόνα τους ή σε συνδυασμό με νευρωνικά μακρο-βραχυπρόθεσμης μνήμης (LSTM) [3], ένα είδος αναδρομικού νευρωνικού ικανό να συγκρατεί χρονική πληροφορία. Με μία λεπτομερή μελέτη στο [61] χρησιμοποιούν 3Δ συνελικτικά δίκτυα ως γενικό εξαγωγέα χαρακτηριστικών (C3D). Ο Montes [37] πετυχαίνει αξιόλογη επίδοση στο πολύ απαιτητικό ActivityNet, χρησιμοποιώντας τα C3D χαρακτηριστικά σαν είσοδο σε ένα LSTM τόσο για αναγνώριση ενεργειών όσο και χρονικό εντοπισμό τους, χωρίς να χρειαστεί μάθηση του νευρωνικού από το οποία προέκυψαν, στο dataset στο οποίο εφαρμόστηκε. Γενικά, η χρήση τους έχει περιοριστεί κυρίως σε ταξινόμηση ολόκληρων βίντεο και όχι αυτόνομων απλών ενεργειών, όπως κοιτάμε σε αυτήν την εργασία.

Εφόσον ένα μεγάλο πρόβλημα των δεδομένων βίντεο είναι η πολύ ακριβή επισημείωση τους, μεγάλο ενδιαφέρον παρουσιάζουν δουλειές που αφορούν μη επιβλεπόμενη μάθηση καλών αναπαραστάσεων για βίντεο. Μια καλή αναπαράσταση, πρέπει είναι γενική για κάθε πρόβλημα και να μην μεταβάλλεται πολύ, με μικρές αλλαγές της εισόδου (αλλαγές

κλίμακας, οπτικής, φωτεινότητας, περιστροφής, μετάφρασης, θόρυβος στο παρασκήνιο ή στο προσκήνιο). Για έναν πλήρη κατάλογο του τι κάνει μια αναπαράσταση καλή, σας προτρέπουμε να κοιτάξετε το [4]. Κάνοντας λανθάνουσα πιθανοτική σημασιολογική ανάλυση (pLSA) στο [39], μαθαίνουν την αντιστοίχιση των κωδικολέξεων που έχουν προκύψει από χειροποίητα τοπικά χαρακτηριστικά σε λανθάνουσες (latent) επισημάνσεις ενεργειών. Στο [60] προτείνεται μάθηση με συνελικτικά επίπεδα ως μέθοδος για εξαγωγής χαρακτηριστικών. Υπό μια διαφορετική οπτική, στο [16] το λάθος ανασχηματισμού ενός συνελικτικού αυτοκωδικοποιητή φανερώνει στατιστικά ανώμαλα καρέ σε ένα βίντεο και παρουσιάζονται σαφώς καλύτερα αποτελέσματα όταν το δίκτυο χρησιμοποιείται σε ανεπεξέργαστα βίντεο από ότι όταν η είσοδος είναι πλέον σύγχρονα χειροποίητα χαρακτηριστικά εικόνας. Χρησιμοποιώντας ότι σε ένα βίντεο διαδοχικά καρέ θα περιέχουν σημασιολογικά παρόμοια πληροφορία στο [15] προτείνεται μια μέθοδος να συμπεριληφθεί αυτή η χρονική συνάφεια στην μάθηση ενός αυτοκωδικοποιητή. Στο [59] φαίνεται αυτό να γίνεται αποτελεσματικά με χρήση αυτοκωδικοποιητών αποτελούμενων από LSTM νευρώνες, που είναι ικανοί να συγκρατήσουν χρονική πληροφορία.

2.2 Αναγνώριση σύνθετων ενεργειών από συμβολικές ροές

Σε πολλές περιπτώσεις αφήνεται στα συστήματα βαθιάς μάθησης να κάνουν πλήρη αναγνώριση σύνθετων ενεργειών μη κάνοντας διαχωρισμό πολύπλοκων και απλών. Όμως, ένα μεγάλο τους πρόβλημα είναι ότι είναι αδιαφανείς διαδικασίες που κάνουν λάθη και η γενικότητα τους επαφίεται στην γενικότητα των δεδομένων τα οποία βλέπουν κατά την φάση της εκπαίδευσης. Γνωρίζουμε όμως ότι κάθε σύνολο δεδομένων υποφέρει από εγγενείς μεροληψίες που είναι αδύνατο να εξαλειφθούν. Σε αυτό το σημείο προτείνουμε τον συγκερασμό τέτοιων αδιαφανών διαδικασιών μάθησης με πιο διαφανείς μεθόδους. Στο πλαίσιο αυτό βλέπουμε συστήματα αναγνώρισης σύνθετων γεγονότων, ένα πεδίο που αφορά την ανίχνευση μοτίβων γεγονότων σε χρονικά δεδομένα. Μερικά παραδείγματα εφαρμογών αποτελούν η αναγνώριση ανθρώπινης δραστηριότητας σε βίντεο [7], διαχείριση μετακινήσεων και συμφόρησης [2], ανίχνευση παραβιάσεων ασφάλειας υπολογιστικών δικτύων [12], παρακολούθηση θαλάσσιων μεταφορών [43].

Στο [9] προτείνεται η χρήση Τυχαίων Μαρκοβιανών Πεδίων, ενώ έχουν χρησιμοποιηθεί επίσης Λανθάνοντα Μαρκοβιανά Μοντέλα (HMM) [67] να αναγνωρίζουν ενέργειες από συμβολοσειρές που έχουν απευθείας παραχθεί από τα χαρακτηριστικά των καρέ. Σαν αναγνώριση συμβολοσειρών βλέπουν το πρόβλημα επίσης στο πιο πρόσφατο [23] και χρησιμοποιούν κανονικές εκφράσεις για την αναγνώριση ενεργειών. Η αναγνώριση ενεργειών μετατρέπεται σε πρόβλημα ομοιότητας γράφων στο [57].

Ωστόσο, προσεγγίσεις βασισμένες στην λογική υπερτερούν παρουσιάζοντας τυπική δηλωτική σημασιολογία, το οποίο είναι μεγάλο προτέρημα σε εφαρμογές που χρειάζεται να εντοπιστούν και επιβεβαιωθούν σημεία της διαδικασίας αναγνώρισης γεγονότων [41, 42]. Οι Tran και Davis [62] αντιμετωπίζουν το πρόβλημα με Λογικά Μαρκοβιανά Δίκτυα [49], όμως χρειάζεται να κατασκευάσουν κανόνες και να τους αποδώσουν βάρη. Το πρόβλημα εξειδικεύεται περισσότερο με την χρήση ειδικών φορμαλισμών για αναγνώριση γεγονότων, όπως ο Λογισμός Γεγονότων [30] που εφαρμόζεται για πραγματικού χρόνου αναγνώριση γεγονότων στο με το RTEC [1] και προσφέρει σύνδεση με μηχανική μάθηση για την εξαγωγή των κανόνων. Στην εργασία [36], οι Μιχελιουδάκης και αλ. λύνουν το πρόβλημα της δημιουργίας των κανόνων στα MLN με δομική μάθηση, χρησιμοποιώντας τα αξιώματα του Λογισμού Γεγονότων. Οι Κατζούρης και αλ. χρησιμοποιεί επαγωγικό λογικό προγραμματισμό για την μάθηση προγραμμάτων Λογισμού Γεγονότων [25] και επεκτείνει με το [26] που είναι ικανό να μάθει θεωρίες με θορυβώδη δεδομένα και θα το δούμε περισσότερο στην συνέχεια της εργασίας.

3 Θεωρητικό υπόβαθρο

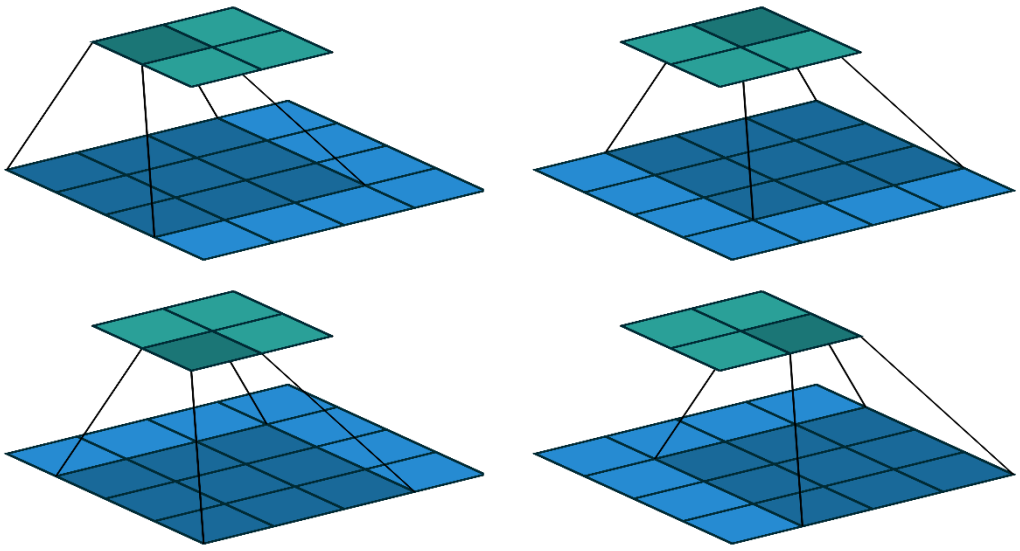
3.1 Συνελικτικά Νευρωνικά Δίκτυα (ΣΝΔ)

Τα ΣΝΔ είναι ένα είδος νευρωνικών δικτύων που έχουν αποδειχθεί χρήσιμα για την όραση υπολογιστών. Είναι ελαφρώς βιολογικά εμπνευσμένα από την λειτουργία των ματιών [13], χρησιμοποιήθηκαν πρωταρχικά επιτυχώς στο [38] , αλλά η ευρεία αποδοχή τους στην ανάλυση εικόνας ήρθε μετά το [31].

Ένα συνελικτικό επίπεδο αποτελείται από ν φίλτρα $f_1, f_2, \dots, f_v$ τα οποία συνελίσσονται χωρικά με την εικόνα εισόδου x, για να δημιουργήσουν στην έξοδο ν δισδιάστατους χάρτες χαρακτηριστικών z:

$$z_k = f_k * x, \kappa \in \{1..v\}$$

Με $*$ εννοούμε την πράξη της συνέλιξης όπως ορίζεται για ΣΝΔ [14]. Στην εικόνα 3.1 μπορούμε να δούμε πως μοιάζει μια συνέλιξη.



Εικόνα 3.1: Εδώ βλέπουμε μία συνέλιξη μιας εισόδου 4x4x1 με ένα φίλτρο 3x3x1. Το αποτέλεσμα είναι ένας χάρτης 2x2. Πηγή[6].

[6] https://github.com/vdumoulin/conv_arithmetic

Είναι ιδιαίτερα χρήσιμα για προβλήματα όρασης, λόγω δύο βασικών περιορισμών που εισάγουν, ώστε να λάβουν υπόψη την χωρική δομή των εικόνων μειώνοντας παράλληλα σημαντικά των αριθμό των εκπαιδεύσιμων παραμέτρων του δικτύου.

- **Τοπική συνδεσιμότητα.** Τα φίλτρα είναι συνήθως πολύ μικρότερα σε διαστάσεις από τις εικόνες στις οποίες εφαρμόζονται, με αποτέλεσμα να καθίστανται ικανά να ανιχνεύουν τοπικά χαρακτηριστικά με νόημα, όπως γραμμές και γωνίες
- **Κοινή χρήση βαρών.** Με την συνέλιξη χρησιμοποιούμε το κάθε φίλτρο σε όλη την εικόνα. Έτσι, μαθαίνουμε μόνο ένα σύνολο παραμέτρων για κάθε τοποθεσία, αντί για πολλά διαφορετικά όπως στην περίπτωση των πλήρως συνδεδεμένων νευρωνικών δικτύων.

Συγκέντρωση (Pooling): Μία πολύ συνήθης πράξη στον χώρο των ΣΝΔ είναι η συγκέντρωση, που πρακτικά συνοψίζει στατιστικά κοντινές εξόδους του προηγούμενου επιπέδου σε μια τετραγωνική περιοχή.

Συγκεκριμένα σε μια περιοχή S ορίζουμε ως συγκέντρωση συνάρτησης $p: R^v \rightarrow R$

$$p_S = p(\, s_i \in S)$$

Συνήθως προτιμάται η συγκέντρωση μεγίστου (max pooling), ωστόσο είναι επίσης δημοφιλής η χρησιμοποίηση σταθμισμένου ή όχι μέσου όρου (average pooling).

Σε κάθε περίπτωση η συγκέντρωση βοηθάει την αναπαράσταση να μένει αμετάβλητη σε μικρές μεταφράσεις της εικόνας εισόδου, μειώνοντας ταυτόχρονα τις διαστάσεις της.

3.1.1 Κανονικοποίηση (Regularization)

Μεγάλα και βαθιά νευρωνικά δίκτυα μπορούν να προσεγγίσουν οσοδήποτε περίπλοκη συνάρτηση. Επομένως, χωρίς τους κατάλληλους περιορισμούς μπορεί να προσαρμοστούν πλήρως στο σύνολο δεδομένων εκπαίδευσης με αποτέλεσμα να μην είναι καλά στην γενικοποιήση σε δοκιμαστικά ή πραγματικά δεδομένα. Αυτό το πρόβλημα ονομάζεται υπερπροσαρμογή (overfitting) και για τα νευρωνικά έχουν προταθεί διάφοροι τρόποι για να ξεπεραστεί.

Κανονικοποίηση L2 Είναι η πιο κλασσική μορφή κανονικοποίησης. Σύμφωνα με αυτή, για κάθε βάρος *w,* προστίθεται ένας όρος $\frac{1}{2}\lambda w^2$ στην συνάρτηση κόστους, για να αποτρέψει το δίκτυο από το να μοντελοποιήσει πλήρως τα δεδομένα εκπαίδευσης. Τότε η συνάρτηση κόστους γίνεται:

$$L(x,y)_{new} = L(x,y) + \frac{1}{2}\lambda w^2$$

Περιορισμός Ενεργοποίησης (Dropout) Μια πιο πρόσφατη προσέγγιση [58] περιλαμβάνει τον περιορισμό ενεργοποίησης. Σε κάθε βήμα της διαδικασίας εκπαίδευσης ένα ποσοστό των

νευρώνων αναγκάζεται να παραμείνει ανενεργό. Έτσι, αποφεύγεται να προσαρμόζονται μεταξύ τους τα διαφορετικά βάρη. Κατά την φάση της δοκιμής (testing), όλοι οι νευρώνες συμμετέχουν στο αποτέλεσμα.

3.1.2 Μεταφερόμενη μάθηση (Transfer learning)

Γενικά, η εκπαίδευση των νευρωνικών είναι μια επίπονα χρονοβόρα διαδικασία που απαιτεί καλούς υπολογιστικούς πόρους και πολλά επισημασμένα δεδομένα που σε πολλές περιπτώσεις δεν είναι διαθέσιμα. Ωστόσο, η από την αρχή εκπαίδευση ενός νευρωνικού δεν είναι πάντα απαραίτητη και στις περιπτώσεις που δεν έχουμε πολλά δεδομένα αποδεικνύεται χειρότερη [24] από κάποιο τύπο μεταφερόμενης μάθησης που χρησιμοποιεί κάποιο προ-εκπαιδευμένο νευρωνικό.

3.1.2.1 Εξαγωγή χαρακτηριστικών (Feature extraction)

Σε αυτό το σενάριο θεωρούμε ένα ήδη εκπαιδευμένο νευρωνικό δίκτυο σε άλλο σύνολο δεδομένων από αυτό στο οποίο θα το χρησιμοποιήσουμε. Τροφοδοτούμε το νευρωνικό με παραδείγματα από το δικό μας σύνολο δεδομένων και επιλέγουμε την έξοδο κάποιου από τα επίπεδα του, ως διάνυσμα χαρακτηριστικών των παραδειγμάτων μας.

Στην συνέχεια αυτά τα χαρακτηριστικά μπορούν να αποτελέσουν είσοδο για κάποιον άλλο ταξινομητή, όπως ένα μοντέλο κ-κοντινότερων γειτόνων ή μια μηχανή διανυσμάτων υποστήριξης (SVM).

3.1.2.2 Προσαρμογή (Fine-tuning)

Σε αυτό το σενάριο θεωρούμε ένα ήδη εκπαιδευμένο νευρωνικό δίκτυο, το οποίο όμως συνεχίζουμε να εκπαιδεύουμε στο σύνολο δεδομένων που χρησιμοποιούμε. Έχει δειχτεί ότι γενικά στα πρώτα συνελικτικά επίπεδα τα δίκτυα μαθαίνουν γενικές αναπαραστάσεις, όποτε συνήθως τα πρώτα επίπεδα αφήνονται ως έχουν και εκπαιδεύονται τα επόμενα. Στις περισσότερες περιπτώσεις τα τελευταία πλήρως συνδεδεμένα κομμάτια του νευρωνικού που είναι εξειδικευμένα στο αρχικό έργο κατάταξης, αρχικοποιούνται εκ νέου. Ανάλογα με το όγκο των δεδομένων που έχουμε στην διάθεση μας θέτουμε μικρότερο ή μεγαλύτερο δείκτη μάθησης και ορίζουμε ως εκπαιδεύσιμα λιγότερα ή περισσότερα επίπεδα.

3.1.3 Οπτική Ερμηνεία Μάθησης

Έχουμε αναφερθεί πολλές φορές μέχρι τώρα στην αδυναμία εξαγωγής ερμηνείας της μάθησης των νευρωνικών, καθώς και της διαδικασίας μέσω της οποίας λαμβάνουν αποφάσεις (αναθέτουν την είσοδο σε κάποια κλάση). Για την ανάπτυξη μιας διαίσθησης στο κομμάτι αυτό, έχουν προταθεί διάφορες απεικονιστικές μέθοδοι.

Σε αυτό το πλαίσιο, μια ενδιαφέρουσα τεχνική που έχει χρησιμοποιηθεί είναι η *t-κατανεμημένη στοχαστική ενσωμάτωση* γειτόνων (t-SNE) [34] που λαμβάνει στην είσοδο τα χαρακτηριστικά από ένα συνελικτικό δίκτυο και τα προβάλλει στον δισδιάστατο χώρο, προσπαθώντας να τηρήσει τις αποστάσεις που είχαν στον πολυδιάστατο χώρο. Έτσι προκύπτουν εικόνες σαν την Εικόνα 3.2.

Εικόνα 3.2: Εικόνες του ImageNet σε 2Δ, χρησιμοποιώντας το t-SNE και χαρακτηριστικά εξαγμένα από ένα εκπαιδευμένο ΣΝΔ με εικόνες του συνόλου δεδομένων CIFAR. Πηγή[7].

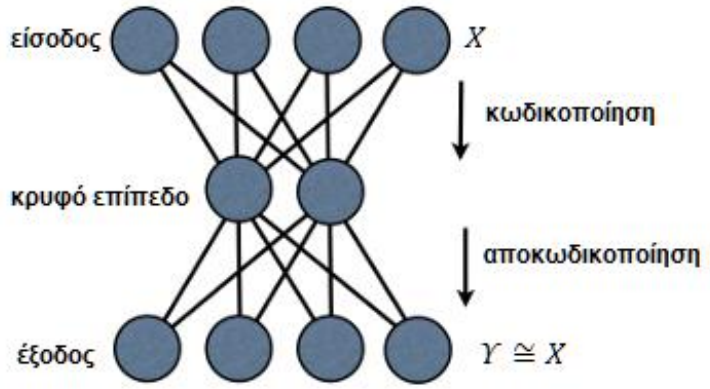
[7] http://cs.stanford.edu/people/karpathy/cnnembed/

3.1.4 Συνελικτικές αρχιτεκτονικές που θα χρησιμοποιήσουμε

3.1.4.1 Συνελικτικός Αυτοκωδικοποιητής (Convolutional Autoencoder)

Ο αυτοκωδικοποιητής είναι ένας είδος νευρωνικού δικτύου που ανήκει στην κατηγορία των μοντέλων μη επιβλεπόμενης ή καλύτερα αυτό-επιβλεπόμενης μάθησης.

Είναι ένα δίκτυο του οποίου οι παράμετροι προσαρμόζονται, επιβάλλοντας η έξοδος Y του να είναι ίδια με την είσοδο X ($X \cong Y$) ή αλλιώς ελαχιστοποιώντας το λάθος ανασχηματισμού $L(X,Y)$, όπως ονομάζεται δηλαδή η συνάρτηση κόστους που αφορά τους αυτοκωδικοποιητές. Ενδιάμεσα στην είσοδο και την έξοδο παρεμβάλλονται κρυφά επίπεδα όπως στην Εικόνα 3.2 που παρουσιάζει ένα πολύ απλό δίκτυο αυτοκωδικοποιητή.



Εικόνα 3.3: Απλό σχήμα αυτοκωδικοποιητή με 1 επίπεδο εισόδου, 1 κρυφό επίπεδο και 1 επίπεδο εξόδου

Για να σιγουρευτούμε ότι δεν θα μάθουμε την ταυτοτική συνάρτηση, θέτουμε τουλάχιστον έναν ή περισσότερους περιορισμούς:

Περιορισμός ενδιάμεσου επιπέδου: Θέτουμε σε κάποιο ενδιάμεσο επίπεδο, αριθμό παραμέτρων μικρότερο από αυτό της εισόδου.

Αραιότητα (Sparsity): Θέτουμε κάποιο περιορισμό στον αριθμό «νευρώνων» που μπορούν να ενεργοποιηθούν μια δεδομένη στιγμή.

Αποτελείται από δύο μέρη:

- Τον κωδικοποιητή που εισάγει τους παραπάνω περιορισμούς που συμπιέζουν την είσοδο
- Τον αποκωδικοποιητή που κάνει την ανάστροφη πράξη από τον κωδικοποιητή προσπαθώντας να ανασχηματίσει την είσοδο στο επίπεδο εξόδου

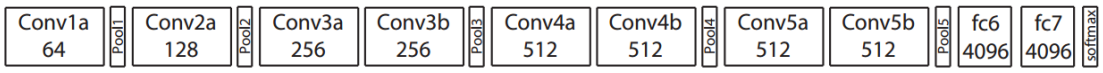
Ο συνελικτικός αυτοκωδικοποιητής έχει ως συστατικά μέρη συνελικτικά επίπεδα και επομένως είναι σχετικός σε προβλήματα που όρασης.

3.1.4.2 Χαρακτηριστικά από 3Δ Συνελικτικό νευρωνικό δίκτυο εκπαιδευμένο στο σύνολο δεδομένων SPORTS1M

Επειδή η εργασία αυτή αφορά την αναγνώριση ενεργειών από βίντεο, είναι σημαντικό η αναπαράσταση που θα επιλέξουμε εκτός από χωρική πληροφορία να περιλαμβάνει επίσης χρονική πληροφορία.

Σε αυτό το πλαίσιο θα κάνουμε χρήση του C3D [61], ενός συνελικτικού δικτύου που για να το κάνει αυτό, χρησιμοποιεί συνελικτικά φίλτρα τριών διαστάσεων. Στην είσοδο λαμβάνει ένα απόσπασμα του βίντεο των 16 καρέ. Έχει χρησιμοποιηθεί κυρίως παράγοντας αναπαραστάσεις σε επίπεδο βίντεο μετά την συγκέντρωση των χαρακτηριστικών των αποσπασμάτων που το αποτελούν. Παρουσιάζει μεγάλη ικανότητα γενίκευσης βγάζοντας πρώτης τάξεως αποτελέσματα σε πολλά απαιτητικά σύνολα δεδομένων [61, 37].

Το δίκτυο C3D αποτελείται από 8 συνελικτικά επίπεδα, 5 συγκεντρωτικά επίπεδα, 2 πλήρως συνδεδεμένα και ένα επίπεδο Softmax για την έξοδο στο τέλος. Τα συνελικτικά επίπεδα έχουν 3x3x3 πυρήνες με βήμα παραθύρου ίσο με 1. Όλα τα συγκεντρωτικά επίπεδα είναι συγκεντρωτικά μεγίστου σε πυρήνες 2x2x2, εκτός από το πρώτο που έχει μέγεθος πυρήνα 2x2x1. Τα δύο πλήρως συνδεδεμένα επίπεδα (fc6 και fc7) έχουν από 4096 νευρώνες το καθένα, ενώ η έξοδος Softmax έχει 487 εξόδους, όσες δηλαδή είναι και οι κλάσεις του SPORTS1M [24] στο οποίο έχει εκπαιδευτεί. Η αρχιτεκτονική του δικτύου φαίνεται στην Εικόνα 3.4, ενώ μερικά παραδείγματα βίντεο του SPORTS1M απεικονίζονται στην Εικόνα 3.5.



Εικόνα 3.4: Δίκτυο C3D. Τα 8 συνελικτικά επίπεδα έχουν το πρόθεμα Conv και ο περιεχόμενος στα κουτιά αριθμός υποδηλώνει των αριθμό των φίλτρων. Τα 5 συγκεντρωτικά επίπεδα μεγίστου παρουσιάζονται με το πρόθεμα Pool. Τα πλήρως συνδεδεμένα επίπεδα υποδηλώνονται με το πρόθεμα fc έχουν και τα 2 από 4096 μονάδες εξόδου. Τέλος ακολουθεί η έξοδος Softmax.

Θεωρούμε ότι επειδή είναι ικανό να κάνει αναγνώριση σύνθετων ενεργειών σε επίπεδο βίντεο, σε κάποιο ενδιάμεσο στάδιο τα χαρακτηριστικά που παράγει μπορεί να είναι ικανή αναπαράσταση για την αναγνώριση απλών ενεργειών. Θα το χρησιμοποιήσουμε για εξαγωγή χαρακτηριστικών, προωθώντας τα δικά μας παραδείγματα και παίρνοντας την έξοδο από το fc6 επίπεδο.

Εικόνα 3.5: Παραδείγματα βίντεο από το SPORTS1M

3.2 Λογισμός Γεγονότων (EC)

Ο Λογισμός Γεγονότων [30] είναι μια τυπική γλώσσα πρώτης τάξης για συμπερασμό γεγονότων και των παρενεργειών τους, η οποία έχει εφαρμοστεί σε διάφορες περιστάσεις [2, 43, 42]. Η απλή διάλεκτος SDEC [26] που χρησιμοποιούμε, αποτελείται από χρονικά σημεία, πράξεις, γεγονότα και ένα σύνολο ανεξαρτήτων-τομέα αξιωμάτων.

Χρονικό σημείο: Ακέραιος ή πραγματικός αριθμός.

Γεγονός: Οντότητα, η οποία μπορεί να αλλάζει την τιμή της σε διαφορετικά χρονικά σημεία.

Πράξη: Συμβάν σε κάποιο χρονικό σημείο που μπορεί να έχει ως αποτέλεσμα την αλλαγή της τιμής ενός ή περισσότερων γεγονότων.

Τα ανεξαρτήτως-τομέα αξιώματα περιγράφουν τυπικά τον νόμο της αδράνειας, σύμφωνα με τον οποίο ένα *γεγονός* επιμένει στον χρόνο από την στιγμή που κάποια *πράξη* την αρχικοποιεί μέχρι την στιγμή που κάποιο άλλο γεγονός την τερματίζει. Παρουσιάζονται στον πίνακα 3.1 μαζί με τα βασικά κατηγορήματα της διαλέκτου SDEC.

Κατηγόρημα	Σημασία
$\boldsymbol{happensAt(E,T)}$	Η πράξη E συμβαίνει στον χρόνο T
$\boldsymbol{holdsAt(F,T)}$	Το γεγονός F παραμένει στην χρονική στιγμή T
$\boldsymbol{initiatedAt(F,T)}$	Την χρονική στιγμή T ξεκινάει μια χρονική περίοδος για την οποία το γεγονός F παραμένει
$\boldsymbol{terminatedAt(F,T)}$	Την χρονική στιγμή T τερματίζει μια χρονική περίοδος για την οποία το γεγονός F παραμένει

Αξιώματα	
$holdsAt(F,T+1) \leftarrow$ $initiatedAt(F,T).$	$holdsAt(F,T+1) \leftarrow$ $holdsAt(F,T), \neg\, terminatedAt(F,T).$

Πίνακας 3.1: Τα βασικά κατηγορήματα και τα ανεξαρτήτως-τομέα αξιώματα στην SDEC

Γενικά, ένα πρόγραμμα EC εμπεριέχει επίσης και ένα σύνολο *τομεοεξαρτώμενων αξιωμάτων* που περιγράφουν πως οι *πράξεις* στις ροές συμβολικών δεδομένων επηρεάζουν τις τιμές των *γεγονότων*. Ωστόσο, η παρασκευή αυτών των αξιωμάτων πρέπει να γίνει με το χέρι, χρειάζεται γνώση του τομέα εφαρμογής και των ιδιαιτεροτήτων του και γενικά είναι μια χρονοβόρα διαδικασία, επιρρεπής σε λάθη και ευαίσθητη στον θόρυβο. Θα θέλαμε επομένως να αποφύγουμε αυτήν την χειρωνακτική διαδικασία χρησιμοποιώντας την σύνδεση που δίνει ο EC με Επαγωγικό Λογικό Προγραμματισμό (ILP), μια μέθοδο μηχανικής μάθησης που επιτρέπει την εξαγωγή λογικών θεωριών με είσοδο θετικά και αρνητικά παραδείγματα. Σε αυτό έρχεται να μας βοηθήσει ένας online αλγόριθμος μάθησης ορισμών γεγονότων (OLED) που είναι ικανός να λειτουργήσει σε συνθήκες θορύβου.

3.2.1 Online μάθηση ορισμών γεγονότων (OLED)

Ο OLED είναι ένα γενικού σκοπού σύστημα Λογικού Προγραμματισμού που στο πλαίσιο της μάθησης θεωριών Λογισμού Γεγονότων (EC) χρησιμοποιεί μια συνάρτηση σκορ σαν ευριστική για να μαθαίνει ξεχωριστά κανόνες initiatedAt και terminatedAt. Μπορεί να μάθει θεωρίες με ένα πέρασμα στα δεδομένα. Χρησιμοποιεί το όριο Hoeffding [18] (Ορισμός 3.1) ως στατιστικό επιχείρημα για την αξιολόγηση, την μάθηση και τον αποκλεισμό κανόνων. Σε αυτή την εργασία τον χρησιμοποιούμε όπως και στην αρχική του παρουσίαση [26] για την μάθηση προγραμμάτων EC.

Ορισμός 3.1 (Όριο Hoeffding). Έστω X μια τυχαία μεταβλητή στο εύρος $[0,1]$ και ο παρατηρούμενος μέσος όρος $\bar{X}$ των τιμών του μετά από n ανεξάρτητες παρατηρήσεις. Τότε, με πιθανότητα $1-\delta$, για τον πραγματικό μέσο όρο $\hat{X}$ της μεταβλητής X ισχύει ότι

$$\hat{X} \in (\bar{X}-\varepsilon, \bar{X}+\varepsilon), \text{ όπου } \varepsilon = \sqrt{\frac{\ln(1/\delta)}{2n}}$$

Η λειτουργία του OLED συνοψίζεται σε τρεις επιμέρους διαδικασίες: την επέκταση θεωρίας, την επέκταση κανόνων και το κλάδεμα κανόνων.

Κατά την επεξεργασία της ροής δεδομένων μετριούνται για κάθε κανόνα και τις ειδικεύσεις του τα αληθή θετικά (TP), ψευδή θετικά (FP), ψευδή αρνητικά (FN) και ο αριθμός παραδειγμάτων στα οποία έχει ήδη αξιολογηθεί. Με τα TP, FP, FN ορίζεται η παρακάτω συνάρτηση σκορ:

Ορισμός 3.2 (Συνάρτηση σκορ). $G(r) = \begin{cases} \frac{TP_r}{TP_r+FP_r}, & \text{αν το } r \text{ είναι κανόνας } initiatedAt \\ \frac{TP_r}{TP_r+FN_r}, & \text{αν το } r \text{ είναι κανόνας } terminatedAt \end{cases}$

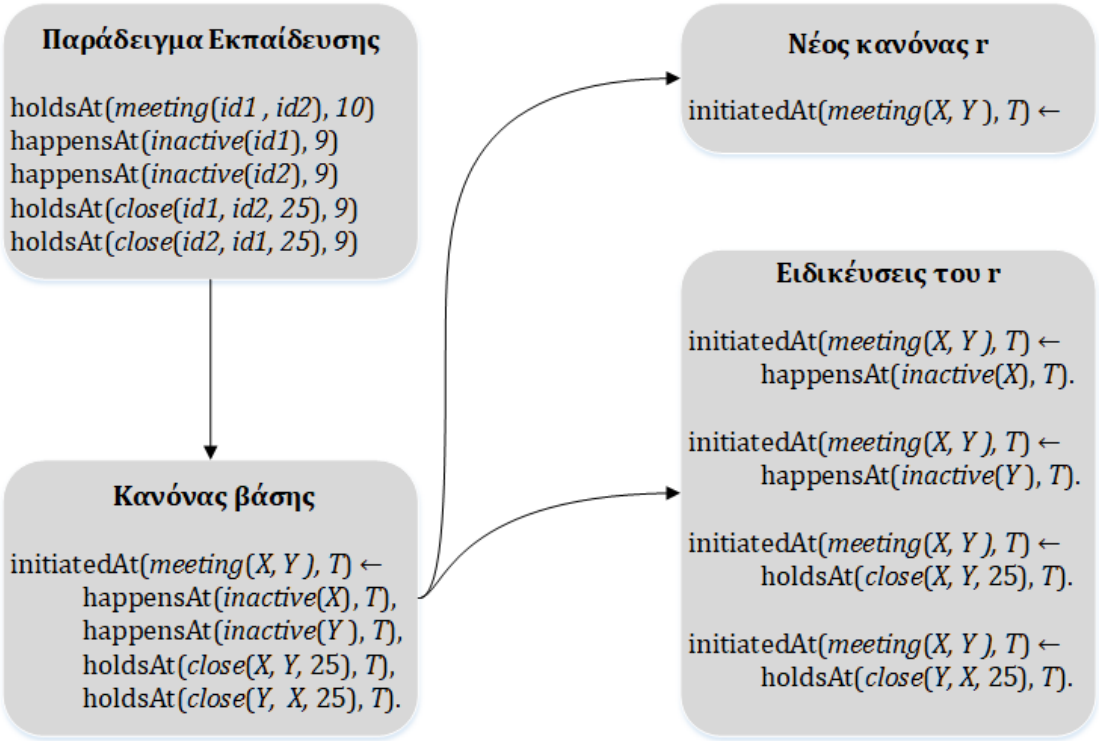
Το όριο Hoeffding επιτρέπει να παρθεί στατιστικά τεκμηριωμένη απόφαση σχετικά με το πότε έχουν συγκεντρωθεί αρκετά παραδείγματα, ώστε κάποια ειδίκευση ενός κανόνα να θεωρείται καλύτερη από κάποια άλλη, οπότε και αντικαθίσταται.

Στις περιπτώσεις, που έχουμε

- FN για initiatedAt κανόνα, δηλαδή κάποιος initiatedAt κανόνας δεν αρχικοποιεί κάποιο γεγονός το οποίο συμβαίνει,
- FP για κάποιο terminatedAt κανόνα, δηλαδή κάποιος terminatedAt κανόνας δεν τερματίζει κάποιο αρχικοποιημένο γεγονός το οποίο δεν συμβαίνει,

περαιτέρω ειδίκευση των κανόνων δεν θα βοηθήσει. Τότε, ο OLED κάνει επέκταση της θεωρίας, ξεκινώντας την μάθηση ενός νέου κανόνα με απαγωγή από παραδείγματα εκπαίδευσης, διαδικασία που συνοψίζεται στην εικόνα 3.6.

Κάποιες φορές σχηματίζονται κακοί κανόνες που δεν μπορούν να βελτιωθούν. Ο OLED υλοποιεί έναν μηχανισμό με τον οποίο ευρίσκονται κανόνες, για τους οποίους το όριο Hoeffding αποφαίνεται ότι είναι στατιστικά απίθανο να βελτιωθούν. Στην συνέχεια εξαλείφονται από την θεωρία, όταν έχουν σκορ μικρότερο από ένα κατώφλι προσδιοριζόμενο από τον χρήστη. Η διαδικασία αυτή ονομάζεται κλάδεμα κανόνων.



Εικόνα 3.6: Παραδείγματα εκπαίδευσης μετατρέπονται με απαγωγή σε κανόνα βάσης, στον οποίο σταθερές έχουν κατάλληλα αντικατασταθεί με ελεύθερες μεταβλητές. Από τον κανόνα βάσης, δημιουργείται ένας κανόνας initiatedAt r και όλες οι ειδικεύσεις του.

4 Υλοποίηση & πειράματα

4.1 Προγραμματιστικά εργαλεία

Για όλα τα πειράματα μας χρησιμοποιήσαμε το περιβάλλον της Python 3. Για την επεξεργασία βίντεο (φόρτωση καρέ του βίντεο, εξαγωγή περικομμάτων) χρησιμοποιήσαμε την διάσημη ανοιχτή βιβλιοθήκη OpenCV [6]. Για ό,τι κάναμε με νευρωνικά δίκτυα (υλοποίηση αυτοκωδικοποιητή, εξαγωγή C3D χαρακτηριστικών) χρησιμοποιήσαμε το σύστημα keras [10] . Όσον αφορά άλλους αλγορίθμους μηχανικής μάθησης (μηχανές διανυσμάτων υποστήριξης, tSNE, προεπεξεργασία), χρησιμοποιήσαμε το αρκετά πλήρες περιβάλλον της βιβλιοθήκης scikit-learn [44] και το βασισμένο στην Python οικοσύστημα του SciPY[22] για τον χειρισμό πινάκων και την οπτικοποίηση των αποτελεσμάτων μας.

Ο κώδικας που παραγάγαμε είναι διαθέσιμος στο διαδίκτυο[8] μαζί με τα πειράματα μας διαθέσιμα για άμεση αναπαραγωγή σε τετράδια jupyter [47].

4.2 Σύνολο Δεδομένων

Το σύνολο δεδομένων που χρησιμοποιούμε σε αυτήν την εργασία είναι το CAVIAR[9]. Αποτελείται από ένα σύνολο 28 βίντεο, που έχουν βιντεοσκοπηθεί στο λόμπι εισόδου του εργαστηρίου INRIA στην Grenoble[10]. Η ανάλυση είναι η μισή του προτύπου PAL (384 x 288 pixels, 25 καρέ ανά δευτερόλεπτο) και η συμπίεση των βίντεο έχει γίνει με MPEG-2. Για όλα τα βίντεο έχει γίνει επισημείωση με το χέρι

1) των δισδιάστατων κουτιών οριοθέτησης που περιέχουν άτομα,
2) των απλών ενεργειών που αυτά εκτελούν:

ανενεργός - inactive	το άτομο είναι εμφανές στο καρέ αλλά δεν κινείται
ενεργός - active	το άτομο είναι εμφανές στο καρέ, κινείται, αλλά δεν μεταφράζεται στην εικόνα

[8] http://users.iit.demokritos.gr/~iprapas/thesis/source_code/
[9] http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/
[10] https://www.inria.fr/

περπατάει - walking — το άτομο είναι εμφανές στο καρέ, κινείται και μεταφράζεται αργά στην εικόνα

τρέχει – running — το άτομο είναι εμφανές στο καρέ, κινείται και μεταφράζεται γρήγορα στην εικόνα

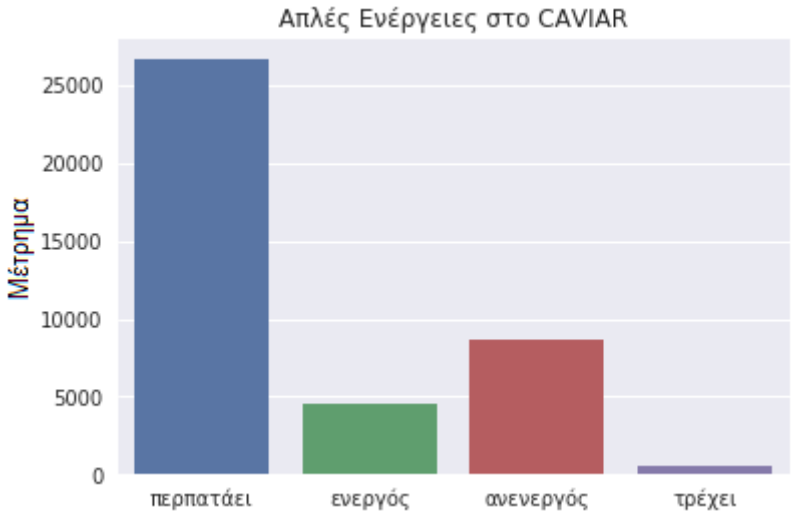
3) αλλά και πιο σύνθετων ενεργειών στις οποίες συμμετέχουν:

συνάντηση - meeting — κάποια άτομα συναντιούνται

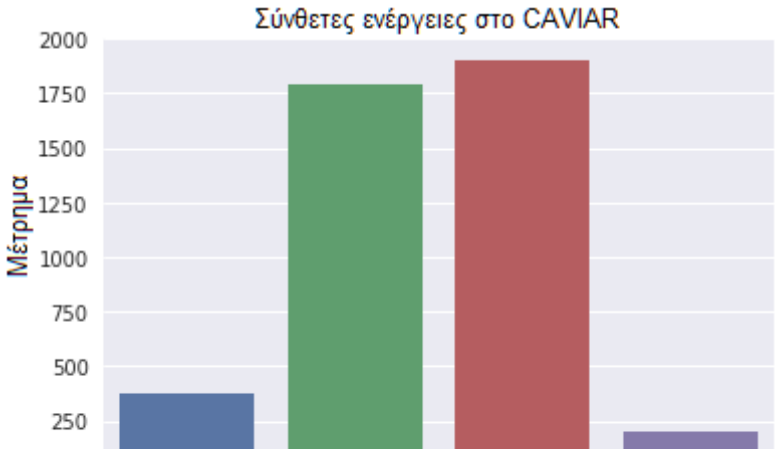
ομαδική κίνηση – moving — κάποια άτομα κινούνται μαζί στον χώρο

πάλη – fighting — κάποια άτομα παλεύουν

αφήνει αντικείμενο - leaving object — το άτομο αφήνει κάποιο αντικείμενο στον χώρο



Εικόνα 4.1: Κατανομή απλών ενεργειών στα βίντεο του CAVIAR



Εικόνα 4.2: Κατανομή σύνθετων ενεργειών στα βίντεο του CAVIAR

Στην εργασία αυτή θεωρούμε την ανίχνευση των ατόμων στον χώρο δεδομένη (κουτιά οριοθέτησης). Η Εικόνα 4.3 δείχνει ένα αντιπροσωπευτικό καρέ των βίντεο μαζί με τις επισημειώσεις για απλές και σύνθετες ενέργειες.

Στο πρώτο μέρος της εργασίας που αφορά την αναγνώριση απλών ανθρώπινων ενεργειών, μαθαίνουμε από ανεπεξέργαστα βίντεο δεδομένα να αναγνωρίζουμε τις απλές ενέργειες που εκτελούν τα άτομα. Στο δεύτερο μέρος, αξιολογούμε κατά πόσο αυτές οι αυτόματα παραγόμενα ενέργειες μπορούν να αντικαταστήσουν την επισημειωμένη αλήθεια για ένα σύστημα λογικού χρονικού συμπερασμού. Η επιλογή του CAVIAR έγινε επειδή περιέχει επισημείωση τόσο για απλές, όσο και για πιο σύνθετες ανθρώπινες ενέργειες και επομένως μπορούμε να το χρησιμοποιήσουμε για μάθηση σε δύο επίπεδα.

Εικόνα 4.3 Παράδειγμα καρέ από το CAVIAR. Με κίτρινο χρώμα βλέπουμε τα κουτιά οριοθέτησης για κάθε άτομο καθώς και την επισημείωση που έχει γίνει για την ενέργεια (περπατάει - walking, ενεργός - active) που εκτελούν. Αντίστοιχα με πράσινο χρώμα βλέπουμε το κουτί οριοθέτησης δύο ατόμων που συμμετέχουν σε μια ομαδική ενέργεια (συνάντηση - meeting)

4.3 Συνελικτικός Αυτοκωδικοποιητής για αναπαράσταση καρέ

4.3.1 Αρχιτεκτονική Δικτύου

Όπως αναφέραμε στο 3.1.2.1, ένας συνελικτικός αυτοκωδικοποιητής αποτελείται από 2 βασικά μέρη, τον κωδικοποιητή και τον αποκωδικοποιητή. Ενώ δεν είναι απαραίτητο, σχεδιάσαμε τους αυτοκωδικοποιητές με μια συμμετρία ανάμεσα σε αυτά τα δύο μέρη, αρχιτεκτονική που συναντάται συχνά.

Δοκιμάσαμε διάφορες αρχιτεκτονικές, αλλά παρουσιάζουμε αυτήν που ποιοτικά βλέπουμε ότι παράγει τον καλύτερο ανασχηματισμό της εισόδου.

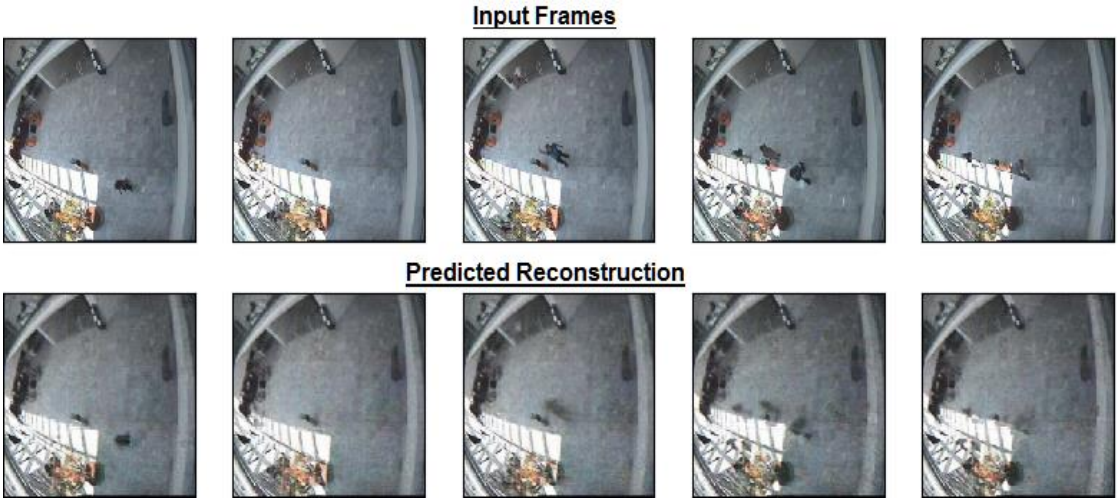
Επίπεδο	**Τύπος Επιπέδου**
Είσοδος	Εικόνα(3x224x224)
1.1	Συνέλιξη2Δ(64, 3x3) ↓
1.2	ΣυγκέντρωσηΜεγίστου(2,2)
2.1	Συνέλιξη2Δ(32, 3x3) ↓
2.2	ΣυγκέντρωσηΜεγίστου(2,2)
3	Συνέλιξη2Δ(16, 3x3) ↓
3’	Συνέλιξη2Δ(16, 3x3) ↑
2.2’	Υπερδειγματοληψία(2,2)
2.1’	Συνέλιξη2Δ(32, 3x3) ↑
1.2’	Υπερδειγματοληψία(2,2)
1.1’	Συνέλιξη2Δ(64, 3x3) ↑
Έξοδος	Εικόνα(3x224x224)

Πίνακας 4.1: Αρχιτεκτονική συνελικτικού αυτοκωδικοποιητή που χρησιμοποιήθηκε στα πειράματά μας. Συνέλιξη2Δ(64, 3x3) σημαίνει 64 φίλτρα μεγέθους 3x3. Το ↓ υποδεικνύει κανονική συνέλιξη ενώ το ↑ υποδεικνύει ανεστραμμένη συνέλιξη. Η ΣυγκέντρωσηΜεγίσου(2,2) υποδιπλασιάζει το μέγεθος της εισόδου σε κάθε διάσταση. Η Υπερδειγματοληψία(2,2) διπλασιάζει το μέγεθος της εισόδου σε κάθε διάσταση.

4.3.2 Εκπαίδευση για ανασχηματισμό καρέ των βίντεο

Χωρίσαμε τα καρέ του CAVIAR σε 90% εκπαιδευτικά και 10% δοκιμαστικά. Για συνάρτηση κόστους χρησιμοποιήσαμε το μέσο τετραγωνικό σφάλμα (Ορισμός 4.1) και ως βελτιστοποιητή της βαθμωτής κατάβασης (gradient descent), τον Adam [28], που υλοποιεί διάφορους μηχανισμούς οι οποίοι επιτρέπουν την εύρεση ενός καλού ελάχιστου της συνάρτησης κόστους. Σε κάθε εποχή κρατάμε 0.01% από τα δεδομένα για επιβεβαίωση και σταματάμε όταν το λάθος επιβεβαίωσης σταματά να βελτιώνεται, κάτι που στην περίπτωση μας έγινε μετά από 28 ώρες τρέχοντας σε μια κάρτα γραφικών NVIDIA GEFORCE 980 GTX (4GB). Μερικά παραδείγματα ανασχηματισμού των καρέ φαίνονται στην Εικόνα 4.4.

Ορισμός 4.1 $MT\Sigma(x, y) = \frac{1}{n}\sum_i |x_i - y_i|^2$



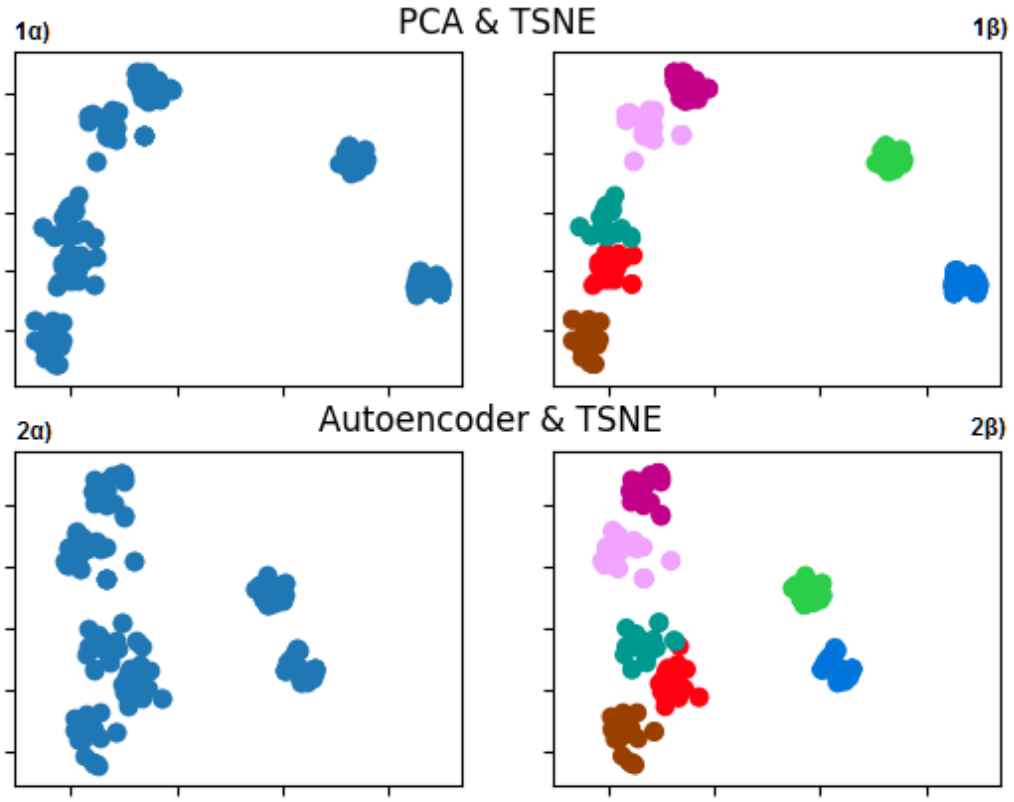
Εικόνα 4.4: Πάνω βλέπουμε τις εικόνες εισόδου και κάτω βλέπουμε τις αντίστοιχες όπως προκύπτουν στην έξοδο του αυτοκωδικοποιητή

4.3.3 Αξιολόγηση Αναπαράστασης

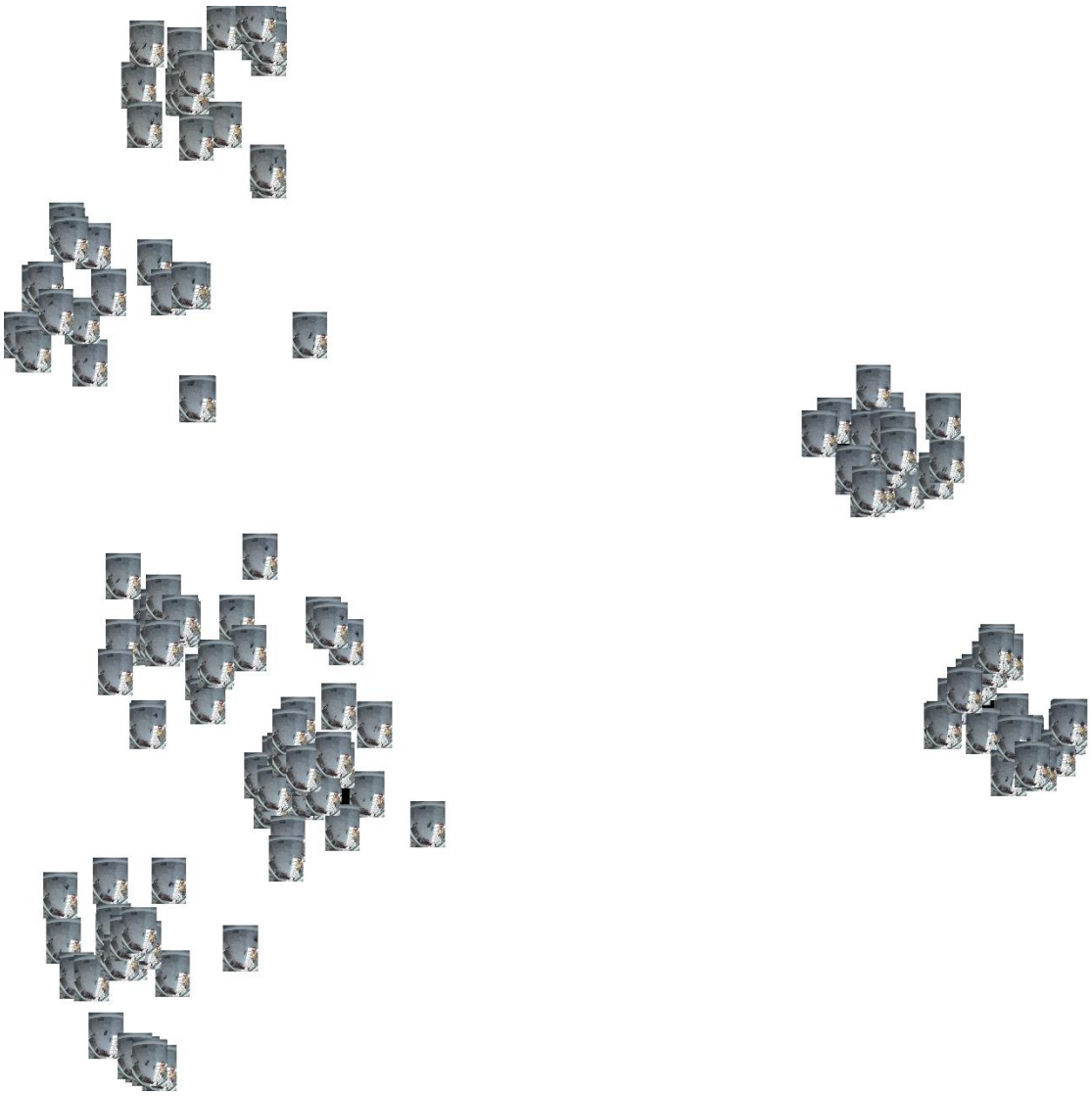
Χρησιμοποιώντας την μέθοδο t-SNE, προβάλλουμε τα κωδικοποιημένα καρέ στον δισδιάστατο χώρο και βλέπουμε καλοσχηματισμένες συστάδες. Χρωματίζοντας τα σημεία του χώρου ανάλογα με το βίντεο στο οποίο ανήκουν τα καρέ που αντιπροσωπεύουν (βλ. Εικόνα 4.5), παρατηρούμε ότι κάθε συστάδα απαρτίζεται από καρέ του ίδιου βίντεο.

Ωστόσο, κάποιες συστάδες βρίσκονται κοντά μεταξύ τους και θα θέλαμε να ελέγξουμε γιατί συμβαίνει αυτό. Χρησιμοποιούμε πάλι την ίδια μέθοδο προβολής των κωδικοποιημένων εικόνων στον δισδιάστατο χώρο και σε κάθε σημείο που προκύπτει απεικονίζουμε το καρέ εισόδου όπως φαίνεται στην εικόνα 4.6. Με μια προσεκτική ματιά μπορούμε να δούμε ότι ο καθοριστικός παράγοντας για το που προβάλλεται ένα καρέ στον δισδιάστατο χώρο, είναι η φωτεινότητα. Αυτό εξηγείται, αφού η φωτεινότητα είναι η μεγαλύτερη συνολική μεταβολή που

παρατηρείται στα δεδομένα του CAVIAR. Υποστηρίζουμε ότι τα βίντεο των οποίων τα καρέ σχηματίζουν κοντινές συστάδες δημιουργήθηκαν κοντά χρονικά.



Εικόνα 4.5: Κάθε σημείο των ανωτέρω διαγραμμάτων αντιπροσωπεύει ένα καρέ κάποιου βίντεο του CAVIAR. Πάνω βλέπουμε τα που έχουν προκύψει από προβολή στον δισδιάστατο χώρο με ανάλυση κύριων συνιστωσών και t-SNE. Κάτω βλέπουμε τα αντίστοιχα διαγράμματα όπως έχουν προκύψει από την κωδικοποίηση του αυτοκωδικοποιητή και προβολή με t-SNE. Δεξιά τα ίδια σημεία των γραφικών αριστερά βάφονται με βάση σε ποιο βίντεο ανήκουν.



Εικόνα 4.6: Με t-SNE έχουμε προβάλει τα κωδικοποιημένα καρέ στον 2διάστατο χώρο με σκοπό να ελέγξουμε ποιοτικά γιατί καρέ διαφορετικών βίντεο μπορεί να βρίσκονται κοντά μεταξύ τους.

Όπως παρουσιάσαμε στην ενότητα 2.1, μια αναπαράσταση εικόνας για να είναι καλή θα πρέπει να παραμένει αμετάβλητη σε μικρές αλλαγές της εισόδου (κλίμακα, οπτική, φωτεινότητα, περιστροφή, μετάφραση, θόρυβος στο παρασκήνιο ή στο προσκήνιο), κάτι που είναι ξεκάθαρο ότι δεν συμβαίνει για την αναπαράσταση που βγαίνει από τον αυτοκωδικοποιητή. Μάλιστα, στην εικόνα 4.5 μπορούμε να δούμε τεράστια ομοιότητα με απλή ανάλυση σε κύριες συνιστώσες (Principal Component Analysis) που μας δείχνει ότι με τις απλές αρχιτεκτονικές που σχεδιάσαμε δεν καταφέραμε να κάνουμε επιτυχή μη γραμμικό αυτοσυσχετισμό της εισόδου [19].

4.3.4 Σύνοψη

Σε αυτήν την ενότητα παρουσιάσαμε την υλοποίηση μας για μια απλή συνελικτικού αρχιτεκτονική αυτοκωδικοποιητή πολλών επιπέδων και αξιολογήσαμε την χρήση του για αναπαράσταση καρέ.

Πολύ νωρίς σε αυτή την προσπάθεια ήρθαμε αντιμέτωποι με πολύ μεγάλο αριθμό σχεδιαστικών επιλογών που αφορούσαν τα είδη των επιπέδων, το μέγεθος τους, τον τύπο των συνελίξεων σε κάθε επίπεδο, τους περιορισμούς του δικτύου, την συνάρτηση κόστους καθώς και το είδος των δεδομένων εκπαίδευσης. Συνδυάζοντας τα προηγούμενα με μεγάλες απαιτήσεις σε υπολογιστικούς πόρους και χρόνους εκπαίδευσης, αποφασίσαμε να αφήσουμε πίσω την έρευνα σε αυτοκωδικοποιητές και να ακολουθήσουμε πιο ελπιδοφόρες οδούς στην αναζήτηση για μια αναπαράσταση κατάλληλη για αναγνώριση απλών ανθρώπινων ενεργειών. Τα προηγούμενα σε καμία περίπτωση δεν έχουν σκοπό να αποτρέψουν την χρησιμοποίηση αυτοκωδικοποιητών γενικά ως αναπαράσταση καρέ, αλλά κρίναμε στην περίπτωσή μας ότι για να γίνει αυτό θα ξεφεύγαμε από τον στόχο της εργασίας.

4.4 Χαρακτηριστικά C3D για αναγνώριση απλών ανθρώπινων ενεργειών

4.4.1 Εξαγωγή χαρακτηριστικών

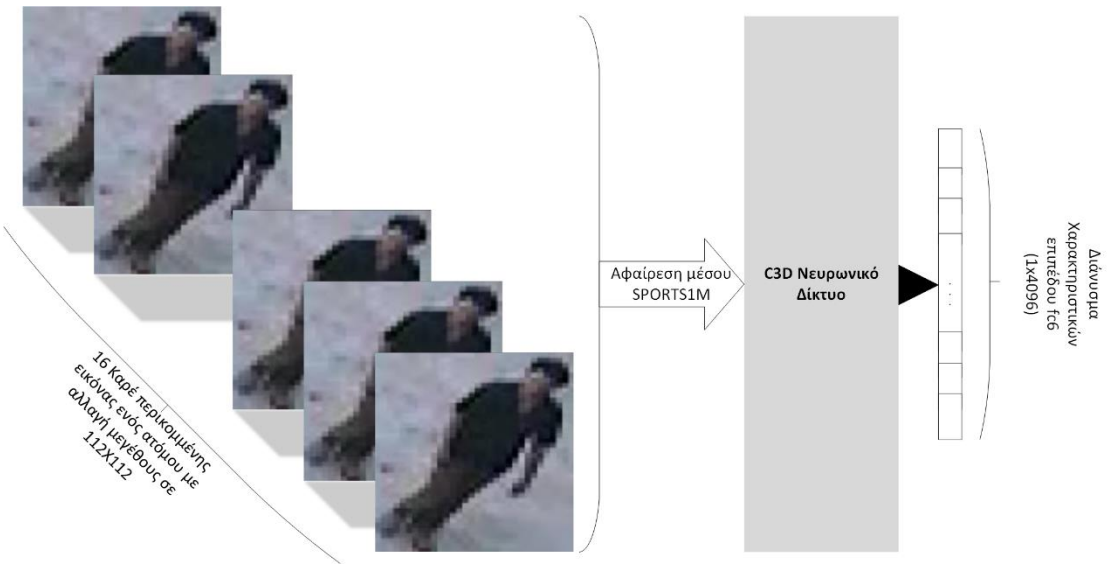
Το C3D νευρωνικό [61] επιλέχθηκε σαν εξαγωγέας χαρακτηριστικών λόγω καλής απόδοσης σε προηγούμενες δουλειές [37, 61] .Στην είσοδο το νευρωνικό δέχεται ένα κλιπ αποτελούμενο από 16 καρέ με μέγεθος 3(κανάλια χρώματος RGB) x 112x112 pixel. Τροφοδοτώντας το νευρωνικό με τα δικά μας παραδείγματα, παίρνουμε ως αναπαράσταση κλιπ (16 διαδοχικά καρέ), τα χαρακτηριστικά όπως προκύπτουν στην έξοδο του πρώτου πλήρως συνδεδεμένου επιπέδου (fc6) του.

Το προεκπαιδευμένο νευρωνικό στο SPORTS1M είχε γίνει αρχικά διαθέσιμο στην γλώσσα Caffe [21] για βαθιά μάθηση. Εμείς το χρησιμοποιούμε όπως το έχει μεταφέρει ο Alberto Montes[11] για χρήση με την βιβλιοθήκη keras [10] σε python.

Για κάθε βίντεο, θεωρούμε τα κουτιά οριοθέτησης των ατόμων δεδομένα και εξάγουμε περικομμένα κλιπ των 16 καρέ στα οποία αυτά εμφανίζονται. Αλλάζουμε το μέγεθος του κάθε καρέ ώστε να μπορέσει να αποτελέσει είσοδο στο C3D νευρωνικό και τροφοδοτούμε το κλιπ σε αυτό. Όπως προτείνουν και οι εμπνευστές του C3D, ως μοναδική προ επεξεργασία,

[11] https://github.com/albertomontesg/keras-model-zoo

αφαιρούμε από κάθε κλιπ τον μέσο όρο κάθε καναλιού χρώματος όπως είχε υπολογιστεί για το SPORTS1M. Έτσι, κάθε διάνυσμα χαρακτηριστικών που προκύπτει στην έξοδο του νευρωνικού, αντιπροσωπεύει 16 συνεχόμενα καρέ στα οποία ένα άτομο εμφανίζεται. Τα προηγούμενα φαίνονται στην εικόνα 4.7.

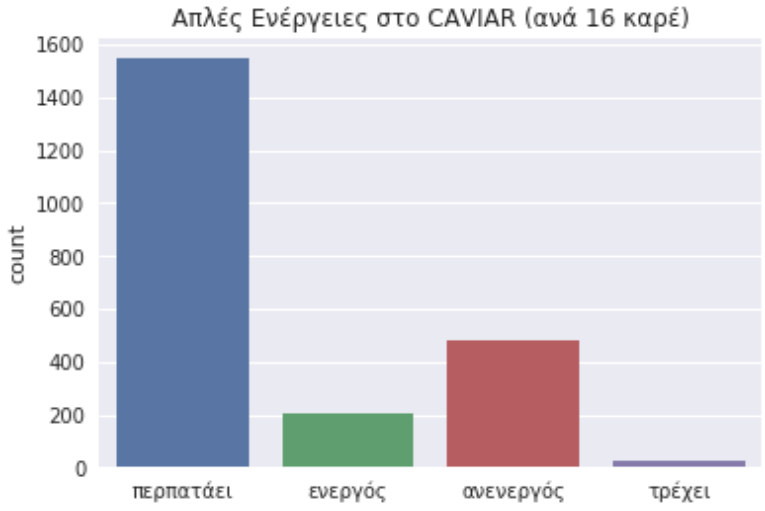


Εικόνα 4.7: Εξαγωγή χαρακτηριστικών από βίντεο του CAVIAR με το C3D νευρωνικό

4.4.2 Ανισορροπία κλάσεων και μετρικές αξιολόγησης

Ρίχνοντας μια ματιά στην κατανομή των απλών ενεργειών όπως απεικονίζονται στην Εικόνα 4.1 είναι ξεκάθαρο ότι υπάρχει μεγάλη ανισορροπία στην ποσότητα παραδειγμάτων που έχουμε για κάθε κλάση ενέργειας.

Εξάγουμε τα ένα διάνυσμα C3D χαρακτηριστικών ανά 16 καρέ στα οποία εμφανίζεται ένα άτομο. Επιλέγουμε να μην λάβουμε υπόψη μας τις 16άδες μη καθαρών καρέ, δηλαδή των κλιπ που αποτελούνται από καρέ στα οποία η ενέργεια του ατόμου δεν είναι σταθερή. Τότε η κατανομή των παραδειγμάτων που έχουμε φαίνεται στην Εικόνα 4.8.



Εικόνα 4.8: Κατανομή καθαρών απλών ενεργειών ανά 16 καρέ στο CAVIAR

Λόγω αυτής της ανισορροπίας δεν θα ήταν ταιριαστό να χρησιμοποιήσουμε το ποσοστό ορθής κατάταξης (accuracy) ως μετρική αξιολόγησης. Με έναν κατηγοριοποιητή ο οποίος θα επέστρεφε πάντα την κλάση *περπατάει* θα φτάναμε accuracy της τάξης του 67%. Σε τέτοιες περιπτώσεις προτείνεται η χρήση των μετρικών *ακρίβειας (precision), ανάκλησης (recall)* και *f1-σκορ* που ορίζεται ως ο αρμονικός μέσος όρος της *ακρίβειας* και της *ανάκλησης*

Ο ορισμός των παραπάνω μεγεθών γίνεται στον Πίνακα 4.2

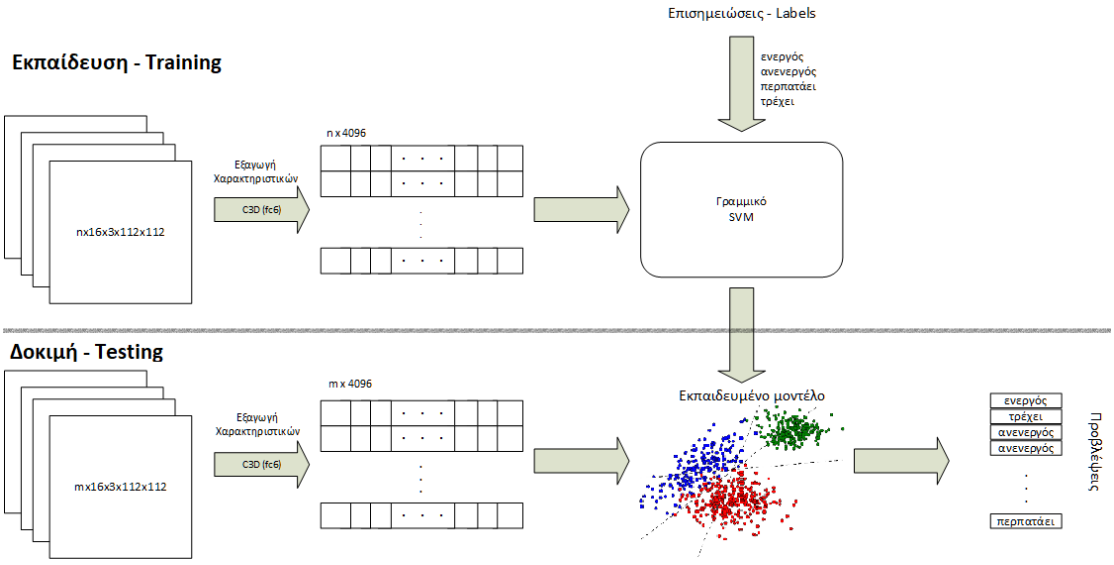
	Πρόβλεψη Θετική	**Πρόβλεψη Αρνητική**
Παράδειγμα Θετικό	Αληθές Θετικό True Positive (TP)	Ψευδές Αρνητικό False Negative (FN)
Παράδειγμα Αρνητικό	Ψευδές Θετικό False Positive (FP)	Αληθές Αρνητικό True Negative (TN)

Ακρίβεια (Precision)	$\frac{\sum TP}{\sum TP + \sum FP}$
Ανάκληση (Recall)	$\frac{\sum TP}{\sum TP + \sum FN}$
$\boldsymbol{f_1 - \sigma\kappa o\rho}$	$\frac{2 * Ακρίβεια * Ανάκληση}{Ακρίβεια + Ανάκληση}$

Πίνακας 4.2: Ορισμός ακρίβειας, ανάκλησης, f1-σκορ

4.4.3 Κατηγοριοποίηση

Για την κατηγοριοποίηση των χαρακτηριστικών στις κλάσεις που μας ενδιαφέρουν (ενεργός, ανενεργός, περπατάει, τρέχει) χρησιμοποιούμε μηχανές διανυσμάτων υποστήριξης (SVM) με γραμμικό πυρήνα, όπως γίνεται και στις περισσότερες περιπτώσεις με τα C3D χαρακτηριστικά [61] και μέθοδο ένας-εναντίον-ενός για την απόφαση, αφού έχουμε πρόβλημα πολλών κλάσεων. Η χρήση των SVM συνοψίζεται στην εικόνα 4.9.



Εικόνα 4.9: Κατηγοριοποίηση με SVM

Πείραμα 1ο. Εξάγουμε χαρακτηριστικά από όλα τα περικόμματα που περιέχουν άτομα σε κάθε βίντεο του CAVIAR και χρησιμοποιούμε μόνο εκείνα για τα οποία υπάρχει σταθερή επισημείωση για 16 συνεχόμενα καρέ ($\frac{16\ καρέ}{25\ καρέ/second} = 0.64s$). Μετά από ένα τυχαίο ανακάτεμα, κάνουμε 10πλή διασταυρωμένη επικύρωση (10 Fold Cross Validation) χωρίζοντας σε 90% δεδομένα εκπαίδευσης (training) και 10% δεδομένα δοκιμής (testing) σε κάθε δίπλωση. Έτσι, δοκιμάζουμε την μέθοδο σε όλα τα δεδομένα, εκπαιδεύοντας το SVM κάθε φορά στα υπόλοιπα. Υπολογίζουμε τα μικρο-μεγέθη ακρίβειας, ανάκλησης και f1-σκορ. Αυτό σημαίνει ότι αθροίζουμε τα αληθή θετικά, αληθή αρνητικά, ψευδή θετικά, ψευδή αρνητικά κάθε δίπλωσης της διασταυρωμένης επικύρωσης και μετράμε τα μεγέθη ακρίβειας, ανάκλησης και f1-σκορ με βάση τα συνολικά τους αθροίσματα. Τα αποτελέσματα συνοψίζονται στους Πίνακες 4.3, 4.4.

Κλάση	**Ακρίβεια**	**Ανάκληση**	f_1**-σκορ**	**#Παραδειγμάτων**
Ανενεργός	0.91	0.94	0.92	479
Ενεργός	0.79	0.65	0.71	203
Περπατάει	0.94	0.97	0.96	1547
Τρέχει	0.00	0.00	-	24
Σταθμισμένος μέσος/σύνολο	0.91	0.92	0.92	2253

Πίνακας 4.3: Πείραμα 1ο. Πίνακας αποτελεσμάτων των μικρο-μεγεθών ακρίβειας, ανάκλησης και f1-σκορ

		Προβλέψεις			
	Κλάση	**Ανενεργός**	**Ενεργός**	**Περπατάει**	**Τρέχει**
Αλήθεια	**Ανενεργός**	**448**	12	19	0
	Ενεργός	24	**131**	48	0
	Περπατάει	20	23	**1502**	2
	Τρέχει	0	0	**24**	0

Πίνακας 4.4: Πείραμα 1ο. Μήτρα σύγχυσης των διαφορετικών κλάσεων απλών ενεργειών

Τα αποτελέσματα που βγάζουμε, φαίνονται να είναι άμεσα συνδεδεμένα τον αριθμό παραδειγμάτων που έχουμε διαθέσιμα. Για την κλάση *τρέχει* παρατηρούμε ότι δεν έγινε ούτε μια σωστή πρόβλεψη. Κατά τα άλλα με μια πρώτη ματιά τα αποτελέσματα φαίνονται πολύ καλά. Όμως χωρίζοντας τυχαία σε εκπαιδευτικά και δοκιμαστικά σύνολα, δεν έχουμε λάβει υπόψη την μεγάλη συσχέτιση που έχουν διαδοχικά καρέ βίντεο. Υπάρχει με αυτό τον τρόπο μεγάλη πιθανότητα να συναντήσουμε στο εκπαιδευτικό σύνολο, κλιπ πολύ κοντά χρονικά σε αυτά που χρησιμοποιούμε για δοκιμή. Επομένως, η παραπάνω κατηγοριοποίηση ενδεχομένως να είναι χρήσιμη σε περιπτώσεις που θα θέλαμε να καλύψουμε κενά μιας μη εξαντλητικής χειροκίνητης επισημείωσης, αλλά δεν είναι απαραίτητα αντιπροσωπευτική για πραγματικά σενάρια που απαιτείται μεγάλη γενίκευση σε βίντεο που δεν έχουν εμφανιστεί στα δεδομένα εκπαίδευσης.

Πείραμα 2ο. Σύμφωνα με τα παραπάνω, τα αποτελέσματα του 1ου πειράματος που εκτελέσαμε δεν είναι απόλυτα έμπιστα. Επομένως κρίνεται επιτακτικό, να κάνουμε εκπαίδευση και δοκιμή σε διαφορετικά βίντεο. Χρησιμοποιούμε την τεχνική Άσε-Ένα-Έξω (Leave-One-Out), σύμφωνα με την οποία κάθε βίντεο χρησιμοποιείται για δοκιμή ενώ όλα τα υπόλοιπα χρησιμοποιούνται για εκπαίδευση σε μια μορφή Κπλής διασταυρωμένης επικύρωσης (K fold cross validation) με K να είναι ίσο με τον αριθμό των βίντεο. Τα αποτελέσματα των μικρο-μεγεθών ακρίβειας, ανάκλησης και f1-σκορ για κάθε κλάση φαίνονται στους Πίνακες 4.5, 4.6.

Κλάση	**Ακρίβεια**	**Ανάκληση**	**f_1-σκορ**	**#Παραδειγμάτων**
Ανενεργός	0.80	0.55	0.65	479
Ενεργός	0.47	0.39	0.42	203
Περπατάει	0.86	0.97	0.91	1547
Τρέχει	0.00	0.00	-	24
Σταθμισμένος μέσος/σύνολο	0.8	0.82	0.8	2253

Πίνακας 4.5: Πείραμα 2ο. Πίνακας αποτελεσμάτων των μικρο-μεγεθών ακρίβειας, ανάκλησης και f1-σκορ

		Προβλέψεις			
Αλήθεια	Κλάση	**Ανενεργός**	**Ενεργός**	**Περπατάει**	**Τρέχει**
	Ανενεργός	**263**	72	144	0
	Ενεργός	45	**80**	78	0
	Περπατάει	19	26	**1499**	2
	Τρέχει	0	0	**24**	0

Πίνακας 4.6: Πείραμα 2ο. Μήτρα σύγχυσης των διαφορετικών κλάσεων απλών ενεργειών

Και σε αυτή την περίπτωση παρατηρούμε ότι η απόδοση του μοντέλου μας εξαρτάται σε μεγάλο βαθμό από τον αριθμό διαθέσιμων παραδειγμάτων. Έτσι, για την κλάση *περπατάει* που έχουμε τα περισσότερα παραδείγματα βλέπουμε ότι το μοντέλο τα πηγαίνει πολύ καλά πετυχαίνοντας 0.91 f1-σκορ, ενώ για την κλάση *τρέχει* που έχουμε μόνο 24 παραδείγματα δεν καταφέρνει να κατηγοριοποιήσει σωστά ούτε ένα. Ωστόσο, κατηγοριοποιούνται όλα τα παραδείγματα *τρέξιμου* σαν *περπάτημα*, που θεωρούμε μικρότερο λάθος από κατηγοριοποίηση στις άλλες κλάσεις. Αποδίδουμε, λοιπόν, ένα μέρος του προβλήματος εσφαλμένης κατηγοριοποίησης στην **ανισορροπία των κλάσεων** στο dataset του CAVIAR και κυρίως στην απουσίας επαρκούς αριθμού παραδειγμάτων για τις **υποεκπροσωπούμενες κλάσεις**.

Για την κλάση *ανενεργός* έχουμε αρκετά καλή ακρίβεια, δηλαδή δεν μπερδεύονται παραδείγματα άλλων κλάσεων με αυτή, ωστόσο παρατηρούμε ότι η ανάκληση πέφτει πολύ λόγω του μεγάλου αριθμού παραδειγμάτων της κλάσης αυτής που μπερδεύονται με *περπάτημα*. Αυτό θεωρούμε ως παράλογο λάθος, μιας και διαισθητικά οι δύο αυτές κλάσεις θα έπρεπε να είναι εμφανώς διαχωρίσιμες. Ελέγχοντας καλύτερα συγκεκριμένα ποια παραδείγματα μπερδεύονται, καταλαβαίνουμε ότι είναι **θέμα χαμηλής ανάλυσης** του περικόμματος που περιέχει το άτομο. Το μέγεθος των περικομμάτων δεν είναι σταθερό, αλλά το νευρωνικό περιμένει σταθερά καρέ 112x112 στην είσοδο, γεγονός που μας αναγκάζει σε μεγέθυνση πολύ μικρών περικομμάτων. Με αυτόν τον τρόπο, ελάχιστο τρεμόπαιγμα της κάμερας ή ανεπαίσθητες κινήσεις μεταφράζονται σε πολύ μεγάλη κίνηση των pixel στην είσοδο του νευρωνικού. Καταλήγουμε στο συμπέρασμα, ότι επειδή τα νευρωνικά δίκτυα είναι στατικά μοντέλα, πριν την επιλογή τους για εξαγωγή χαρακτηριστικών είναι επιτακτικό να γίνεται καλή μελέτη για τις επιδράσεις που έχει η μετατροπή παραδειγμάτων σε είσοδο για αυτά.

Για την κλάση *ενεργός* δεν παρατηρούμε κάποιο συγκεκριμένο μοτίβο και θεωρούμε λογικό ότι με τον συνδυασμό των παραπάνω προβλημάτων, την μπερδεύουμε τόσο με την υπερεκπροσωπούμενη κλάση *περπάτημα*, όσο και με την κλάση *ανενεργός*.

Σύμφωνα, με μια μετα-μελέτη από τους δημιουργούς του CAVIAR, υπάρχει ασυνέπεια της τάξης του 11% στην σημασιολογική επισημείωση των ενεργειών που εκτελούνται, λόγω σύγχυσης που οφείλεται στα παρακάτω:

- Οι παρατηρητές επηρεάζονται από την δικιά τους ερμηνεία του τι συμβαίνει σε μια σκηνή.
- Υπάρχει εγγενής αμφισημία για κάποιες συμπεριφορές.
- Υπάρχουν μικρές διαφορές (τυπική απόκλιση = 0.65s) μεταξύ της ανίχνευσης αλλαγών από διαφορετικούς παρατηρητές.

Ποιοτικά, παρατηρούμε ότι κάποια λάθη του μοντέλου μας πράγματι θα μπορούσαν να θεωρηθούν σωστά σημασιολογικά υπό διαφορετικές ερμηνείες. Τα βίντεο του CAVIAR, με την ανίχνευση και την ανά καρέ κατηγοριοποίηση των δραστών που έχουμε εξάγει βρίσκονται διαθέσιμα στην σελίδα της εργασίας[12].

Σύγκριση με καλύτερα αποτελέσματα (state-of-the-art). Τα καλύτερα αποτελέσματα [48] στην αναγνώριση απλών ανθρώπινων ενεργειών στο σύνολο δεδομένων του CAVIAR αναφέρουν:

- 98% γενικό ποσοστό κατηγοριοποίηση (total accuracy).
- τουλάχιστον 95% σωστή κατηγοριοποίηση ανά διαφορετική κατηγορία ενέργειας.

Με μια πρώτη ματιά φαίνονται πολύ καλύτερα από αυτά που βγάζουμε εμείς, ωστόσο παρατηρούμε ότι **δεν είναι συγκρίσιμα**, γιατί:

- Θεωρούν δεδομένη την γνώση της γεωμετρίας του χώρου
- Χρησιμοποιούν χαρακτηριστικά εξαγμένα από την θέση των κουτιών οριοθέτησης που δεν γενικεύουν σε διαφορετικές γωνίες κάμερας
- Δεν αναφέρεται ο τρόπος με τον οποίο έχουν χωρίσει σε σύνολα εκπαίδευσης/δοκιμής και αν αυτά περιέχουν ή όχι διαφορετικά βίντεο. Όπως είδαμε στην σύγκριση του 1^{ου} με το 2^{ο} πείραμα που εκτελέσαμε, αυτό μπορεί να επηρεάσει σημαντικά την απόδοση του μοντέλου.

Το CAVIAR είναι ένα παλιό σύνολο δεδομένων, διαμορφωμένο για το πιο ειδικό πρόβλημα της επιτήρησης (surveillance), σε σύγκριση με το γενικότερο που κοιτάμε σε αυτήν την εργασία

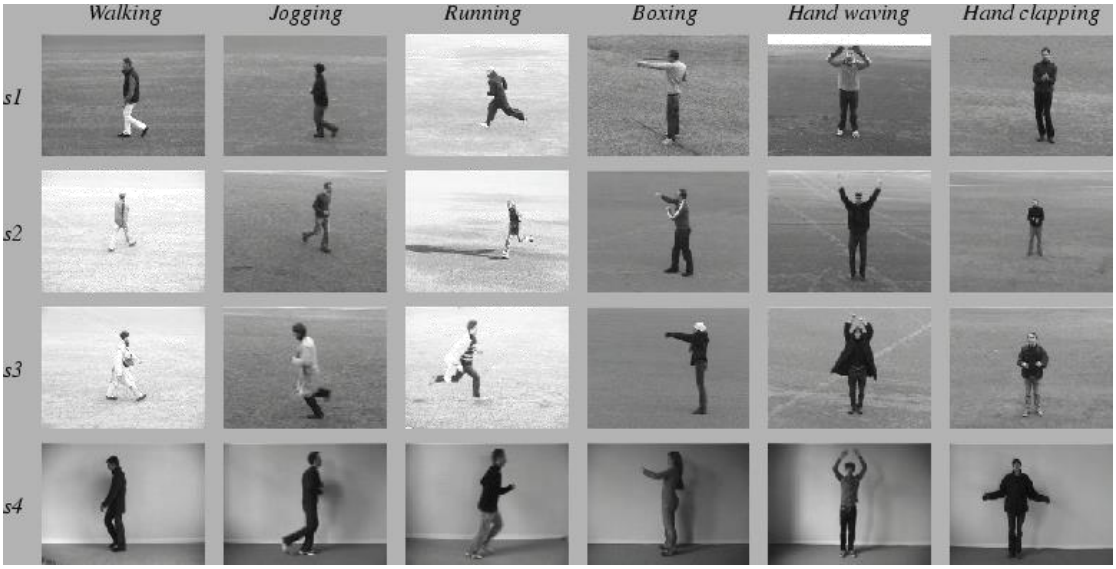
[12] http://users.iit.demokritos.gr/~iprapas/thesis/video_results

της αναγνώρισης ανθρώπινων ενεργειών,. Αυτό λύνουν στο [48], με χαρακτηριστικά εξειδικευμένα στο πρόβλημα της επιτήρησης, σε αντίθεση με εμάς που χρησιμοποιούμε χαρακτηριστικά γενικά, κατάλληλα για ταξινόμηση μεγάλης κλίμακας. Καλές πρακτικές μάθησης από βίντεο που τώρα είναι καθιερωμένες, όταν είχε βγει το συγκεκριμένο dataset και όταν αναφέρθηκαν τα συγκεκριμένα αποτελέσματα, ήταν ακόμα ζήτημα έρευνας.

4.4.4 Αξιολόγηση των C3D χαρακτηριστικών για αναγνώριση απλών ανθρώπινων ενεργειών στο KTH

Σε αυτό το σημείο θα θέλαμε να αξιολογήσουμε περαιτέρω τα C3D χαρακτηριστικά για την αναγνώριση απλών ανθρώπινων ενεργειών. Επιλέγουμε το dataset του KTH [52] για ανθρώπινες ενέργειες, που έχει χρησιμοποιηθεί κατά κόρον και είναι λιγότερο απαιτητικό από το CAVIAR, αφού περιέχει βίντεο σε ελεγχόμενο περιβάλλον και δεν υποφέρει από τα προβλήματα ανάλυσης και ανισορροπίας κλάσεων που συναντάμε σε ρεαλιστικά σενάρια ανάλυσης ενεργειών [8]. Επίσης τα άτομα εκτελούν μια συγκεκριμένη προκαθορισμένη απλή ενέργεια καθόλη την διάρκεια του βίντεο στο οποίο συμμετέχουν, χωρίς να υπάρχουν περιθώρια αμφισημίας.

Το KTH περιέχει 6 τύπους ανθρώπινων ενεργειών (*περπάτημα*, *τζόκινγκ*, *τρέξιμο*, *σκιαμαχία*, *χαιρέτισμα* και *παλαμάκια*) πραγματοποιούμενες διαρκώς στο βίντεο από 25 διαφορετικά άτομα σε 4 διαφορετικά σενάρια (σε εξωτερικό χώρο s1, σε εξωτερικό χώρο με αλλαγές κλίμακας s2, σε εξωτερικό χώρο με διαφορετικά ρούχα s3 και σε εσωτερικό χώρο s4). Η κάμερα είναι στατική και το φόντο ομογενές, ενώ υπάρχουν μεγάλες διαφορές φωτεινότητας και κλίμακας από βίντεο σε βίντεο . Παραδείγματα ενεργειών του KTH φαίνονται στην εικόνα 4.10.



Εικόνα 4.10: Παρουσίαση ενεργειών του συνόλου δεδομένων KTH

Εξάγουμε τα C3D χαρακτηριστικά όπως και στην προηγούμενη ενότητα και χρησιμοποιούμε τον χωρισμό δεδομένων εκπαίδευσης, επαλήθευσης, εκπαίδευσης που προτείνουν και οι δημιουργοί [52] της βάσης με 8 άτομα για εκπαίδευση, 8 άλλα για επαλήθευση και άλλα 9 για δοκιμή. Συγκεκριμένα, κάνοντας εκπαίδευση στο σύνολο εκπαίδευσης και αξιολογώντας στο σύνολο επαλήθευσης επιλέγουμε τις υπερπαραμέτρους του SVM που θα χρησιμοποιήσουμε για αξιολόγηση στο σύνολο δοκιμής. Αντί για κατηγοριοποίηση σε επίπεδο βίντεο, κάνουμε κατηγοριοποίηση σε επίπεδο κλιπ (16 καρέ). Έτσι καταλήγουμε στα αποτελέσματα των πινάκων 4.7, 4.8.

Κλάση	**Ακρίβεια**	**Ανάκληση**	f_1**-σκορ**	**#Παραδειγμάτων**
Σκιαμαχία	0.77	0.87	0.82	810
Παλαμάκια	0.80	0.68	0.74	757
Χαιρέτισμα	0.82	0.81	0.76	929
Περπάτημα	0.65	0.68	0.67	766
Τζόκινγκ	0.42	0.42	0.42	478
Τρέξιμο	0.51	0.67	0.58	338
Σταθμισμένος μέσος/σύνολο	0.70	0.69	0.70	4078

Πίνακας 4.7: : Πίνακας αποτελεσμάτων των μικρο-μεγεθών ακρίβειας, ανάκλησης και f1-σκορ στο KTH

		Προβλέψεις					
	Κλάση	**Σκιαμαχία**	**Παλαμάκια**	**Χαιρέτισμα**	**Τζόκινγκ**	**Τρέξιμο**	**Περπάτημα**
Αλήθεια	**Σκιαμαχία**	**704**	15	34	2	5	50
	Παλαμάκια	108	**516**	112	0	3	18
	Χαιρέτισμα	94	111	**660**	2	3	59
	Τζόκινγκ	0	0	0	**200**	154	124
	Τρέξιμο	0	0	0	86	**225**	27
	Περπάτημα	3	0	0	187	52	**524**

Πίνακας 4.8: Μήτρα σύγχυσης των διαφορετικών κλάσεων του KTH

Εύκολα διαπιστώνει κανείς από την μήτρα σύγχυσης του 4.8 ότι περισσότερο μπερδεύονται μεταξύ τους οι κλάσεις

- σκιαμαχία, παλαμάκια, χαιρέτισμα
- τζόκινγκ, τρέξιμο, περπάτημα.

Αυτό γίνεται πιο εμφανές αν γενικεύσουμε τις κλάσεις *σκιαμαχία, παλαμάκια, χαιρέτισμα* σε μια ενιαία κλάση *ενεργός*, και αντίστοιχα της κλάσεις *τζόκινγκ, τρέξιμο και περπάτημα* σε μία ενιαία κλάση *κίνηση* και στην συνέχεια επιχειρήσουμε να επανεκπαιδεύσουμε το μοντέλο μας. Σε αυτής την περίπτωση καταλήγουμε στους πίνακες 4.9, 4.10.

Κλάση	**Ακρίβεια**	**Ανάκληση**	**f_1-σκορ**	**#Παραδειγμάτων**
Ενεργός	1.00	0.96	0.98	2496
Κίνηση	0.94	1.00	0.97	1582
Σταθμισμένος μέσος/σύνολο	0.98	0.98	0.98	4078

Πίνακας 4.9: Πίνακας αποτελεσμάτων των μικρο-μεγεθών ακρίβειας, ανάκλησης και f1-σκορ με γενικοποιημένες κλάσεις του KTH

		Προβλέψεις	
Αλήθεια	Κλάση	**Ενεργός**	**Κίνηση**
	Ενεργός	**2402**	94
	Κίνηση	1	**1581**

Πίνακας 4.10: Πείραμα 2ο. Μήτρα σύγχυσης με γενικοποιημένες κλάσεις του KTH

Με τις γενικοποιήμενες κλάσεις ενεργός και κίνηση, παρατηρούμε σχεδόν τέλειο διαχωρισμό από τον κατηγοριοποιητή μας. Έτσι, δικαιολογείται σε ένα βαθμό η επιλογή των χαρακτηριστικών C3D για κατηγοριοποιήση των ενεργειών του CAVIAR που ανήκουν κυρίως σε τέτοιου είδους κλάσεις.

Τα αποτελέσματα στην βιβλιογραφία για το KTH, αφορούν κυρίως κατηγοριοποίηση σε επίπεδο βίντεο και γενικά έχουν κάποια στάδια προεπεξεργασίας που λείπουν από την δική μας μελέτη, όπως αφαίρεση παρασκηνίου και ανίχνευση ατόμων. συγκεντρώνουμε τα χαρακτηριστικά ανά βίντεο παίρνοντας τον μέσο όρο των χαρακτηριστικών που έχουμε εξάγει και εν συνεχεία κάνουμε L2 κανονικοποίηση. Η σύγκριση συνολικής σωστής κατηγοριοποίησης (accuracy) παρουσιάζεται στον πίνακα 4.11. Σε αυτή την περίπτωση, το accuracy είναι καλή μετρική της απόδοσης, αφού σε επίπεδο βίντεο οι κλάσεις είναι ίσα κατανεμημένες.

Συγκρίνοντας την επίδοση των C3D χαρακτηριστικών (βλ πίνακα 4.11), με άλλα χαρακτηριστικά που έχουν χρησιμοποιηθεί για το KTH, βλέπουμε ότι υπολείπονται σημαντικά. Καταφέρνουν όμως αξιοσέβαστη επίδοση, χωρίς κάποια προεπεξεργασία, όπως ανίχνευση ατόμων ή αφαίρεση παρασκηνίου. Καταλήγουμε ότι όσον αφορά την αναγνώριση απλών ενεργειών τα C3D χαρακτηριστικά είναι ικανά να συλλαμβάνουν μεγάλες διαφορές στις κινήσεις, ωστόσο υπολείπονται σε ότι αφορά τον λεπτεπίλεπτο διαχωρισμό τους. Επίσης, παρουσιάζουν πρόβλημα στο να συγκρατούν πληροφορία για την ταχύτητα των εμπλεκόμενων ατόμων στο βίντεο. Λείπει η πολύ μεγάλη κίνηση που εμφανίζεται στο SPORTS1M, αλλά επίσης και οι μεγάλες διαφορές χρώματος ανάμεσα σε διαφορετικές κλάσεις και έτσι δεν επιτρέπεται ομαλή μεταφορά μάθησης.

Μέθοδος	**Accuracy (%)**
ActionBank [50]	- / 98.2
iDT [65]	- / 95.0
Snippets [51]	90.9 / 92.7
Niebles και αλ. [39]	- / 81.5
Δικιά μας με C3D	0.69 / 72.4
Schuldt και αλ.[52]	- / 71.7
Yan Ke και αλ.[27]	- / 63.0

Πίνακας 4.11: Σύγκριση με άλλα αποτελέσματα σε ταξινόμηση ενεργειών στο KTH. Στην στήλη Accuracy, αριστερά της διαχωριστικής γραμμής '/' φαίνεται η κατηγοριοποίηση σε επίπεδο καρέ όπου είναι διαθέσιμη και δεξιά της η κατηγοριοποίηση σε επίπεδο βίντεο.

4.4.5 Σύνοψη

Σε αυτό το κεφάλαιο παρουσιάσαμε τα πειραματικά μας αποτελέσματα όσον αφορά την χρήση των C3D χαρακτηριστικών για αναγνώριση απλών ανθρώπινων ενεργειών. Αρχικά, αξιολογούμε την επίδοση στην βάση του CAVIAR και επισημαίνουμε την σημασία σε προβλήματα όρασης, η εκπαίδευση και η δοκιμή να γίνεται σε διαφορετικά βίντεο. Αξιολογούμε περαιτέρω αν τα C3D είναι κατάλληλη αναπαράσταση για την αναγνώριση ανθρώπινων ενεργειών δοκιμάζοντας τα στην πλέον χρησιμοποιημένη για αυτό το σκοπό βάση του KTH. Από όσο γνωρίζουμε, αυτή η εργασία είναι η πρώτη που επιχειρεί κάτι τέτοιο. Καταλήγουμε ότι δεν είναι τόσο ταιριαστά σε αυτήν την περίπτωση, κυρίως λόγω της βάσης δεδομένων που έχει χρησιμοποιηθεί για την αρχική εκπαίδευση του νευρωνικού.

Βλέποντας ότι το μοντέλο έχει σχεδόν τέλεια απόδοση κατηγοριοποιώντας στις κλάσεις *ενεργός* και *κίνηση* στο KTH, δικαιολογούμε την επιλογή της μεθόδου για κατηγοριοποίηση στο CAVIAR, που περιέχει τέτοιου είδους κλάσεις. Επιβεβαιώνουμε παράλληλα ότι πολλές λάθος κατηγοριοποιήσεις οφείλονται στην εισαγωγή θορύβου κατά την αλλαγή μεγέθους πολύ μικρών περικομμάτων και όχι στην αδυναμία των χαρακτηριστικών να γενικεύσουν.

Γενικά, η εξαγωγή χαρακτηριστικών από ένα προεκπαιδευμένο νευρωνικό ανήκει στην κατηγορία μεταφερόμενης μάθησης. Για να είναι αξιόλογη η μεταφορά μάθησης θα πρέπει το dataset στο οποίο έχει γίνει η αρχική εκπαίδευση να παρουσιάζει ομοιότητες με αυτό από το

οποίο εξάγουμε χαρακτηριστικά. Πολλά από τα προβλήματα που αντιμετωπίζουμε με τα C3D χαρακτηριστικά οφείλονται σε αυτήν την αστοχία. Συγκεκριμένα, το SPORTS1M στο οποίο έχει εκπαιδευτεί το νευρωνικό που χρησιμοποιούμε παρουσιάζει εξαιρετικά περισσότερη κίνηση [37] από ότι το CAVIAR στο οποίο το δοκιμάζουμε. Ενώ το SPORTS1M περιέχει ασυνήθιστες και πολύπλοκες δραστηριότητες που δεν συναντώνται στην καθημερινή ζωή [54, 55] , το CAVIAR περιέχει πολύ απλές ενέργειες. Στο πείραμα με το KTH αναγνωρίζουμε και τα 2 παραπάνω προβλήματα. Βεβαίως αναγνωρίζουμε και το γεγονός ότι το C3D νευρωνικό είχε αρχικά εκπαιδευτεί για πιο χοντροκομμένη αναγνώριση σε επίπεδο βίντεο και όχι σε επίπεδο κλιπ. Ένας άλλος παράγοντας είναι ότι το πρόβλημα μας είχε ως προϋπόθεση την ανίχνευση των ατόμων που σε πολλές περιπτώσεις γινόταν στην άκρη του πεδίου της κάμερας. Έτσι, λόγω χαμηλής ανάλυσης του περικόμματος, με αλλαγή μεγέθους μεγεθύναμε αισθητά τον θόρυβο στην είσοδο του νευρωνικού.

4.5 OLED για αναγνώριση σύνθετων ενεργειών

4.5.1 Λεπτομέρειες Υλοποίησης

Με την τεχνική άσε-ένα-έξω της προηγούμενης ενότητας, κάθε βίντεο χρησιμοποιείται για δοκιμή όταν όλα τα υπόλοιπα χρησιμοποιούνται για εκπαίδευση. Έτσι, παράγουμε νέες επισημειώσεις των απλών ενεργειών για όλα τα βίντεο εξαντλητικά με την ταξινόμηση που παρουσιάσαμε στο 4.4.3. Από εδώ και πέρα, θα αναφερόμαστε στις επισημειώσεις που έχουν προκύψει από την παραπάνω ταξινόμηση ως *αυτόματα παραγμένες* και σε αυτές που δίνονται από το σύνολο δεδομένων του CAVIAR, ως *πραγματικές ή μη αυτόματες*. Αξιολογούμε την χρήση των αυτόματα παραγμένων επισημειώσεων για τις απλές ενέργειες και συγκρίνουμε με την επίδοση που παίρνουμε χρησιμοποιώντας τις μη αυτόματες επισημειώσεις, για αναγνώριση σύνθετων ενεργειών.

Σε αυτό το μέρος της εργασίας θέλουμε να χρησιμοποιήσουμε τις αυτόματα παραγόμενες χαμηλού επιπέδου ενέργειες (XEE) ως είσοδο στον OLED που παρουσιάσαμε στην ενότητα 3.2.1, για να μάθουμε θεωρίες Λογισμού Γεγονότων που θα μας επιτρέψουν να αναγνωρίσουμε υψηλού επιπέδου ενέργειες (YEE). Όπως και ο συγγραφέας του OLED, θα προσπαθήσουμε να μάθουμε θεωρίες μόνο για τις YEE *συνάντηση* (meeting) και *ομαδική κίνηση* (moving), γιατί οι υπόλοιπες εμφανίζονται σε πολύ μικρό παραδειγμάτων.

Ο κώδικας για τον OLED είναι ελεύθερα διαθέσιμος[13]. Για να τον χρησιμοποιήσουμε με αυτόματα παραγμένα δεδομένα, τα μετατρέπουμε σε γλώσσα Λογισμού Γεγονότων όπως και στο [1] είχαν κάνει με τα πραγματικά δεδομένα.

[13] https://github.com/nkatzz/OLED

Σύμφωνα με την μέθοδο που χρησιμοποιείται και στο [26], χωρίζουμε το σύνολο των δεδομένων σε χρονικές περιόδους $holdAt$ και $\neg\ holdsAt$ για την ενέργεια που μας ενδιαφέρει και μετά από 10πλή διασταυρωμένη επικύρωση, υπολογίζουμε τα μικρο-μεγέθη ακρίβειας, ανάκλησης και f_1-σκορ. Κατά την αναπαραγωγή των αποτελεσμάτων χρησιμοποιούμε τις συνιστώμενες για το πρόβλημα, παραμέτρους, αλλά μειώνουμε το κατώφλι κλαδέματος σε 0.5 (από 0.7) για την *συνάντηση* και 0.4 (από 0.5) για την *ομαδική κίνηση*.

Αφού στην περίπτωση μας για να βρεθούν καλές θεωρίες πρέπει να μειώσουμε το κατώφλι κλαδέματος κανόνων, συνεχίζουμε με μια μελέτη της επίδρασης αυτής της παραμέτρου σε κάθε περίπτωση.

4.5.2 Αποτελέσματα

Ενέργεια	**Δεδομένα**	**TPs**	**FPs**	**FNs**	**Ακρίβεια**	**Ανάκληση**	f_1
Συνάντηση	Πραγματικά	2750	226	844	**0.92**	0.77	**0.84**
	Αυτόματα	**2787**	1159	807	0.71	**0.78**	0.74
Ομαδική Κίνηση	Πραγματικά	**4700**	3314	1583	**0.59**	**0.75**	**0.66**
	Αυτόματα	3967	5773	2315	0.41	0.63	0.50

Πίνακας 4.12: Πίνακας αποτελεσμάτων συνολικών αληθών θετικών, αληθών αρνητικών, αληθών αρνητικών και των μικρο-μεγεθών ακρίβειας, ανάκλησης και f1-σκορ.

Στον πίνακα 4.12 παρουσιάζονται τα αποτελέσματα του πειράματός μας. Όπως ήταν αναμενόμενο, τα αποτελέσματα που παίρνουμε με τα αυτόματα παραγμένα δεδομένα, είναι σε κάθε περίπτωση χειρότερα από εκείνα που παίρνουμε όταν χρησιμοποιούμε τα πραγματικά. Σε αυτό φταίει ο αρκετά αυξημένος αριθμός των FP που είναι επόμενο αποτέλεσμα του θορύβου που εισάγει το μοντέλο του προηγούμενο επιπέδου.

Για την ενέργεια της *συνάντησης*, τα αποτελέσματα που παίρνουμε είναι σε μεγάλο βαθμό συγκρίσιμα. Αυτό έχει να κάνει και με την φύση της συγκεκριμένης σύνθετης ενέργειας που παρουσιάζει λιγότερο θόρυβο [2].

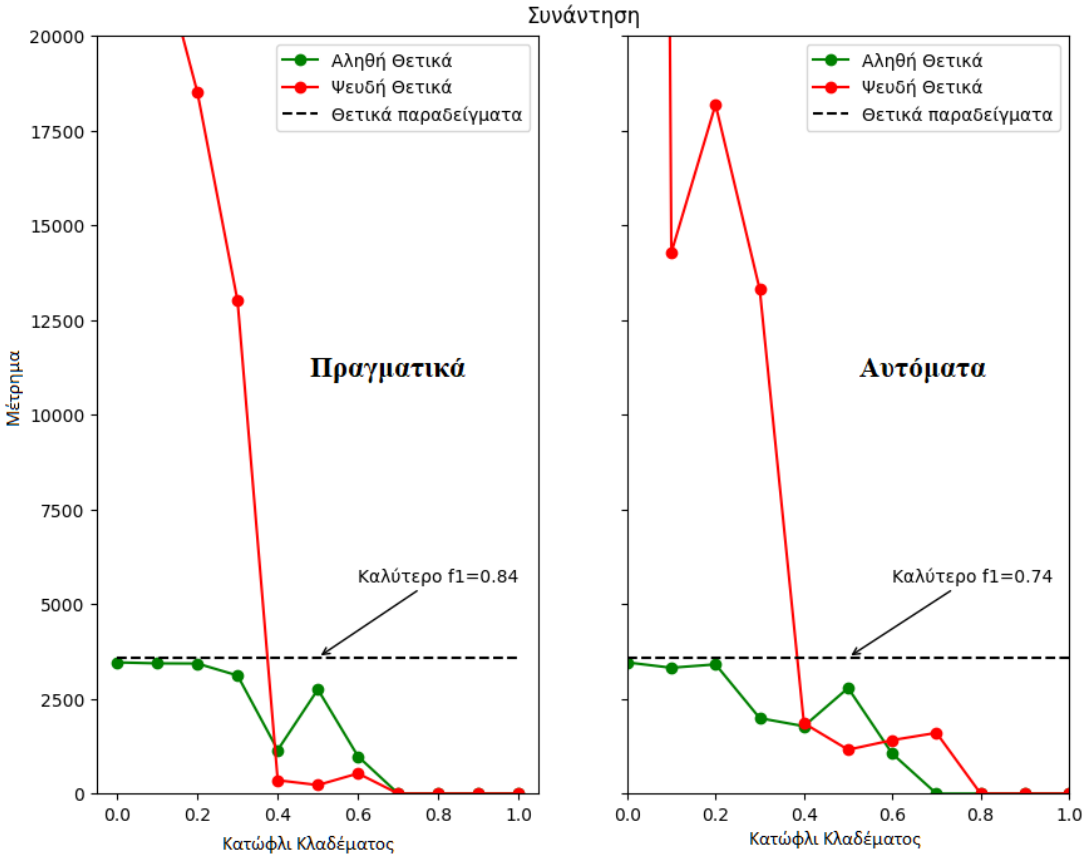
Στην *ομαδική κίνηση* βλέπουμε αρκετά χαμηλότερη επίδοση λόγω του σχεδόν διπλάσιου αριθμού FP. Εξηγείται από το είδος του θορύβου που παίρνουμε στις ΧΕΕ. Βλέποντας τους καλύτερους κανόνες που σχηματίζονται με τα πραγματικά δεδομένα παρατηρούμε ότι τόσο η ΧΕΕ *ενεργός*, όσο και η ΧΕΕ *τρέξιμο* εμφανίζονται πολλές φορές στους κανόνες terminatedAt της ομαδικής κίνησης. Ωστόσο, η ΧΕΕ *τρέξιμο* μπερδεύεται πάντα με *περπάτημα* και η ΧΕΕ *ενεργός* σε έναν μεγάλο βαθμό. Έτσι, σε πολλές περιπτώσεις δεν επιτρέπεται ο έγκαιρος τερματισμός, αφού με τα αυτόματα παραγμένα δεδομένα δεν είναι

δυνατό να υπάρχουν κανόνες τερματισμού με την ΧΕΕ *τρέξιμο*, ενώ οι κανόνες που περιέχουν την ΧΕΕ ενεργός είναι πιο θορυβώδεις.

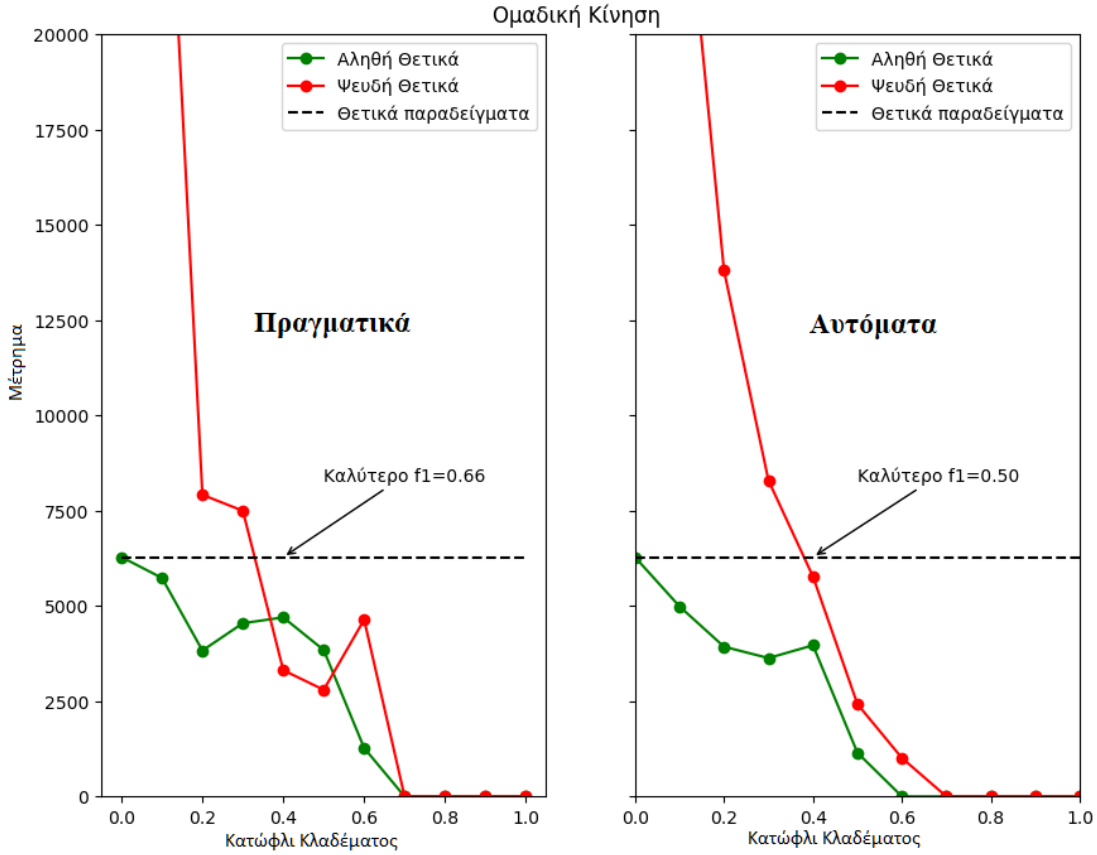
Ποιοτικά ελέγχοντας τις θεωρίες, παρατηρούμε ότι υπάρχουν πολλές περιπτώσεις που παράγονται ίδιοι κανόνες με τα πραγματικά και τα αυτόματα δεδομένα. Σε όλες τις περιπτώσεις αυτοί οι κανόνες έχουν μικρότερο σκορ στα αυτόματα παραγμένα δεδομένα από ότι στα πραγματικά. Στην συνέχεια, η παραγόμενη θεωρία πετυχαίνει μικρότερο f1-σκορ στα δοκιμαστικά δεδομένα. Αυτό μας επιτρέπει να συμπεράνουμε ότι το σκορ που χρησιμοποιείται για τους κανόνες παρέχει άμεση σύνδεση με την απόδοση της τελικής θεωρίας και επομένως είναι δικαιολογημένη και πρακτικά η επιλογή της για ευριστική.

Τα διαγράμματα των εικόνων 4.11, 4.12 συνοψίζουν την επίδραση του κατωφλίου κλαδέματος κανόνων στην δημιουργία θεωριών. Εύκολα διαπιστώνουμε την σημασία της μελέτης αυτής της παραμέτρου, αφού είναι εμφανές το γεγονός ότι η απόδοση των παραγόμενων θεωριών είναι πολύ ευαίσθητη σε αλλαγές της. Θυμίζουμε ότι το κλάδεμα γίνεται με βάση το σκορ των κανόνων όπως ορίστηκε στον Ορισμό 3.2 της ενότητας 3.2.1. Κατώφλι κλαδέματος 0.7 σημαίνει ότι όσοι κανόνες έχουν σκορ μικρότερο του 0.7 και είναι στατιστικά απίθανο να βελτιωθούν, εξαλείφονται από την θεωρία.

Γενικά, βλέπουμε ότι τόσο για τα πραγματικά, όσο για τα αυτόματα παραγμένα δεδομένα παρατηρείται το ίδιο μοτίβο με τις διαφορές που περιγράψαμε προηγουμένως (ότι δηλαδή στην περίπτωση των αυτόματων δεδομένων έχουμε περισσότερα FP). Κατά τα άλλα, ποιοτικά οι γραφικές μοιάζουν τόσο για τα διαφορετικά είδη των δεδομένων, όσο και για τις διαφορετικές ενέργειες που μελετάμε.



Εικόνα 4.11: Διαγράμματα συνολικών αληθών θετικών, ψευδών θετικών, ψευδών αρνητικών για την ενέργεια της *συνάντησης* μετά από 10πλή Διασταυρωμένη Επικύρωση για διάφορες τιμές κατωφλίου κλάδεματος κανόνων 0.0..1.0. Αριστερά με τα *πραγματικά* δεδομένα, δεξιά με τα *αυτόματα παραγμένα*.



Εικόνα 4.12: Διαγράμματα συνολικών αληθών θετικών, ψευδών θετικών, ψευδών αρνητικών για την ενέργεια της *ομαδικής κίνησης* μετά από 10πλή Διασταυρωμένη Επικύρωση για τιμές κατωφλίου κλαδέματος κανόνων 0.0..1.0. Αριστερά με τα *πραγματικά* δεδομένα, δεξιά με τα *αυτόματα παραγμένα.*

Συγκεκριμένα:

- όταν έχουμε ελάχιστο ή καθόλου κλάδεμα, παρατηρείται πολύ υψηλός αριθμός FP, ενώ τα TP είναι ίσα ή σχεδόν ίσα με τον συνολικό αριθμό θετικών παραδειγμάτων. Κοιτώντας τις θεωρίες που παράγονται σε αυτές τις περιπτώσεις, βλέπουμε ότι αυτό οφείλεται στο γεγονός ότι έχουμε πολύ «ανοιχτούς» initiatedAt κανόνες που υπερισχύουν έναντι των terminatedAt, στην περίπτωση ενεργοποιηθούν ταυτόχρονα.
- καθώς το κλάδεμα μεγαλώνει, τα FP πέφτουν, αφού οι κανόνες initiatedAt εξειδικεύονται. Με παράμετρο κλάδεμα μεγαλύτερη από 0.7 σε καμία από τις περιπτώσεις δεν σχηματίζεται κανόνας initiatedAt, γεγονός που οδηγεί σε μηδενικά TP.
- ένα ενδιαφέρον μοτίβο που παρουσιάζεται σε όλες τις γραφικές είναι ότι τα TP παρουσιάζουν ένα τοπικό ελάχιστο πριν παρουσιάσουν τοπικό μέγιστο στην τιμή κλαδέματος που έχουμε και το καλύτερο f1-σκορ. Από μόνη της η εξάλειψη κανόνων initiatedAt όμως, θα έπρεπε να φέρνει διαρκώς λιγότερα TP. Ωστόσο, η εξήγηση του

παραπάνω μοτίβου έρχεται από το γεγονός ότι μαζί τους εξαλείφονται και οι χαμηλού σκορ terminatedAt που τερματίζουν λανθασμένα τα σύνθετα γεγονότα .

4.5.3 Σύνοψη

Σε αυτήν την ενότητα χρησιμοποιούμε τον OLED, ένα σύστημα μάθησης θεωριών Λογισμού Γεγονότων, με ατελή δεδομένα που έχουν προκύψει από κάποιον ταξινομητή στο προηγούμενο επίπεδο. Κάνουμε σύγκριση με δεδομένα που έχουν προκύψει με χειροκίνητη επισημείωση καθώς και παρουσιάζουμε αναλυτικά τους λόγους που βλέπουμε διαφορές στην απόδοση αναγνώρισης σύνθετων γεγονότων σε κάθε περίπτωση. Οι παραγόμενες θεωρίες μοιάζουν στις δύο περιπτώσεις και είναι συγκρίσιμες σε απόδοση , αλλά καταλήγουμε ότι κάνοντας μάθηση σε δύο επίπεδα αποκομμένα μεταξύ τους, τα λάθη του πρώτου επιπέδου μάθησης μεταφράζονται σε θόρυβο στην είσοδο του δεύτερου.

Τέλος, μελετούμε αναλυτικά την σημαντική επίδραση που έχει η επιλογή της παραμέτρου κλαδέματος για την απόδοση της τελικής θεωρίας και αναγνωρίζουμε ένα μοτίβο που ακολουθεί και μπορεί να βοηθήσει στην αναζήτηση μιας καλής τιμής της.

5 *Επίλογος*

5.1 Σύνοψη και συμπεράσματα

Σε αυτή την εργασία μελετήθηκε ένα σενάριο μάθησης δύο επιπέδων για αναγνώριση ανθρώπινων ενεργειών σε βίντεο. Στο πρώτο επίπεδο, εξάγουμε Βαθιά Μαθημένα χαρακτηριστικά από ένα προηγμένο συνελικτικό νευρωνικό δίκτυό τα οποία χρησιμοποιούμε για να κάνουμε ταξινόμηση απλών ενεργειών με μηχανές διανυσμάτων υποστήριξης. Αξιολογώντας τα C3D χαρακτηριστικά για αναγνώριση απλών ενεργειών και καταλήξαμε ότι δεν ήταν ταιριαστά, λόγω προβλημάτων που αφορούν σε μεταφορά μάθησης από εκπαίδευση στο SPORTS1M, αλλά και του στατικού μεγέθους της εισόδου που περιμένει το νευρωνικό. Στο δεύτερο επίπεδο, χρησιμοποιούμε την ταξινόμηση του προηγούμενου επιπέδου για τις απλές ενέργειες για να μάθουμε θεωρίες Λογισμού Γεγονότων που μας επιτρέπουν να αναγνωρίσουμε πιο σύνθετες ανθρώπινες ενέργειες με απόδοση συγκρίσιμη με την περίπτωση που η είσοδος των απλών ενεργειών έχει προκύψει με χειρωνακτική επισημείωση.

Θεωρούμε σημαντικό το πέρασμα από πλήρως αδιαφανείς αρχιτεκτονικές, σε περισσότερο διαφανείς που θα επιτρέπουν ερμηνεία των αποφάσεων τους. Ένας τρόπος για να γίνει αυτό είναι με συστήματα Λογικού Προγραμματισμού, που στηρίζονται στα γερά θεμέλια της Μαθηματικής Λογικής. Σε αυτήν την εργασία προτείναμε να συνδυάσουμε την αδιαφανή αποτελεσματικότητα των τεχνητών νευρωνικών δικτύων με την διαφανή συμπερασματολογία του Λογισμού Γεγονότων. Δεν σταθήκαμε επιτυχείς στην ομαλή σύνδεση τους, ωστόσο φέραμε τις δύο φιλοσοφίες πιο κοντά σε μία μελλοντική τους συνέργεια.

Θα ήθελα να ολοκληρώσω, λέγοντας ότι αισθάνομαι πολύ τυχερός που στα πλαίσια της διπλωματική μου μελέτησα ένα πολύ μεγάλο εύρος του τομέα της μηχανικής μάθησης, ο οποίος στις μέρες μας είναι μας είναι πιο επίκαιρος από ποτέ.

5.2 Μελλοντικές επεκτάσεις

Παραθέτουμε παρακάτω μερικά προβλήματα που συναντήσαμε και θα θέλαμε να αντιμετωπίσουμε, καθώς και θέματα για μελλοντική έρευνα:

- Αντί για κατηγοριοποίηση των διαφορών ενεργειών, θα θέλαμε να κάνουμε μη επιβλεπόμενη συσταδοποίηση (unsupervised clustering). Στην συνέχεια, σε κάθε συστάδα θα ανατεθεί ένα διακριτό σύμβολο. Τα συστατικά κάθε συστάδας θα λάβουν το αντίστοιχο σύμβολο και η παραγόμενη ροή συμβόλων θα αποτελέσει είσοδο στον OLED. Έτσι, θα μάθουμε θεωρίες λογισμού γεγονότων για τις σύνθετες ενέργειες, με απλές ενέργειες που θα έχουν προκύψει με μη επιβλεπόμενο τρόπο.
- Τα προβλήματα στην μεταφορά μάθησης που συναντήσαμε με τα C3D χαρακτηριστικά, θα μπορούσαν ενδεχομένως να εξαλειφθούν με χρήση χειροποίητων χαρακτηριστικών (hand-crafted features) για αναπαράσταση βίντεο.
- Θα θέλαμε να προσαρμόσουμε το δίκτυο του C3D, με μάθηση στο CAVIAR, και να αφήσουμε το νευρωνικό να κάνει όλη την ταξινόμηση χωρίς να χρειάζεται να εξάγουμε χαρακτηριστικά. Λόγω λίγων δεδομένων θα θέλαμε να πειραματιστούμε με τεχνικές συνθετικής αύξησης οπτικών δεδομένων για βίντεο.
- Εκτελώντας αυτόματη ανίχνευση των ανθρώπων στα βίντεο θα ήμαστε σε θέση να δούμε όλη την διαδικασία της αναγνώρισης ανθρώπινων ενεργειών.
- Αντί για συνδυασμό δύο αποκομμένων συστημάτων, θέλουμε να επιτύχουμε καλύτερη συγχώνευση της μάθηση θεωριών Λογισμού Γεγονότων με την παραγωγή των ροών χαμηλού επιπέδου ενεργειών.

6 *Βιβλιογραφία*

[1] Alexander Artikis, Marek Sergot, and Georgios Paliouras. Run-time composite event recognition. In *Proceedings of the 6th ACM International Conference on Distributed Event-Based Systems*, pages 69–80. ACM, 2012.

[2] Alexander Artikis, Marek Sergot, and Georgios Paliouras. An event calculus for event recognition. *IEEE Transactions on Knowledge and Data Engineering*, 27(4):895–908, 2015.

[3] Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia, and Atilla Baskurt. Sequential deep learning for human action recognition. In *International Workshop on Human Behavior Understanding*, pages 29–39. Springer, 2011.

[4] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[5] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1395–1402. IEEE, 2005.

[6] Gary Bradski and Adrian Kaehler. *Learning OpenCV: Computer vision with the OpenCV library*. " O'Reilly Media, Inc.", 2008.

[7] William Brendel, Alan Fern, and Sinisa Todorovic. Probabilistic event logic for interval-based event recognition. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3329–3336. IEEE, 2011.

[8] Jose M CHAQUET, Enrique J CARMONA, and Antonio FERNANDEZ-CABALLERO. A survey of video datasets for human action and activity recognition. *Computer vision and image understanding*, 117(6):633–659, 2013.

[9] Wongun Choi, Khuram Shahid, and Silvio Savarese. Learning context for collective activity recognition. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3273–3280. IEEE, 2011.

[10] François Chollet et al. Keras, 2015.

[11] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[12] Christophe Dousson and Pierre Le Maigat. Chronicle recognition improvement using temporal focusing and hierarchization. In *IJCAI*, volume 7, pages 324–329, 2007.

[13] Kunihiko Fukushima and Sei Miyake. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pages 267–285. Springer, 1982.

[14] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[15] Ross Goroshin, Joan Bruna, Jonathan Tompson, David Eigen, and Yann LeCun. Unsupervised feature learning from temporal data. *arXiv preprint arXiv:1504.02518*, 2015.

[16] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 733–742, 2016.

[17] Samitha Herath, Mehrtash Harandi, and Fatih Porikli. Going deeper into action recognition: A survey. *arXiv preprint arXiv:1605.04988*, 2016.

[18] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

[19] Nathalie Japkowicz, Stephen Jose Hanson, and Mark A Gluck. Nonlinear autoassociation is not equivalent to pca. *Neural computation*, 12(3):531–545, 2000.

[20] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2013.

[21] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014.

[22] Eric Jones, Travis Oliphant, and Pearu Peterson. {SciPy}: open source scientific tools for {Python}. 2014.

[23] Manohar Karki, Saikat Basu, Robert DiBiano, Supratik Mukhopadhyay, Jerry Weltman, and Malcolm Stagg. A symbolic framework for recognizing activities in full motion surveillance videos.

[24] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In

Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 1725–1732, 2014.

[25] Nikos Katzouris, Alexander Artikis, and Georgios Paliouras. Incremental learning of event definitions with inductive logic programming. *Machine Learning*, 100(2-3):555–585, 2015.

[26] Nikos Katzouris, Alexander Artikis, and Georgios Paliouras. Online learning of event definitions. *arXiv preprint arXiv:1608.00100*, 2016.

[27] Yan Ke, Rahul Sukthankar, and Martial Hebert. Efficient visual event detection using volumetric features. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 166–173. IEEE, 2005.

[28] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[29] Alexander Klaser, Marcin Marszaek, and Cordelia Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC 2008-19th British Machine Vision Conference*, pages 275–1. British Machine Vision Association, 2008.

[30] Robert Kowalski and Marek Sergot. A logic-based calculus of events. *New generation computing*, 4(1):67–95, 1986.

[31] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[32] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[33] David G Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999.

[34] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[35] Pyry Matikainen, Martial Hebert, and Rahul Sukthankar. Trajectons: Action recognition through the motion analysis of tracked features. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 514–521. IEEE, 2009.

[36] Evangelos Michelioudakis, Anastasios Skarlatidis, Georgios Paliouras, and Alexander Artikis. Oslα: Online structure learning using background knowledge axiomatization. In *European Conference of Machine Learning and Knowledge Discovery in Databases*, 2016.

[37] Alberto Montes, Amaia Salvador, and Xavier Giro-i Nieto. Temporal activity detection in untrimmed videos with recurrent neural networks. *arXiv preprint arXiv:1608.08128*, 2016.

[38] Back-Propagation Network. Handwritten digit recognition with. 1989.

[39] Juan Carlos Niebles, Hongcheng Wang, and Li Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International journal of computer vision*, 79(3):299–318, 2008.

[40] David Nister and Henrik Stewenius. Scalable recognition with a vocabulary tree. In *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, volume 2, pages 2161–2168. Ieee, 2006.

[41] Adrian Paschke. Eca-ruleml: An approach combining eca rules with temporal interval-based kr event/action logics and transactional update logics. *arXiv preprint cs/0610167*, 2006.

[42] Adrian Paschke and Alexander Kozlenkov. Rule-based event processing and reaction rules. In *International Workshop on Rules and Rule Markup Languages for the Semantic Web*, pages 53–66. Springer, 2009.

[43] Kostas Patroumpas, Alexander Artikis, Nikos Katzouris, Marios Vodas, Yannis Theodoridis, and Nikos Pelekis. Event recognition for maritime surveillance. In *EDBT*, pages 629–640, 2015.

[44] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

[45] Florent Perronnin and Christopher Dance. Fisher kernels on visual vocabularies for image categorization. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[46] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. *Computer Vision–ECCV 2010*, pages 143–156, 2010.

[47] M Ragan-Kelley, F Perez, B Granger, T Kluyver, P Ivanov, J Frederic, and M Bussonier. The jupyter/ipython architecture: a unified view of computational research, from interactive exploration to communication and publication. In *AGU Fall Meeting Abstracts*, volume 1, page 07, 2014.

[48] Pedro Canotilho Ribeiro, José Santos-Victor, and P Lisboa. Human activity recognition from video: modeling, feature selection and classification architecture. In *Proceedings of International Workshop on Human Activity Recognition and Modelling*, pages 61–78. Citeseer, 2005.

[49] Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine learning*, 62(1):107–136, 2006.

[50] Sreemanananth Sadanand and Jason J Corso. Action bank: A high-level representation of activity in video. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1234–1241. IEEE, 2012.

[51] Konrad Schindler and Luc Van Gool. Action snippets: How many frames does human action recognition require? In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[52] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: A local svm approach. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 32–36. IEEE, 2004.

[53] Paul Scovanner, Saad Ali, and Mubarak Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th ACM international conference on Multimedia*, pages 357–360. ACM, 2007.

[54] Shikhar Shrestha. Learning deep representations for human activity in video.

[55] Gunnar A Sigurdsson, Santosh Divvala, Ali Farhadi, and Abhinav Gupta. Asynchronous temporal fields for action recognition. *arXiv preprint arXiv:1612.06371*, 2016.

[56] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[57] Muralikrishna Sridhar, Anthony G Cohn, and David C Hogg. Unsupervised learning of event classes from video. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, pages 1631–1638. AAAI Press, 2010.

[58] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[59] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *International Conference on Machine Learning*, pages 843–852, 2015.

[60] Graham Taylor, Rob Fergus, Yann LeCun, and Christoph Bregler. Convolutional learning of spatio-temporal features. *Computer Vision–ECCV 2010*, pages 140–153, 2010.

[61] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4489–4497, 2015.

[62] Son Tran and Larry Davis. Event modeling and recognition using markov logic networks. *Computer vision–ECCV 2008*, pages 610–623, 2008.

[63] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 588–595, 2014.

[64] Chunyu Wang, Yizhou Wang, and Alan L Yuille. An approach to pose-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 915–922, 2013.

[65] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Dense trajectories and motion boundary descriptors for action recognition. *International journal of computer vision*, 103(1):60–79, 2013.

[66] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3551–3558, 2013.

[67] Junji Yamato, Jun Ohya, and Kenichiro Ishii. Recognizing human action in time-sequential images using hidden markov model. In *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR'92., 1992 IEEE Computer Society Conference on*, pages 379–385. IEEE, 1992.