Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Αγρονομων Τοπογραφων Μηχανικων
Εργαστηριο Τηλεπισκοπησης

Αναγνώριση και Ταξινόμηση Δράσεων σε Βίντεο Προ-κλινικών Πειραμάτων με Τεχνικές Μηχανικής Μάθησης

Διπλωματικη Εργασια

Αλεξανδρος Βυθουλκας

Αθήνα, Ιούλιος 2019

National Technical University of Athens
School of Rural and Surveying Engineering
Remote Sensing Laboratory

Action Recognition and Classification in Pre-Clinical Experiment Videos with Machine Learning Techniques

Thesis Project

Alexandros Vythoulkas

Athens, July 2019

Αναγνώριση και Ταξινόμηση Δράσεων σε Βίντεο Προ-κλινικών Πειραμάτων με Τεχνικές Μηχανικής Μάθησης

Διπλωματικη Εργασια

Αλέξανδρος Βυθούλκας

Επιβλέπων: Κωνσταντίνος Καράντζαλος
Αναπληρωτής Καθηγητής ΕΜΠ

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 5η Ιουλίου 2019.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Κωνσταντίνος Καράντζαλος	Χριστίνα Δάλλα	Αναστάσιος Δουλάμης
Αναπληρωτής Καθηγητής ΕΜΠ	Αναπληρώτρια Καθηγήτρια ΕΚΠΑ	Επίκουρος Καθηγητής ΕΜΠ

Αθήνα, Ιούλιος 2019

RSLab
Remote Sensing Laboratory
National Technical University of Athens
√ Sensing √ Analytics √ Monitoring

Υπεύθυνη Δήλωση
Βεβαιώνω ότι είμαι συγγραφέας αυτής της πτυχιακής εργασίας, και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην πτυχιακή εργασία. Επίσης έχω αναφέρει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επίσης, βεβαιώνω ότι αυτή η πτυχιακή εργασία προετοιμάστηκε από εμένα προσωπικά ειδικά για τις απαιτήσεις του προγράμματος σπουδών του Τμήματος Αγρονόμων Τοπογράφων Μηχανικών του Εθνικού Μετσόβιου Πολυτεχνείου.

(Υπογραφή)

.....
Αλέξανδρος Βυθούλκας

Περίληψη

Η ραγδαία εξέλιξη του κλάδου της μηχανικής μάθησης, λόγω της ανάπτυξης της βαθιάς μάθησης, έχει σημάνει μεγάλες αλλαγές στην αυτοματοποίηση της επίλυσης διαφόρων προβλημάτων, που μέχρι πρότινος απαιτούσαν ώρες ανθρώπινης επαναλαμβανόμενης εργασίας. Στην παρούσα διπλωματική, σκοπός είναι η αυτοματοποίηση της αναγνώρισης της συμπεριφοράς, κατά την δοκιμασία εξαναγκασμένης κολύμβησης σε επιμύες, με τεχνικές μηχανικής μάθησης. Το πείραμα αυτό αποτελεί ένα σύνηθες μέσο για τη μελέτη της επίδρασης αντικαταθλιπτικών φαρμάκων. Πρόκειται για την τοποθέτηση επιμυών σε κυλίνδρους με νερό για τη μέτρηση του χρονικού διαστήματος ακινησίας, κολύμβησης και αναρρίχησης του υποκειμένου. Για την πραγματοποίηση επιβλεπόμενης ταξινόμησης, χρησιμοποιήθηκε dataset με βίντεο 8 ωρών περιλαμβανομένων των αντίστοιχων αληθών τιμών δύο ειδικών παρατηρητών. Έπειτα από διόρθωση και επεξεργασία του dataset, υλοποιήθηκαν μοντέλα εκτίμησης της συμπεριφοράς από την ανάλυση δεδομένων των βίντεο. Για την πρόβλεψη της συμπεριφοράς σχεδιάστηκαν και εφαρμόστηκαν τόσο συμβατικές τεχνικές αναγνώρισης συμπεριφοράς όσο και τεχνικές βαθιάς μάθησης με τη χρήση νευρωνικών δικτύων. Αρχικά εφαρμόστηκε ο αλγόριθμος πυκνών τροχιών, για την εξαγωγή χωροχρονικών περιγραφών, κωδικοποίηση με Fisher Vectors και ταξινόμηση τους με Μηχανές Διανυσμάτων Υποστήριξης. Στη συνέχεια με χρήση της αρχιτεκτονικής τεχνητών νευρωνικών δικτύων Inflated 3D, κατάλληλη για αναγνώριση δράσης, πραγματοποιήθηκε βελτιστοποίηση παραμέτρων. Ακόμη σχεδιάστηκαν και βελτιστοποιήθηκαν αρχιτεκτονικές με συνδυασμούς δισδιάστατων συνελικτικών δικτύων για εξαγωγή χωρικών χαρακτηριστικών και ανατροφοδοτούμενων δικτύων για τη συσχέτιση τους στο χρονικό πεδίο. Η ταξινόμηση των περιγραφών του αλγορίθμου πυκνών τροχιών, επιτυγχάνει ικανοποιητική ακρίβεια στις δύο κυρίαρχες κατηγορίες του dataset. Τα δίκτυα LSTM παρουσιάζουν αντίστοιχης ποιότητας αποτελέσματα. Βελτίωση στο πρόβλημα της ανισορροπίας των κατηγοριών, σημειώνει η αρχιτεκτονική Inflated 3D, με τη χρήση προεκπαιδευμένων βαρών, με ευστοχία 82% στα δείγματα που συμφωνούν οι ειδικοί παρατηρητές. Τα σφάλματα του μοντέλου εντοπίζονται στις περιπτώσεις όπου υπάρχει αβεβαιότητα και για τους παρατηρητές, σε αντίστοιχο βαθμό. Σαν αποτέλεσμα, υλοποιήθηκε ένα πλήρως αυτοματοποιημένο σύστημα για την ανίχνευση της συμπεριφοράς των επιμύων στη δοκιμασία εξαναγκασμένης κολύμβησης.

Λέξεις Κλειδιά

βαθιά μάθηση, μηχανική μάθηση, όραση υπολογιστών, βίντεο, επίμυες, βιολογικά πειράματα, ταξινόμηση βίντεο, ταξινόμηση συμπεριφοράς, αναγνώριση συμπεριφοράς, νευροεπιστήμη, ψυχοφαρμακολογία

Abstract

The rapid evolution of machine learning science, due to the development of deep learning, has led to major changes in the automation of solving various problems, which until recently required hours of human repetive work. The purpose of this diploma thesis is to automate the forced swimming test in rats. This experiment is a common tool for studying the effect of antidepressant drugs. The rats are placed in water cylinders to measure the immobility, swimming and climbing time of the subject. To perform a supervised classification, a 8-hour dataset was used, including the corresponding ground truth values of two expert observers. After correcting and preprocessing the dataset, prediction models for the of behavior recognition of the rats were implemented. For behavior prediction, both conventional behavioral recognition techniques and deep learning techniques were designed and implemented. Initially, the Dense Trajectories algorithm was applied to extract spatio-temporal descriptions, Fisher Vectors coding, and Classification with Support Vector Machines.Then, using Inflated 3D Artificial Neural Network, suitable for action recognition, parameter optimization was performed. Architectures have also been designed and optimized with combinations of two-dimensional convolutional networks for the extraction of spatial features and recurrent networks for their correlation in time domain.The classification of the dense trajectories algorithm descriptions achieves satisfactory accuracy in the two predominant categories of the dataset. LSTM networks show similar quality results. Inflated 3D, using pre-trained weights, improves the problem of class imbalance, achieving 82% accuracy. The model errors are detected in cases where there is uncertainty for observers, to an equivalent rate. As a result, a fully automated system for detecting the behavior of rats in the forced swimming test was implemented.As a result, a fully automated system for detecting the behavior of rats in the forced swimming test was implemented.

Keywords

deep learning, machine learning, computer vision, video, action recognition, biology experiments, video classification, behavior recognition, rat, neuroscience

Ευχαριστίες

Θα ήθελα καταρχήν να ευχαριστήσω τον καθηγητή κ. Καράντζαλο για την επίβλεψη αυτής της εργασίας και για την ευκαιρία που μου έδωσε να εκπονήσω το θέμα της παρούσας διπλωματικής. Ακόμη τους Χ. Δάλλα και Ν. Κόκρα για την παροχή του συνόλου των δεδομένων των βιολογικών πειραμάτων και την βοήθεια τους. Επίσης ευχαριστώ τους Α. Ψάλτα και Β. Τσιρώνη για την καθοδήγησή και το χρόνο που μου παρείχαν. Τέλος θα ήθελα να ευχαριστήσω την οικογένεια και τους φίλους μου για την στήριξη και την ηθική συμπαράσταση που μου προσέφεραν όλα αυτά τα χρόνια.

Περιεχόμενα

Μέρος Ι

Εισαγωγή

Κεφάλαιο 1

Ιστορική Αναδρομή

Τα τελευταία χρόνια με την ανάπτυξη των τεχνητών νευρωνικών δικτύων έχει σημανθεί μεγάλη εξέλιξη στον τομέα της μηχανικής μάθησης. Απ᾽ το 2012 η αρχιτεκτονική AlexNet [1] άλλαξε τις ισορροπίες στην ταξινόμηση εικόνων και έστρεψε την προσοχή της κοινότητας της μηχανικής μάθησης στα νευρωνικά συνελικτικά δίκτυα. Ακολούθησαν βελτιωμένες αρχιτεκτονικές όπως τα δίκτυα VGG [2], GoogleNet [3] και ResNet [4] οι οποίες καθιέρωσαν τον όρο βαθιά μάθηση (Deep Learning). Οι αρχιτεκτονικές αυτές χάραξαν το δρόμο για την αντιμετώπιση πολλών διαφορετικών προβλημάτων πέραν της ταξινόμησης εικόνων, όπως η ανίχνευση αντικειμένων, η κατάτμηση εικόνων και η αναγνώριση δράσεων. Ιδιαίτερα η αναγνώριση δράσεων ή γενικότερα η ταξινόμηση βίντεο, αποτελεί ένα πολυσύνθετο πρόβλημα καθώς εμπλέκει μια παραπάνω διάσταση, αυτή του χρόνου.

Στο πρόβλημα της αναγνώρισης δράσεων μέχρι και το 2014, τεχνολογία αιχμής αποτελούσαν ρηχές προσεγγίσεις, όπου μια συνήθης τακτική ήταν η εξαγωγή χωροχρονικών χαρακτηριστικών, η κωδικοποίηση και ταξινόμηση τους. Πρωτοποριακές λύσεις έδωσαν οι μελέτες των Laptev [5] και Dollar [6] για την μελέτη και ανάπτυξη της εξαγωγής χωροχρονικών χαρακτηριστικών. Τεχνολογία αιχμής στις συμβατικές προσεγγίσεις αποτέλεσαν τα χαρακτηριστικά πυκνών τροχιών [7] [8], οι οποίες εξάγουν πυκνά χαρακτηριστικά διαφόρων ειδών ανά πάσα χρονική στιγμή, τα οποία κωδικοποιούνται με Bag of Words ή διανύσματα Fisher και τελικά ταξινομούνται με μηχανές διανυσμάτων υποστήριξης.

Μία απ᾽ τις πρώτες πετυχημένες προσπάθειες αναγνώρισης δράσης με βαθιά μάθηση αποτέλεσε η δουλεία των Simmoyan και Zisserman το 2014 [9], σχεδιάζοντας την αρχιτεκτονική δύο ροών. Με κεντρική ιδέα ότι ο άνθρωπος αντιλαμβάνεται μια δράση ως συνδυασμό αφενός των αντικειμένων που βλέπει στο οπτικό του πεδίο και αφετέρου της κίνησης των αντικειμένων αυτών, ανέπτυξαν μία αρχιτεκτονική που διαχωρίζεται σε δύο ροές, η πρώτη για εξαγωγή χαρακτηριστικών σε εικόνες RGB και η δεύτερη για εξαγωγή χαρακτηριστικών σε εικόνες οπτικής ροής, όπου η τελική εκτίμηση της δράσης προκύπτει από τον συνδυασμό των εκτιμήσεων των δύο ροών. Η ιδέα αυτή χρησιμοποιήθηκε από διαφορετικές προσεγγίσεις του προβλήματος όπως με ανατροφοδοτούμενα δίκτυα [10] [11] καθώς και 3D συνελικτικά δίκτυα [12] [13].

Σαν αποτέλεσμα, οι προσεγγίσεις αυτές βελτίωσαν την αναγνώριση δράσεων ώστε να χρησιμοποιηθούν σε πραγματικές εφαρμογές. Μια ανάγκη που μπορούν να καλύψουν αποτελεί η ταξινόμηση βιολογικών πειραμάτων, μια επαναλαμβανόμενη εργασία χρονοβόρα και απαιτητική για τους ερευνητές.

Ένα τέτοιο παράδειγμα είναι η **δοκιμασία εξαναγκασμένης κολύμβησης** [14]. Το πείραμα αυτό χρησιμοποιείται για την μελέτη αντικαταθλιπτικών ουσιών. Το πρω-

τόκολλο του πειράματος ορίζει την τοποθέτηση επιμύων σε δεξαμενή με νερό για 5 λεπτά, απ᾽ την οποία δεν μπορούν να δραπετεύσουν. Ο Porsolt απέδειξε ότι όταν οι επιμύες έχουν λάβει αντικαταθλιπτικές ουσίες, έχουν πιο ενεργητική συμπεριφορά προσπαθώντας να αποδράσουν απ᾽ τη δεξαμενή, και αντιθέτως η ακινησία τους αποτελεί ένδειξη παραίτησης. Έτσι το πείραμα είναι μια συνήθης διαδικασία για τη μελέτη και αξιολόγηση αντικαταθλιπτικών ουσιών.

Κεφάλαιο 2

Αντικείμενο της διπλωματικής

Στόχος της παρούσας διπλωματικής είναι η δημιουργία ενός ολοκληρωμένου συστήματος ταξινόμησης της συμπεριφοράς των επιμυών στη δοκιμασία εξαναγκασμένης κολύμβησης. Η διαδικασία αυτή πρόκειται να πραγματοποιηθεί με επιβλεπόμενη ταξινόμηση. Για την επίτευξη αυτού του στόχου αναγκαία είναι η εύρεση, διόρθωση και προεπεξεργασία ενός dataset που να περιλαμβάνει μεγάλο όγκο δειγμάτων της δοκιμασίας εξαναγκασμένης κολύμβησης, καθώς και των αντίστοιχων πραγματικών προβλέψεων για κάθε ένα καθορισμένο χρονικό διάστημα του βίντεο. Η ταξινόμηση αφορά τις καταστάσεις που μπορεί να βρεθεί ο επίμυς την ώρα του πειράματος και έχουν βιολογικό ενδιαφέρον για τους μελετητές στον τομέα της ψυχοφαρμακολογίας.

Στη συνέχεια πρόκειται να μελετηθούν, σχεδιαστούν και να υλοποιηθούν μοντέλα ανάλυσης και πρόβλεψης της συμπεριφοράς μέσω βίντεο. Για το λόγο αυτό γίνεται υλοποίηση τεχνικών Όρασης Υπολογιστών με τη χρήση κατασκευασμένων χαρακτηριστικών προς ταξινόμηση αλλά και σχεδιασμός διαφόρων μοντέλων τεχνητών νευρωνικών δικτύων. Σ' αυτό το σημείο είναι αναγκαία η βελτιστοποίηση παραμέτρων του dataset και των μοντέλων. Με βάση τα παραπάνω, αποτελέσματα πρόκειται να γίνει αξιολόγηση και σύγκριση όλων των τεχνικών και παραμέτρων που πραγματοποιήθηκαν.

Κεφάλαιο 3

Οργάνωση του τόμου

Η διπλωματική οργανώνεται ως εξής:

- Στο 2^o μέρος του τεύχους αναλύεται το θεωρητικό υπόβαθρο που θα χρησιμοποιηθεί για την υλοποίηση της πρακτικής εφαρμογής της εργασίας. Συγκεκριμένα γίνεται παρουσίαση και ανάλυση του βιολογικού πειράματος της δοκιμασίας εξαναγκασμένης κολύμβησης, τεχνικών όρασης υπολογιστών. Ακόμη παρουσιάζεται το θεωρητικό υπόβαθρο των τεχνητών νευρωνικών δικτύων. Τέλος γίνεται ανάλυση των σχετικών τεχνικών αναγνώρισης δράσης με μεθόδους κατασκευασμένων χαρακτηριστικών και βαθιάς μάθησης.
- Στο 3^o μέρος παρουσιάζεται το σύνολο δεδομένων που χρησιμοποιήθηκε και την επεξεργασία που υπέστη. Ακόμη την αναλυτική υλοποίηση των τεχνικών που χρησιμοποιήθηκαν, την βελτιστοποίηση των παραμέτρων και τα αποτελέσματα τους. Τέλος πραγματοποιείται σύγκριση των αποτελεσμάτων.
- Στο 4^o μέρος γίνεται σύνοψη και αξιολόγηση της εργασίας. Ακόμη προτείνονται κατευθύνσεις για την περαιτέρω διερεύνηση και βελτίωση του θέματος και των αποτελεσμάτων.

Μέρος II

Θεωρητικό Υπόβαθρο

Κεφάλαιο 1

Βιολογικό Πείραμα Δοκιμασίας Εξαναγκασμένης Κολύμβησης

Ένα βασικό πρόβλημα για την ανακάλυψη νέων αντικαταθλιπτικών φαρμάκων είναι η δυσκολία εύρεσης μοντέλων για την μελέτη της καταθλιπτικής συμπεριφοράς και της φαρμακευτικής της θεραπείας. Ένα τέτοιο μοντέλο είναι η δοκιμασία εξαναγκασμένης κολύμβησης και χρησιμοποιείται συχνά για την μελέτη αντικαταθλιπτικών ουσιών.

1.1 Αρχικό Πρωτόκολλο

Το πείραμα εξαναγκασμένης κολύμβησης, όπως ορίστηκε απ' τον Porsolt [15] το 1977, περιλαμβάνει την τοποθέτηση ενός μυ [14] ή επίμυ [16] σε κύλινδρο από πλεξιγκλάς, με νερό ύψους 15 cm στους 25 ^{o}C, απ' την οποία είναι αδύνατο να εξέλθει. Στη συγκεκριμένη περίπτωση θα αναφερθεί το μοντέλο που αφορά τους επιμύες. Συνηθίζεται οι επιμύες να τοποθετούνται σε πρώτη φάση για 15 λεπτά στον κύλινδρο. Παρατηρείται ότι προσπαθούν για περίπου 2-3 λεπτά να εξέλθουν απ' τον κύλινδρο, και για τα υπόλοιπα λεπτά στέκονται ακίνητα, κάνοντας τις ελάχιστες κινήσεις για να διατηρούν το κεφάλι τους πάνω απ' το νερό. Ο Porsolt ερμηνεύει αυτή την συμπεριφορά ως απελπισία των επιμυών και εγκατάλειψη της προσπάθειας που καταβάλουν να εξέλθουν απ' τον κύλινδρο, λόγω της κατανόησης του γεγονότος ότι δεν είναι δυνατόν να εξέλθουν. Τα ίδια ποντίκια τοποθετούνται δεύτερη φορά στον κύλινδρο, μια μέρα μετά. Στο σημείο αυτό έγινε η παρατήρηση ότι αν έχουν χορηγηθεί συγκεκριμένες αντικαταθλιπτικές ουσίες, μειώνεται αρκετά ο χρόνος της ακινησίας. Τέτοιες ουσίες είναι οι αναστολείς της μονοαμινοξειδάσης, τα τρικυκλικά αντικαταθλιπτικά και οι αναστολείς επαναπρόσληψης σεροτονίνης. Η μετρήσεις αφορούσαν την ποσοτικοποίηση της κίνησης και της ακινησίας των επιμύων, παρατηρώντας την κυρίαρχη απ' τις 2 συμπεριφορές σε διαστήματα των 5 δευτερολέπτων.

1.2 Τροποποιημένο Πρωτόκολλο

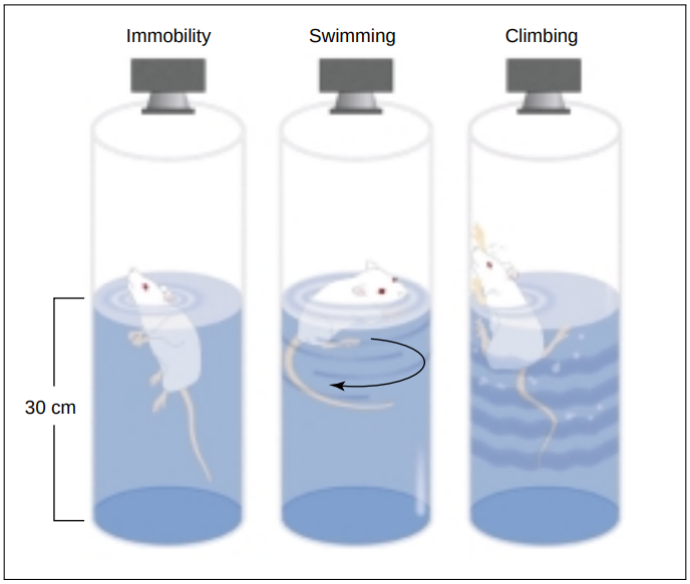
Το 1995 [17] αναδιατυπώθηκε η διαδικασία του πειράματος ώστε να γίνεται επιπλέον διάκριση μεταξύ των νοραδρενεργικών φάρμακων και των σεροτονινεργικών αντικαταθλιπτικών μέσω της διαφοροποίησης της κινητικής συμπεριφοράς σε δύο νέες διακριτές συμπεριφορές. Κατά τη διαδικασία αυτή, οι επιμύες τοποθετούνται σε

κύλινδρο με μεγαλύτερο ύψος νερού ώστε να μην μπορούν να αγγίξουν τον πυθμένα και οι ενεργητικές συμπεριφορές προσδιορίζονται ως κολύμβηση ή αναρρίχηση. Η συμπεριφορά της κολύμβησης αυξάνεται με χορήγηση σεροτονινεργικών αντικαταθλιπτικών και η αναρρίχηση με νοραδρενεργικές ουσίες αντίστοιχα.

1.3 Κατηγορίες Ενδιαφέροντος

Οι κατηγορίες ενδιαφέροντος για το πείραμα εξαναγκασμένης κολύμβησης, σύμφωνα με το πρωτόκολλο που αναφέρθηκε, είναι:

- Ακινησία (Immobility): Η κατάσταση κατά την οποία ο επίμυς είναι απολύτως ακίνητος, ή κάνει τις ελάχιστες κινήσεις ώστε να επιπλέει στο νερό.
- Κολύμβηση (Swimming): Η κατάσταση όπου ο επίμυς κινείται, συνήθως οριζοντίως, κινούμενος κατά πλάτος του κυλίνδρου.
- Αναρρίχηση (Climbing): Η κατάσταση κατά την οποία ο επίμυς, με κινήσεις των εμπρόσθιων ποδιών, κινείται κατα μήκος του κυλίνδρου, προσπαθώντας να εξέλθει απ᾽ τον κύλινδρο.
- Τίναγμα Κεφαλής (Head Shaking): Χαρακτηριστική κίνηση των επιμύων, για την αποβολή του νερού απ το κεφάλι τους.
- Κατάδυση (Diving): Βουτιά του επιμύος εντός του νερού, στρέφοντας τη φορά του σώματος του.



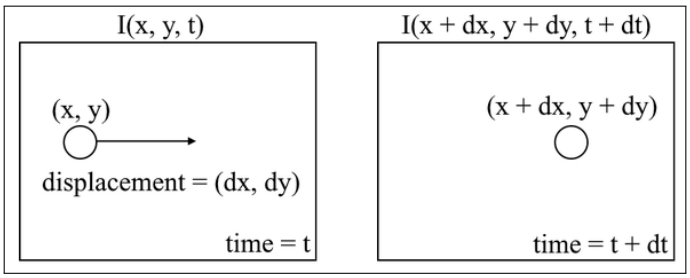
Σχήμα 1.1: Οι 3 βασικές κατηγορίες της Δοκιμασίας Εξαναγκασμένης Κολύμβησης
Πηγή: *Assessing antidepressant activity in rodents: recent developments and future needs [18]*

Κεφάλαιο 2

Τεχνικές Όρασης Υπολογιστών

2.1 Εικόνες Οπτικής Ροής

Οι εικόνες οπτικής ροής αφορούν την κίνηση των αντικειμένων μεταξύ διαδοχικών στιγμιοτύπων ενός βίντεο. Προκύπτει λόγω της κίνησης των αντικειμένων ή την κίνηση της κάμερας. Μαθηματικά το πρόβλημα του εντοπισμού μπορεί να οριστεί ως : $I(x, y, t) = I(x + \delta x, y + \delta y, t + \delta t)$, $I(x, y, t)$ η ένταση της εικόνας συναρτήσει της θέσης και της χρονικής στιγμής, όπως φαίνεται και στην εικόνα 2.1



Σχήμα 2.1: Ορισμός του προβλήματος της οπτικής ροής
Πηγή: *blog.nanonets.com/optical-flow/*

Η αραιή οπτική ροή αφορά των εντοπισμό χαρακτηριστικών σημείων, όπως γωνίες και ακμές της εικόνας, και την παρακολούθηση τους σε διαδοχικά στιγμιότυπα. Αντιθέτως η πυκνή οπτική ροή αφορά την εκτίμηση των διανυσμάτων κίνησης ολόκληρης της εικόνας, δηλαδή όλων των εικονοστοιχείων. Στα πλαίσια της παρούσας διπλωματικής θα εξεταστεί η εκτίμηση πυκνής οπτικής ροής.

Σχήμα 2.2: Παράδειγμα εκτίμησης της αραιής οπτικής ροής στα αριστερά και πυκνής ροής στα δεξιά
Πηγή: *blog.nanonets.com/optical-flow/*

2.1.1 Οπτική Ροή Farneback

Η οπτική ροή του Farneback [19], αποτελεί μια τυπική εκτίμηση πυκνής οπτικής ροής, συχνά χρησιμοποιούμενη. Η θέση κάθε σημείου ορίζεται με τη χρήση τετραγωνικού πολυωνύμου της μορφής $f_1(\mathbf{x}) = \mathbf{x}^T\mathbf{A}_1\mathbf{x} + \mathbf{b}_1^T\mathbf{x} + c_1$, όπου A συμμετρικός πίνακας, b διάνυσμα και c βαθμωτός αριθμός. Οι συντελεστές προσδιορίζονται με συνόρθωση ελαχίστων τετραγώνων. Αντίστοιχα για την δεύτερη εικόνα θα ισχύει ότι:

$$f_2(\mathbf{x}) = f_1(\mathbf{x} - \mathbf{d})$$

Επομένως προκύπτει:

$$\begin{aligned} f_2(\mathbf{x}) = f_1(\mathbf{x} - \mathbf{d}) &= (\mathbf{x} - \mathbf{d})^T\mathbf{A}_1(\mathbf{x} - \mathbf{d}) + \mathbf{b}_1^T(\mathbf{x} - \mathbf{d}) + c_1 \\ &= \mathbf{x}^T\mathbf{A}_1\mathbf{x} + (\mathbf{b}_1 - 2\mathbf{A}_1\mathbf{d})^T\mathbf{x} + \mathbf{d}^T\mathbf{A}_1\mathbf{d} - \mathbf{b}_1^T\mathbf{d} + c_1 \\ &= \mathbf{x}^T\mathbf{A}_2\mathbf{x} + \mathbf{b}_2^T\mathbf{x} + c_2 \end{aligned}$$

Αν γίνει εξίσωση των συντελεστών των τετραγωνικών πολυωνύμων έχουμε:

$$\begin{aligned} \mathbf{A}_2 &= \mathbf{A}_1 \\ \mathbf{b}_2 &= \mathbf{b}_1 - 2\mathbf{A}_1\mathbf{d} \\ c_2 &= \mathbf{d}^T\mathbf{A}_1\mathbf{d} - \mathbf{b}_1^T\mathbf{d} + c_1 \end{aligned}$$

Και εφόσον ο A είναι αντιστρέψιμος έχουμε:

$$\mathrm{d} = -\frac{1}{2}\mathrm{A}_1^{-1}\left(\mathrm{b}_2 - \mathrm{b}_1\right)$$

Προφανώς η συνθήκη αυτή δεν γίνεται να ισχύει για ισχύει για ολόκληρο το σήμα της εικόνας, καθώς δεν υπάρχει καθολική μετάθεση. Επομένως η καθολική πολυωνυμική εξίσωση μετατρέπεται σε τοπική με συντελεστές $\mathbf{A}_1(\mathbf{x}), \mathbf{b}_1(\mathbf{x})$, και $c_1(\mathbf{x})$. Ακόμη η συνθήκη $\mathbf{A}_1 = \mathbf{A}_2$ πρακτικά δεν ισχύει οπότε εκτιμάται ως:

$$\mathbf{A}(\mathbf{x}) = \frac{\mathbf{A}_1(\mathbf{x}) + \mathbf{A}_2(\mathbf{x})}{2}$$

Τέλος ορίζοντας:

$$\Delta\mathrm{b}(\mathrm{x}) = -\frac{1}{2}\left(\mathrm{b}_2(\mathrm{x}) - \mathrm{b}_1(\mathrm{x})\right)$$

Έχουμε ότι:

$$\mathbf{A}(\mathbf{x})\mathbf{d}(\mathbf{x}) = \Delta\mathbf{b}(\mathbf{x})$$

1 όπου το d(x) πλέον έχει τοπική ισχύ και όχι καθολική.

Τέλος για την βελτίωση της ακρίβειας μπορούμε να εφαρμόσουμε αυτή τη συνθήκη για όλη την γειτονική περιοχή και όχι για κάθε pixel ξεχωριστά ελαχιστοποιώντας τη σχέση:

$$\sum_{\Delta\mathbf{x}\in I} w(\boldsymbol{\Delta x})\|\mathbf{A}(\mathbf{x} + \Delta\mathbf{x})\mathbf{d}(\mathbf{x}) - \Delta\mathbf{b}(\mathbf{x} + \Delta\mathbf{x})\|^2$$

όπου $w(\Delta x)$, συνάρτηση βαρών των γειτονικών σημείων. Άρα το πεδίο οπτικής ροής τελικά είναι:

$$\mathbf{d}(\mathbf{x}) = \left(\sum w\mathbf{A}^T\mathbf{A}\right)^{-1}\sum w\mathbf{A}^T\ \mathbf{b}$$

Σχήμα 2.3: Παράδειγμα εκτίμησης οπτικής ροής με τον αλγόριθμο Farneback
Πηγή: *Βιβλιοθήκη OpenCV*

2.1.2 Οπτική Ροή TV-L1

Έναν διαφορετικό τρόπο εκτίμησης της οπτικής ροής αποτελεί ο αλγόριθμος TV-L1 [20]. Η λειτουργία του αλγορίθμου βασίζεται στην ελαχιστοποίηση μιας συνάρτησης που περιλαμβάνει έναν όρο πληροφοριών με τη χρήση της L1 νόρμας και έναν όρο κανονικοποίησης με τη χρήση της διακύμανσης της οπτικής ροής. Αρχικά θεωρείται η σύμβαση της συνέπειας της φωτεινότητας (brightness constancy assumption) ως:

$$\frac{d}{dt}I(x(t), y(t), t) = 0$$

όπου $I(x(t), y(t), t)$ το βίντεο και (x(t), y(t)) η τροχιά ενός σημείου της εικόνας. Εφαρμόζοντας τον κανόνα της αλυσίδας:

$$\nabla I \cdot (\dot{x}, \dot{y}) + \frac{\partial}{\partial t} I = 0$$

Ακόμη ορίζεται ως ταχύτητα των τροχιών:

$$\mathbf{u}(x, y) = (u_1(x, y), u_2(x, y))$$

και προκύπτει η λεγόμενη δέσμευση της οπτικής ροής:

$$\nabla I \cdot \mathbf{u} + \frac{\partial}{\partial t} I = 0$$

Για κάθε σημείο της εικόνας, η εξίσωση αυτή έχει 2 άγνωστες μεταβλητές, τις συνιστώσες της ταχύτητας u. Επομένως το σύστημα δεν έχει μοναδική λύση. Για την επίλυση αυτού του προβλήματος χρησιμοποιείται η χρήση ενός όρου εξομάλυνσης για τον εξαναγκασμό της κανονικοποίησης του u. Τελικά η επίλυση πραγματοποιείται με την ελαχιστοποίηση της συνάρτησης ενέργειας που προκύπτει από το άθροισμα της μεταβλητότητας του u και του όρου L1.

$$E(\mathbf{u}) = \int_{\Omega} |\nabla u_1| + |\nabla u_2| + \lambda |\rho(\mathbf{u})|$$

Η διαδικασία ελαχιστοποίησης για την εύρεση του u πραγματοποιείται για διαφορετικές κλίμακες εικόνας. Αρχικά υπολογίζεται το διάνυσμα u για μεγάλες κλίμακες οι οποίες αποτελούν αρχικές τιμές για τις μικρότερες κλίμακες. Έτσι σταδιακά γίνεται ακριβέστερος προσδιορισμός του διανύσματος u.

Σχήμα 2.4: Παράδειγμα εκτίμησης οπτικής ροής με τον αλγόριθμο TV-L1
Πηγή: *http://demo.ipol.im/*

2.2 Περιγραφείς Χαρακτηριστικών Βίντεο

Όπως στην ταξινόμηση εικόνων έτσι και στην αναγνώριση δράσης, ο πιο τυπικός τρόπος για την ταξινόμηση βίντεο είναι η εξαγωγή κατασκευασμένων χαρακτηριστικών.

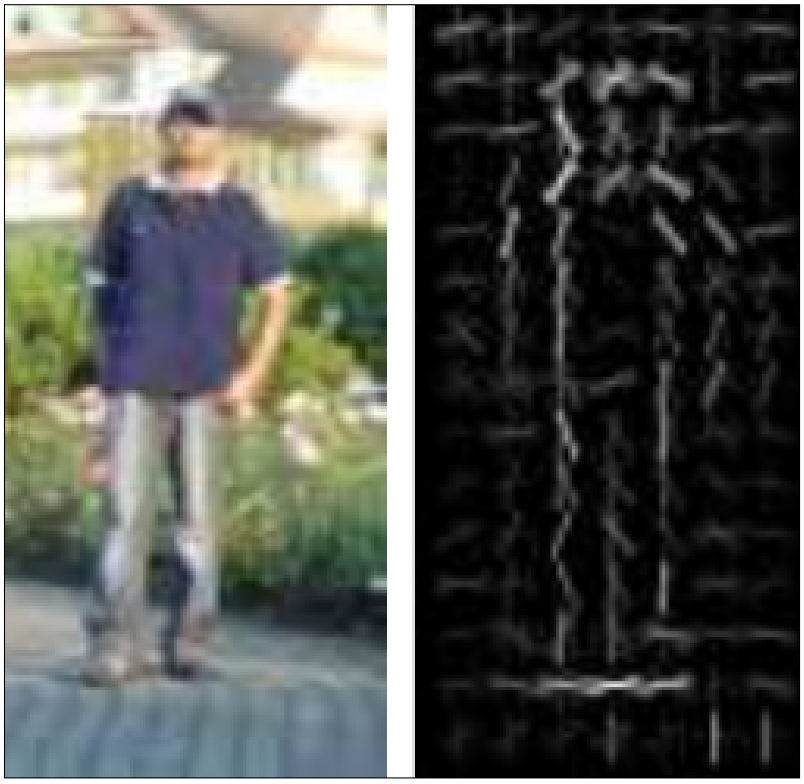
2.2.1 Περιγραφέας HoG

Ο περιγραφέας HoG [21] αφορά την εξαγωγή ιστογραμμάτων κατευθυνόμενων παραγώγων (Histograms of Oriented Gradients). Πρόκειται για την αναπαράσταση κάθε μέρους της εικόνας με τις κατευθύνσεις των ακμών του. Βασίζεται στον υπολογισμό κανονικοποιημένων τοπικών ιστογραμμάτων των βαθμίδων των κατευθύνσεων σε πυκνό κάναβο της εικόνας. Χρησιμοποιείται για την περιγραφή αντικειμένων της εικόνας. Ο αλγόριθμος HoG ακολουθεί τα εξής βήματα:

- Κανονικοποίηση των χρωμάτων της εικόνας κα διόρθωση gamma.

- Υπολογισμός των κλίσεων της εικόνας οριζοντίως και κατακορύφως με συνελικτικά φίλτρα της μορφής [-1 0 1].
- Υπολογισμός της διεύθυνση και το μέτρου της κλίσης για κάθε σημείο της εικόνας.
- Διαίρεση της εικόνας σε μικρές περιοχές - κελιά.
- Για κάθε κελί υπολογισμός των ιστογραμμάτων των διευθύνσεων της κλίσης. Συνηθίζεται η ομαδοποίηση των τιμών σε 8 διευθύνσεις.
- Οι περιοχές ομαδοποιούνται σε μεγαλύτερα blocks και τα ιστογράμματα τους συνενώνονται σε ενιαίο περιγραφέα. Έτσι μειώνεται ο θόρυβος της αναπαράστασης του κάθε μπλοκ.

Ο σχεδιασμός αυτού του αλγορίθμου πραγματοποιήθηκε κυρίως για τον εντοπισμό ανθρώπινου σώματος με τροφοδότηση των αναπαραστάσεων σε ταξινομητή SVM. Ωστόσο συχνά χρησιμοποιείται και σε διαφορετικές εφαρμογές για τον εντοπισμό αντικειμένων σε εικόνες.



Σχήμα 2.5: Παράδειγμα του περιγραφέα HoG
Πηγή: *Histograms of Oriented Gradients for Human Detection [21]*

2.2.2 Περιγραφέας HoF

Σε αντίθεση με τον περιγραφέα HoG που αναπαριστά το είδος των αντικειμένων μιας εικόνας, ο περιγραφές HoF καλείται να αναπαραστήσει την κίνηση. Ο αλγόριθμος HoF πρόκειται για ιστόγραμμα οπτικής ροής (Histogram of Optical Flow) και

σχεδιάστηκε απ᾽ τον Laptev κα.[22]. Αρχικά υπολογίζονται περιγραφείς ιστογραμμάτων χωροχρονικών όγκων, στο πεδίο της οπτικής ροής, σε κάθε σημείο ενδιαφέροντος. Έπειτα ο κάθε όγκος υποδιαιρείται σε περαιτέρω κυβοειδή για κάθε ένα απ᾽ τα οποία υπολογίζεται το ιστόγραμμα της οπτικής ροής. Τα ιστογράμματα αυτά κανονικοποιούνται και συγκεντρώνονται δημιουργώντας διανυσματικές περιγραφές.

2.2.3 Περιγραφέας MBH

Οι περισσότεροι περιγραφείς της κίνησης επηρεάζονται απ᾽ την κίνηση της κάμερας. Αυτό δεν είναι επιθυμητό και δυσχεραίνει την διαδικασία της ταξινόμησης μέσω περιγραφών. Για το λόγο αυτό σημαντικός είναι ο σχεδιασμός περιγραφέων, οι οποίοι προσπαθούν να αποδώσουν την κίνηση των αντικειμένων και όχι της κάμερας. Μια τέτοια προσπάθεια αποτελεί ο περιγραφέας MBH.

Ο περιγραφέας MBH(Motion Boundary Histograms) [23] αποτελεί αναπαράσταση της κίνησης ζεύγους εικόνων, μέσω του προσδιορισμού των ιστογραμμάτων των ορίων της κίνησης. Ο αλγόριθμος αρχικά υπολογίζει ανεξάρτητα τις κλίσεις των εικόνων των δύο συνιστωσών της κίνησης. Έπειτα γίνεται υπολογισμός των διευθύνσεων και των μεγεθών της κλίσης για κάθε συνιστώσα για την κατασκευή ιστογραμμάτων κατευθύνσεων σε κάθε μικρή περιοχή της εικόνας. Τελικά προκύπτουν δυο περιγραφές, μία για κάθε συνιστώσα της κίνησης.

Σχήμα 2.6: Παράδειγμα του περιγραφέα MBH
Πηγή: *Human Detection Using Oriented Histograms of Flow and Appearance [23]*

2.3 Διανύσματα Fisher

Μετά την επιλογή περιγραφέων για την εξαγωγή χαρακτηριστικών σε εικόνες η βίντεο, απαιτείται η κωδικοποίηση τους. Η διαδικασία αυτή συνήθως υλοποιείται με τη χρήση σάκου λέξεων (Bag of Words). Ωστόσο όπως αναφέρεται στην δημοσίευση Action Recognition With Improved Trajectories [8], η κωδικοποίηση με Διανύσματα Fisher (Fisher Vectors), φαίνεται να μπορεί να αποδώσει καλύτερα τα κωδικοποιημένα χαρακτηριστικά προς ταξινόμηση.

Σε γενικές γραμμές η λειτουργία των Fisher Vectors [24, 25] χωρίζεται σε δυο βήματα. Αρχικά υπολογίζεται Gaussian Mixture Model (GMM), για την μοντελοποίηση των κατανομών των περιγραφών εικόνων ή βίντεο. Στη συνέχεια τα Fisher Vectors κωδικοποιούν τις κλίσεις της λογαριθμικής πιθανοφάνειας των χαρακτηριστικών σύμφωνα με τις παραμέτρους του GMM.

Έστω ότι X = $\{x_1, x_2, x_t\}$. τα ν-διάστατα χαρακτηριστικά. Με βάση αυτά τα χαρακτηριστικά εκτιμώνται οι παράμετροι του GMM, οι οποίες είναι τα βάρη , οι μέσοι όροι , και οι μεταβλητότητες . Για τον προσδιορισμό του Fisher Vector, υπολογίζονται οι κλίσεις της λογαριθμικής πιθανότητας για τις παραμέτρους του GMM ως εξής:

$$\nabla_{\alpha_k} \log p(X) = \sum_{i=1}^{t} \nabla_{\alpha_k} \log p\left(x_t\right)$$

$$\nabla_{\mu_k} \log p(X) = \sum_{i=1}^{t} \nabla_{\mu_k} \log p\left(x_t\right)$$

$$\nabla_{\sigma_k} \log p(X) = \sum_{i=1}^{t} \nabla_{\sigma_k} \log p\left(x_t\right)$$

Τελικά, το Fisher Vector προκύπτει από τη συγκέντρωση των τριών διανυσμάτων ως εξής:

$$FV = [\nabla_{\alpha_k} \log p(X), \nabla_{\mu_k} \log p(X), \nabla_{\sigma_k} \log p(X)]$$

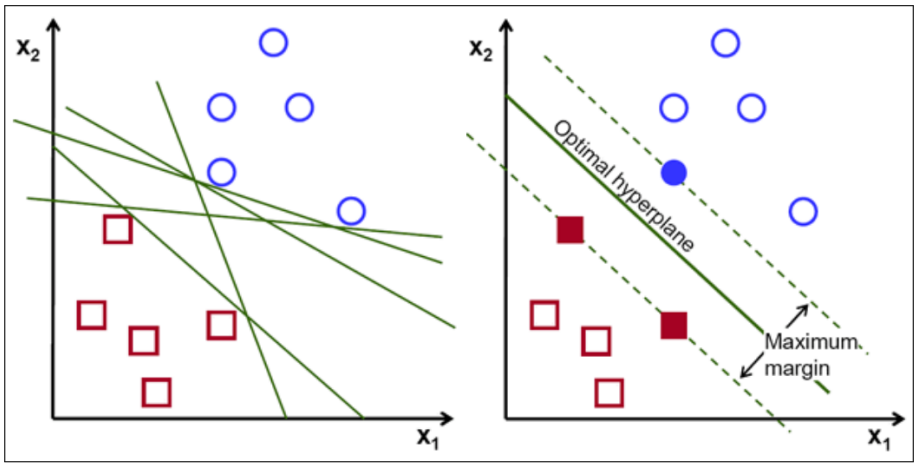
[26]

Σε αντίθεση με την τεχνική Bag of Words, τα οποία κωδικοποιούν τα στατιστικά πρώτης τάξης, τα Fisher Vectors κωδικοποιούν επιπλέον στατιστικά δεύτερης τάξης. Για το λόγο αυτό αποτελούν μια αξιόπιστη μέθοδο κωδικοποίησης.

Κεφάλαιο 3

Μηχανές Διανυσμάτων Υποστήριξης (SVM)

Οι μηχανές διανυσμάτων υποστήριξης (Support Vector Machine - SVM), είναι ένας αλγόριθμος που χρησιμοποιείται για παλινδρόμηση και ταξινόμηση αντικειμένων. Στην περίπτωση της επιβλεπόμενης ταξινόμησης, ο αλγόριθμος προσπαθεί να βρει ένα υπερεπίπεδο σε ένα ν-διάστατο χώρο, όπου ν ο αριθμός των χαρακτηριστικών, το οποίο διαχωρίζει γραμμικά τα χαρακτηριστικά των δειγμάτων δύο διαφορετικών πληθυσμών. Επομένως πρόκειται για μια διαδικασία εύρεσης του βέλτιστου υπερεπιπέδου, το οποίο απέχει κατά το μέγιστο από τα δείγματα των δύο πληθυσμών.



Σχήμα 3.1: Η λειτουργία του αλγορίθμου SVM
Πηγή: *towardsdatascience.com*

Έστω ότι έχουμε δεδομένα εκπαίδευσης της μορφής:

$$(\vec{x}_1, y_1), \ldots, (\vec{x}_n, y_n)$$

όπου y_i μπορεί να πάρει τις τιμές -1, 1, οι οποίες αφορούν τις 2 πιθανές κλάσεις της ταξινόμησης στις οποίες ανήκουν τα διανύσματα $\vec{x}_i$. Στόχος του αλγορίθμου είναι η εύρεση του βέλτιστου υπερεπιπέδου της μορφής:

$$\vec{w} \cdot \vec{x} - b = 0$$

Για ένα γραμμικώς διαχωρίσιμο πρόβλημα σε ένα κανονικοποιημένο δείγμα, λύση του προβλήματος αποτελεί ο προσδιορισμός των οριακών υπερπεπιπέδων που έχουν τη μέγιστη απόσταση μεταξύ τους, της μορφής:

$$\vec{w} \cdot \vec{x} - b = 1$$

$$\vec{w} \cdot \vec{x} - b = -1$$

Γεωμετρικά τα δύο υπερεπίπεδα έχουν απόσταση $\frac{2}{\|\vec{w}\|}$, άρα βέλτιση λύση αποτελεί η ελαχιστοποίηση του $\|\vec{w}\|$.

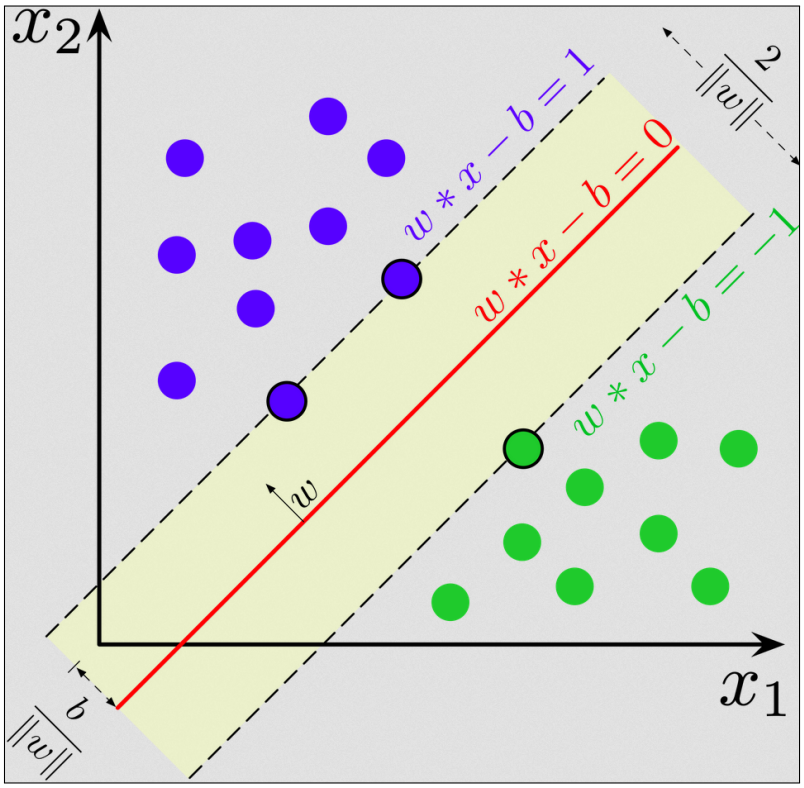
Ωστόσο στην πράξη συνήθως τα δείγματα δεν είναι γραμμικώς διαχωρίσιμα. Για το λόγο αυτό γίνεται χρήση της συνάρτησης κόστους Hinge:

$$\max\left(0, 1 - y_i \left(\vec{w} \cdot \vec{x}_i - b\right)\right)$$

Επομένως, σε αυτή την περίπτωση λύση του προβλήματος της ταξινόμησης αποτελεί η ελαχιστοποίηση του:

$$\left[\frac{1}{n}\sum_{i=1}^{n} \max\left(0, 1 - y_i \left(\vec{w} \cdot \vec{x}_i - b\right)\right)\right] + \lambda\|\vec{w}\|^2$$

όπου η παράμετρος λ καθορίζει την ευχέρεια στην "χαλάρωση" των ορίων. Μια απειροελάχιστη τιμή του λ δίνει αντίστοιχα αποτελέσματα με την λειτουργία του SVM σε ένα γραμμικώς διαχωρίσιμο πρόβλημα.



Σχήμα 3.2: SVM - Υπολογισμός του βέλτιστου υπερεπιπέδου
Πηγή: *wikipedia.com*

Υπάρχουν υλοποιήσεις του αλγορίθμου για μη γραμμικά προβλήματα καθώς και για την χρήση του για ταξινόμηση άνω των 2 κατηγοριών.

Κεφάλαιο 4

Τεχνητά Νευρωνικά Δίκτυα

Τα τελευταία χρόνια, η εξέλιξη των τεχνητών νευρωνικών δικτύων, έχει σημάνει επανάσταση στην αντιμετώπιση προβλημάτων ταξινόμησης και παλινδρόμησης. Τα τεχνητά νευρωνικά έχουν εμπνευστεί από βιολογικά μοντέλα της λειτουργίας του ανθρώπινου εγκεφάλου και προσπαθούν να προσομοιώσουν τη λειτουργία της σκέψης για την λήψη μιας απόφασης. Οι εφαρμογή τους ανθεί σε τομείς όπως η όραση υπολογιστών και η επεξεργασία φυσικής γλώσσας. Ένα απ᾽ τα βασικά πλεονεκτήματα των νευρωνικών δικτύων αποτελεί η εξάλειψη της ανάγκης δημιουργίας χειροποίητων χαρακτηριστικών, μέσω της μάθησης αναπαραστάσεων (Represantation Learning). Χάρη σε αυτό το στοιχείο τους η κατασκευή περιγραφητών δεν χρειάζεται πλέον να εξαρτάται απ᾽ το εκάστοτε πρόβλημα, άρα η προσέγγιση για την λύση κάθε προβλήματος παρόμοιας φύσης έχει κοινή μεθοδολογία. Στα πλαίσια της διπλωματικής θα γίνει αναφορά σε θέματα που αφορούν επιβλεπόμενη ταξινόμηση για εικόνες και βίντεο.

4.1 Τεχνητός Νευρώνας

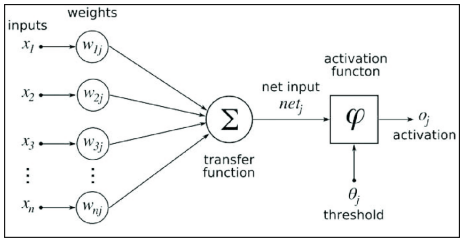
Το πιο στοιχειώδες αντικείμενο ενός νευρωνικού δικτύου είναι ο τεχνητός νευρώνας (perceptron). Το μοντέλο του τεχνητού νευρώνα περιγράφεται από τη μαθηματική σχέση:

$$y_k = f(\sum_{i=0}^{N} x_{ki} w_{ki})$$

όπου y_k η έξοδος του νευρώνα, f η συνάρτηση ενεργοποίησης, w_{ki} το συναπτικό βάρος και x_{ki} η είσοδος του νευρώνα. Στον k-οστό νευρώνα υπάρχει ένα συναπτικό βάρος w_{k0} με ιδιαίτερη σημασία, το οποίο καλείται bias και η τιμή της εισόδου του είναι πάντα $x_{k0} = 1$.

Ιδιαίτερη σημασία έχει η συνάρτηση ενεργοποίησης f καθώς επηρεάζει άμεσα τη λειτουργία του νευρώνα. Μπορεί να είναι γραμμική ή μη γραμμική. Συνήθεις συναρτήσεις ενεργοποίησης αποτελούν οι:

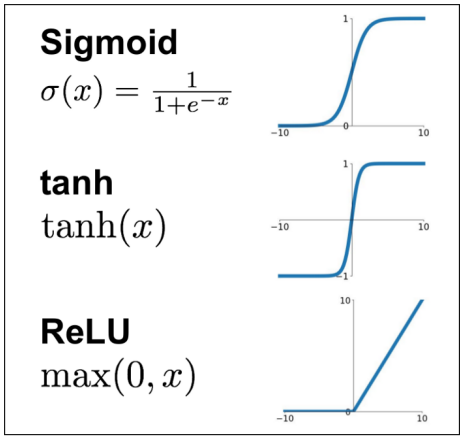
- Γραμμική: $f(x) = cx$ με άκρα το άπειρο
- Σιγμοειδής: $f(x) = \frac{1}{1+e^{-x}}$ με άκρα το 0, +1
- Υπερβολική Εφαπτομένη(tanh): $f(x) = \tanh(x) = \frac{2}{1+e^{-2x}} - 1 = 2\,\text{sigmoid}(2x) - 1$, με άκρα το -1, +1



Σχήμα 4.1: Διαγραμματική αναπαράσταση του τεχνητού νευρώνα
Πηγή: *commons.wikimedia.org*

- ReLu: $f(x) = \max(0, x)$ με άκρα το 0 και το άπειρο.
- Softmax: $f(x)_i = \frac{e^{x_i}}{\sum_{j=1}^{K} e^{x_j}}$, η οποία αποτελεί γενίκευση της σιγμοειδούς και χρησιμοποιείται για την κανονικοποίηση ενός διανύσματος σε κατανομή πιθανότητας.

Η χρήση των συναρτήσεων ενεργοποίησης μετατρέπει τον νευρώνα σε μη γραμμικό, χρήζοντας τον κατάλληλο για την λύση μη γραμμικών προβλημάτων ταξινόμησης.



Σχήμα 4.2: Κοινές συναρτήσεις ενεργοποίησης
Πηγή: *Shruti Jadon, medium.com*

4.2 Πλήρως Συνδεδεμένα Δίκτυα

Η πιο απλή μορφή ενός τεχνητού νευρωνικού δικτύου είναι αυτή του πλήρως συνδεδεμένου δικτύου (Fully connected Network). Τα πλήρως συνδεδεμένα δίκτυα απαρτίζονται από στρώσεις (layers). Το πρώτο layer αποτελεί την είσοδο του δικτύου, το τελευταίο την έξοδο, και τα ενδιάμεσα ονομάζονται "κρυφά" layers. Κάθε layer απαρτίζεται από τεχνητούς νευρώνες, διαφορετικού πλήθους σε κάθε layer ανάλογα με τις ανάγκες κάθε εφαρμογής. Κάθε νευρώνας συνδέεται με τους όλους τους νευρώνες του προηγούμενου και επόμενου layer. Θεωρητικά όσο περισσότερα layers τόσο περιπλοκότερα χαρακτηριστικά δύναται να μάθει ένα μοντέλο νευρωνικού δικτύου. Το πλήθος των κρυφών layers, ονομάζεται και βάθος της αρχιτεκτονικής, απ' το οποίο προκύπτει και η έννοια βαθιά μάθηση (deep learning).

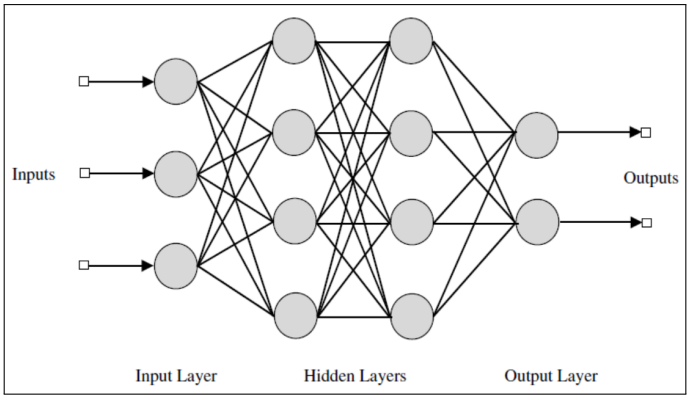
Το δίκτυο αποτελείται από συνδέσεις μεταξύ των νευρώνων, οι οποίες μεταφέρουν τις εξόδους των νευρώνων layer i στις εισόδους των διαδοχικών τους νευρώνων που

ανήκουν σε layer j. Κάθε σύνδεση αποτελείται από ένα βάρος w_{ij}, καθώς και το λεγόμενο bias a_{ij} το οποίο μεταθέτει την συνάρτηση ενεργοποίησης.

Κάθε νευρώνας που ανήκει σε layer i δέχεται σαν είσοδο το άθροισμα όλων των προηγούμενων νευρώνων του προηγούμενου layer j. Η σχέση της διάδοσης είναι η:

$$p_j(t) = \sum_i o_i(t) w_{ij} + w_{0j}$$

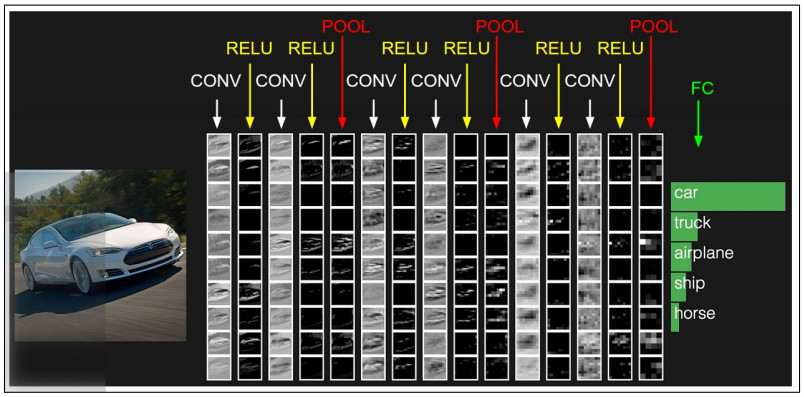
όπου $p_j(t)$ η έξοδος του νευρώνα j και $o_i(t)$ οι έξοδοι των προκάτοχων νευρώνων.



Σχήμα 4.3: Διάγραμμα ενός πλήρως συνδεδεμένου νευρωνικού δικτύου
Πηγή: *cse22-iiith.vlabs.ac.in*

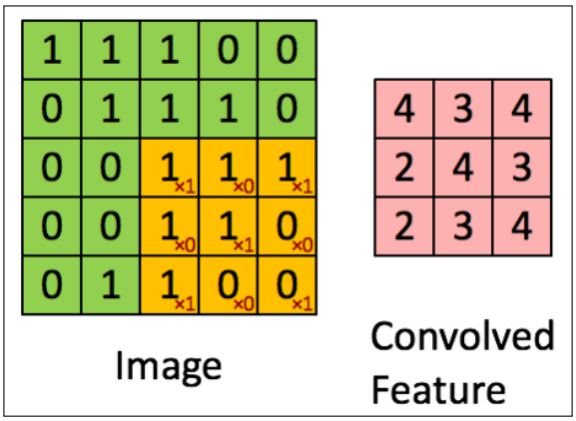
4.3 Συνελικτικά Δίκτυα

Τα πλήρως συνδεδεμένα δίκτυα, δεν έχουν τη δυνατότητα να εκφράσουν πλήρως, σχέσεις γειτνίασης σε δεδομένα πινάκων, όπως οι εικόνες και τα βίντεο. Για το λόγο αυτό σε δεδομένα που εμπλέκουν εικόνες χρησιμοποιούνται τα συνελικτικά δίκτυα. Πρόκειται για νευρώνες που έχουν τη μορφή φίλτρων συνέλιξης, όπου οι τιμές κάθε φίλτρου παίζουν τον ρόλο εκπαιδεύσιμων βαρών. Επομένως κατά την εκπαίδευση συνελικτικών νευρωνικών δικτύων (CNNs), εκπαιδεύονται φίλτρα τα οποία εξάγουν σταδιακά αναπαραστάσεις - χαρακτηριστικά για την εικόνα.



Σχήμα 4.4: Οπτικοποίηση των συνελικτικών φίλτρων μιας αρχιτεκτονικής συνελικτικών δικτύων
Πηγή: *Stanford CS231n*

Μια αρχιτεκτονική με συνελικτικά δίκτυα, παράγει χαρακτηριστικά με ιεραρχικό τρόπο, από τα γενικότερα στοιχεία μιας εικόνας, όπως γωνίες και ακμές, στα πιο ειδικά χαρακτηριστικά, όπως σχήματα, αντικείμενα κλπ. Όσο βαθύτερη είναι μια αρχιτεκτονική, τόσο καλύτερα δύναται να "μάθει" να εξάγει αναλυτικά χαρακτηριστικά, τα οποία συσχετίζουν την είσοδο με την έξοδο του νευρωνικού δικτύου. Στην κορυφή της αρχιτεκτονικής, τα χαρακτηριστικά πινάκων μετατρέπονται σε διανύσματα, και ταξινομούνται συνήθως με ένα ή περισσότερα πλήρως συνδεδεμένα layers, στο τέλος του δικτύου.



Σχήμα 4.5: Ο υπολογισμός της εξαγωγής ενός συνελικτικού φίλτρου
Πηγή: *towardsdatascience.com*

Όπως και στα πλήρως συνδεδεμένα δίκτυα, έτσι και στα συνελικτικά, χρησιμοποιείται συνάρτηση ενεργοποίησης, με συνηθέστερη τη ReLu.

Τα βασικά στοιχεία - παράμετροι των συνελικτικών δικτύων είναι:

- Το πλήθος των συνελικτικών φίλτρων: σε κάθε layer επιλέγεται το πλήθος των φίλτρων τα οποία θα εξάγουν χαρακτηριστικά για τις εικόνες. Όσο μεγαλύτερο είναι το πλήθος, τόσο περισσότερα διαφορετικά χαρακτηριστικά πρόκειται να εξάγει το συνελικτικό layer, ωστόσο δυσχεραίνεται η διαδικασία της εκπαίδευσης και ανεβαίνει εκθετικά το επεξεργαστικό κόστος και ο χρόνος της διαδικασίας της εκπαίδευσης.

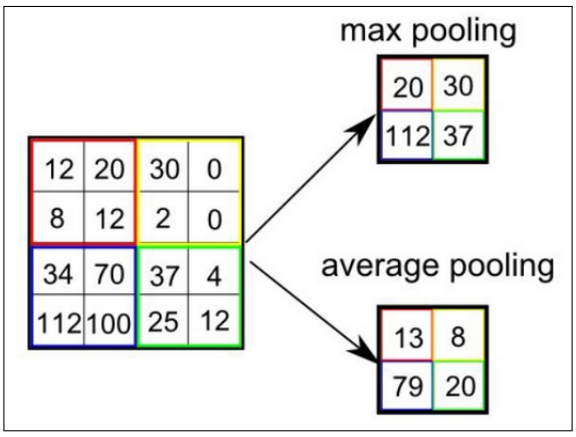
- Το μέγεθος του παραθύρου: το μέγεθος του παραθύρου των συνελικτικών φίλτρων, επηρεάζει το μέγεθος των εξαγόμενων χαρακτηριστικών και συνηθίζεται να είναι τετραγωνικό με τιμές 3x3 εώς 7x7.

- Το βήμα της εφαρμογής των συνελικτικών φίλτρων (stride): Τα συνελικτικά φίλτρα συνηθίζεται να εφαρμόζονται με μοναδιαία μετάθεση στην εικόνα, ωστόσο χρήσιμο είναι να χρησιμοποιηθεί διάστημα 2 μεταξύ των φίλτρων στα αρχικά layers, σαν μια μορφή υπόδειγματοληψίας, για τη μείωση του επεξεργαστικού κόστους της διαδικασίας.

- Το padding πρόκειται για την προσθήκη μηδενικών στα άκρα της εικόνας. Φυσιολογικά με κάθε συνελικτικό layer, η διαστάσεις τις εικόνας μειώνονται, ωστόσο με την προσθήκη μηδενικών στα άκρα τις εικόνας διατηρούνται σταθερές.

Μετά από κάθε συνελικτικό layer, συνηθίζεται η χρήση Pooling layer. Πρόκειται για φίλτρο το οποίο διατηρεί τις μέγιστες (Max Pooling) ή μέσες τιμές (Average

Pooling) το οποίο μειώνει το μέγεθος των εξαγόμενων πληροφοριών του δικτύου. Με αυτό τον τρόπο μειώνεται ο όγκος της πληροφορίας, άρα και ο χρόνος εκπαίδευσης και πρόβλεψης των νευρωνικών δικτύων. Συνηθίζεται η χρήση Max Pooling μιας και θεωρητικά διατηρεί τις πληροφορίες με τη μέγιστη σημασία σε κάθε έξοδο ενός συνελικτικού νευρώνα.

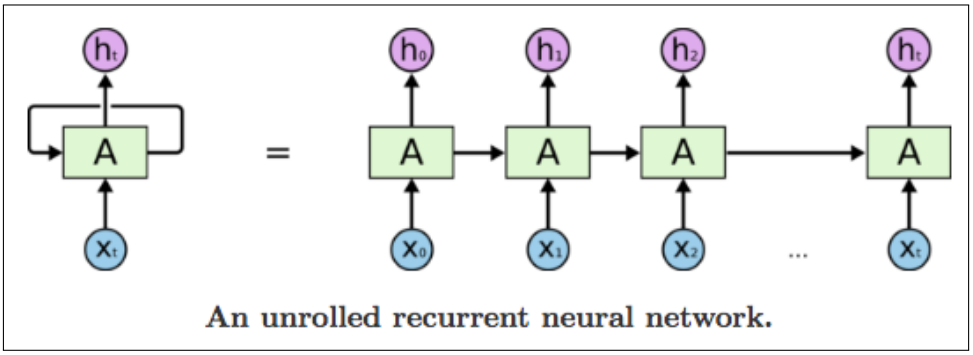


Σχήμα 4.6: Η λειτουργία ενός Max Pooling layer
Πηγή: *towardsdatascience.com*

Συνήθως η έννοια των συνελικτικών δικτύων, αφορά 2D φίλτρα, ωστόσο σε δεδομένα τριών διαστάσεων όπως τα βίντεο, αντίστοιχη λειτουργία έχουν τα 3D συνελικτικά δίκτυα και τα 3D Max Pooling.

4.4 Ανατροφοδοτούμενα Δίκτυα

Όλες οι παραπάνω αρχιτεκτονικές που αναφέρθηκαν, έχουν το μειονέκτημα ότι εξετάζουν κάθε είσοδο ενός δικτύου χωρίς να διατηρούν οποιαδήποτε πληροφορία απ᾽ τις προηγούμενες χρονικά εισόδους. Για το λόγο αυτό δεν είναι κατάλληλα για μοντέλα που δέχονται διαδοχικές χρονικά εισόδους με σκοπό την συσχέτιση τους για την εξαγωγή κάποιας πληροφορίας ή πρόβλεψης. Τέτοια περίπτωση αποτελούν τα βίντεο. Για την αναπαράσταση ενός βίντεο, κάθε στιγμιότυπο που εισάγεται, πρέπει να συσχετίζεται με πληροφορίες από τα προηγούμενα στιγμιότυπα, ώστε να δημιουργηθεί μια αντιπροσωπευτική αναπαράσταση για ολόκληρο το βίντεο. Τη λύση δίνουν τα ανατροφοδοτούμενα δίκτυα.



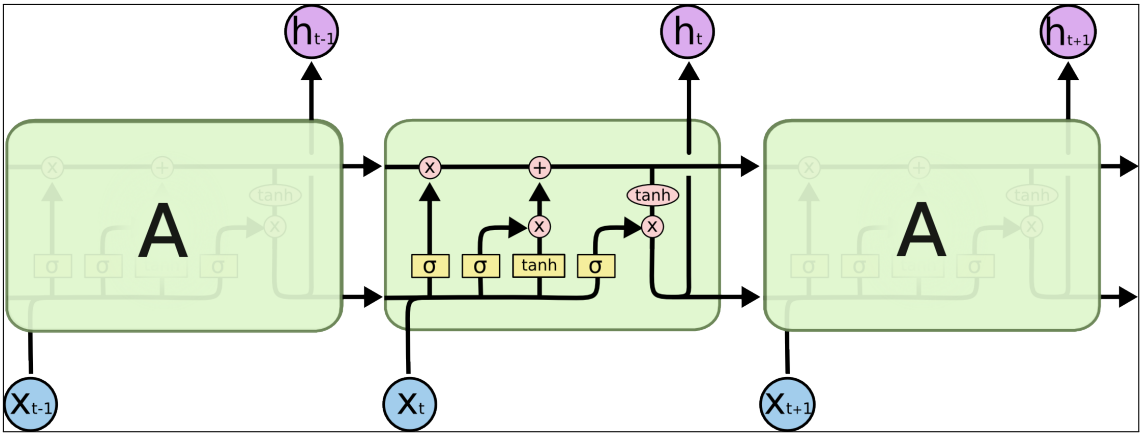
Σχήμα 4.7: Οπτικοποίηση ανατροφοδοτούμενου νευρώνα
Πηγή: *colah.github.io*

Τα ανατροφοδοτούμενα δίκτυα(Recurrent Neural Networks, RNN), έχουν μια μορφή μνήμης που αποθηκεύει τις πληροφορίες από τις προηγούμενες εισόδους του δικτύου. Τα χαρακτηριστικά που εξάγουν αφορούν ένα συνδυασμό της τωρινής εισόδου και των προηγουμένων. Έτσι κάθε ανατροφοδοτούμενος νευρώνας, δέχεται δύο εισόδους, την είσοδο σε χρόνο t και την παρελθοντική μνήμη που προέκυψε σε χρόνο $t-1$. Η κρυφή μνήμη των ανατροφοδοτούμενων νευρώνων μπορεί μαθηματικά να περιγραφεί ως:

$$h_t = f(Wx_t + Uh_{t-1})$$

όπου f η συνάρτηση ενεργοποίησης, W το βάρος της εισόδου και U το βάρος της κρυφής μνήμης του προηγούμενου βήματος. Στα RNN συνηθίζεται η χρήση της $tanh$ ως συνάρτησης ενεργοποίησης.

Η πιο συνηθισμένη μορφή ανατροφοδοτούμενων νευρώνων είναι το LSTM (Long Short-Term Memory). Αποτελεί μια ιδιαίτερη μορφή ανατροφοδοτούμενου δικτύου σχεδιασμένη για την μάθηση μακροχρόνιων συσχετίσεων.



Σχήμα 4.8: Διάγραμμα ροής της λειτουργίας του LSTM
Πηγή: *colah.github.io*

Βασικό στοιχεία του LSTM είναι το C_t (Cell State) που αποτελεί την "κατάσταση" του κελιού του LSTM. Μεταφέρει πληροφορίες των προηγούμενων χρονικών βημάτων στο παροντικό βήμα. Οι πληροφορίες μου μεταφέρει μπορούν να αλλάξουν απ' την παροντική είσοδο. Το πόση πληροφορία θα διατηρηθεί καθορίζεται από την πύλη forget (forget gate):

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right)$$

Στη συνέχεια υπολογίζεται το νέο $\tilde{C}_t$ το οποίο προκύπτει με βάση την έξοδο του προηγούμενο χρονικού βήματος h_{t-1}και την τωρινή είσοδο x_t:

$$\tilde{C}_t = \tanh\left(W_C \cdot [h_{t-1}, x_t] + b_C\right)$$

Το πόσο θα επηρεάσει το τελικό C_t καθορίζεται από το την πύλη εισόδου (input gate) της μορφής:

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right)$$

Άρα τελικά προκύπτει το νέο cell state ως:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Στη συνέχεια η πύλη εξόδου o_t αποφασίζει τα μέρη της πληροφορίας του C_t θα προαχθούν στην τελική έξοδο:

$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right)$$

Ο υπολογισμός της εξόδου h_t του LSTM τελικά καθορίζεται ως εξής:

$$h_t = o_t * \tanh\left(C_t\right)$$

[27]

Αξίζει να αναφερθεί και η αρχιτεκτονική GRU [28], που αποτελεί παραλλαγή του LSTM, αλλά με απλούστερη λειτουργία. Χρησιμοποιεί μόνο δυο πύλες, την πύλη ενημέρωσης (update gate), και επαναφοράς (reset gate).

Κεφάλαιο 5

Εκπαίδευση Νευρωνικών Δικτύων

Για να είναι χρήσιμο ένα μοντέλο τεχνητών νευρωνικών δικτύων, θα πρέπει σε πρώτη φάση να πραγματοποιηθεί διαδικασία εκπαίδευσης, με διαδικασία επιβλεπόμενης ταξινόμησης. Με αυτό τον τρόπο είναι δυνατόν να προσδιοριστούν τα βάρη του μοντέλου ώστε να εκτιμούν την σωστή πρόβλεψη για κάθε είσοδο. Η διαδικασία της εκπαίδευσης γίνεται με την προς τα πίσω διάδοση (back-propagation). Σε γενικές γραμμές, πρέπει να οριστεί ένας αντιπροσωπευτικός δείκτης της απόδοσης του μοντέλου και με επαναληπτικές εκτιμήσεις και αναπροσδιορισμούς των βαρών πραγματοποιείται η βελτιστοποίηση του μοντέλου.

5.1 Υποσύνολα του Dataset

Βασικός κίνδυνος της εκπαίδευσης αποτελεί η αποτυχία γενίκευσης του μοντέλου καθώς και η υπερπροσαρμογή στα δεδομένα εκπαίδευσης (overfitting). Όσο μικρότερο είναι το dataset, τόσο πιο εύκολο είναι να αναγνωρίσει συγκεκριμένα πρότυπα των δεδομένων που δέχεται, όπως ο θόρυβος και το παρασκήνιο, και να τα συσχετίσει με την έξοδο. Για το λόγο αυτό τα η ακρίβεια που πετυχαίνει το μοντέλο στα δεδομένα με τα οποία εκπαιδεύεται, σε καμία περίπτωση δεν αποτελεί αντιπροσωπευτική μετρική για την πραγματική ακρίβεια του μοντέλου.

Επομένως, συνηθίζεται το dataset να χωρίζεται σε 2 υποσύνολα. Το υποσύνολο εκπαίδευσης (train), και το υποσύνολο επικύρωσης (validation). Ο διαχωρισμός τους μπορεί να προκύψει με τυχαίο τρόπο, αρκεί να διατηρούν τις πραγματικές αναλογίες μεταξύ των κλάσεων των δειγμάτων. Σε καμία περίπτωση τα δύο υποσύνολα δεν πρέπει να έχουν επικαλύψεις δειγμάτων. Συνηθίζεται το validation υποσύνολο να καταλαμβάνει το 20% - 40% του συνολικού dataset. Τα δείγματα που ανήκουν στο υποσύνολο validation, δεν συμμετέχουν στην διαδικασία την εκπαίδευσης, παρά μόνο χρησιμοποιούνται για την εξαγωγή στατιστικών στοιχείων για την πρόοδο του μοντέλου εκτίμησης.

Ωστόσο ένα μοντέλο εκτός από τα βάρη των νευρώνων, καθορίζεται και από τις υπερπαραμέτρους με τις οποίες σχεδιάστηκε, όπως είναι ο αριθμός των layer, το πλήθος των νευρώνων κλπ. Για τον προσδιορισμό των κατάλληλων παραμέτρων πραγματοποιείται διαδικασία βελτιστοποίησης τους, όπως με την κατασκευή κανάβου τυχαίων συνδυασμών των υπερπαραμέτρων και την δοκιμή τους στο υποσύνολο validation. Κατά τη διαδικασία αυτή, ενώ το validation δεν συμμετέχει στην διαδικασία της εκπαίδευσης, τελικά επηρεάζει τις υπερπαραμέτρους οι οποίες καθορίζονται απ᾽ αυτό.

Επομένως στην περίπτωση της βελτιστοποίησης των υπερπαραμέτρων, για τον προσδιορισμό απόλυτα αντικειμενικών στατιστικών της πρόβλεψης του μοντέλου, χρειάζεται η δημιουργία και ενός τρίτου υποσυνόλου ελέγχου (test). Το test υποσύνολο, αποτελεί έναν απολύτως αντικειμενικό παράγοντα ακρίβειας του μοντέλου. Συνηθίζεται διαχωρισμός της μορφής 60% train - 20% validation - 20% test. Για την αποφυγή της μείωσης των δειγμάτων εκπαίδευσης, είναι δυνατή η ένωση των υποσυνόλων train και validation, όταν ολοκληρωθεί η διαδικασία του σχεδιασμού του μοντέλου.

5.2 Αρχικοποίηση των βαρών

Για να ξεκινήσει η διαδικασία τις εκπαίδευσης, αναγκαίος είναι ο καθορισμός αρχικών τιμών για τα βάρη του μοντέλου. Η αρχικοποίηση με μηδενικές τιμές, υπάρχει περίπτωση να δυσχεραίνει σημαντικά την εκπαίδευση. Δύο δημοφιλείς τεχνικές για την αρχικοποίηση των βαρών αποτελούν αυτή του Glorot [29] και του He [30].

Ωστόσο, αν γίνεται χρήση γνωστής αρχιτεκτονικής, η οποία έχει εκπαιδευτεί σε γνωστά dataset, σπουδαία βελτίωση της ακρίβειας του μοντέλου θα αποτελέσει η χρήση των προεκπαιδευμένων βαρών της, ως αρχικές τιμές του νέου μοντέλου. Η διαδικασία αυτή ονομάζεται μεταφορά μάθησης (transfer learning). Βοηθάει στην γενίκευση του παραγόμενου μοντέλου, ιδιαίτερα στην περίπτωση ελλιπούς μεγέθους δειγμάτων εκπαίδευσης. Ακόμη, τα βάρη των αρχικών layer μπορούν να καθοριστούν ως μη εκπαιδεύσιμα. Η πρακτική αυτή είναι αποδοτική, καθώς μια βαθιά αρχιτεκτονική στα πρώτα layers εξάγει χαμηλού επιπέδου χαρακτηριστικά, όπως στην περίπτωση των εικόνων γωνίες και ακμές, ενώ στα επόμενα layers υψηλότερου επιπέδου χαρακτηριστικά όπως σχήματα, αντικείμενα κλπ. Άρα η γνώση που έχει αποκτήσει ένα μοντέλο εκπαιδευόμενο σε ένα πλήρες dataset ίδιου τύπου δεδομένων, μπορεί να χρησιμοποιηθεί ακέραια και αμετάβλητη για την επίλυση ενός διαφορετικού προβλήματος ταξινόμησης.

5.3 Η συνάρτηση κόστους

Για την εκπαίδευση ενός νευρωνικού δικτύου χρειάζεται ο καθορισμός ενός δείκτη σφάλματος του μοντέλου. Για το λόγο αυτό χρησιμοποιείται η συνάρτηση κόστους (loss function). Η συνάρτηση αυτή προσδιορίζει ποσοτικά, το μέγεθος της αστοχίας του μοντέλου σε σχέση με τα δείγματα με τα οποία τροφοδοτήθηκε. Επομένως, το σημείο στο οποίο ελαχιστοποιείται η συνάρτηση κόστους, αποτελεί τη βέλτιστη λύση του μοντέλου. Η ακρίβεια του μοντέλου επηρεάζεται από τις παραμέτρους του, οι οποίες προσδιορίζονται με κριτήριο της ελαχιστοποίηση της loss function. Συνήθεις συναρτήσεις κόστους αποτελούν οι:

- Η συνάρτηση μέσου απόλυτου σφάλματος ορίζεται ως:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

όπου n ο αριθμός των δειγμάτων, x_i η πραγματική τιμή του δείγματος i και y_i η εκτιμώμενη τιμή του δείγματος i. Η ερμηνεία της είναι σαφής μιας και πρόκειται για για το άθροισμα των απόλυτων τιμών των αποκλίσεων $e_i = y_i - x_i$.

- Η συνάρτηση μέσου τετραγωνικού σφάλματος:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2$$

Στην περίπτωση του μέσου τετραγωνικού σφάλματος, η απόκλιση ορίζεται ως το τετράγωνο της διαφοράς της εκτιμώμενης τιμής y_i με την πραγματική τιμή x_i. Η χρήση της είναι πιο συνήθης μιας και έχει την λογική ερμηνεία της εκθετικής αύξησης της συνάρτησης κόστους, με την αύξηση της απόκλισης e_i, λόγω της τετραγωνικής δύναμης.

- Η συνάρτηση Cross-Entropy όπου για δυαδικά προβλήματα ταξινόμησης ορίζεται ως:

$$CE = -\sum_{i=1}^{C'=2} x_i \log(y_i) = -x_1 \log(y_1) - (1 - x_1)\log(1 - y_1)$$

Όπου C ο αριθμός των κλάσεων. Πρόκειται για την μέτρηση της απόστασης δύο κατανομών, την κατανομή των εκτιμήσεων του μοντέλου και την κατανομή των πραγματικών τιμών των κατηγοριών.

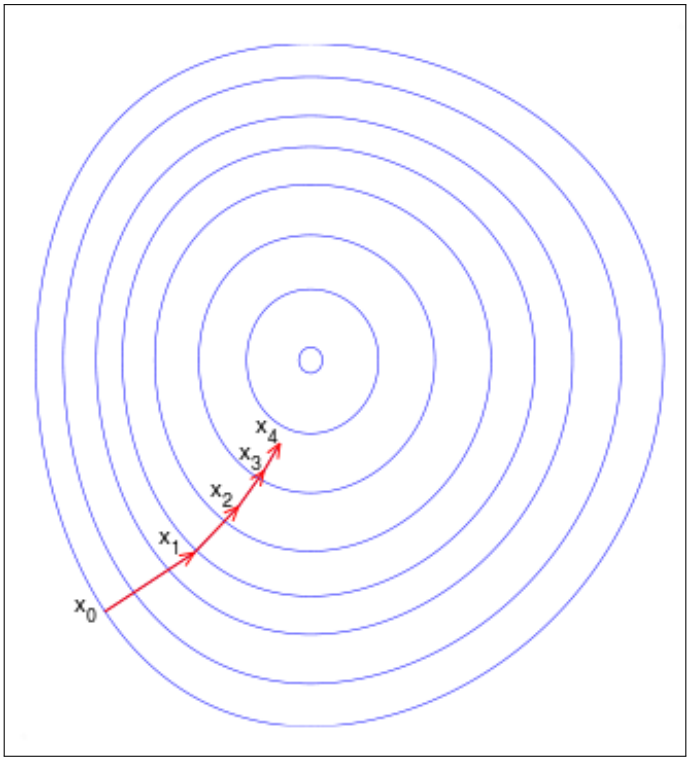
- Γενίκευση της αποτελεί η Categorical Cross-Entropy (CEE): Χρησιμοποιείται για ταξινόμηση σε περισσότερες από δυο κλάσεις και δέχεται αληθείς τιμές One-Hot κωδικοποίησης:

$$CCE = -\sum_{i=1}^{C} x_{o,c} \log(y_{o,c})$$

5.4 Εκπαίδευση Μοντέλου

Η ελαχιστοποίηση της συνάρτησης κόστους, μπορεί να γίνει κατανοητή στην περίπτωση των 2 παραμέτρων, σαν ένας τρισδιάστατος χώρος με υψώματα και κοιλάδες, όπου αναζητείται το κατώτερο σημείο του. Κατάλληλος για τη λύση αυτού του προβλήματος είναι ο αλγόριθμος Gradient Descent. Ο αλγόριθμος αυτός είναι μια επαναληπτική διαδικασία βελτιστοποίησης που προσεγγίζει το ελάχιστο της συνάρτησης κόστους, κάνοντας μικρά βήματα προς την αρνητική κλίση (ανάδελτα) της loss function ως προς τις παραμέτρους της, τα βάρη του μοντέλου. Με αυτό τον τρόπο υπολογίζεται η αρνητική κλίση για τις υπάρχουσες παραμέτρους και στη συνέχεια πολλαπλασιάζονται με τον βαθμό μάθησης (learning rate), προσθέτοντας σε αυτές τις προηγούμενες παραμέτρους. Το ίδιο συμβαίνει και στην περίπτωση n παραμέτρων. Στην περίπτωση των νευρωνικών δικτύων συνήθως χρησιμοποιείται ο αλγόριθμος Stochastic Gradient Descent [31]. Κατά την εφαρμογή του SGD δημιουργούνται τυχαίες ομάδες δειγμάτων (batches) και οι αλλαγές των βαρών προκύπτουν για κάθε batch. Το γεγονός αυτό καθιστά δυνατή την εφαρμογή του αλγορίθμου σε έναν υπολογιστή, μιας και συνήθως είναι αδύνατο να φορτωθεί στη μνήμη του το σύνολο του dataset και να υπολογιστούν οι αρνητικές κλίσεις για όλα τα δείγματα ταυτοχρόνως.

Βελτιώσεις του SGD αποτελούν οι αλγόριθμοι βελτιστοποίησης Adam [32] καθώς και ο Stochastic Gradient Descent με Momentum [33].



Σχήμα 5.1: Ο αλγόριθμος Gradient Descent
Πηγή: *wikipedia.org*

Για τον υπολογισμό της αρνητικής κλίσης πραγματοποιείται η διαδικασία της αντίστροφης διάδοσης (backpropagation). Κατά τη διαδικασία αυτή, με βάση τον κανόνα της αλυσίδας, υπολογίζονται σταδιακά οι μερικές παράγωγοι της συνάρτησης κόστους ως προς τα βάρη. Συνολικά, για να ολοκληρωθεί μια εποχή της μάθησης, κατά τον αλγόριθμο SGD, γίνονται τα εξής βήματα:

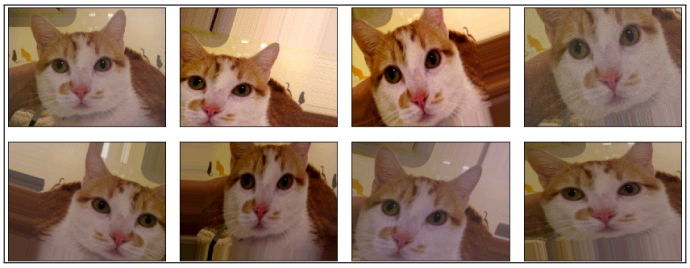
- Πρόβλεψη των εκτιμήσεων ενός batch με τα υπάρχοντα βάρη.
- Υπολογισμός της loss function για το συγκεκριμένο batch.
- Υπολογισμός της κλίσης για κάθε παράμετρο με το backpropagation.
- Ενημέρωση των βαρών με βάση τη διόρθωση που προκύπτει από την αρνητική κλίση πολλαπλασιαζόμενη με το learning rate.
- Επανάληψη των παραπάνω για κάθε batch και κάθε εποχή.

5.5 Regularization των Μοντέλων

Για την εκπαίδευση ενός επιτυχημένου μοντέλου αναγκαία είναι η εφαρμογή τεχνικών Regularization, για την αποφυγή του overfitting. Χαρακτηριστικό παράδειγμα είναι ένα δίκτυο που δέχεται ελάχιστα δεδομένα μάθησης και είναι αναμενόμενο να αναγνωρίσει τα συγκεκριμένα πρότυπα με τα οποία τροφοδοτείται, όπως ο θόρυβος και τα στοιχεία υποβάθρου, χωρίς τελικά να αναγνωρίσει την πραγματική συσχέτιση μεταξύ εισόδου και εξόδου. Για το λόγο αυτό χρειάζεται μεγάλος όγκος δειγμάτων για

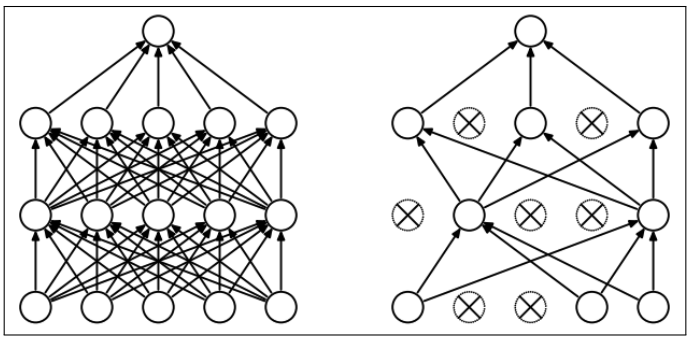
την εξασφάλιση της σωστής εκπαίδευσης ενός μοντέλου. Ωστόσο η δημιουργία δεδομένων προς εκπαίδευση συνήθως απαιτεί ανθρώπινη εργασία .είναι χρονοβόρα και επίπονη. Για το λόγο αυτό έχουν αναπτυχθεί τεχνικές περιορισμού του overfitting. Οι πιο συνηθισμένες τεχνικές είναι:

- Επαύξηση Δεδομένων (Data Augmentation): καλείται η δημιουργία τεχνητών δεδομένων, μέσω της παραποίησης των αρχικών δεδομένων εκπαίδευσης. Στις εικόνες και τα βίντεο, συνηθισμένες τεχνικές augmentation είναι οι γεωμετρικοί μετασχηματισμοί όπως μετάθεση, στροφή, κλίμακα και ο αφινικός μετασχηματισμός. Ακόμη ραδιομετρικοί μετασχηματισμοί όπως η αλλαγή της απόχρωσης, της έντασης η του κορεσμού της εικόνας.



Σχήμα 5.2: Παράδειγμα τυχαίων μετασχηματισμών με σκοπό το augmentation
Πηγή: *way2vat.com*

- Η τεχνική Dropout: Το Dropout [34] αποτελεί μια συνηθισμένη τεχνική αποφυγής του overfitting. Πρόκειται για την απενεργοποίηση ενός ποσοστού τυχαίων νευρώνων σε κάποιο layer, ώστε να αποφεύγονται συστηματικές συσχετίσεις των ίδιων προτύπων.



Σχήμα 5.3: Οι συνδέσεις των νευρώνων με την εφαρμογή dropout
Πηγή: *Rinat Maksutov, medium.com*

- Batch Normalization [35]: Πρόκειται για την κανονικοποίηση κάθε ομάδας δεδομένων που εισάγεται στο δίκτυο κατά τη διάρκεια της εκπαίδευσης, ώστε να έχουν μηδενικό μέσο όρο και μοναδιαία τυπική απόκλιση. Η τεχνική αυτή επιταχύνει την εκπαίδευση και βοηθάει στην αποφυγή του overfitting.

Κεφάλαιο 6

Τεχνικές Αναγνώρισης Δράσης

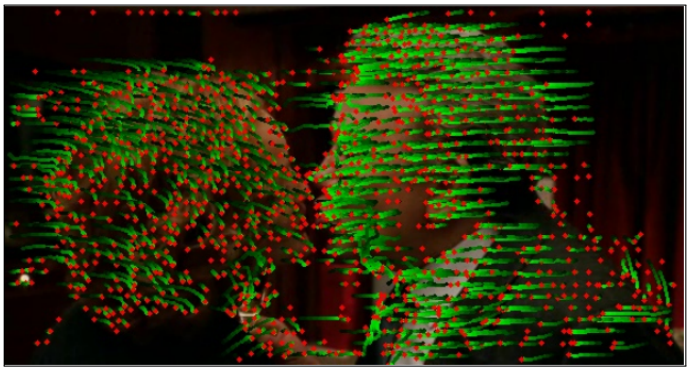
Η αναγνώρισης δράσης αποτελεί ένα σημαντικό πρόβλημα του τομέα της Όρασης Υπολογιστών. Χρησιμοποιείται για παρακολούθηση σε πραγματικό χρόνο, έλεγχο της κυκλοφορίας, διαχείριση καταστάσεων έκτακτης ανάγκης κλπ. Οι πρώτες σημαντικές προσεγγίσεις για την επίλυση του προβλήματος έγιναν το 2005 [6, 5] με κοινό στοιχείο την εξαγωγή αραιών τρισδιάστατων χωροχρονικών χαρακτηριστικών, και την ταξινόμηση τους για την ανίχνευση της συμπεριφοράς. Συνήθως οι περιγραφείς των χαρακτηριστικών αποτελούσαν γενικεύσεις περιγραφέων εικόνων, όπως οι Extended Harris [5], 3D-SIFT [36], HOG3D [37] κλπ. Ωστόσο η χρονική 1D διάσταση έχει διαφορετικά χαρακτηριστικά απ᾽ την 2D χωρική διάσταση της εικόνας. Μια συνήθης τεχνική για εξαγωγή βελτιωμένων χωροχρονικών χαρακτηριστικών είναι ο εντοπισμός της τροχιάς χαρακτηριστικών σημείων, με κάποια τεχνική παρακολούθησης όπως ο KLT Tracker [38]. Μια βελτιωμένη προσέγγιση είναι οι πυκνές τροχιές σημείων [7, 8], σε αντίθεση με τα αραιά χαρακτηριστικά σημεία. Μετά την εξαγωγή τον χαρακτηριστικών ακολουθεί η κωδικοποίηση τους με πιο συνήθεις μεθόδους τους σάκους χαρακτηριστικών και τα Fisher Vectors [24], και τέλος η ταξινόμηση τους με την εκπαίδευση κάποιου ταξινομητή όπως ο SVM classifier. Στην διπλωματική αυτή μελετήθηκε και εφαρμόστηκε η τεχνική Βελτιωμένων Πυκνών Τροχιών με κωδικοποίηση Fisher Vectors και ταξινόμηση μέσω μηχανής διανυσματικής υποστήριξης (SVM).

6.1 Αναγνώριση Δράση με τον Αλγόριθμο Πυκνών Τροχιών

Την πιο πετυχημένη μέθοδο αναγνώρισης δράσης, πριν την επανάσταση που έφερε η βαθιά μάθηση, αποτελούσε η αναπαράσταση του βίντεο μέσω της παρακολούθησης της τροχιάς πυκνών σημείων. Η τεχνική αυτή εκφράστηκε αρχικά στην δημοσίευση του Heng Wang κλ. "Action Recognition by Dense Trajectories" το 2011 [7]. Το 2013 υπήρξε βελτίωση της στην δημοσίευση "Action Recognition with Improved Trajectories" [8].

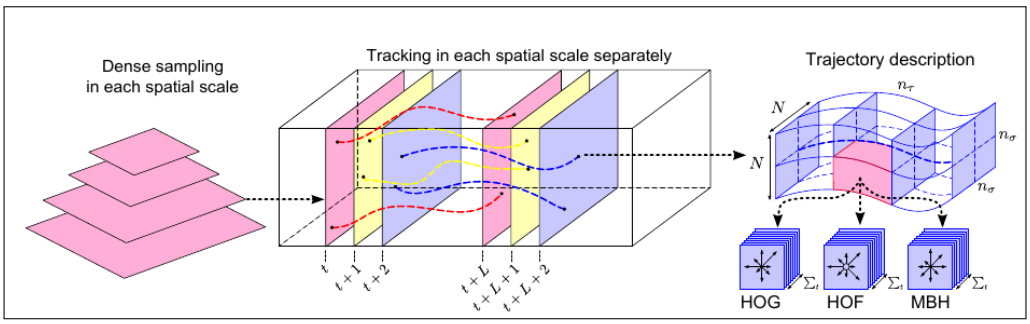
Το πρώτο βήμα για την υλοποίηση της μεθόδου, είναι ο ορισμός ενός κανάβου σημείων (πχ. 5 pixels). Τα σημεία αυτά παρακολουθούνται στα επόμενα καρέ εικόνων, σχηματίζοντας πυκνές τροχιές. Για την αποδέσμευση από την παράμετρο της κλίμακας, τα σημεία παρακολουθούνται σε 8 διαφορετικές κλίμακες της εικόνας. Η ταύτιση των σημείων σε διαδοχικά καρέ του βίντεο υλοποιείται με φίλτρο διαμέσου στην οπτική ροή των εικόνων, όπως περιγράφεται στην εξίσωση: $P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * \omega)|_{(\overline{x}_t, \overline{y}_t)}$, όπου M είναι το φίλτρο διαμέσου της οπτικής ροής της εικόνας. Η οπτική ροή υπολογίζεται συνήθως με τη μέθοδο

Färneback [19]. Για τη μείωση των χονδροειδών σφαλμάτων στην συνταύτιση σημείων, η τροχιά κάθε σημείου παρακολουθείται για L καρέ του βίντεο(συνήθως 15), και έπειτα ξεκινάει παρακολούθηση του σημείου εξ' αρχής. Σε ομογενείς περιοχές της εικόνας, τα σημεία δεν παρακολουθούνται. Ακόμη οι τροχιές των σημείων που είναι πολύ μικρές ή πολύ μεγάλες διαγράφονται. Την περιγραφή της τροχιάς, εν' τέλει, αποτελεί ένα διάνυσμα S', που απαρτίζεται από τις διαδοχικές μεταθέσεις ΔP_t ενός σημείου, κανονικοποιημένο σε σχέση με το άθροισμα των μεταθέσεων, όπως περιγράφεται στην εξίσωση: $S' = \frac{(\Delta P_t,\ldots,\Delta P_{t+L-1})}{\sum_{j=t}^{t+L-1}\|\Delta P_j\|}$, όπου ΔP_t το διάνυσμα της μετάθεσης ενός σημείου.



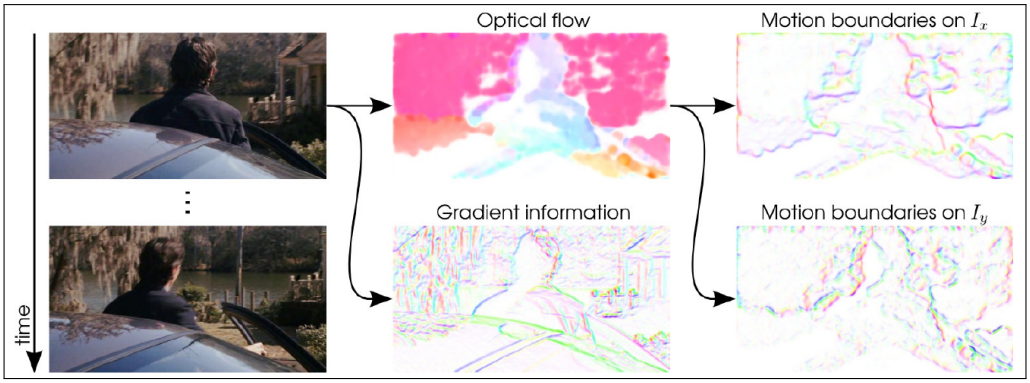
Σχήμα 6.1: Αναπαράσταση των πυκνών τροχιών
Πηγή: *Action Recognition by Dense Trajectories [7]*

Ωστόσο η γεωμετρική απόδοση των τροχιών δεν είναι αρκετή για να αναπαραστήσει εξ' ολοκλήρου ένα βίντεο. Για το λόγο αυτό γίνεται εξαγωγή επιπλέον περιγραφών για το βίντεο. Έτσι, για κάθε τροχιά που έχει χρονικό μήκους L, και χωρικών διαστάσεων NxN pixels, σχηματίζεται ένας όγκος με διαστάσεις LxNxN. Ο όγκος χωρίζεται με κάναβο διαστάσεων $n_\sigma \times n_\sigma \times n_\tau$, όπου για κάθε περιοχή του κανάβου υπολογίζονται οι περιγραφές HOG [21], HOF [22] και MBH [23].



Σχήμα 6.2: Διαγραμματική αναπαράσταση του υπολογισμού των περιγραφών του αλγορίθμου Πυκνών Τροχιών
Πηγή: *Action Recognition by Dense Trajectories [7]*

Η περιγραφή του HOG (Ιστόγραμμα κλίσεων) πετυχαίνει την αναπαράσταση των στατικών αντικειμένων του βίντεο, ενώ του HOF (Ιστόγραμμα της οπτικής ροής) την αναπαράσταση των κινήσεων. Το πρόβλημα που προκύπτει είναι ότι η περιγραφή του HOF δεν μπορεί να εξαλείψει την κίνηση της κάμερας. Για το λόγο αυτό χρησιμοποιείται και η περιγραφή του MBH (Ιστόγραμμα ορίων της κίνησης), η οποία αποτελείται από δύο συνιστώσες, μία για τα οριζόντια και μια για τα κατακόρυφα στοιχεία της οπτικής ροής.



Σχήμα 6.3: Οπτικοποίηση των πληροφοριών που καταγράφουν οι περιγραφές των HOG, HOF και MBH
Πηγή: *Action Recognition by Dense Trajectories [7]*

Τέλος χρησιμοποιείται σάκος χαρακτηριστικών (bag-of-features) στον οποίο κωδικοποιούνται οι περιγραφές των τροχιών, HOG, HOF και MBH και στη συνέχεια ταξινομούνται με μη γραμμικό SVM.

Το 2013 δημοσιεύτηκαν βελτιώσεις για τον αλγόριθμο πυκνών τροχιών. Οι βασικές αλλαγές αφορούν:

- Χρήση αλγορίθμου εντοπισμού σώματος για εστίαση των τροχιών στις ανθρώπινες δραστηριότητες.
- Εκτίμηση της κίνησης της κάμερας μέσω αραιής συνταύτισης σημείων και εκτίμησης της ομογραφείας μεταξύ διαδοχικών καρέ.
- Κωδικοποίηση με Fisher Vectors, αντί για bag-of-words.

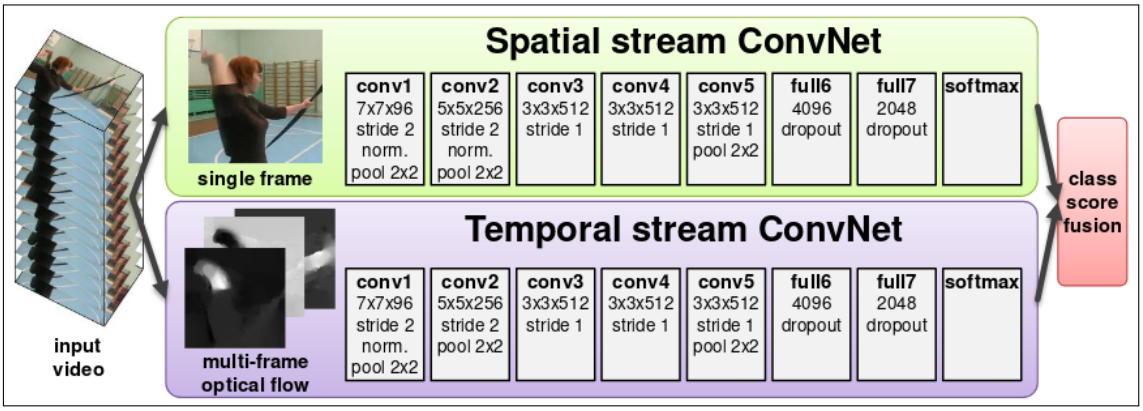
6.2 Αναγνώριση δράσης με Νευρωνικά Δίκτυα

Χάρη στην ραγδαία εξέλιξη των νευρωνικών δικτύων και κυρίως της ταξινόμησης εικόνων, όπως ήταν αναμενόμενο, το 2014 έγιναν οι πρώτες πετυχημένες προσπάθειες αξιοποίησης της βαθιάς μάθησης για την αναγνώριση δράσης. Ένα βασικό πρόβλημα που καλείται να λυθεί στην περίπτωση του video classification με βαθιά μάθηση, είναι η αδυναμία δημιουργίας ενός dataset αντίστοιχου πλήθους δειγμάτων όπως στην περίπτωση της ταξινόμησης εικόνων. Ακόμη κι αν αυτό ήταν δυνατό, θα απαιτούσε υπέρογκο ποσό μνήμης για την αποθήκευση του, καθώς και βδομάδες εκπαίδευσης για τον προσδιορισμό των βαρών του μοντέλου εκτίμησης. Έτσι δυσκολεύεται η διαδικασία ταξινόμησης βίντεο, καθώς τα μοντέλα είναι επιρρεπή σε overfitting. Το πιο δημοφιλές dataset για action recognition τα τελευταία χρόνια είναι το UCF101 [39], με 101 κατηγορίες ανθρώπινων δράσεων, σε 13320 RGB video.

6.2.1 Θεμελιώδεις Τεχνικές Προσεγγίσεις

Την πιο πετυχημένη ρηχή προσέγγιση αποτελούν οι πυκνές τροχιές [8], η οποία πετυχαίνει ακρίβεια 87.9% στο dataset UCF-101. Δουλεία ορόσημο αποτέλεσε η δημοσίευση των Karen Simonyan και Andrew Zisserman [9], με τα συνελικτικά δίκτυα 2 ροών. Η τεχνική αυτή είναι υλοποίηση της ιδέας ότι ο άνθρωπος για να αποκωδικοποιήσει την πληροφορία της όρασης και να αναγνωρίσει μια ανθρώπινη δράση,

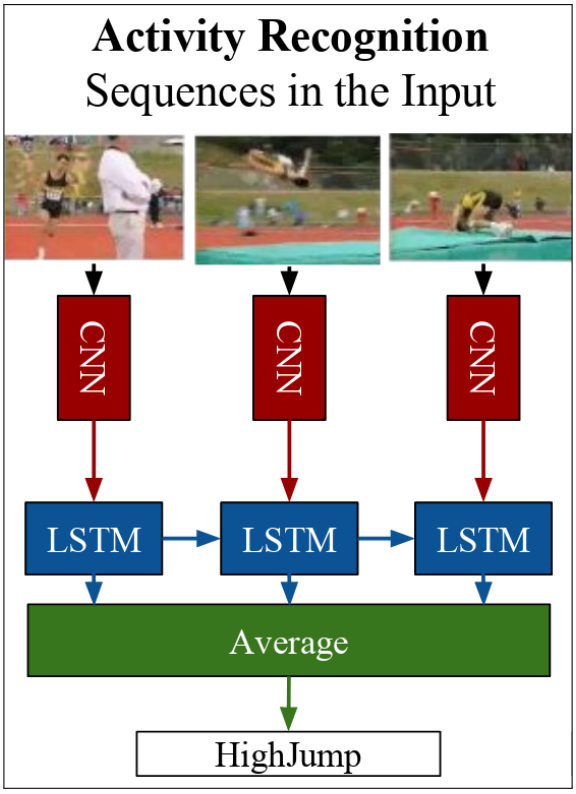
αφενός αντιλαμβάνεται τα αντικείμενα που υπάρχουν στον χώρο και αφετέρου τις κινήσεις που πραγματοποιούνται. Έτσι λοιπόν σχεδιάστηκαν δυο ανεξάρτητα δίκτυα, όπου το ένα αποτελεί ένα τυπικό δίκτυο ταξινόμησης εικόνων, ενώ το δεύτερο αφορά την αναγνώριση και ταξινόμηση της κίνησης. Για τον προσδιορισμό της κίνησης προϋπολογίστηκαν εικόνες οπτικής ροής, οι οποίες τροφοδότησαν τη δεύτερη ροή του μοντέλου. Η είσοδος των εικόνων RGB της πρώτης ροής αφορούσε μόνο μία εικόνα του βίντεο, ενώ της δεύτερης ροής μια ομάδα χρονικά συνεχόμενων εικόνων οπτικής ροής. Πιο συγκεκριμένα, οι εικόνες οπτικής ροής, αναπαρίστανται με 2 κανάλια μετάθεσης κατα x και y. Στη συνέχεια τα L καρέ του βίντεο στιβάζονται σαν κανάλια εικόνας, οπότε τελικά η είσοδος αφορά τανυστή (tensor) μεγέθους $w * h * 2L$ Η τελική εκτίμηση του μοντέλου προκύπτει από το συνδιασμό των δυο ροών μέσω ενός SVM ταξινομητή. Στο σχήμα 6.4 παρουσιάζεται η αρχιτεκτονική των δύο συνελικτικών δικτύων, με 5 συνελικτικά layers και 2 πλήρως συνδεδεμένα layers με χρήση dropout, για τον περιορισμό του overfiting. Η τεχνική αυτή πετυχαίνει 88% στο UCF-101, βελτιώνοντας κατά 0.1% την υπάρχουσα ακρίβεια και κυρίως, αποδεικνύοντας ότι η χρήση deep learning μπορεί να βελτιώσει τις υπάρχουσες ακρίβειες στην αναγνώριση δράσης.



Σχήμα 6.4: Η αρχιτεκτονική 2 ροών για αναγνώριση δράσης
Πηγή: *Two-Stream Convolutional Networksfor Action Recognition in Videos [9]*

Μια διαφορετική προσέγγιση αποτελεί το δίκτυο LRCN (Long-term Recurrent Convolutional Networks) του Donahue κλ. [11] με τη χρήση συνδιασμού συνελικτικού - LSTM δικτύου. Η αρχιτεκτονική αυτή πρόκειται για ένα συνελικτικό δίκτυο VGG [2] και ακολούθως ένα LSTM layer και ένα πλήρως συνδεδεμένο softmax layer με 101 νευρώνες για την ταξινόμηση των 101 κατηγοριών. Ουσιαστικά, τα συνελικτικά δίκτυα εξάγουν χαρακτηριστικά σε κάθε καρέ διατηρώντας την χρονική διάσταση και στη συνέχεια το LSTM κωδικοποιεί αυτά τα χαρακτηριστικά. Η κωδικοποίηση αυτή τελικά ταξινομείται σε μία από τις κατηγορίες δράσης. Για την τελική εκτίμηση της δράσης, εκπαιδεύτηκε ένα RGB και ένα Optical Flow μοντέλο, και υπολογίστηκε σταθμισμένο βάρος των δύο εκτιμήσεων. Το μοντέλο αυτό πετυχαίνει ακρίβεια 82.3% στο UCF-101.

Την τρίτη θεμελιώδη προσέγγιση αποτελεί η δημοσίευση Learning Spatiotemporal Features with 3D Convolutional Networks [12]. Πρόκειται για εξονυχιστική μελέτη των 3D συνελίξεων για την αναγνώριση δράσης. Το μοντέλο αυτό εξήγαγε χαρακτηριστικά με τη χρήση βαθιάς αρχιτεκτονικής 3D συνελικτικών layers, και στη συνέχεια τα ταξινομούσε με SVM ταξινομητή. Το πρόβλημα αυτής της στρατηγικής, αλλά και γενικότερα των 3D συνελίξεων είναι το επεξεργαστικό κόστος για την εκπαίδευση

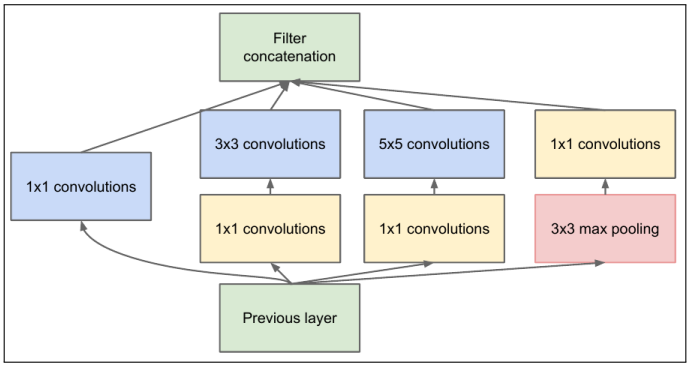


Σχήμα 6.5: Η αρχιτεκτονική LRCN
Πηγή: *Long-term Recurrent Convolutional Networks for Visual Recognition and Description [11]*

3D φίλτρων, που ανεβάζει εκθετικά την επεξεργαστική δύναμη και τον χρόνο που απαιτείται για την εκπαίδευση ενός τέτοιου μοντέλου.

6.2.2 Η αρχιτεκτονική Inflated 3D

Η τελευταία αρχιτεκτονική που θα αναφερθεί είναι το δίκτυο Inflated 3D [13]. Πρόκειται για γενίκευση της αρχιτεκτονικής Inception-v1 [3]για ταξινόμηση εικόνων, σε ταξινομητή βίντεο.

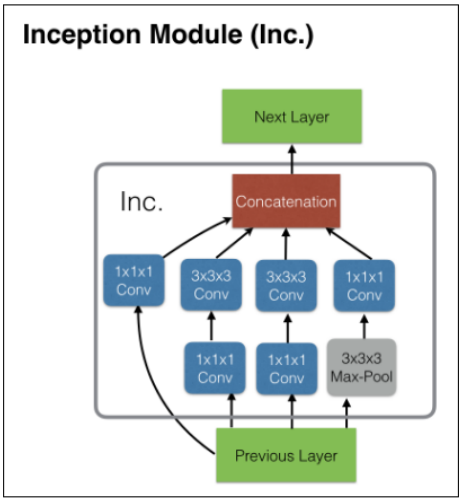


Σχήμα 6.6: Διαγραμματική Αναπαράσταση του Inception Module
Πηγή: *Going Deeper with Convolutions [3]*

Η αρχιτεκτονική Inception [3] αποτελεί ένα πολύ βαθύ δίκτυο με συνελικτικά layers για την ταξινόμηση εικόνων. Εστιάζει τόσο στην ακρίβεια των προβλέψεων,

όσο και στην βελτιστοποίηση της ταχύτητας για την εκπαίδευση και πρόβλεψη του δικτύου. Δομικό στοιχείο της αποτελεί το Inception Module. Πρόκειται για το βασικό αρχιτεκτονικό στοιχείο του δικτύου, το οποίο ευθύνεται για την εξαγωγή χαρακτηριστικών σε διαφορετικές κλίμακες. Όπως φαίνεται και στο σχήμα 6.6, εξάγονται χαρακτηριστικά με τρεις διαφορετικές διαστάσεις τετραγωνικών παραθύρων μεγέθους 1, 3 και 5. Επιπλέον γίνεται χρήση $1 * 1$ συνελίξεων για τη μείωση των διαστάσεων των χαρακτηριστικών που εισέρχονται. Τελικά τα φίλτρα αυτά συνδέονται και τροφοδοτούν τα επόμενα layers.

Η αρχιτεκτονική Inception αποφεύγει τη χρήση πλήρως συνδεδεμένων δικτύων στην κορυφή του δικτύου, περιορίζοντας σημαντικά τις παραμέτρους του μοντέλου. Η δουλεία αυτή [13] , βασίζεται στην ιδέα ότι η εξέλιξη της ταξινόμησης εικόνων

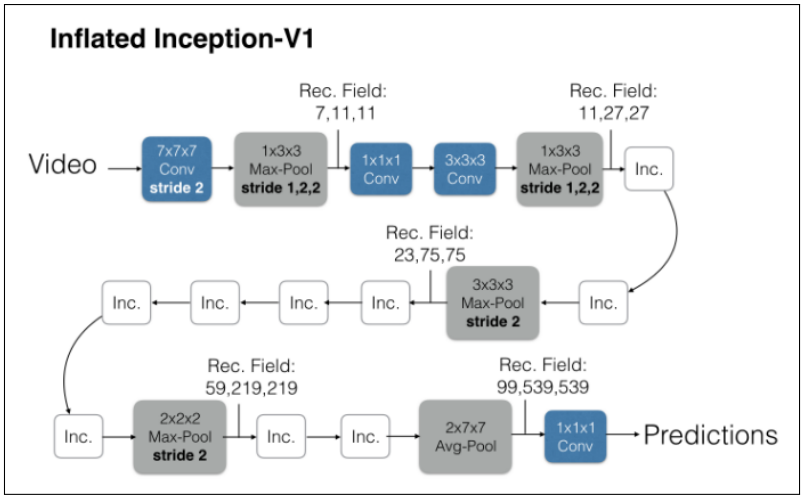


Σχήμα 6.7: Διαγραμματική Αναπαράσταση του Inception 3D Module
Πηγή: *Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset [13]*

πρέπει να αξιοποιηθεί στο έπακρο, και θα ωφελήσει την ταξινόμηση βίντεο. Έτσι λοιπόν τα 2D Convolutional, και Max Pooling layers της αρχιτεκτονικής Inception σε 3D, χρίζοντας την κατάλληλη για ταξινόμηση βίντεο. Με τη χρήση 3D Convolutions, καθίσταται δυνατή η αναπαράσταση του βίντεο με την εξαγωγή χωροχρονικών χαρακτηριστικών.

Σε σχέση με το Inception-v1, έχει γίνει χρήση Batch Normalization [35]. Αρκετή έρευνα πραγματοποιήθηκε όσον αφορά και την χρήση προεκπαιδευμένων βαρών. Τα προεκπαιδευμένα βάρη της αρχιτεκτονικής Inception στο ImageNet [40], διαμορφώθηκαν ώστε να είναι κατάλληλα για χρήση ως αρχικές τιμές στο I3D. Ταυτόχρονα, δημοσιεύεται το dataset Kinetics [41] για αναγνώριση δράσης, με πολύ μεγαλύτερο όγκο δειγμάτων απ᾽ αυτόν του UCF-101, και έπειτα από εκπαίδευση με τη χρήση αυτού, τα βάρη χρησιμοποιούνται ως αρχικές τιμές στο UCF-101. Και πάλι γίνεται εκπαίδευση 2 μοντέλων, με εικόνες RGB και οπτικής ροής, και η τελική πρόβλεψη προκύπτει απ᾽ το μέσο όρο των επιμέρους προβλέψεων.

Σαν αποτέλεσμα, το μοντέλο πετυχαίνει ακρίβεια 98% στο UCF-101 και αποτελεί μέχρι και σήμερα μια από τις καλύτερες τεχνικές αναγνώρισης δράσης.



Σχήμα 6.8: Η αρχιτεκτονική Inflated 3D
Πηγή: *Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset [13]*

Μέρος ΙΙΙ

Μεθοδολογία - Αποτελέσματα

Βασική επιδίωξη της διπλωματικής ήταν η δημιουργία ενός συστήματος αυτόματης ταξινόμησης της Δοκιμασίας Εξαναγκασμένης Κολύμβησης. Πιο συγκεκριμένα, στόχο αποτέλεσε ο σχεδιασμός και ανάπτυξη ενός συστήματος επεξεργασίας, το οποίο δέχεται σαν είσοδο ένα βίντεο που περιλαμβάνει το πείραμα ενδιαφέροντος, και επιστρέφει τον τελικό χρόνο που διήρκησε η κάθε κατηγορία εντός του βίντεο. Για την εκπλήρωση αυτού του στόχου χρειάστηκαν επιμέρους ενέργειες οι οποίες ήταν:

- Εύρεση Dataset με μεγάλο όγκο δεδομένων βίντεο και τις πραγματικές ταξινομήσεις κάθε δευτερολέπτου τους, όπως έχουν προκύψει από την χειροκίνητη ταξινόμηση ειδικών.
- Επιδιόρθωση και Οργάνωση του Dataset σε μορφή κατάλληλη για την αξιοποίηση του με τεχνικές μηχανικής μάθησης.
- Κατασκευή μοντέλων για την πρόβλεψη της κατηγορίας κάθε στιγμιότυπου ή κάθε χρονικού διαστήματος ενός βίντεο.
- Σύγκριση και βελτιστοποίηση αυτών των μοντέλων.
- Ποιοτική και ποσοτική αξιολόγηση των αποτελεσμάτων

Στο κεφάλαιο που ακολουθεί αναλύονται επεξηγηματικά όλες οι παραπάνω ενέργειες, για την κατασκευή ενός τέτοιου μοντέλου.

Κεφάλαιο 1

Προετοιμασία Δεδομένων Εφαρμογής

1.1 Περιγραφή του Dataset

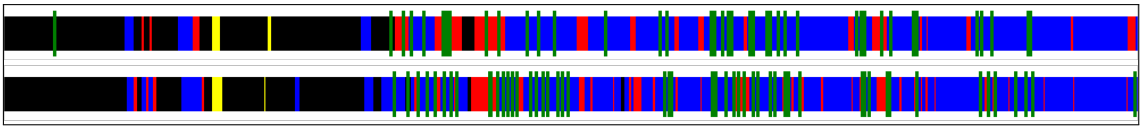
Η δημιουργία μοντέλων εκτίμησης πραγματοποιήθηκε με διαδικασία επιβλεπομένης ταξινόμησης. Για το λόγο αυτό, αναγκαία ήταν η εύρεση ενός dataset με δείγματα των αντικείμενων προς ταξινόμηση και την πραγματική τους κλάση. Στη συγκεκριμένη περίπτωση, για το πείραμα εξαναγκασμένης κολύμβησης, χρησιμοποιήθηκαν βίντεο με πειράματα της δοκιμασίας αυτής, καθώς και τις αντίστοιχες ταξινομήσεις που πραγματοποίησαν ειδικοί, με συνεχείς παρατηρήσεις σε πραγματικό χρόνο για κάθε βίντεο.

Για το λόγο αυτό, ήρθαμε σε συνεργασία με τους διδάκτορες ερευνητές της Ιατρικής σχολής Χριστίνα Δάλλα και Νίκο Κόκρα, οι οποίοι ειδικεύονται στο θέμα και διέθεταν μεγάλο όγκο τέτοιων δεδομένων τα οποία έχουν ταξινομήσει οι ίδιοι. Οι συγκεκριμένοι ερευνητές για της ανάγκες ερευνητικών δραστηριοτήτων [42] σκόραραν χειροκίνητα με τη χρήση της εφαρμογής Kinoscope [43] 105 διαφορετικά πειράματα σε επιμύες Wistar διάρκειας 5 λεπτών. Τα βίντεο έχουν διαστάσεις 1920*1080 και 25 frames ανά second. Για κάθε βίντεο υπάρχουν οι παρατηρήσεις και των 2 ερευνητών. Έπειτα από έλεγχο απορρίφθηκαν 5 βίντεο, λόγω ελλιπών παρατηρήσεων, οπότε τελικά προέκυψαν 100 βίντεο προς χρήση. Οι κατηγορίες ενδιαφέροντος και οι αντίστοιχες κωδικοποιήσεις τους παρουσιάζονται στον πίνακα 1.1.

Κωδικός	Κατηγορία - Ελληνικά	Κατηγορία - Αγγλικά	Χρώμα
0	Ακινησία	Immobility	Μπλε
1	Κολύμβηση	Swimming	Κόκκινο
2	Αναρρίχηση	Climbing	Μαύρο
3	Τίναγμα Κεφαλής	Head Shaking	Πράσινο
4	Κατάδυση	Diving	Κίτρινο

Πίνακας 1.1: Οι κατηγορίες ενδιαφέροντος της Δοκιμασίας Εξαναγκασμένης Κολύμβησης

Οι εκτιμήσεις των ερευνητών δόθηκαν σε μορφή εικόνων png, όπως προκύπτουν απ' την εφαρμογή Kinoscope. Οι εικόνες αυτές έχουν πλάτος 1000 το οποίο αντιστοιχίζεται σε χρονικό διάστημα 5 λεπτών. Επομένως τεχνικά η ακρίβεια των παρατηρήσεων είναι 0.3 sec. Στην πραγματικότητα δεν μπορεί να θεωρηθεί μεγαλύτερη από του ενός sec καθώς επηρεάζεται από τον χρόνο αντίδρασης του παρατηρητή.



Σχήμα 1.1: Παράδειγμα των παρατηρήσεων των 2 ερευνητών σε κοινό βίντεο

Παρακάτω παρουσιάζονται σε πίνακες ορισμένα χρήσιμα χαρακτηριστικά για το dataset:

Κατηγορία	**Παρατηρητής 1**		**Παρατηρητής 2**	
	sec	Ποσοστιαία	sec	Ποσοστιαία
Immobility	12308	41.17	11962	40.02
Swimming	4821	16.13	4918	16.45
Climbing	11092	37.11	11468	38.36
Head Shaking	1632	5.46	1504	5.03
Diving	39	0.13	40	0.13
Σύνολο	29892	100	29892	100

Πίνακας 1.2: Οι συχνότητα της κάθε κατηγορίας ανά παρατηρητή

Όπως προκύπτει απ᾽ τα στατιστικά στοιχεία του πίνακα 1.2 υπάρχει μεγάλη ανισορροπία μεταξύ των κλάσεων, καθώς η κατηγορίες climbing και immobility καταλαμβάνουν το 78% του dataset, ενώ οι 3 υπόλοιπες το 22%. Επιπλέον, ιδιαίτερα η κατηγορία Diving αποτελεί μόλις το 0.13% του dataset.

Συμφωνία Παρατηρητών	**σε sec**	**Ποσοστιαία**
	23110	77.31
Σύγχυση κατηγοριών	**σε sec**	**Ποσοστιαία**
Immobility με Swimming	2740	9.17
Climbing με Swimming	1674	5.60
Immobility με Climbing	845	2.83
Λοιπές Αστοχίες	1523	5.10

Πίνακας 1.3: Πίνακας σύγχυσης μεταξύ των 2 παρατηρητών

Στο σχήμα 1.1, αλλά και σύμφωνα με τον πίνακα 1.3, οι παρατηρητές συμφωνούν κατά 77% μεταξύ τους, και φαίνεται να συγχέουν κυρίως τις κλάσεις Immobility με Swimming, και έπειτα Immobility με Climbing. Σημειώνεται ακόμη, ότι η κατηγορία Head Shaking, διαρκεί περίπου 0.5 - 1 sec, γεγονός που καθιστά δύσκολο το συγχρονισμό της στο βίντεο λόγω του χρόνου αντίδρασης των παρατηρηρών. Επομένως αναμένεται δυσκολία στην μάθηση αυτής της κατηγορίας.

1.2 Προεπεξεργασία του Dataset

Το dataset που χρησιμοποιήθηκε υπέστη μια σειρά τροποποιήσεων ώστε να έρθει σε μορφή κατάλληλη για αξιοποίηση του σε διαδικασίες επιβλεπόμενης ταξινόμησης. Οι τροποποιήσεις αυτές παρουσιάζονται παρακάτω με την αντίστοιχη σειρά:

1. Με την θέαση των βίντεο του dataset, παρατηρήθηκε η ύπαρξη του φαινομένου Interlacing, καθώς η εγγραφή του βίντεο δεν πραγματοποιήθηκε με progressive

σάρωση όπως συνηθίζεται. Ουσιαστικά, η σάρωση Interlaced, πρόκειται για την αποθήκευση 2 διαφορετικών συνεχόμενων χρονικά σκηνών στο ίδιο καρέ του βίντεο. Το αποτέλεσμα αυτό προκύπτει όταν η κάμερα που χρησιμοποιείται αποθηκεύει την πρώτη σκηνή στις μονές γραμμές κάθε εικόνας και την δεύτερη σκηνή στις ζυγές γραμμές. Έτσι δημιουργείται ένα ανεπιθύμητο εφέ όπως φαίνεται στην εικόνα 1.2, σε κινούμενα αντικείμενα όπως η ουρά του επιμύ. Για την μετατροπή του βίντεο σε progressive σάρωση, συντάχθηκε κώδικας shell με χρήση της βιβλιοθήκης ffmpeg. Όπως είναι λογικό, εκτός απ᾽ την αφαίρεση αυτής της μορφής θορύβου, προκύπτει και διπλασιασμός των frames ανά second, από 25 σε 50.

Σχήμα 1.2: Μετατροπή της σάρωσης Interlaced σε Progressive

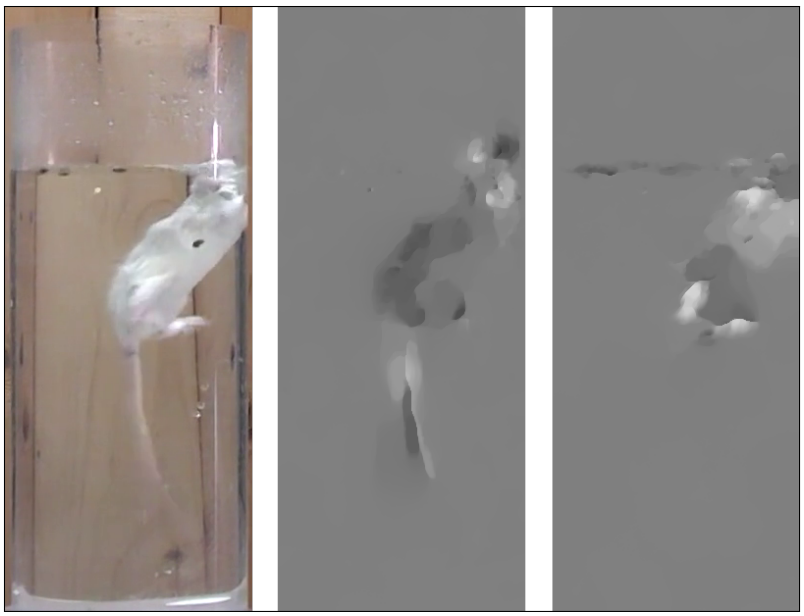
2. 18 από τα βίντεο, στην αρχική τους μορφή, κατέγραφαν 2 πειράματα ταυτόχρονα. Αυτά τα βίντεο κόπηκαν χωρικά με χειροκίνητο τρόπο με τη χρήση της εφαρμογής Handbreak, σε 2 επιμέρους βίντεο, ώστε το καθένα να περιλαμβάνει ένα μοναδικό πείραμα. Όπως έχει ήδη αναφερθεί, τελικά προέκυψαν 100 βίντεο, για 100 ανεξάρτητα πειράματα.

3. Τα βίντεο περιλάμβαναν στο μεγαλύτερο μέρος το παρασκήνιο του εργαστηρίου. Το γεγονός αυτό, επηρεάζει την επεξεργαστική ισχύ που χρειάζεται για την εκπαίδευση του μοντέλου και την πρόβλεψη της ταξινόμησης αλλά και την ποιότητα των αποτελεσμάτων. Για το λόγο αυτό, τα βίντεο κόπηκαν ώστε να παραμείνει μόνο η δεξαμενή. Όπως είναι λογικό, οι διαστάσεις των βίντεο μειώθηκαν περίπου στο μισό τους, και πλέον δεν διατηρούν σταθερή αναλογία ύψους - πλάτους.

4. Το βασικότερο πρόβλημα της προεπεξεργασίας του dataset, αποτέλεσε το γεγονός ότι οι ερευνητές ξεκινούσαν τις παρατηρήσεις τους μερικά δευτερόλεπτα αφότου ξεκινούσε το βίντεο. Επομένως υπήρξε για κάθε βίντεο ένα σφάλμα μετάθεσης στη διάσταση του χρόνου. Για την επίλυση αυτού του προβλήματος, υλοποιήθηκε αλγόριθμος στη γλώσσα python για την προβολή του βίντεο και την ταυτόχρονη εκτύπωση των παρατηρήσεων των 2 ερευνητών. Έπειτα από δοκιμές στην αλλαγή του συγχρονισμού των παρατηρήσεων, τελικά διορθώθηκαν κατά το βέλτιστο και εξήχθησαν οι συγχρονισμένες παρατηρήσεις σε

αρχείο text. Η διαδικασία αυτή πραγματοποιήθηκε για κάθε βίντεο και για τους 2 παρατηρητές ανεξάρτητα. Ωστόσο το γεγονός αυτό αποτέλεσε μια πηγή σφάλματος για τη διαδικασία την εκπαίδευσης, ιδιαίτερα σε σύντομες χρονικά κατηγορίες όπως το Head Shake.

5. Για την εκπαίδευση μοντέλων νευρωνικών δικτύων, η είσοδος των δεδομένων πρέπει να δίνεται σε μορφή εικόνων και όχι βίντεο. Για το λόγο αυτό, τα βίντεο μετατράπηκαν σε εικόνες png, διατηρώντας στο μέγιστο την χρονική και χωρική ακρίβεια των βίντεο. Ταυτόχρονα εξήχθησαν και εικόνες οπτικής ροής με τη χρήση του αλγορίθμου TV-L1 [20]. Σε αυτό το σημείο αναφέρεται ότι η διαδικασία αυτή πραγματοποιήθηκε επαναληπτικά, για διαφορετικά frames ανά second, μιας και αποτελούν μια παράμετρο που δεν θα μπορούσε να αλλάξει μετά την εξαγωγή των εικόνων οπτικής ροής. Έτσι τελικά, εξήχθησαν εικόνες RGB και εικόνες οπτικής ροής για 50, 25, 16.6, 12.5 και 10 frames ανά second. Οι εικόνες οπτικής ροής, αποθηκεύονται ως 2 συνιστώσες της μετάθεσης του κάθε pixel, της δεύτερης εικόνας σε σχέση με την πρώτη. Η μία συνιστώσα αφορά την οριζόντια μετάθεση του κάθε pixel και η δεύτερη τις κατακόρυφη μετάθεση. Για την απόδοση της κίνησης σε εικόνα, η μεσαία ραδιομετρική τιμή 128, αφορά την μη ύπαρξη κίνησης, ενώ η αρνητική και θετική μετάθεση αποδίδεται με τιμές προς το 0 και το 255 αντίστοιχα, αναλογικά με το μέγεθος της μετάθεσης, όπως φαίνεται και στο σχήμα 1.3. Ο αλγόριθμος αυτός πραγματοποιήθηκε σε γλώσσα python με τη χρήση της βιβλιοθήκης OpenCV και χρησιμοποιήθηκαν οι προεπιλεγμένες παράμετροι της συνάρτησης DualTVL1OpticalFlow.



Σχήμα 1.3: Το frame ενός βίντεο στα αριστερά, οι 2 συνιστώσες της οπτικής ροής κατά x και y στα δεξιά

6. Σε κάθε διαδικασία επιβλεπόμενης ταξινόμησης, και ιδιαιτέρως στην εκπαίδευση νευρωνικών δικτύων, το dataset πρέπει να χωριστεί σε 2 επιμέρους sets, τα λεγόμενα training και test set. Έτσι, το dataset χωρίστηκε σε 2 υποσύνολα, με απολύτως τυχαίο τρόπο. Ο διαχωρισμός έγινε σε επίπεδο βίντεο. Από τα 100 βίντεο τα 20 επιλέχτηκαν για το υποσύνολο test και τα υπόλοιπα 80 ως train. Δεν θεωρήθηκε αναγκαίο για τις ανάγκες τις διπλωματικής, η ύπαρξη

ενός τρίτου υποσυνόλου validation.

Οι παραπάνω διαδικασίες πραγματοποιήθηκαν για όλα τα βίντεο, ανεξαρτήτως μοντέλου εκτίμησης. Στη συνέχεια, ανάλογα με τις ανάγκες κάθε μοντέλου, πραγματοποιήθηκαν περαιτέρω ενέργειες προεπεξεργασίας, πριν τη τροφοδοσία των εισόδων στο μοντέλο. Αυτές θα αναφερθούν στη συνέχεια.

1.3 Παραμετροποίηση του Dataset

Η προεπεξεργασία έγινε στο dataset με τρόπο ώστε να είναι δυνατή η παραμετροποίηση όλων των στοιχείων που ενδέχεται να επηρεάσουν τη διαδικασία της εκπαίδευσης. Το dataset μετά την βασική προεπεξεργασία αποτελείται από τις εισόδους των μοντέλων, δηλαδή τα frames του κάθε βίντεο, όπως έχει ήδη ειπωθεί σε 10, 12.5, 16.6, 25 και 50 frames ανά second.

Υπενθυμίζεται ότι τα δεδομένα εκπαίδευσης είναι πεντάλεπτα βίντεο, και αναγκαία είναι η ταξινόμηση τουλάχιστον κάθε δευτερολέπτου αυτών. Η πιο συνήθης διαδικασία αναγνώρισης δράσης, αφορά βίντεο τα οποία περιλαμβάνουν μοναδική δράση. Ωστόσο, στο συγκεκριμένο dataset, όπως έχει αναφερθεί, κάθε βίντεο διαρκεί 5 λεπτά και περιλαμβάνει συνεχείς αλλαγές μεταξύ όλων των κατηγοριών. Για το λόγο αυτό γίνεται η βασική παραδοχή ότι η ταξινόμηση θα διενεργηθεί σε τμήματα του βίντεο με σταθερό χρονικό διάστημα μεταξύ τους, τα οποία είναι απολύτως ανεξάρτητα.

Οι αληθείς τιμές (ground truth) βρίσκονται σε αρχεία text, και αναφέρονται στα 50 frames ανά second, δυο για κάθε βίντεο λόγω των δυο διαφορετικών παρατηρητών. Σύμφωνα με τα παραπάνω, υλοποιήθηκαν αλγόριθμοι για τη φόρτωση αυτών των δεδομένων καθορίζοντας τις εξής παραμέτρους, οι οποίες έπειτα βελτιστοποιήθηκαν:

- Τα frames ανά second (fps): Πρόκειται για την χρονική ακρίβεια των βίντεο. Ανάλογα με τα fps που επιλέγονται για την εκπαίδευση κάθε μοντέλου, υπολογίζονται αυτομάτως και οι αντίστοιχες αληθείς τιμές, πχ για 25 fps παραλείπεται κάθε δεύτερη τιμή των αληθών τιμών. Εκτιμάται ότι με χαμηλά fps (πχ. 10), δεν θα είναι δυνατή η αναγνώριση γρήγορων δράσεων από το μοντέλο, ενώ αντιθέτως με πολύ ψηλά fps (πχ 50), θα δυσχεραίνεται η χρονική συσχέτιση των frames του βίντεο, λόγω της υπέρογκης πληροφορίας, καθώς και θα καθυστερεί ιδιαίτερα η διαδικασία της εκπαίδευσης και πρόβλεψης.

- Το είδος της εικόνας: Εφόσον έχει πραγματοποιηθεί εξαγωγή εικόνων rgb και optical flow, είναι δυνατή η χρήση οποιουδήποτε τύπου δεδομένων απ᾽ τα δύο, καθώς και συνδυασμός τους.

- Το μέγεθος της εικόνας: Το αρχικό μέγεθος της εικόνας πρόκειται να επηρεάσει και πάλι την ταχύτητα εκπαίδευσης και εκτίμησης. Ωστόσο με πολύ μικρό μέγεθος ενδέχεται να δυσχεραίνεται η αναγνώριση λεπτομερών χωρικών προτύπων από το μοντέλο.

- Το χρονικό διάστημα κάθε βίντεο προς εκτίμηση: Πρόκειται για το πόσα δευτερόλεπτα ή πόσα frames θα αποτελούν το κάθε τμήμα του βίντεο, που θα τροφοδοτεί την είσοδο των μοντέλων. Με μικρές τιμές οι δράσεις των επιμυών δεν θα είναι αναγνωρίσιμες, ωστόσο με μεγάλες τιμές ενδέχεται να συμπεριλαμβάνονται πάνω από μία κατηγορίες συμπεριφοράς στο ίδιο βίντεο.

- Η ταυτότητα των εξόδων: Όπως έχει αναφερθεί, υπάρχουν αληθείς τιμές από δύο διαφορετικούς παρατηρητές. Επομένως είναι δυνατή η χρήση μόνο του ενός, ή και μόνο των σημείων στα οποία συμφωνούν οι δυο παρατηρητές. Επιπλέον είναι δυνατή η χρήση των αληθών τιμών του ενός απ᾽ τους δύο και η ενίσχυση του βάρους του κάθε δείγματος εφόσον συμφωνεί και ο δεύτερος.

Συνολικά, κατασκευάστηκε αλγόριθμος που δέχεται όλες τις παραπάνω παραμέτρους, και αυτομάτως επιστρέφει τις εισόδους που ζητήθηκαν, καθώς και τις κατάλληλες εξόδους. Όλες οι παραπάνω παράμετροι βελτιστοποιήθηκαν. Στην περίπτωση των μοντέλων βαθιάς μάθησης, πραγματοποιήθηκε επιπλέον augmentation κατά τη στιγμή της φόρτωσης των βίντεο στο μοντέλο. Συγκεκριμένα:

- Πιθανότητα 50% καθρεφτισμού κατά των άξονα x, για κάθε βίντεο που εισάγεται στο μοντέλο
- Τυχαία χωρική αποκοπή (crop) της τάξης του 10% του βίντεο στις διαστάσεις x, y των εικόνων.

Οι παράμετροι του augmentation, και ιδιαιτέρως του crop, επιλέχτηκαν με κριτήριο την μη αποκοπή σημαντικών σημείων του χώρου. Ακόμη κατά την φόρτωση των εικόνων σε μοντέλα νευρωνικών δικτύων, οι τιμές των εικόνων που έχουν οριακές τιμές 0 - 255, μειώνονται κατά 128 και στη συνέχεια διαιρούνται με το 128, ώστε να έχουν μέσο όρο κοντά στο 0, και όρια κοντά στο -1, +1.

1.4 Δείκτες Αξιολόγησης

Για την αξιολόγηση των αποτελεσμάτων κάθε μοντέλου στο συγκεκριμένο dataset, έγινε υπολογισμός των εξής στατιστικών στοιχείων:

- Πίνακας σύγχυσης (Confusion Matrix): Πρόκειται για πίνακα διαστάσεων $k * k$, όπου k ο αριθμός των κλάσεων της ταξινόμησης. Σε κάθε στοιχείο του πίνακα f_{ij}, συμπληρώνεται το πλήθος των δειγμάτων που εκτιμήθηκαν απ΄ το μοντέλο ως κατηγορία i, με πραγματική κατηγορία του δείγματος j. Η διαγώνιος $i = j$ αναφέρει τις πετυχημένες εκτιμήσεις του μοντέλου, ενώ πάνω απ᾽ τη διαγώνιο αναφέρονται οι False Negative προβλέψεις, και κάτω οι False Positive.

Πίνακας Σύγχυσης	**Πρόβλεψη Μοντέλου**		
Αληθής Τιμή		Positive	Negative
	Positive	f_{11} TP	f_{12} FN
	Negative	f_{21} FP	f_{22} TN

Πίνακας 1.4: Ορισμός του Πίνακα Σύγχυσης

TP = True Positive (Αληθές Θετικό)

FP = False Positive (Ψευδές Θετικό)

TN = True Negative (Αληθές Αρνητικό)

FN = True Negative (Ψευδές Αρνητικό)

- Ευστοχία(Accuracy):

$$\frac{TP + TN}{TP + TN + FP + FN}$$

- Ανάκληση (Recall):

$$\frac{TP}{TP + FN}$$

- Ακρίβεια (Precision):

$$\frac{TP}{TP + FP}$$

- F-Score:

$$\frac{2 * TP}{2 * TP + FP + FN}$$

Η ανάκληση, ευστοχία και το F-score, υπολογίστηκαν σε επίπεδο κλάσης, μάκρο-στατιστικών (ο αστάθμιστος μέσος όρος ενός στατιστικού δείκτη για κάθε κατηγορία) και σταθμισμένων στατιστικών (σταθμισμένος μέσος όρος κάθε στατιστικού δείκτη αναλογικά με τη συχνότητα κάθε κατηγορίας).

Κεφάλαιο 2

Εφαρμογή Μοντέλων Αναγνώρισης Δράσης

Στο κεφάλαιο αυτό περιγράφεται η μεθοδολογία και τα αποτελέσματα των τριών βασικών μοντέλων εκτίμησης της Δοκιμασίας Εξαναγκασμένης Κολύμβησης. Σκοπός ήταν να προκύψει μια αρχική εκτίμηση της αξιοπιστίας του κάθε μοντέλου, ώστε με βάση αυτή να υλοποιηθούν περαιτέρω διορθώσεις και βελτιστοποιήσεις. Οι τεχνικές που εφαρμόστηκαν είναι:

- Ο Αλγόριθμος Πυκνών Τροχιών
- Αρχιτεκτονική Νευρωνικών Δικτύων CNN - LSTM
- Αρχιτεκτονική Inflated 3D

Για την διαδικασία σύγκρισης των μοντέλων επιλέχθηκαν οι εξής παράμετροι:

- Δείγματα Dataset : Αυτά για τα οποία οι παρατηρητές έρχονται σε συμφωνία, καθώς θεωρούνται πιο αξιόπιστα
- Μέγεθος Εικόνων Βίντεο: 200x100
- Frames / sec: 12.5
- Χρονικό Διάστημα Δειγμάτων: 1.33 sec (16 frames)

2.1 Εφαρμογή Αλγορίθμου Πυκνών Τροχιών

Ο αλγόριθμος πυκνών τροχιών [8] αποτελεί μία απ᾽ τις πιο πετυχημένες τεχνικές ταξινόμησης βίντεο με κατασκευασμένα χαρακτηριστικά. Για το λόγο αυτό επιλέχτηκε για την εφαρμογή του στο συγκεκριμένο πρόβλημα. Η χρήση του γίνεται για λόγους επιστημονικής εμβάθυνσης και πληρότητας. Οι παράμετροι του δεν βελτιστοποιήθηκαν για τις ανάγκες της συγκεκριμένης διπλωματικής εργασίας, όπως έγινε στην περίπτωση των μοντέλων βαθιάς μάθησης.

2.1.1 Μεθοδολογία

Η λειτουργία του αλγορίθμου περιγράφεται στο Θεωρητικό Υπόβαθρο στην Ενότητα 6.1. Για την εξαγωγή χαρακτηριστικών χρησιμοποιήθηκε η επίσημη υλοποίηση

του αλγορίθμου Improved Dense Trajectories στη γλώσσα C++. Ακόμη επιλέχτηκαν οι προτεινόμενες παράμετροι μήκος τροχιάς $L = 15$ frames, βήμα δειγματοληψίας $W = 5$ pixels, μέγεθος γειτονιάς $= 32$ pixels, χωρικά κελιά $n_{xy} = 2$ cells και χρονικά κελιά $n_t = 3$ cells. Συνολικά εξάγει 5 διαφορετικούς περιγραφείς: περιγραφέας τροχιάς, HoG, HoF, MBHx, MBHy. Για τη λειτουργία του αλγορίθμου χρειάστηκε η διαδικασία μεταγλώττισης της βιβλιοθήκης OpenCV, με υποστήριξη για τις βιβλιοθήκες CUDA.

Πρώτο βήμα για την υλοποίηση του αλγορίθμου, αποτελεί η εκτίμηση Gaussian Mixture Model (GMM). Για την εκτίμηση του χρειάζεται η εξαγωγή χαρακτηριστικών για μερικά δείγματα του dataset. Για την εκτίμηση του GMM χρησιμοποιήθηκε η βιβλιοθήκη της python, yael. Η διαδικασία πραγματοποιήθηκε για GMM με γκαουσιανές $k = 128$. Επιλέχτηκε αριθμός δειγμάτων τέτοιος ώστε να προκύψουν χαρακτηριστικά της τάξης του $1000 * k = 128000$, όπως προτείνει η βιβλιοθήκη.

Επόμενο βήμα ήταν η κωδικοποίηση όλων των χαρακτηριστικών των βίντεο, με Fisher Vectors. Και σε αυτή τη περίπτωση χρησιμοποιήθηκε η υλοποίηση της βιβλιοθήκης yael, με βάση το GMM που εκτιμήθηκε.

Τέλος, τα εξαγώμενα κωδικοποιημένα χαρακτηριστικά πρέπει να ταξινομηθούν. Για το λόγο αυτό έγινε ταξινόμηση με Support Vector Machines, με γραμμικό πυρήνα, συνάρτηση κόστους hinge, με την υλοποίηση OneVsRestClassifier κατάλληλη για ταξινόμηση σε προβλήματα πολλών κατηγοριών, της βιβλιοθήκης scikit-learn της python.

Για την υλοποίηση του συνόλου της διαδικασίας εξαγωγής, κωδικοποίησης και ταξινόμησης χαρακτηριστικών, προσαρμόστηκε σχετικός κώδικας [44], που βρέθηκε στη σελίδα github.com.

2.1.2 Αποτελέσματα

Μετά την δημιουργία του μοντέλου ταξινόμησης, πραγματοποιήθηκαν εκτιμήσεις στο υποσύνολο test του dataset. Με τη χρήση των εκτιμήσεων και των πραγματικών τιμών των δειγμάτων, εξήχθησαν στατιστικά στοιχεία τα οποία παρουσιάζονται στους παρακάτω πίνακες 2.1, 2.2.

Confusion Matrix		**Predicted**				
		Immobility	**Swimming**	**Climbing**	**Head Sh.**	**Diving**
True	**Immobility**	**1749**	86	91	24	0
	Swimming	**281**	222	155	12	0
	Climbing	95	41	**1750**	3	0
	Head Shake	70	14	44	**64**	0
	Diving	0	1	**2**	0	1

Πίνακας 2.1: Πίνακας Σύγχυσης των αποτελεσμάτων του αλγορίθμου Improved Dense Trajectories

Όπως φαίνεται στον πίνακα σύγχυσης, βασικό πρόβλημα αποτελεί η κατηγορία swimming, μιας και τα δείγματα που την αντιπροσωπεύουν ταξινομούνται περισσότερο ως Immobility και λιγότερο ως την πραγματική τους κατηγορία. Μεγάλο μέρος των δειγμάτων ταξινομείται και ως Climbing. Εκτιμάται ότι το πρόβλημα αυτό δημιουργείται λόγω της ανισσοροπίας μεταξύ των κλάσεων, μιάς και υπάρχουν 3 φορές λιγότερα δείγματα Swimming σε σχέση με τις άλλες 2 κυρίαρχες κατηγορίες. Ωστόσο, το πρόβλημα του διαχωρισμού της κατηγορίας Swimming είναι από τη φύση του δύσκολο, μιας και αποτελεί ενδιάμεσο στάδιο των κατηγοριών Immobility και Climbing,

με τα διαχωριστικά όρια να είναι πολύ λεπτά σε πολλές περιπτώσεις.

	Precision	**Recall**	**F1-Score**	**Support**
Class Statistics				
Immobility	0.80	0.90	**0.84**	1950
Swimming	0.61	0.33	0.43	670
Climbing	0.86	0.93	**0.89**	1889
Head Shake	0.62	0.33	0.43	192
Diving	1.00	0.25	0.40	4
Overall Statistics				
Accuracy		0.80		4705
Macro Average	0.78	0.55	0.60	4705
Weighted Average	0.79	0.80	0.79	4705

Πίνακας 2.2: Στατιστικά στοιχεία των αποτελεσμάτων του μοντέλου Πυκνών Τροχιών

Όπως προκύπτει απ᾽ τα στατιστικά στοιχεία του πίνακα 2.2, φαίνεται το μοντέλο εκτίμησης να αναγνωρίζει ικανοποιητικά τις 2 κυρίαρχες κατηγορίες Climbing και Immobility, και να δυσκολεύεται ιδιαίτερα στην αναγνώριση των υπολοίπων δράσεων. Αυτό αποτυπώνεται στο recall οπού οι τιμές των υπόλοιπων κατηγοριών είναι μικρότερες του 35%. Η κατηγορία Diving, αν και ιδιαίτερα χαρακτηριστική σε σχέση με τις υπόλοιπες, δεν καταφέρνει να αναγνωριστεί, μάλλον λόγω των ελλιπών δειγμάτων εκπαίδευσης.

Σε γενικές γραμμές το μοντέλο εκτίμησης του αλγορίθμου Improved Dense Trajectories, φαίνεται να πάσχει λόγω της ανισορροπίας των 5 κλάσεων, χαρακτηρίζοντας το μεγαλύτερο μέρος των περισσότερων κατηγοριών ως Immobility και Climbing.

2.2 Εφαρμογή μοντέλων Νευρωνικών Δικτύων

Στο κεφάλαιο αυτό παρουσιάζονται αναλυτικά οι αρχιτεκτονικές νευρωνικών δικτύων που σχεδιάστηκαν ή εφαρμόστηκαν για την πρόβλεψη της συμπεριφοράς των επιμυών. Τα μοντέλα υλοποιήθηκαν στην python με τη χρήση της βιβλιοθήκης υψηλού επιπέδου keras [45] που αξιοποιεί τη βιβλιοθήκη χαμηλού επιπέδου tensorflow [46]. Αρχικά σχεδιάστηκε μοντέλο CNN - LSTM και στη συνέχεια εφαρμόστηκε η σύνθετη αρχιτεκτονική Inflated 3D, με προεκπαιδευμένα βάρη.

Για την εκπαίδευση των μοντέλων των νευρωνικών δικτύων, σχεδιάστηκε και υλοποιήθηκε αλγόριθμος στην γλώσσα python για την παραμετροποίηση των στοιχείων που έχουν αναφερθεί στην ενότητα 1.3. Η εκπαίδευση των μοντέλων πραγματοποιήθηκε σε 2 μηχανήματα με κάρτες γραφικών Nvidia 960 και Nvidia 1050 Ti. Η χρήση GPU για την εκπαίδευση μοντέλων βαθιάς μάθησης επιταχύνει σημαντικά την εκπαίδευση και πρόβλεψη των μοντέλων σε σχέση με την επεξεργαστική μονάδα CPU ενός υπολογιστή.

Πρώτα βήμα είναι η κλήση συνάρτησης, με τις παραμέτρους που ζητούνται στο εκάστοτε πείραμα, για την δημιουργία πινάκων με γραμμές τα δείγματα της εκπαίδευσης και στήλες τις θέσεις που βρίσκονται τα αρχεία για το κάθε δείγμα, τα ονόματα των εικόνων που πρόκειται να φορτωθούν και να σχηματίσουν το κάθε δείγμα, την αληθή τιμή του δείγματος καθώς και την τιμή βάρους για το συγκεκριμένο δείγμα. Η επιστροφή της συνάρτησης αυτής καθορίζεται από τις παραμέτρους που ζητούνται όπως τα frames / sec, την επιλογή παρατηρητή, το χρονικό διάστημα που θα

απαρτίζει κάθε δείγμα κλπ. Στη συνέχεια υλοποιήθηκε κλάση τροφοδότη (generator) συμβατού με τη βιβλιοθήκη keras. Η ανάγκη για την δημιουργία generator είναι εμφανής, καθώς τα δείγματα δεν μπορούν να φορτωθούν εξ' αρχής στη μνήμη της GPU, μιας και το υπολογιστικό κόστος θα καθιστούσε αδύνατη τη διαδικασία της εκπαίδευσης. Επιπλέον ο generator δίνει την δυνατότητα προσθήκης augmentation με την ύπαρξη τυχαιότητας στον σχηματισμό κάθε δείγματος. Ο ρόλος του generator είναι να τροφοδοτεί το δίκτυο με τις εισόδους, τις εξόδους και τα βάρη των δειγμάτων ενός batch, κάθε φορά που καλείται από την βασική συνάρτηση εκπαίδευσης του μοντέλου. Έτσι η CPU αναλαμβάνει την κατασκευή των batches, και η GPU την εκπαίδευση του μοντέλου, σε παράλληλο χρόνο.

Κάθε εποχή την εκπαίδευσης αφορά την τροφοδότηση του μοντέλου με όλα τα δείγματα. Στο τέλος κάθε εποχής τα δείγματα ανακατεύονται με τυχαία σειρά, ώστε να εξασφαλιστεί η μη αναγνώριση συστηματικών προτύπων που αφορούν τη σειρά των δειγμάτων. Το μέγεθος του batch αναγκαστικά επιλέχτηκε από 4 έως 8, καθώς οι κάρτες γραφικών που χρησιμοποιήθηκαν, με μνήμη 4GB, δεν ήταν δυνατόν να πραγματοποιήσουν εκπαίδευση μοντέλων αναγνώρισης δράσης με μεγαλύτερα μεγέθη batch. Ο optimizer που επιλέχτηκε, είναι ο αλγόριθμος Adam. Ακόμη ως συνάρτηση κόστους χρησιμοποιήθηκε η categorical cross entropy. Στα αρχικά πειράματα χρησιμοποιούνται οι εικόνες Optical Flow για την είσοδο των μοντέλων, μιας και περιλαμβάνουν εξ' αρχής πληροφορίες για την κίνηση των επιμυών. Κριτήριο για την λήξη του κάθε πειράματος αποτέλεσε η τιμή loss του υποσυνόλου test, αλλά και προσαρμόστηκε σε ορισμένες περιπτώσεις με βάση την εξέλιξη των στατιστικών accuracy του υποσυνόλου train και test. Το πείραμα σταματούσε όταν η τιμή loss του υποσυνόλου test, δεν παρουσίασε βελτίωση για ικανό αριθμό εποχών ώστε να εξασφαλίζεται η πλήρης εκπαίδευση κάθε μοντέλου. Κάθε πείραμα διήρκησε από μισή έως 2 μέρες μέχρι την ολοκλήρωση της εκπαίδευσης. Για κάθε εποχή της εκπαίδευσης, αποθηκεύτηκαν τα τρέχοντα βάρη του μοντέλου και τα στατιστικά στοιχεία των αποτελεσμάτων, ώστε να μην υπάρχει ανάγκη συνεχούς επίβλεψης της εκπαίδευσης και παράλληλα να είναι δυνατή η επιλογή οποιουδήποτε μοντέλου κριθεί ως βέλτιστο, ακόμη και στην περίπτωση της υπερπροσαρμογής (overfit).

Το υποσύνολο test, κατά τη διαδικασία του ελέγχου των εκτιμήσεων στο τέλος κάθε εποχής, καθορίστηκε να έχει επικάλυψη μεταξύ των δειγμάτων, σε επίπεδο χρόνου, τέτοιο ώστε να σχηματίζεται ένα δείγμα κάθε 0.5 sec.

2.2.1 Αρχιτεκτονική CNN - LSTM

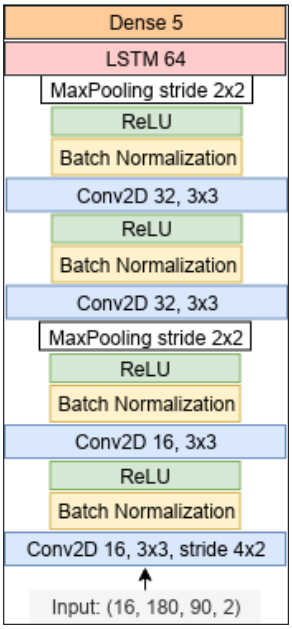
2.2.1.1 Μεθοδολογία

Αρχικά σχεδιάστηκε αρχιτεκτονική με συνδυασμό συνελικτικών και ανατροφοδοτούμενων δικτύων για την αναγνώριση δράσης. Τα ανατροφοδοτούμενα δίκτυα επιλέχθηκαν, καθώς η αρχιτεκτονική τους διατηρεί μια μορφή μνήμης, την οποία χρησιμοποιεί, ώστε να συσχετίσει τα παρελθοντικά με το παρόν χρονικό βήμα. Για αυτό το λόγο, τα δίκτυα αυτά επιλέγονται για την αναπαράσταση του βίντεο. Θα χρησιμοποιηθεί η αρχιτεκτονική LSTM, καθώς διακρίνεται στη μάθηση μακροχρόνιων εξαρτήσεων.

Στην περίπτωση του βίντεο, τα αρχικά layers καθορίστηκαν ως συνελικτικά ώστε να εξάγουν χωρικά χαρακτηριστικά, ανεξάρτητα για κάθε χρονικό βήμα. Έπειτα η έξοδος του συνελικτικού δικτύου τροφοδοτεί τα LSTM, για την συσχέτιση των χαρακτηριστικών στην διάσταση του χρόνου. Είναι δυνατή τόσο η ταξινόμηση κάθε

στιγμιότυπου ενός βίντεο με εξαγωγή της αναπαράστασης κάθε χρονικού βήματος των LSTM (sequence to sequence) όσο και η ταξινόμηση ολόκληρου του βίντεο με εξαγωγή της αναπαράστασης του τελευταίου χρονικού βήματος για κάθε δείγμα. Επιλέχθηκε η δεύτερη τεχνική, για να καθίσταται δυνατή η σύγκριση με τα υπόλοιπα μοντέλα.

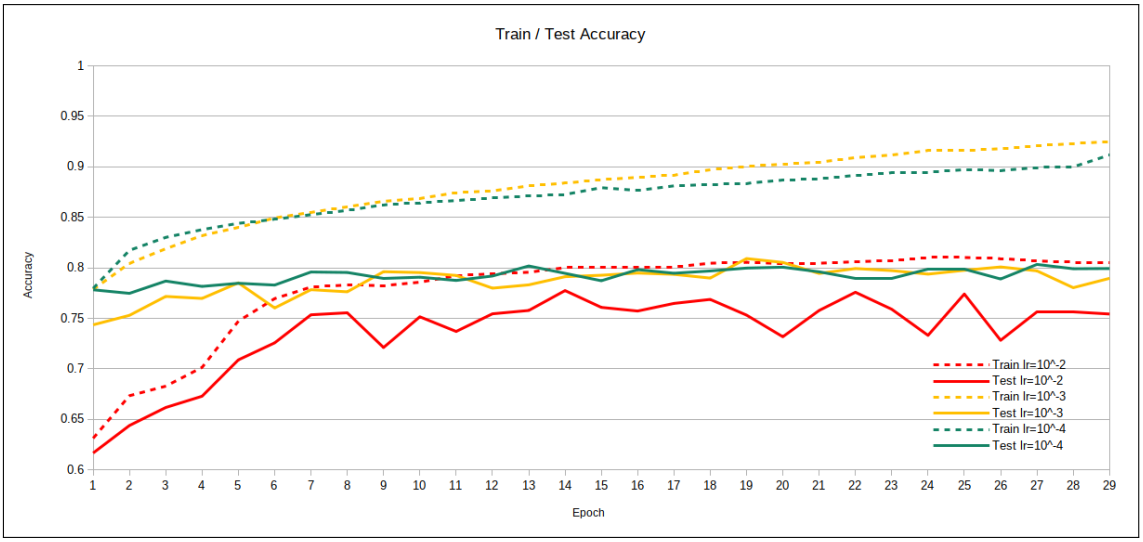
Όσον αφορά τον σχεδιασμό της αρχιτεκτονικής, το πρώτο μέρος που αφορά την εξαγωγή χαρακτηριστικών των εικόνων, αποτελείται από δύο Convolutional Layers 16 φίλτρων, με Batch Normalization, την συνάρτηση ReLU ως συνάρτηση ενεργοποίησης, και ένα Max Pooling Layer. Στη συνέχεια, επαναλαμβάνεται η ίδια σύσταση Layers με 32 συνελικτικά φίλτρα. Τέλος για την χρονική συσχέτιση των χαρακτηριστικών που εξήχθησαν, χρησιμοποιείται ένα LSTM Layer. Ακολουθεί η ταξινόμηση της εξόδου του τελευταίου χρονικού βήματος του LSTM, με ένα Dense Layer 5 νευρώνων με τη συνάρτηση ενεργοποίησης Softmax. Αναλυτικά, η αρχιτεκτονική παρουσιάζεται στο σχήμα 2.1



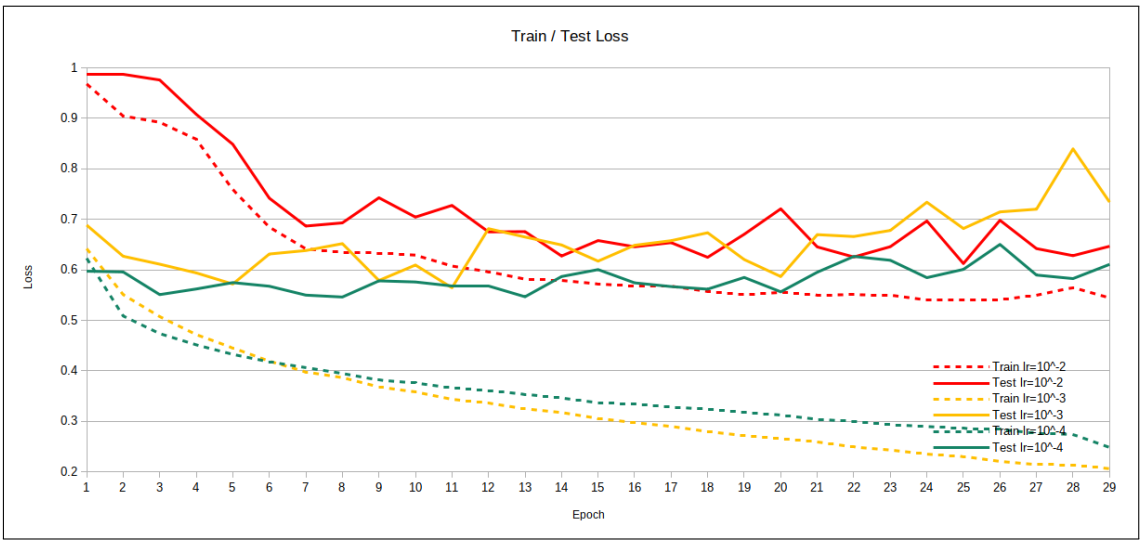
Σχήμα 2.1: Διαγραμματική Αναπαράσταση αρχιτεκτονικής CNN - LSTM

Για την εξασφάλιση μιας πετυχημένης εκπαίδευσης, πραγματοποιήθηκαν δοκιμές για τον ρυθμό εκμάθησης (learning rate), του αλγορίθμου βελτιστοποίησης Adam. Συγκεκριμένα, δοκιμάστηκαν οι τιμές 10^{-2}, 10^{-3}, 10^{-4}, και τα αποτελέσματα του accuracy και του loss, στα υποσύνολα train και test, ανά εποχή, παρουσιάζονται στα διαγράμματα, 2.2, 2.3,

Παρατηρήθηκε ότι η εκπαίδευση με learning rate = 10^{-2} δυσχεραίνεται , καθώς το accuracy του υποσυνόλου train, σταματάει να σημειώνει βελτίωση, μετά την πέμπτη εποχή, που φτάνει στην τιμή 80%. Αντιθέτως, για τις δύο μικρότερες τιμές learning rate, το accuracy φτάνει τιμές άνω του 90%. Τελικά θα επιλεχτεί η τιμή $10 - -4$ του learning rate, καθώς επιτυγχάνει σταθερά το μικρότερο loss για το υποσύνολο test, αλλά και ταχύτερα. Όπως είναι αναμενόμενο, μετά από ένα σημείο, ενώ η τιμή του train accuracy αυξάνεται, το test accuracy μένει σταθερό ή και μειώνεται. Αυτό συμβαίνει γιατί το μοντέλο αρχίζει να μαθαίνει χαρακτηριστικά των δεδομένων μάθησης, όπως ο θόρυβος και το παρασκήνιο, τα οποία δεν αποτελούν χρήσιμα χαρακτηριστικά για την ταξινόμηση του συγκεκριμένου προβλήματος. Για το λόγο αυτό



Σχήμα 2.2: Διάγραμμα accuracy ανά, εποχή της αρχιτεκτονικής CNN - LSTM



Σχήμα 2.3: Διάγραμμα loss ανά εποχή, της αρχιτεκτονικής CNN - LSTM

επιλέχθηκε η 13η εποχή, καθώς έχει το μικρότερο test loss και το μεγαλύτερο test accuracy της εκπαίδευσης, προτού το μοντέλο κάνει υπερπροσαρμογή.

2.2.1.2 Αποτελέσματα

Εφόσον επιλέχτηκε το κατάλληλο learning rate, και η βέλτιστη εποχή της εκπαίδευσης, στη συνέχεια παρουσιάζονται τα αποτελέσματα που προέκυψαν, στους πίνακες 2.3, 2.4.

Confusion Matrix		**Predicted**				
		Immobility	**Swimming**	**Climbing**	**Head Sh.**	**Diving**
True	**Immobility**	**3432**	122	210	54	0
	Swimming	**567**	532	167	23	0
	Climbing	262	178	**3218**	12	0
	Head Shake	127	40	43	**149**	0
	Diving	0	0	**5**	2	1

Πίνακας 2.3: Πίνακας Σύγχυσης των αποτελεσμάτων της αρχιτεκτονικής CNN - LSTM

	Precision	**Recall**	**F1-Score**	**Support**
Class Statistics				
Immobility	0.78	0.90	**0.84**	3818
Swimming	0.61	0.41	0.49	1289
Climbing	0.88	0.88	**0.88**	3670
Head Shake	0.62	0.42	0.50	359
Diving	1.00	0.25	0.40	8
Overall Statistics				
Accuracy		0.80		9144
Macro Average	0.78	0.55	0.59	9144
Weighted Average	0.79	0.80	0.79	9144

Πίνακας 2.4: Στατιστικά στοιχεία των αποτελεσμάτων της αρχιτεκτονικής CNN - LSTM

Όπως προκύπτει, το συνολικό accuracy του μοντέλου είναι 80%, όπως και στην περίπτωση του αλγορίθμου Πυκνών Τροχιών. Υπάρχει βελτίωση 9% ως προς το recall του Swimming, που αποτελεί το βασικότερο πρόβλημα του μοντέλου, λόγω της σημασίας της κατηγορίας. Ακόμη βελτίωση της τάξης του 9% στο recall παρουσιάζει και η κατηγορία Head Shake. Η κατηγορία Diving πετυχαίνει και πάλι recall 25%. Συνολικά παρουσιάζεται μια μικρή βελτίωση σε σχέση με το μοντέλο των πυκνών τροχιών, ωστόσο οι κατηγορίες με ελλιπή δείγματα και σε αυτήν την περίπτωση έχουν χαμηλό recall της τάξης του 40%.

2.2.2 Αρχιτεκτονική Inflated 3D

2.2.2.1 Μεθοδολογία

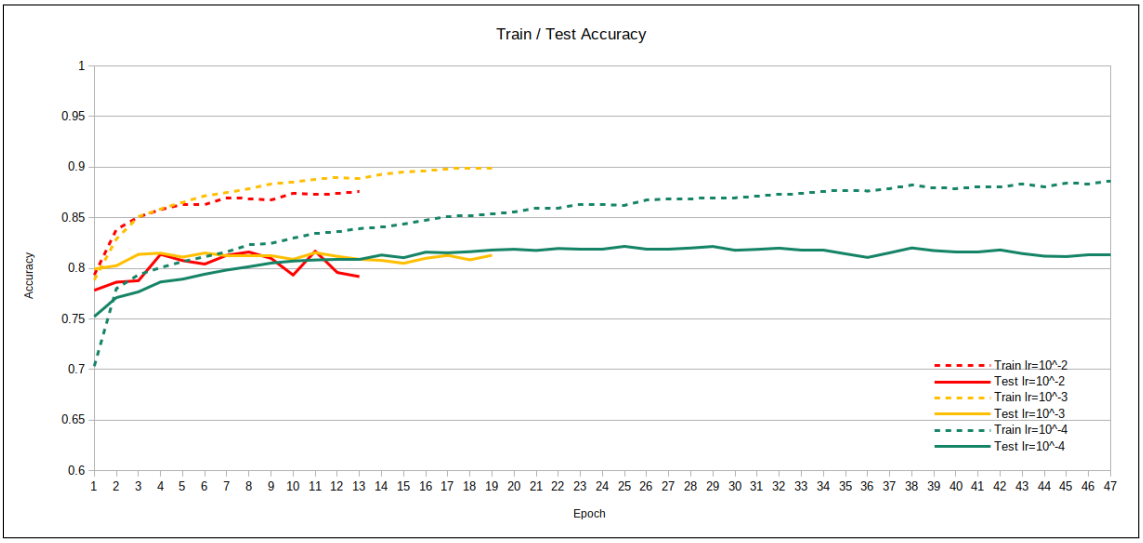
Η αρχιτεκτονική Inflated 3D [13], αποτελεί μία αρχιτεκτονική βαθιάς μάθησης, για την αναγνώριση δράσης. Θεμέλιο της αρχιτεκτονικής, αποτελεί η αξιοποίηση της ραγδαίας εξέλιξης και επιτυχίας των αρχιτεκτονικών ταξινόμησης εικόνων, όπως η αρχιτεκτονική Inception V1 (GoogleNet) [3]. Πρόκειται για γενίκευση της αρχιτεκτονικής Inception V1 στο τρισδιάστατο χωρο-χρονικό πεδίο, χρήζοντας την κατάλληλη για αναγνώριση δράσης. Η λειτουργία της περιγράφεται στο Θεωρητικό Υπόβαθρο στην Ενότητα 6.2.2.

Η αρχιτεκτονική αυτή επιλέχτηκε λόγω της αξιοπιστίας της, την δυνατότητα της για εξαγωγή ποικίλων χαρακτηριστικών διαφορετικής κλίμακας και την ύπαρξη προεκπαιδευμένων βαρών στα dataset Kinetics[41] και ImageNet [1], σε εικόνες RGB και Optical Flow. Η χρήση προεκπαιδευμένων βαρών δεν επιτρέπει σημαντικές αλλαγές στην αρχιτεκτοκινή, παρά μόνο στις παραμέτρους του dataset.

Χρησιμοποιήθηκε υλοποίηση του μοντέλου για την βιβλιοθήκη keras [47], περιλαμβανομένων των προεκπαιδευμένων βαρών. Η χρήση προεκπαιδευμένων βαρών προσφέρει υπάρχουσα γνώση αναγνώρισης δράσης στο μοντέλο, καθώς και ταχύτερη εκπαίδευση και πρόβλεψη. Επιπλέον, μια αρχιτεκτονική τέτοιου βάθους, δεν θα ήταν δυνατό να χρησιμοποιηθεί χωρίς την ύπαρξη προεκπαιδευμένων βαρών, λόγω του υπολογιστικού κόστους. Χάρη στην ύπαρξη βαρών, με τον καθορισμό τους σε συγκεκριμένα layers ως μη εκπαιδευόμενα, η χρήση μιας τέτοιας αρχιτεκτονικής τέτοιου βάθους γίνεται εφικτή.

Η αρχιτεκτονική Inflated 3D, χρησιμοποιήθηκε όπως ακριβώς περιγράφεται από τους δημιουργούς της, μιας και οποιαδήποτε αλλαγή θα καθιστούσε αδύνατη τη χρήση των προεκπαιδευμένων βαρών. Η διαγραμματική της αναπαράσταση παρουσιάζεται στο παράρτημα Β΄. Τα βάρη σε όλη την αρχιτεκτονική μέχρι και το τελευταίο 3D Inception Module, καθορίστηκαν ως μη εκπαιδευόμενα. Μοναδική διαφοροποίηση σε σχέση με την αρχιτεκτονική των δημιουργών ήταν η προσθήκη Dropout απόρριψης του 60% των νευρώνων, μετά το 3D Global Average Pooling layer. Απ᾽ αυτό το layer και έπειτα, τα βάρη καθορίστηκαν ως εκπαιδευόμενα.

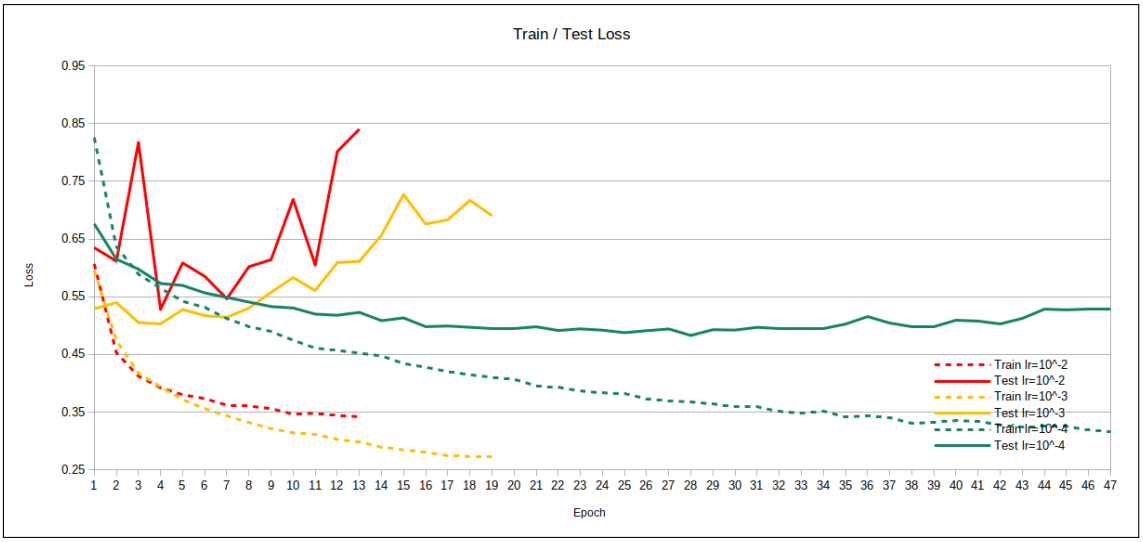
Πραγματοποιήθηκαν τρία πειράματα με διαφορετικό learning rate, του αλγορίθμου βελτιστοποίησης Adam. Τα διαγράμματα για τα υποσύνολα test / train, για τους δείκτες accuracy, loss, παρουσιάζονται στις εικόνες 2.4, 2.5.



Σχήμα 2.4: Διάγραμμα accuracy ανά, εποχή της αρχιτεκτονικής Inflated 3D

Όπως και στην περίπτωση της αρχιτεκτονικής CNN - LSTM, έτσι και εδώ, το learning rate 10^{-2} κρίθηκε ακατάλληλο, καθώς, το train accuracy δεν μπορεί να ξεπεράσει το 82%. Τα δυο υπόλοιπα πειράματα δίνουν αντίστοιχα αποτελέσματα. Επιλέχθηκε ως learning rate η τιμή 10^{-3}, καθώς η διάρκεια εκπαίδευσης είναι πολύ συντομότερη σε σχέση με την τιμή 10^{-4}, ενώ πετυχαίνουν το ίδιο αποτέλεσμα Accuracy στο υποσύνολο test.

Ως βέλτιστη εποχή του πειράματος με learning rate = 10^{-3} θα επιλεχτεί η $5η$



Σχήμα 2.5: Διάγραμμα loss ανά εποχή, της αρχιτεκτονικής Inflated 3D

καθώς στη συνέχεια υπήρξε αύξηση του train accuracy με ταυτόχρονη μείωση του test accuracy.

2.2.2.2 Αποτελέσματα

Τα αποτελέσματα για το learning rate και τη βέλτιστη εποχή της εκπαίδευσης που επιλέχτηκαν, παρουσιάζονται στους πίνακες 2.5, 2.6.

Confusion Matrix		**Predicted**				
		Immobility	**Swimming**	**Climbing**	**Head Sh.**	**Diving**
True	**Immobility**	**3218**	203	324	67	0
	Swimming	354	**626**	291	27	0
	Climbing	101	105	**3461**	21	2
	Head Shake	62	34	55	**210**	0
	Diving	0	0	3	0	**4**

Πίνακας 2.5: Πίνακας Σύγχυσης των αποτελεσμάτων της αρχιτεκτονικής Inflated 3D

	Precision	**Recall**	**F1-Score**	**Support**
Class Statistics				
Immobility	0.86	0.84	**0.85**	3812
Swimming	0.65	0.48	0.55	1298
Climbing	0.84	0.94	**0.88**	3690
Head Shake	0.65	0.58	**0.61**	361
Diving	0.67	0.57	**0.62**	7
Overall Statistics				
Accuracy		0.82		9168
Macro Average	0.73	0.68	0.70	9168
Weighted Average	0.81	0.82	0.81	9168

Πίνακας 2.6: Στατιστικά στοιχεία των αποτελεσμάτων της αρχιτεκτονικής Inflated 3D

Παρατηρείται σαφής βελτίωση σε σχέση με τα προηγούμενα μοντέλα εκτίμησης. Το accuracy του μοντέλου βρίσκεται στο 82%, 2% πάνω απ' τα δυο προηγούμενα μοντέλα. Οι δυο κυρίαρχες κατηγορίες έχουν ικανοποιητικό recall και accuracy. Η κατηγορία swimming επιτυγχάνει recall 48%, ενώ οι κατηγορίες Head Shake και Diving, σημειώνουν recall άνω του 55%.

2.3 Σχολιασμός Αποτελεσμάτων

Εφόσον πραγματοποιήθηκαν τα αρχικά πειράματα για τα τρία μοντέλα εκτίμησης της Δοκιμασίας Εξαναγκασμένης Κολύμβησης, θα γίνει σχολιασμός και σύγκριση των αποτελεσμάτων. Από πλευράς συνολικού Accuracy των μοντέλων, η αρχιτεκτονική CNN - LSTM και ο αλγόριθμος Πυκνών Τροχιών πέτυχαν 80%, ενώ η αρχιτεκτονική Inflated 3D 82%, όπως παρουσιάζεται στον πίνακα 2.7 Η απόκλιση του Accuracy δεν κρίνεται σημαντική μεταξύ των τριών μοντέλων.

	Model			
Accuracy	Dense Trajectories	0.80		
	CNN – LSTM	0.80		
	Inflated 3D	0.82		
		Precision (%)	Recall (%)	F1 - Score (%)
Macro Average	Dense Trajectories	**0.78**	0.55	0.60
	CNN – LSTM	**0.78**	0.55	0.59
	Inflated 3D	0.73	**0.68**	**0.70**
Weighted Average	Dense Trajectories	0.79	0.80	0.79
	CNN – LSTM	0.79	0.80	0.79
	Inflated 3D	**0.81**	**0.82**	**0.81**

Πίνακας 2.7: Πίνακας σύγκρισης των συνολικών στατιστικών των τριών μοντέλων

Class	**Model**	Precision (%)	Recall (%)	F1 - Score (%)
Immobility	Dense Trajectories	0.80	0.90	0.84
	CNN – LSTM	0.78	0.90	0.84
	Inflated 3D	0.86	0.84	**0.85**
Swimming	Dense Trajectories	0.61	0.33	0.43
	CNN – LSTM	0.61	0.41	0.49
	Inflated 3D	0.65	0.48	**0.55**
Climbing	Dense Trajectories	0.86	0.93	**0.89**
	CNN – LSTM	0.88	0.88	0.88
	Inflated 3D	0.84	0.94	0.88
Head Shake	Dense Trajectories	0.62	0.33	0.43
	CNN – LSTM	0.62	0.42	0.50
	Inflated 3D	0.65	0.58	**0.61**
Diving	Dense Trajectories	0.100	0.25	0.43
	CNN – LSTM	0.100	0.25	0.40
	Inflated 3D	0.67	0.57	**0.62**

Πίνακας 2.8: Πίνακας σύγκρισης των στατιστικών των τριών μοντέλων ανά κατηγορία

Σε επίπεδο κατηγοριών, όπως φαίνεται στον πίνακα 2.8, σε όλες τις περιπτώσεις υπήρξε ικανοποιητική ταξινόμηση των δυο κυριάρχων κατηγοριών, Immobility και Climbing, που αποτελούν το 80% του dataset, πετυχαίνοντας Precision και Recall άνω του 85%. Η κατηγορία Swimming πέτυχε 33% με χαρακτηριστικά του αλγορίθμου Πυκνών Τροχιών, 41% με την αρχιτεκτονική CNN - LSTM και 49% με την αρχιτεκτονική Inflated 3D. Όπως προέκυψε απ᾽ τους πίνακες σύγχυσης, σε όλες τις περιπτώσεις ένα σημαντικό μέρος των δειγμάτων της κατηγορίας Swimming, ταξινομήθηκε ως Immobility, και ένα μικρότερο ως Climbing. Τον βέλτιστο διαχωρισμό

μεταξύ των 3 βασικών κατηγοριών πέτυχε το μοντέλο Inflated 3D. Αντίστοιχα και για τις κατηγορίες Head Shake και Diving, η αρχιτεκτονική Inflated 3D αναγνώρισε σωστά το 60% των δειγμάτων τους, τη στιγμή που τα υπόλοιπα 2 μοντέλα σημείωναν Recall κάτω του 45%.

Για τους παραπάνω λόγους, ως καταλληλότερη αρχιτεκτονική επιλέχθηκε η Inflated 3D. Το βάθος της, η ύπαρξη προεκπαιδευμένων βαρών και η εξαγωγή χαρακτηριστικών διαφορετικού μεγέθους ενδέχεται να ευθύνονται για την επιτυχία της σε σχέση με τα υπόλοιπα μοντέλα. Ωστόσο οι αποκλίσεις του Accuracy της τάξης του 2%, τριών δομικά απολύτως διαφορετικών μοντέλων, κρίνεται πολύ μικρή.

Όπως αναφέρεται στο κεφάλαιο 1.1, οι παρατηρητές συμφωνούν κατά 77%. Για την καλύτερη αξιολόγηση των αποτελεσμάτων, το μοντέλο Inflated 3D που προέκυψε, χρησιμοποιήθηκε για την πρόβλεψη όλων των δειγμάτων του dataset, και όχι μόνο στα δείγματα για τα οποία οι δυο παρατηρητές βρίσκονται σε συμφωνία. Έπειτα, έγινε σύγκριση αυτών των προβλέψεων με τον πρώτο παρατηρητή. Με τον τρόπο αυτό, έγινε εφικτή η σύγκριση των συγχύσεων των δύο παρατηρητών, σε σχέση με τις συγχύσεις ενός παρατηρητή και του μοντέλου. Τα αποτελέσματα παρουσιάζονται στον πίνακα 2.9.

	Παρατ. 1 - Παρατ. 2	Παρατ. 1 - Μοντέλο I3D
Συμφωνία Προβλέψεων	77%	73%

Κατηγορίες Σύγχυσης	Παρατ.1 - Παρατ.2	Παρατ. 1 - Μοντέλο I3D
Immobility με Swimming	40%	40%
Climbing με Swimming	24%	27%
Immobility με Climbing	14%	21%
Λοιπές Αστοχίες	22%	15%

Πίνακας 2.9: Σύγκριση συμφωνίας - σύγχυσης παρατηρητών και μοντέλου I3D

Όπως προκύπτει, οι αστοχίες του μοντέλου, παρουσιάζουν αρκετή αρμονία σε σχέση με τις συγχύσεις των παρατηρητών. Επομένως, η αδυναμία των τριών μοντέλων να ξεπεράσουν τη δεδομένη ευστοχία, φαίνεται να οφείλεται στη δυσκολία της ταξινόμησης της Δοκιμασίας Εξαναγκασμένης Κολύμβησης.

2.4 Έλεγχοι - Δοκιμές

Στην προσπάθεια εύρεσης αδυναμιών των μοντέλων, πραγματοποιήθηκε έλεγχος για τη δυνατότητα γενίκευσης του μοντέλου σε διαφορετικούς επιμύες και παρασκήνια. Όπως έχει ήδη αναφερθεί, για την δημιουργία αντικειμενικού υποσυνόλου test, τα δείγματα χωρίστηκαν με βάση τα πειράματα απ᾽ τα οποία προέρχονται, εξασφαλίζοντας ότι δεν υπήρξαν κοινοί επιμύες και πειράματα και στα δύο υποσύνολα ταυτοχρόνως. Για τον έλεγχο της ικανότητας γενίκευσης, το dataset αυτή τη φορά διαχωρίστηκε απολύτως τυχαία, επιλέγοντας το 20% των δειγμάτων ως test, ανεξαρτήτως του πειράματος προέλευσης. Ο έλεγχος έγινε με την αρχιτεκτονική Inflated 3D και προέκυψε Accuracy 84%. Η αύξηση κατά 2% σε σχέση με το αρχικό μοντέλο κρίθηκε ότι δεν υποδηλώνει αδυναμία γενίκευσης λόγω διαφοροποιήσεων των επιμυών και της θέσης της κάμερας κάθε πειράματος. Επομένως, θεωρείται ότι το μοντέλο εκτίμησης έχει ήδη καταφέρει σε μεγάλο βαθμό να γενικεύσει τις κατηγορίες ενδιαφέροντος, ανεξαρτήτως του πειράματος και των διαφοροποιήσεων που προκαλούν.

Εκτιμάται ότι τα στατιστικά αυτά δεν μπορούν να σημειώσουν σημαντική βελτίωση για το δεδομένο dataset, ανεξαρτήτως μοντέλου. Όπως έχει ήδη αναφερθεί, ο διαχωρισμός μεταξύ των κατηγοριών στο συγκεκριμένο πρόβλημα, σε ορισμένες περιπτώσεις είναι αρκετά αβέβαιος, γι᾽ αυτό και οι δυο ειδικοί παρατηρητές έρχονται σε συμφωνία για το 77% των συνολικών δειγμάτων του dataset.

Στη λογική αυτή, έγινε προσπάθεια διόρθωσης των κατηγοριών των δειγμάτων του dataset. Τα βίντεο χωρίστηκαν σε επιμέρους τμήματα των 1.3 δευτερολέπτων και έπειτα από θέαση τους, ταξινομήθηκαν στην κατάλληλη, σύμφωνα με τον συγγραφέα, κατηγορία. Μετά τη διόρθωση αυτή, το μοντέλο Inflated 3D πέτυχε Accuracy της τάξης του 91%. Ωστόσο έπειτα από ποιοτικό έλεγχο της ταξινόμησης, με παράλληλη απεικόνιση του βίντεο και της ταξινόμησης κάθε στιγμιότυπου, το συγκεκριμένο μοντέλο κρίθηκε χειρότερο των προηγουμένων ποιοτικά. Αυτό αποδεικνύει ότι το πρόβλημα της ταξινόμησης της Δοκιμασίας Εξαναγκασμένης Κολύμβησης είναι αρκετά σύνθετο, και δεν μπορεί εύκολα να ταξινομηθεί χειροκίνητα από έναν ανειδίκευτο παρατηρητή.

Κεφάλαιο 3

Βελτιστοποίηση Υπερπαραμέτρων

Στη συνέχεια, έγινε προσπάθεια βελτίωσης των μοντέλων νευρωνικών δικτύων CNN - LSTM και Inflated 3D, μέσω της βελτιστοποίησης των υπερπαραμέτρων τους.

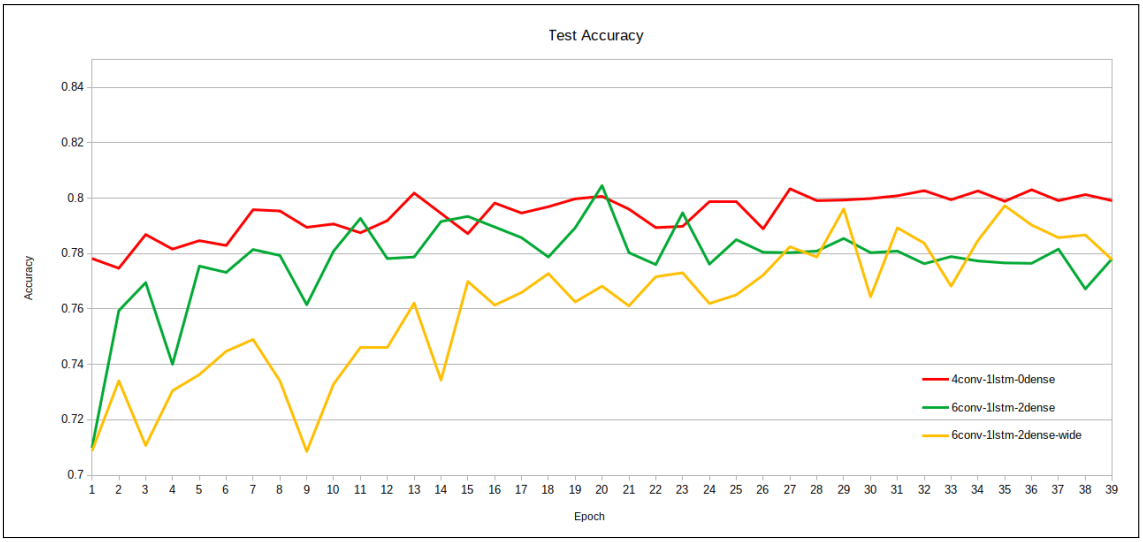
3.1 Αρχιτεκτονική CNN - LSTM

Για την αρχιτεκτονική CNN - LSTM, πραγματοποιήθηκαν πειράματα με διαφορετικό πλήθος CNN - LSTM layers, δηλαδή με ενίσχυση του βάθους της αρχιτεκτονικής. Ακόμη δοκιμάστηκε η αύξηση των νευρώνων των συνελικτικών και LSTM layers, το λεγόμενο πλάτος ή χωρητικότητα της αρχιτεκτονικής. Τέλος δοκιμάστηκε η προσθήκη Dense Layers στο τέλος του δικτύου. Ως learning rate του αλγορίθμου βελτιστοποίησης Adam, επιλέχθηκε η τιμή 10^{-4}, όπως προέκυψε απ᾽ το αρχικό πείραμα CNN - LSTM. Η επιλογή της βέλτιστης εποχής της ταξινόμησης γίνεται με αντίστοιχα κριτήρια του αρχικού πειράματος.

3.1.1 Πειράματα με 1 LSTM Layer

Οι πρώτες δοκιμές πραγματοποιήθηκαν με τις αρχιτεκτονικές που παρουσιάζονται στα σχήματα Γ΄.1, Γ΄.4, Γ΄.3. Δοκιμάστηκαν αρχιτεκτονικές παρόμοιας λογικής με αυτή του αρχικού πειράματος, αλλά βαθύτερες και με περισσότερη χωρητικότητα. Στη συγκεκριμένη περίπτωση εξετάστηκαν αρχιτεκτονικές με 1 LSTM layer. Το αρχικό πείραμα από εδώ και στο εξής θα αναφέρεται κωδικοποιημένα ως ”4conv-1lstm-0dense”. Τα δυο νέα δίκτυα αποτελούνται από έξι Convolutional Layers, στην συνέχεια ένα LSTM και τρία Dense Layers. Οι δύο αρχιτεκτονικές είναι πανομοιότυπες με μόνη διαφορά ότι η δεύτερη είναι πιο πλατιά δηλαδή αποτελείται από περισσότερα συνελικτικά φίλτρα και περισσότερους απλούς και ανατροφοδοτούμενους νευρώνες. Το πρώτο απ᾽ τα δυο νέα δίκτυα αναφέρεται ως ”conv6-1lstm-2dense” ενώ το δεύτερο ως ”conv6-1lstm-2dense-wide”.

Τα αποτελέσματα του Accuracy του υποσυνόλου test, των 3 αρχιτεκτονικών παρουσιάζονται στο σχήμα 3.1.



Σχήμα 3.1: Διαγράμμα Accuracy των αρχιτεκτονικών με 1 LSTM για τα υποσύνολο test

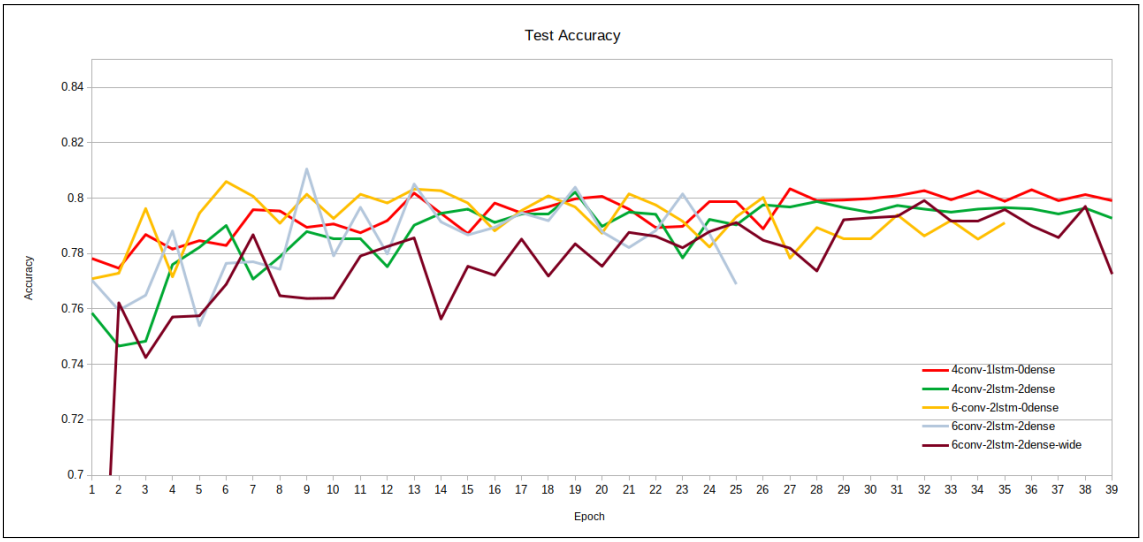
Όπως φαίνεται στο διάγραμμα 3.1 η αύξηση του βάθους της συγκεκριμένης αρχιτεκτονικής, δεν φαίνεται να ωφελεί την γενίκευση του μοντέλου. Μάλιστα σε συνδυασμό με την αύξηση του πλάτους, σημειώνονται ακόμη χειρότερα αποτελέσματα.

3.1.2 Πειράματα με 2 LSTM Layers

Στη συνέχεια, σχεδιάστηκαν παρόμοιες αρχιτεκτονικές, αλλά με τη χρήση 2 LSTM layers. Αρχικά σχεδιάστηκε μια αρχιτεκτονική Γ΄.2 με τέσσερα Convolution Layers και δυο LSTM , η οποία αναφέρεται ως 4conv-2lstm-0dense. Οι υπόλοιπες τρεις αρχιτεκτονικές που υλοποιήθηκαν έχουν έξι Convolutional Layers . Η πρώτη από αυτές Γ΄.5 δεν διαθέτει επιπλέον Dense Layers πέρα από αυτό της ταξινόμησης και αναφέρεται ως 6conv-2lstm-0dense. Η επόμενες δύο περιλαμβάνουν δύο επιπλέον Dense Layers, με την δεύτερη να έχει περισσότερο πλάτος και αναφέρονται ως 6conv-2lstm-2dense και 6conv-2lstm-2dense-wide αντίστοιχα. Επίσης παρουσιάζονται στο παράρτημα, στα σχήματα Γ΄.6 Γ΄.7 αντίστοιχα. Το accuracy του υποσυνόλου test για κάθεμια από τις παραπάνω αρχιτεκτονικές παρουσιάζεται στο σχήμα 3.2.

Για να είναι δυνατή η χρήση διαδοχικών LSTM Layers πρέπει τα Layers, πλην του τελευταίου, να εξάγουν την κωδικοποίηση κάθε χρονικού βήματος. Οι έξοδοι αυτές, ανά χρονικό βήμα, τροφοδοτούν το διάδοχο LSTM Layer, έχοντας ήδη κωδικοποιήσει την χρονική συσχέτιση.

Παρακάτω παρουσιάζονται τα αποτελέσματα των στατιστικών accuracy και loss για τα 2 υποσύνολα του dataset ανά εποχή. Στο ίδιο σχήμα παρατίθεται και το αντίστοιχο διάγραμμα του αρχικού πειράματος για λόγους σύγκρισης.



Σχήμα 3.2: Διαγράμμα Accuracy των αρχιτεκτονικών με 2 LSTM για τα υποσύνολο test

Στο σχήμα 3.2, παρατηρείται ότι στις αρχιτεκτονικές με 4 Convolutional Layers, η προσθήκη δεύτερου LSTM Layer δεν βελτιώνει τα αποτελέσματα. Ακόμη για άλλη μια φορά φαίνεται ότι η αύξηση του πλάτους της αρχιτεκτονικής φέρει τα χειρότερα αποτελέσματα σε σχέση με τα υπόλοιπα πειράματα. Ωστόσο οι άλλες δύο βαθύτερες αρχιτεκτονικές, με ή χωρίς επιπλέον Dense Layers, παρουσιάζουν βελτίωση της τάξη του 0.5 - 1 % σε σχέση με το αρχικό πείραμα.

Για την καλύτερη σύγκριση του αρχικού πειράματος 4conv-1lstm-0dense με την αρχιτεκτονική 6conv-2lstm-2dense, παρατίθενται επιπλέον τα στατιστικά στοιχεία της κατά την ένατη εποχή, που πετυχαίνει Accuracy 81%:

	Precision	**Recall**	**F1-Score**	**Support**
Class Statistics				
Immobility	0.81	0.90	**0.85**	2927
Swimming	0.67	0.30	0.42	989
Climbing	0.84	0.93	**0.88**	2830
Head Shake	0.61	0.50	0.55	272
Diving	0	0	0	6
Overall Statistics				
Accuracy		0.81		7024
Macro Average	0.59	0.53	0.54	7024
Weighted Average	0.80	0.81	0.79	7024

Πίνακας 3.1: Στατιστικά στοιχεία των αποτελεσμάτων του μοντέλου 6conv-2lstm-2dense

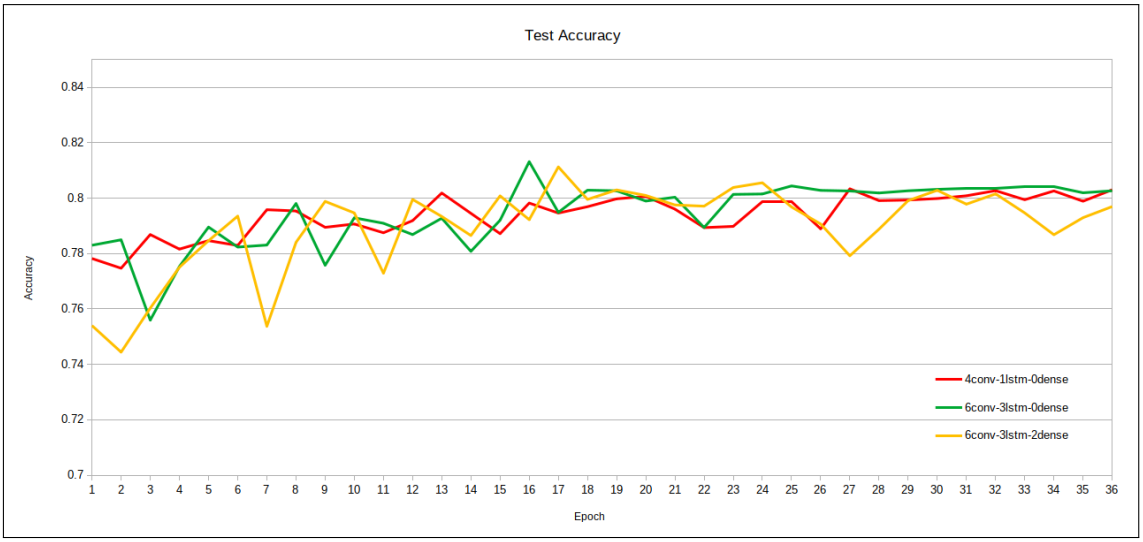
Όπως προκύπτει, το μοντέλο αυτό ενώ στατιστικά σημειώνει κάποιες βελτιώσεις, σε καμία περίπτωση δεν αποτελεί ποιοτική βελτίωση των αποτελεσμάτων, μιας και αναγνωρίζει ικανοποιητικά μόνο τις 2 βασικές κατηγορίες, ενώ δεν αναγνωρίζει καθόλου την κατηγορία Diving και πετυχαίνει 30% Recall στην κατηγορία Swimming.

Επομένως η αρχιτεκτονική του αρχικού πειράματος, εξακολουθεί να θεωρείται η πιο αξιόπιστη αρχιτεκτονική CNN - LSTM, για τα έως τώρα πειράματα.

3.1.3 Πειράματα με 3 LSTM Layers

Ακολουθούν δυο νέες αρχιτεκτονικές με 3 LSTM Layer. Και οι δυο διαθέτουν 6 Convolutional layers και είναι αρχιτεκτονικές ίδιου πλάτους. Η πρώτη (Γ΄.9) διαθέτει επιπλέον 2 Dense layers και θα αναφέρεται ως 6conv-3lstm-2dense ενώ η δεύτερη (Γ΄.8) περιλαμβάνει μόνο το βασικό Dense layer της ταξινόμησης και αναφέρεται ως 6conv-3lstm-0dense.

Τα διαγράμματα του Accuracy ανά εποχή, του υποσυνόλου test, για τις 2 αυτές αρχιτεκτονικές, καθώς και αυτή του αρχικού πειράματος, παρουσιάζονται στο σχήμα 3.3



Σχήμα 3.3: Διαγράμμα Accuracy των αρχιτεκτονικών με 3 LSTM για τα υποσύνολο test

Όπως φαίνεται στο διάγραμμα οι τρεις αρχιτεκτονικές φαίνονται πολύ κοντά στην ευστοχία που επιτυγχάνουν. Για πιο αξιόπιστη σύγκριση θα ελεγχθούν τα στατιστικά στοιχεία για την $16^{η}$ εποχή της αρχιτεκτονικής 6conv-3lstm-0dense.

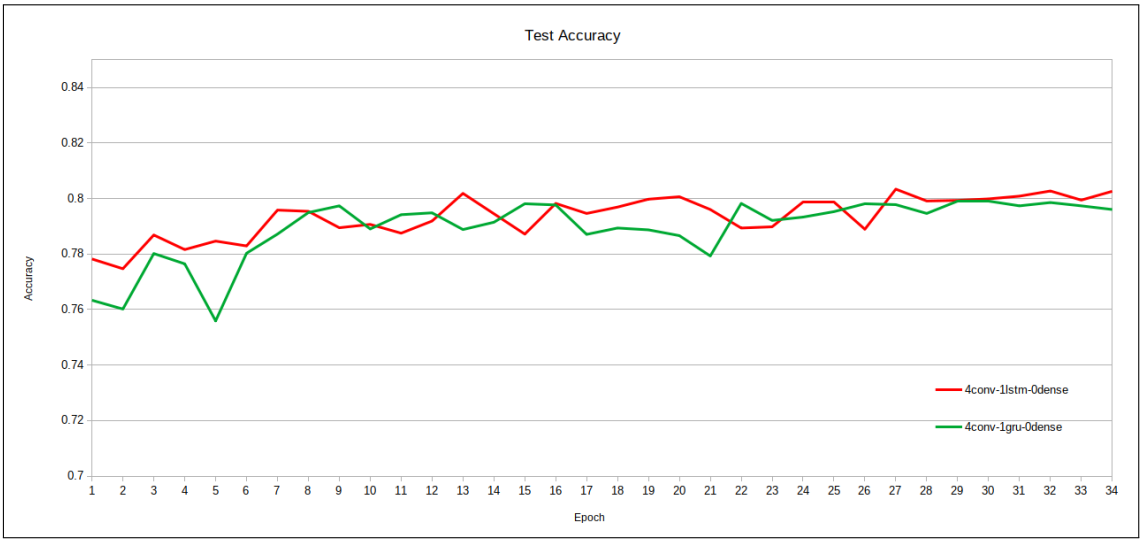
	Precision	**Recall**	**F1-Score**	**Support**
Class Statistics				
Immobility	0.86	0.84	**0.85**	2927
Swimming	0.58	0.50	0.53	989
Climbing	0.84	0.93	**0.88**	2830
Head Shake	0.71	0.42	0.53	272
Diving	1	0.17	0.29	6
Overall Statistics				
Accuracy		0.81		7024
Macro Average	0.80	0.57	0.62	7024
Weighted Average	0.80	0.81	0.80	7024

Πίνακας 3.2: Στατιστικά στοιχεία των αποτελεσμάτων του μοντέλου 6conv-3lstm-0dense

Σε αυτή τη περίπτωση, υπάρχει βελτίωση της τάξης του 9% στο Recall της κατηγορίας Swimming και μικρή πτώση για το Recall των κατηγοριών Immobility και Climbing. Ποιοτικά η αρχιτεκτονική θεωρείται καλύτερη, με μικρή διαφορά, σε σχέση με το αρχικό πείραμα.

3.1.4 Πείραμα με GRU Layer

Τέλος πραγματοποιήθηκε δοκιμή για αντικατάσταση του LSTM layer με GRU. Χρησιμοποιήθηκε αρχιτεκτονική πανομοιότυπη με το αρχικό πείραμα 4conv-1lstm-0dense, με μόνη διαφορά την αντικατάσταση του LSTM με GRU. Η αρχιτεκτονική CNN - GRU φαίνεται αναλυτικά στο σχήμα Γ΄.10, του Παραρτήματος. Τα διαγράμματα για τη σύγκριση των δύο αρχιτεκτονικών παρουσιάζονται στο σχήμα 3.4



Σχήμα 3.4: Διαγράμμα Accuracy των αρχιτεκτονικών LSTM και GRU για τα υποσύνολο test

Όπως φαίνεται στο σχήμα 3.4, οι δύο αρχιτεκτονικές έχουν παρόμοια συμπεριφορά, με την αρχιτεκτονική LSTM να έχει ελάχιστη υπεροχή. Επομένως δεν θεωρείται ότι η αρχιτεκτονική GRU μπορεί να φέρει βελτίωση των αποτελεσμάτων.

3.1.5 Συμπεράσματα

Σύμφωνα με τα παραπάνω πειράματα, δεν κρίνεται ότι η αρχιτεκτονική CNN - LSTM έχει τη δυνατότητα βελτίωσης των αποτελεσμάτων, συγκριτικά με την αρχιτεκτονική Inflated 3D. Για το λόγο αυτό, θα γίνει περαιτέρω προσπάθεια βελτιστοποίησης των παραμέτρων του dataset στην αρχιτεκτονική I3D.

3.2 Αρχιτεκτονική Inflated 3D

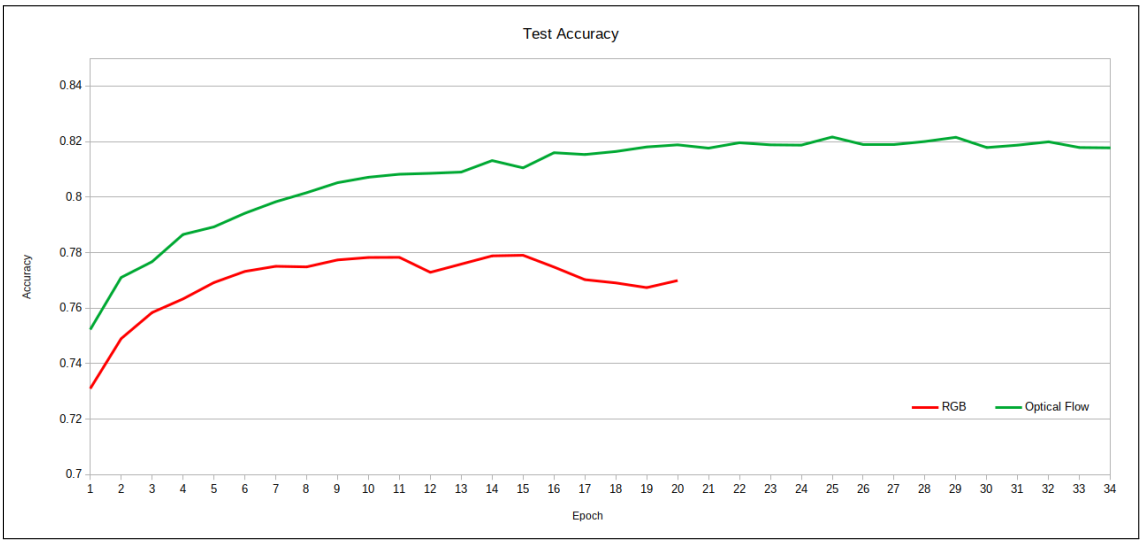
Η βελτιστοποίηση των παραμέτρων, κανονικά, πραγματοποιείται για διαφορετικούς συνδυασμούς αυτών, μιας και ενδέχεται να υπάρχουν συσχετίσεις μεταξύ τους που επηρεάζουν την τελική ακρίβεια. Ωστόσο κάτι τέτοιο δεν ήταν δυνατόν να πραγματοποιηθεί, αφενός λόγω του χρονικού περιορισμού, μιας και κάθε πείραμα διαρκεί έως 2 μέρες, και αφετέρου δεν κρίνεται αναγκαίο για τις ανάγκες τις παρούσας διπλωματικής. Επομένως η βελτιστοποίηση των παραμέτρων πραγματοποιήθηκε ανεξάρτητα για κάθε παράμετρο. Επιλέχθηκε μια λογική σειρά, για τον περιορισμό ενδεχόμενων συσχετίσεων μεταξύ των παραμέτρων. Κριτήριο για την βελτιστοποίηση των παραμέτρων θα αποτελέσει το βέλτιστο accuracy κάθε πειράματος. Η σειρά με την οποία πραγματοποιήθηκε η βελτιστοποίηση, καθώς και οι αρχικές τιμές κάθε παραμέτρου παρουσιάζονται παρακάτω:

1. Επιλογή Παρατηρητή και βαρών δειγμάτων: αληθείς τιμές Παρατηρητή 1
2. Είδος Εικόνας: Optical Flow
3. Μέγεθος εικόνας: πλάτος $x = 100$, ύψος $y = 200$
4. Χρονικό Διάστημα κάθε δείγματος: 1 sec
5. frames / sec : 25 fps

Ως learning rate χρησιμοποιήθηκε η τιμή 10^{-3}. όπως προέκυψε απ' το αρχικό πείραμα του Inflated 3D. Ακόμη με την ίδια λογική που περιγράφτηκε στο αρχικό πείραμα, επιλέγεται η βέλτιστη εποχή των πειραμάτων, απ΄ την οποία προκύπτει και παρουσιάζεται το accuracy του υποσυνόλου test.

3.2.1 Είδος Εικόνων

Στο συγκεκριμένο πείραμα εξετάζεται ενδεχόμενη βελτίωση, με τροφοδότηση του δικτύου με εικόνες RGB, αντί για Optical Flow.



Σχήμα 3.5: Διάγραμμα βελτιστοποίησης Μεγέθους εικόνας

Όπως φαίνεται στο διάγραμμα 3.5, υπάρχει σαφής υπεροχή των εικόνων Optical Flow, καθώς πετυχαίνει 4% μεγαλύτερο Accuracy. Επομένως συνεχίστηκε η χρήση εικόνων Optical Flow.

3.2.2 Επιλογή Παρατηρητή - Βαρών

Αρχικά πραγματοποιήθηκαν πειράματα για την επιλογή του παρατηρητή, καθώς και την επιλογή βαρών για τα δείγματα. Όσον αφορά την επιλογή του παρατηρητή πραγματοποιήθηκαν πειράματα αλλάζοντας τις αληθείς τιμές των υποσυνόλων train και test με όλους τους δυνατούς συνδιασμούς και προέκυψαν τα εξής αποτελέσματα:

Είναι φανερό ότι όταν οι αληθείς τιμές του υποσυνόλου train προέρχονται απ' τον Παρατηρητή 1, σε κάθε περίπτωση βελτιώνεται το accuracy του υποσυνόλου test. Για το λόγο αυτό οι αληθείς τιμές του train επιλέχτηκαν να είναι από τον παρατηρητή 1.

Επιλογή Παρατηρητή		**Train**	
		Παρατηρητής 1	Παρατηρητής 2
Test	Παρατηρητής 1	**73.21**%	71.86%
	Παρατηρητής 2	**70.36**%	69.49%

Πίνακας 3.3: Αποτελέσματα accuracy ανά παρατηρητή

Η πτώση του test accuracy σε σχέση με τα αρχικά πειράματα δικαιολογείται, καθώς με αυτό τον τρόπο δεν διατηρούνται κατά την εκπαίδευση και επαλήθευση, μόνο τα πειράματα κατά τα οποία οι δυο παρατηρητές βρίσκονται σε συμφωνία, αλλά το σύνολο των δειγμάτων.

Ωστόσο μπορεί να είναι βοηθητική η αύξηση των βαρών των δειγμάτων στις περιπτώσεις που ο παρατηρητής 2 συμφωνεί με τον παρατηρητή 1. Τα αποτελέσματα για τα βάρη παρουσιάζονται στον παρακάτω πίνακα: Όπως προκύπτει απ᾽ τον πίνακα

Επιλογή Βαρών	Accuracy (%)
1	73.05
1.5	73.16
2	73.29
3	**73.94**
4	73.30

Πίνακας 3.4: Αποτελέσματα της βελτιστοποίησης του Augmentation

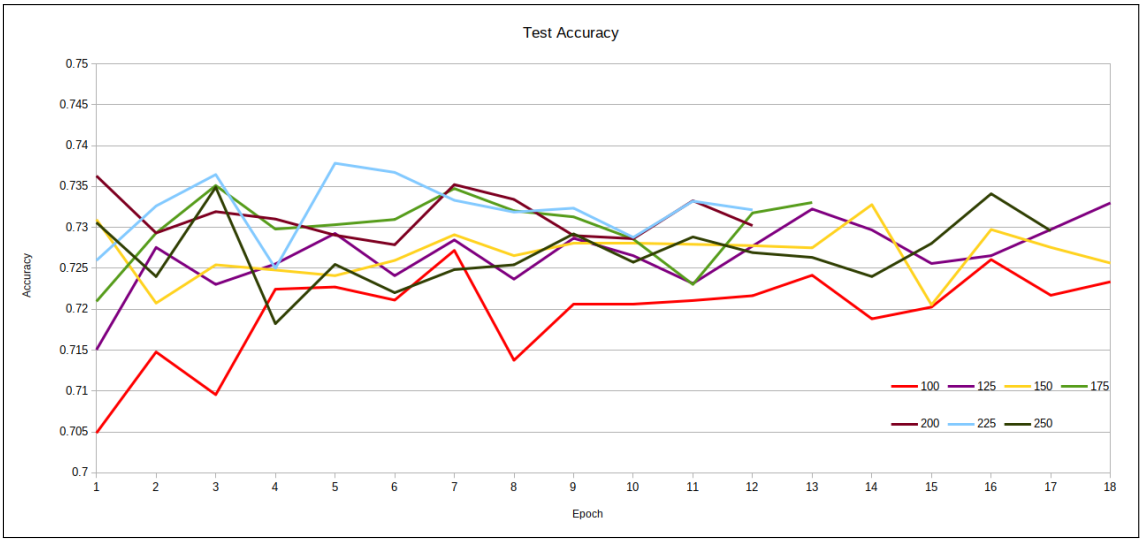
3.4, η ενίσχυση του βάρους των δειγμάτων φαίνεται να δίνει βελτίωση της τάξης του 1%. Για το λόγο αυτό, συνολικά, επιλέχθηκε για κάθε δείγμα στο οποίο οι 2 παρατηρητές συμφωνούν να έχει βάρος 3, ενώ για αυτά που διαφωνούν, επιλέγεται η αληθής τιμή του παρατηρητή 1, με βάρος 1.

3.2.3 Μέγεθος Εικόνας

Στη συνέχεια πραγματοποιήθηκαν πειράματα σχετικά με το βέλτιστο μέγεθος της εικόνας. Οι εικόνες έχουν αναλογία πλάτους / ύψους περίπου 1/2. Για το λόγο αυτό έγινε μετασχηματισμός της εικόνας τέτοιος ώστε με καθοριζόμενο πλάτος, να προκύπτει διπλάσιο ύψος. Η διαδικασία αυτή αποτελεί μια μορφή augmentation, μιας και προκύπτει ένας διαφορετικός αφινικός μετασχηματισμός για κάθε βίντεο. Σημειώνεται ότι η εικόνα που τροδοδοτεί το δίκτυο, τελικά έχει διαστάσεις 10% μικρότερες απ᾽ το καθορισμένο μέγεθος, λόγω του augmentation. Οι τιμές του πλάτους x που επιλέχτηκαν και τα αποτελέσματα παρουσιάζονται παρακάτω:

Μέγεθος Εικόνας	Accuracy (%)
100	72.72
125	73.33
150	73.28
175	73.51
200	73.63
225	**73.78**
250	73.49

Πίνακας 3.5: Αποτελέσματα της βελτιστοποίησης του μεγέθους εικόνας

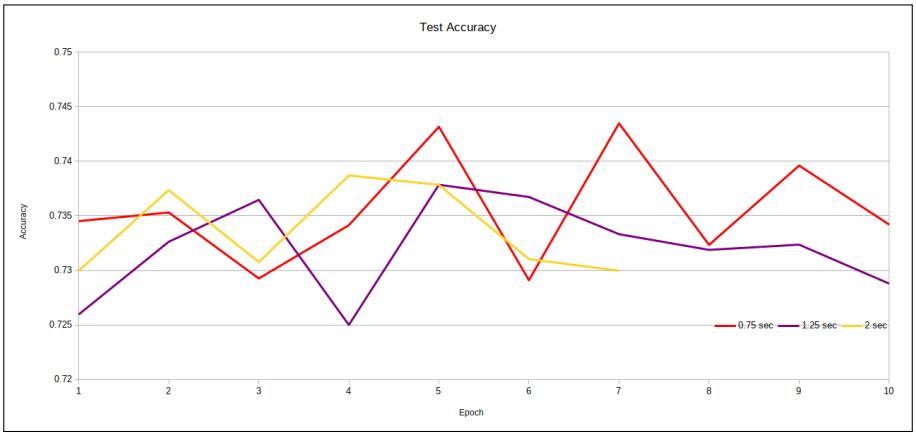


Σχήμα 3.6: Διάγραμμα βελτιστοποίησης Μεγέθους εικόνας

Προέκυψε λοιπόν, ότι υπήρξαν μικρές βελτιώσεις τις τάξης του 1% με μεγαλύτερο μέγεθος εικόνας. Για το λόγο αυτό επιλέχθηκε ως τελικό μέγεθος της εικόνας: πλάτος 225 - ύψος 450.

3.2.4 Χρονικό Διάστημα Δειγμάτων

Για τη χρονική διάρκεια του κάθε δείγματος, προέκυψαν τα εξής αποτελέσματα: Παρατηρείται ότι το ελάχιστο δυνατό μέγεθος φέρει τα καλύτερα αποτελέσματα



Σχήμα 3.7: Διάγραμμα βελτιστοποίησης χρονικού διαστήματος των δειγμάτων

Χρονικό Διάστημα	Accuracy (%)
0.75 sec	**74.55**
1.25 sec	73.78
2 sec	73.67

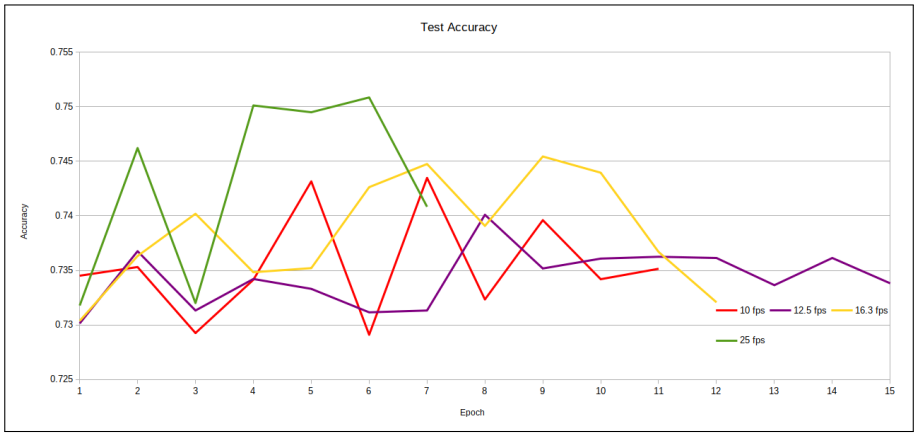
Πίνακας 3.6: Αποτελέσματα της βελτιστοποίησης του χρονικού διαστήματος των δειγμάτων

σχετικά με την ακρίβεια του μοντέλου. Η βελτίωση αυτή ενδεχομένως, οφείλεται στο

γεγονός ότι όσο μικρότερος είναι το χρονικό διάστημα του κάθε δείγματος, τόσο μικρότερη είναι η πιθανότητα να περιλαμβάνει περισσότερες από μία κατηγορίες σ' αυτό το χρόνο. Επιλέχθηκε λοιπόν, τα δείγματα να αποτελούνται από χρονικό διάστημα 0.75 sec.

3.2.5 Frames / sec

Τέλος έγινε έλεγχος για την κατάλληλη τιμή των frames per second. Τα αποτελέσματα φαίνονται παρακάτω:



Σχήμα 3.8: Διάγραμμα βελτιστοποίησης frames / sec

frames / sec	Accuracy (%)
10 fps	74.35
12.5 fps	74.01
16.6 fps	74.54
25 fps	**75.08**

Πίνακας 3.7: Αποτελέσματα της βελτιστοποίησης των frames / sec

Επομένως σύμφωνα με τον πίνακα 3.7, η αύξηση των frames / sec έφερε μικρές βελτιώσεις στην ακρίβεια της τάξης του 1%. Άρα επιλέχθηκε η τιμής των 25 fps. Σημειώνεται ότι δεν έγινε πείραμα για τα 50 fps, καθώς δεν ήταν δυνατόν με τη χρήση της δεδομένης κάρτα γραφικών λόγω περιορισμού της μνήμης της.

Για την απεικόνιση και αξιολόγιση των τελικών αποτελεσμάτων, τελικά επιλέγεται η αρχιτεκτονική Inflated 3D, με τις υπερπαραμέτρους που προέκυψαν, μέσω τις βελτιστοποίησης.

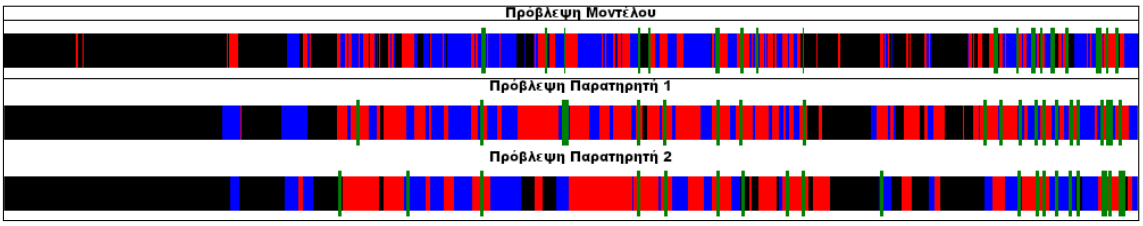
Κεφάλαιο 4

Τελικά Αποτελέσματα - Συζήτηση

Έπειτα από αρκετές δοκιμές στις μεθόδους και τις παραμέτρους τους, επιλέχτηκε το μοντέλο Inflated 3D, ως καταλληλότερο για την αυτόματη ταξινόμηση της Δοκιμασίας Εξαναγκασμένης Κολύμβησης. Στο παρόν κεφάλαιο, γίνεται απεικόνιση των αποτελεσμάτων με το συγκεκριμένο μοντέλο, για την ποιοτική και ποσοτική αξιολόγηση τους. Ως κριτήριο για την αξιολόγηση θα χρησιμοποιηθούν τα βίντεο του υποσυνόλου test.

4.1 Παραγωγή Εικόνων Χρονοσειρών

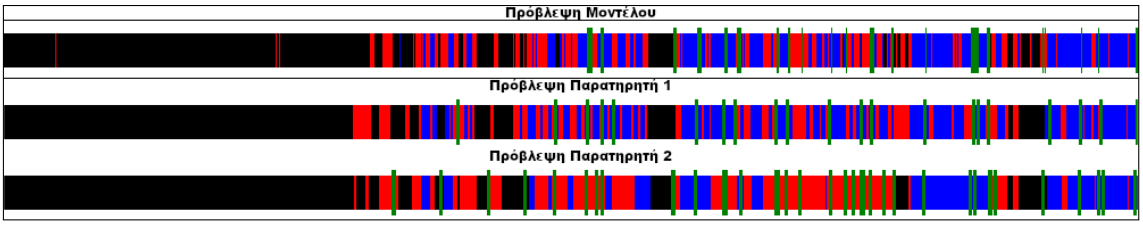
Όπως έχει ήδη αναφερθεί, οι παρατηρήσεις των βίντεο, δόθηκαν σε μορφή εικόνων χρονοσειρών, όπου ο άξονας x συμβολίζει τον χρόνο και το χρώμα την κατηγορία. Για λόγους ομοιομορφίας και σύγκρισης των αποτελεσμάτων, τα αποτελέσματα της εκτίμησης του μοντέλου σε κάθε βίντεο, αποθηκεύτηκαν με script στην python, σε πανομοιότυπο πρότυπο με αυτό της Εφαρμογής Kinoscope. Μαζί με τις εικόνες της εκτίμησης του μοντέλου, παρατίθενται κατακορύφως και οι εικόνες των παρατηρήσεων των δύο ειδικών. Ακόμη για κάθε βίντεο, αναγράφεται και ο συνολικό χρόνος σε δευτερόλεπτα που διήρκησε η κάθε κατηγορία εντός του βίντεο. Ορισμένα δείγματα των βίντεο του συνόλου test, παρουσιάζονται παρακάτω, ενώ τα υπόλοιπα βρίσκονται στο παράρτημα Δ΄.



Σχήμα 4.1: Αποτελέσματα Πειράματος 4C-2012-F15

Video: 4C-2012-F15	Immobility	Swimming	Climbing	Head Shake	Diving
Μοντέλο I3D	86	63	139	15	0
Παρατηρητής 1	70	106	105	22	0
Παρατηρητής 2	64	91	126	22	0

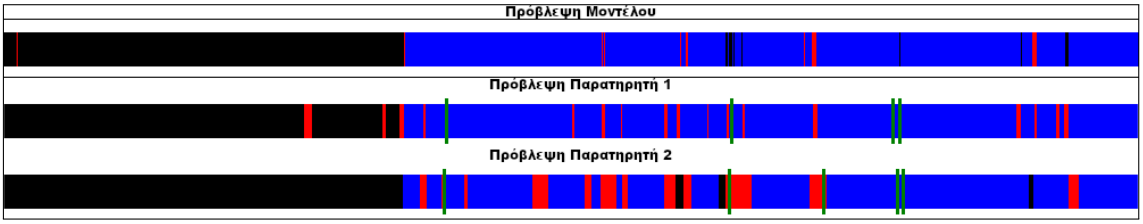
Πίνακας 4.1: Πίνακας συνολικών δευτερολέπτων ανά κατηγορία πειράματος 4C-2012-F15



Σχήμα 4.2: Αποτελέσματα Πειράματος 4C-2012-F19

Video: 4C-2012-F19	Immobility	Swimming	Climbing	Head Shake	Diving
Μοντέλο I3D	86	66	133	16	0
Παρατηρητής 1	92	62	128	22	0
Παρατηρητής 2	62	77	130	37	0

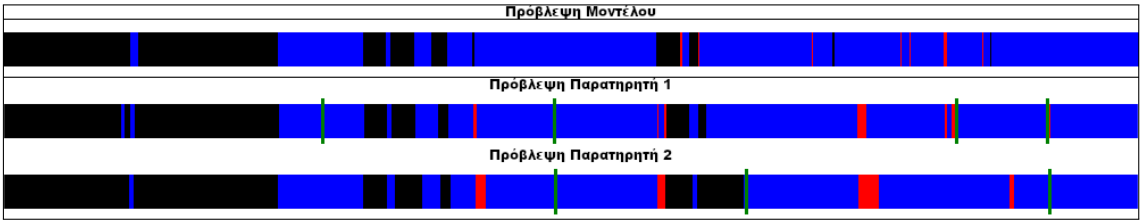
Πίνακας 4.2: Πίνακας συνολικών δευτερολέπτων ανά κατηγορία πειράματος 4C-2012-F19



Σχήμα 4.3: Αποτελέσματα Πειράματος 4C-F4

Video: 4C-F4	Immobility	Swimming	Climbing	Head Shake	Diving
Μοντέλο I3D	187	5	109	0	0
Παρατηρητής 1	179	15	102	0	0
Παρατηρητής 2	152	32	111	5	0

Πίνακας 4.3: Πίνακας συνολικών δευτερολέπτων ανά κατηγορία πειράματος 4C-F4



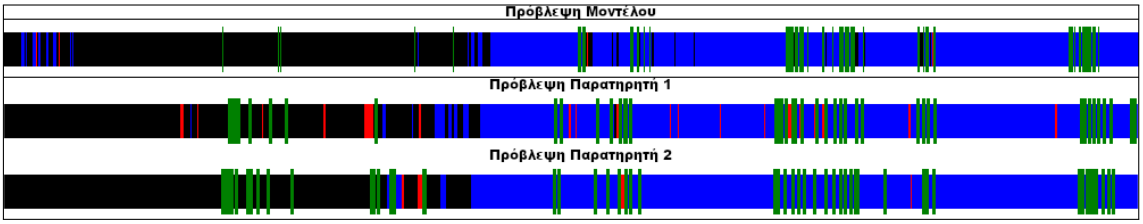
Σχήμα 4.4: Αποτελέσματα Πειράματος 4C-F7

Video: 4C-F7	Immobility	Swimming	Climbing	Head Shake	Diving
Μοντέλο I3D	204	3	98	0	0
Παρατηρητής 1	196	6	94	4	0
Παρατηρητής 2	179	11	107	3	0

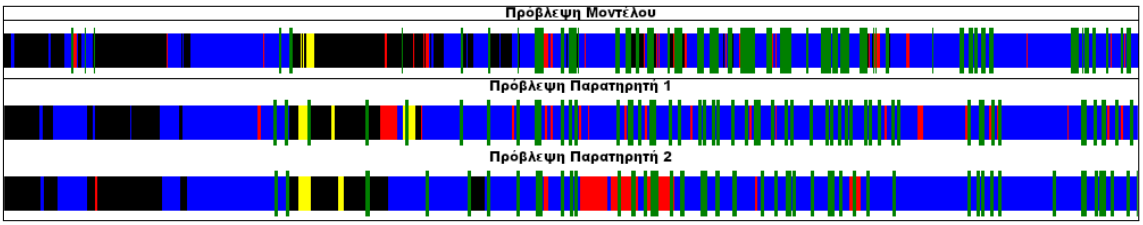
Πίνακας 4.4: Πίνακας συνολικών δευτερολέπτων ανά κατηγορία πειράματος 4C-F7

Video: 4C-M5	Immobility	Swimming	Climbing	Head Shake	Diving
Μοντέλο I3D	159	2	135	25	0
Παρατηρητής 1	144	11	107	44	0
Παρατηρητής 2	148	3	104	53	0

Πίνακας 4.5: Πίνακας συνολικών δευτερολέπτων ανά κατηγορία πειράματος 4C-M5



Σχήμα 4.5: Αποτελέσματα Πειράματος 4C-M5



Σχήμα 4.6: Αποτελέσματα Πειράματος 4C-M21

Video: 4C-M21	Immobility	Swimming	Climbing	Head Shake	Diving
Μοντέλο I3D	144	13	85	58	4
Παρατηρητής 1	178	16	53	55	6
Παρατηρητής 2	171	21	59	52	4

Πίνακας 4.6: Πίνακας συνολικών δευτερολέπτων ανά κατηγορία πειράματος 4C-M21

Όπως διαπιστώνεται, τα αποτελέσματα φαίνονται αρκετά ικανοποιητικά και παραπλήσια με τους δύο παρατηρητές, σαν συνολική εικόνα. Οι κατηγορίες Climbing και Immobility του μοντέλου έχουν παρόμοια αποτελέσματα με τους δυο παρατηρητές. Σε ορισμένες περιπτώσεις το μοντέλο δυσκολεύεται να εντοπίσει την κατηγορία Swimming, μειώνοντας συστηματικά ελαφρώς τον συνολικό χρόνο της. Αντίστοιχα και η κατηγορία Head Shake, που αποτελεί το βασικότερο πρόβλημα του μοντέλου. Τέλος η πρόβλεψη της κατηγορίας Diving έχει απόλυτα σωστή συμπεριφορά, για το μοναδικό βίντεο στο οποίο εμφανίζεται.

Μέρος IV

Επίλογος

Κεφάλαιο 1

Συμπεράσματα

Στην διπλωματική αυτή, πραγματεύτηκε το πρόβλημα της αναγνώρισης συμπεριφοράς επιμυών, για τη Δοκιμασία Εξαναγκασμένης Κολύμβησης. Το πείραμα αυτό έχει ιδιαίτερο ενδιαφέρον στην μελέτη και σχεδίαση αντικαταθλιπτικών φαρμάκων. Η ταξινόμηση γίνεται για περίπου κάθε δευτερόλεπτο του πειράματος, σε πέντε διαφορετικές κατηγορίες ενδιαφέροντος. Τη δεδομένη στιγμή, υπάρχουν εμπορικές εφαρμογές υψηλού κόστους, για την αυτοματοποίηση του προβλήματος αυτού, και συνήθως, έπειτα από καθορισμό παραμέτρων για την ευαισθησία των αλγορίθμων. Για το λόγο αυτό, η διαδικασία ταξινόμησης πολλές φορές συμβαίνει με χειροκίνητη εργασία από ειδικούς παρατηρητές. Για την επίλυση του προβλήματος αυτού, βρέθηκαν τα κατάλληλα δεδομένα εκπαίδευσης. Τα δεδομένα αυτά υπέστησαν τις κατάλληλες επεξεργασίες και διορθώσεις για την δημιουργία dataset που μπορεί να χρησιμοποιηθεί για την περαιτέρω ανάπτυξη και βελτίωση μοντέλων εκτίμησης συμπεριφοράς σε βίντεο Προ-κλινικών πειραμάτων.

Για το συγκεκριμένο πρόβλημα, κριτήριο της ταξινόμησης αποτελεί η κίνηση των επιμυών και όχι τα χαρακτηριστικά της μορφής τους, γι᾽ αυτό και εξήχθησαν και χρησιμοποιήθηκαν εικόνες οπτικής ροής. Η σύγκριση και αξιολόγηση των αποτελεσμάτων, έγινε δυνατή με την εφαρμογή τριών απολύτως διαφορετικών μοντέλων εκτίμησης. Αρχικά πραγματοποιήθηκε εξαγωγή χαρακτηριστικών των βίντεο με τον Βελτιωμένο Αλγόριθμο Πυκνών Τροχιών, κωδικοποίηση τους με Fisher Vectors και Ταξινόμηση με Μηχανές Διανυσμάτων Υποστήριξης. Η διαδικασία αυτή πέτυχε ευστοχία 80%, αλλά με χαμηλο recall (ανάκληση) σε συγκεκριμένες κατηγορίες ενδιαφέροντος. Συγκεκριμένα, οι δυο κυρίαρχες κατηγορίες ενδιαφέροντος, επιτυγχάνουν F1-Score άνω του 85%, ωστόσο οι υπόλοιπες τρεις μικρότερο του 45%. Στη συνέχεια σχεδιάστηκε μοντέλο με δίκτυο συνελικτικών νευρώνων για την εξαγωγή χωρικών χαρακτηριστικών στο δισδιάστατο πεδίο του χώρου, και τροφοδότηση τους σε LSTM για την δημιουργία χρονικών συσχετίσεων των χαρακτηριστικών αυτών. Η αρχιτεκτονική αυτή είχε αντίστοιχα αποτελέσματα, με ευστοχία 80%, αλλά σημείωσε μερική βελτίωση στο F1-Score των πασχόντων κατηγοριών, αυξάνοντας το κατά 5% σε σχέση με τον Αλγόριθμο Πυκνών Τροχιών. Πραγματοποιήθηκε προσπάθεια για εύρεση καταλληλότερης αρχιτεκτονικής CNN - LSTM, με αυξομειώσεις στο μήκος και πλάτος της, καθώς και τη χρήση ανατροφοδοτούμενων νευρώνων GRU, χωρίς ιδιαίτερες βελτιώσεις στα αποτελέσματα. Τέλος, εφαρμόστηκε η αρχιτεκτονική Inflated 3D που αξιοποιεί τις 3D συνελίξεις για την εξαγωγή χωροχρονικών χαρακτηριστικών, ενώ διαθέτει και προεκπαιδευμένα βάρη. Έδωσε τα καλύτερα αποτελέσματα με ευστοχία 82%, ενώ βελτίωσε σημαντικά τα προβλήματα που προκύπτουν από την ανισορροπία των κλάσεων. Το F1-Score των δύο κυριάρχων κατηγοριών παρέμεινε

στο 85%, ενώ των υπολοίπων κυμάνθηκε στις τιμές 55-62%, δηλαδή βελτίωση της τάξης του 10% σε σχέση με τα προηγούμενα δύο μοντέλα.

Για το λόγο αυτό πραγματοποιήθηκε προσπάθεια βελτιστοποίησης των υπερπαραμέτρων στην αρχιτεκτονική αυτή, που ενίσχυσε την ευστοχία της κατά 1.5%, κυρίως με την αύξηση του μεγέθους των εικόνων και των στιγμιοτύπων ανά δευτερόλεπτο. Για την εκπαίδευση του τελικού μοντέλου χρειάστηκαν 2 μέρες, ενώ προβλέπει τις κατηγορίες κάθε πειράματος 5 λεπτών, σε 3 λεπτά. Στο αρχικό πείραμα με μικρότερο μέγεθος εικόνας και χρονική ακρίβεια, για την εκπαίδευση του μοντέλου χρειάστηκαν 6 ώρες, ενώ προέβλεπε τις κατηγορίες ενός βίντεο, σε μισό λεπτό. Τα εναπομένοντα σφάλματα του μοντέλου, βρίσκονται σε αντιστοιχία με τις διαφωνίες των δύο παρατηρητών. Φαίνεται λοιπόν, να οφείλονται σε σφάλματα του dataset, λόγω της αυξημένης δυσκολίας της χειροκίνητης ταξινόμησης του προβλήματος, καθώς και την έλλειψη σωστού συγχρονισμού των βίντεο με τις παρατηρήσεις.

Σαν αποτέλεσμα υλοποιήθηκε σύστημα το οποίο με την είσοδο βίντεο της Δοκιμασίας Εξαναγκασμένης Κολύμβησης, εξάγει αυτομάτως τη συμπεριφορά του επιμύ κάθε μισό δευτερόλεπτο, ταξινομώντας μικρά χρονικά διαστήματα στις πέντε κατηγορίες ενδιαφέροντος.

Κεφάλαιο 2

Μελλοντικές Επεκτάσεις

Κατά την εκπόνηση της διπλωματικής, υπήρξαν ιδέες για την περαιτέρω βελτίωση του προβλήματος που θα μπορούσαν μελλοντικά να βελτιώσουν τα υπάρχοντα αποτελέσματα.

Για την ταξινόμηση των δειγμάτων, δυνατή είναι η παράλληλη τροφοδότηση του δικτύου με τον ήχο των αντίστοιχων δειγμάτων. Η συνεισφορά του ήχου στην ταξινόμηση θα μπορούσε να βελτιώσει σημαντικά την ευστοχία της ταξινόμησης, καθώς σύμφωνα με την κίνηση των επιμυών, προκύπτουν χαρακτηριστικοί ήχοι του νερού.

Ακόμη, αξίζει να εξεταστεί η τροφοδότηση των νευρωνικών δικτύων με εικόνες Αφαίρεσης Παρασκηνίου (Background Subtraction), αντί της οπτικής ροής. Εκτός από τη μείωση του χρόνου της εξαγωγής τους, θα μπορούσε να επιτευχθεί και βελτίωση των αποτελεσμάτων, καθώς το συγκεκριμένο πείραμα διατηρεί σταθερή κάμερα και η εξαγωγή του προσκηνίου θα περιέγραφε ικανοποιητικά τις κινήσεις των επιμυών.

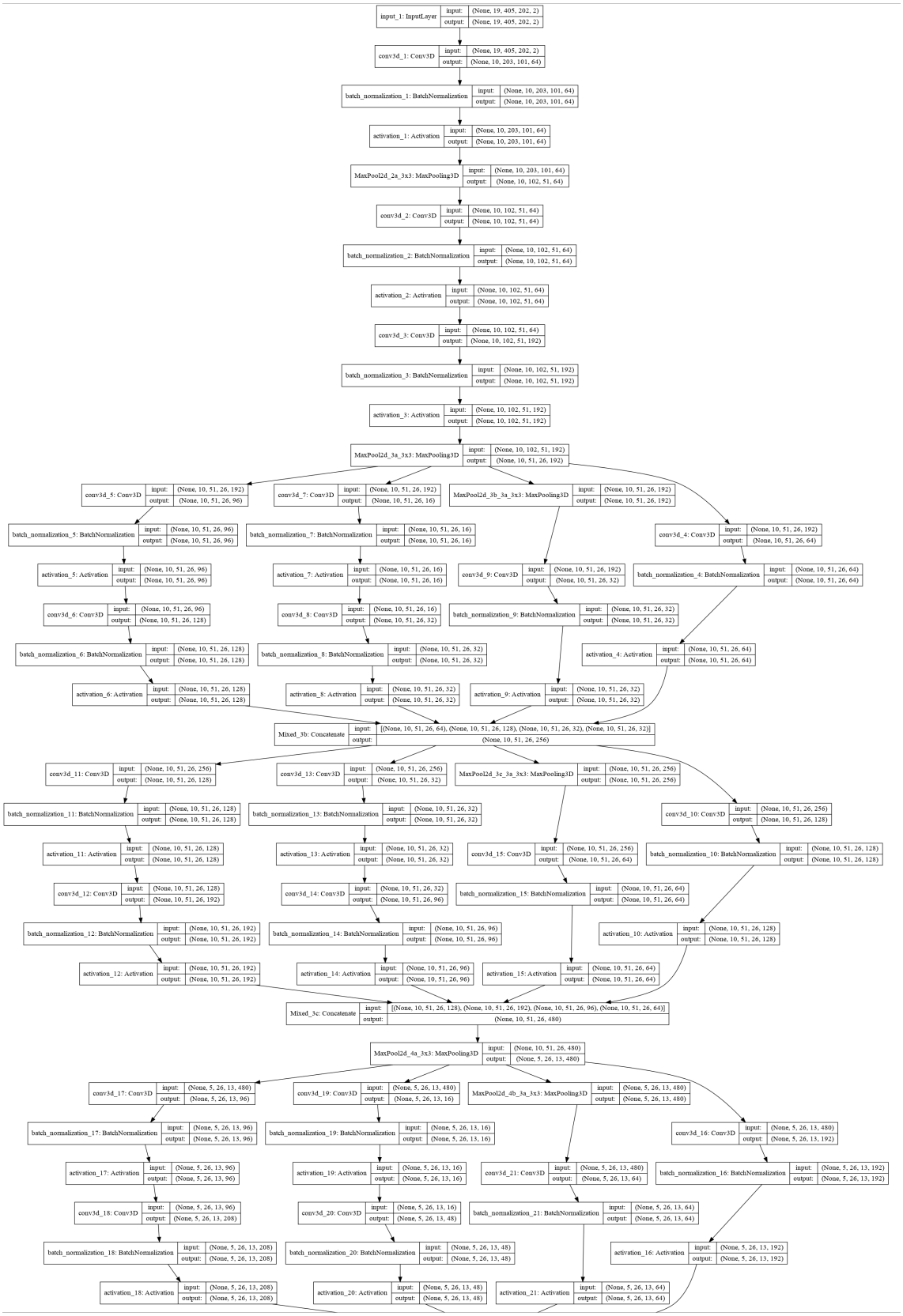
Στη συγκεκριμένη διπλωματική, εξετάζονται τεχνικές για την ταξινόμηση αποκομμένων βίντεο. Αυτό σημαίνει ότι κάθε βίντεο περιλαμβάνει μόνο μία κατηγορία. Ωστόσο υπάρχει προσπάθεια για ανάπτυξη τεχνικών που διαχωρίζουν το βίντεο σε επιμέρους τμήματα με βάση τα διαφορετικά χαρακτηριστικά κάθε τμήματος, και ταξινομούν το κάθε ένα απ' αυτά. Το πρόβλημα αυτό ονομάζεται αναγνώριση δράσης συνεχών βίντεο (Untrimmed / Continuous Video Action Recognition). Η εφαρμογή του στη συγκεκριμένη εφαρμογή έχει ιδιαίτερο ερευνητικό ενδιαφέρον.

Ένα απ' τα βασικά προβλήματα που αντιμετωπίστηκαν, είναι η ανισορροπία των κλάσεων. Δυνατή είναι η ενίσχυση των βαρών των δειγμάτων σπανιότερων κατηγοριών, η εξαναγκαστική τροφοδότηση του κάθε batch με δείγματα αυτών των κατηγοριών, καθώς και άλλες τεχνικές όπως η συνάρτηση κόστους Focal Loss.

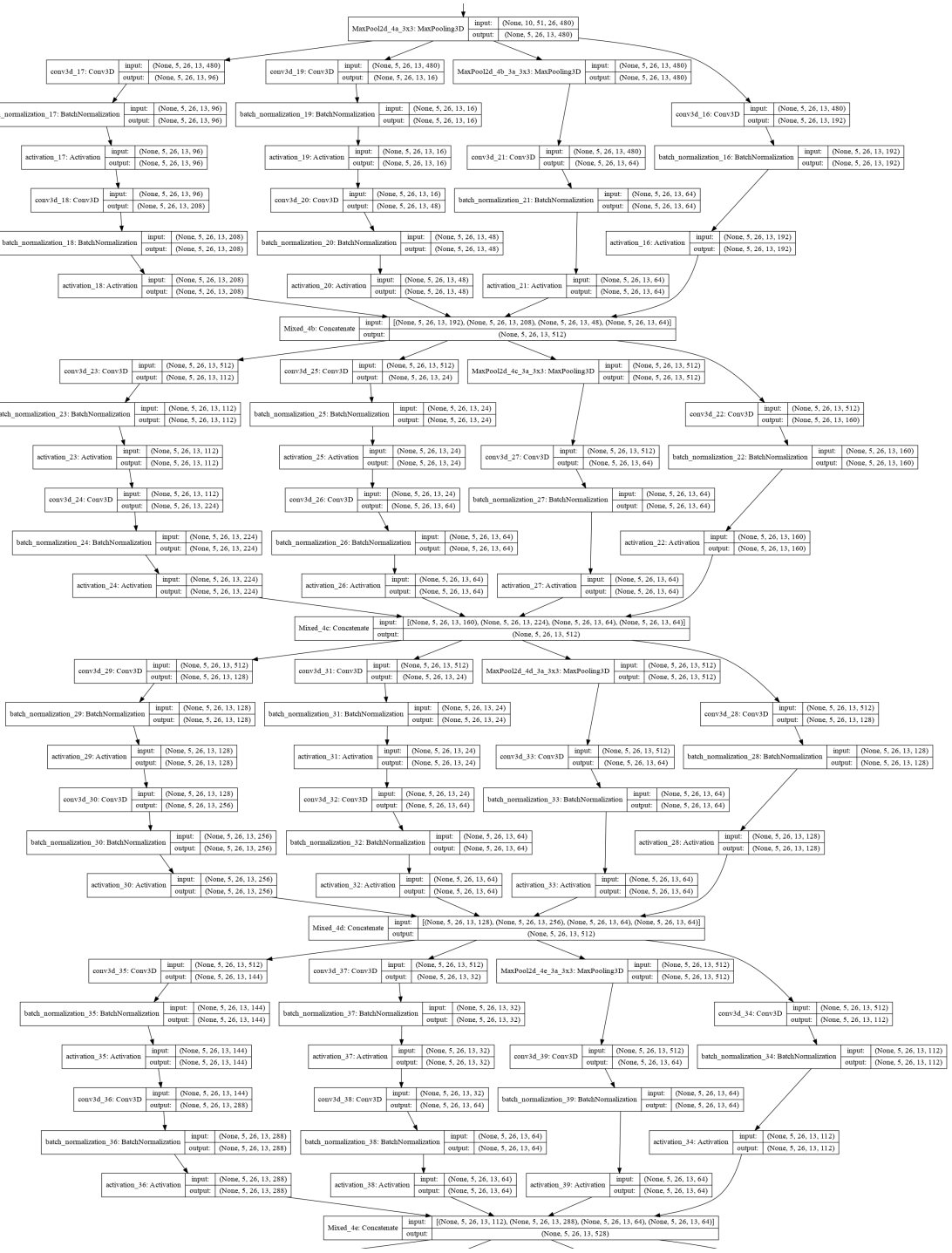
Τέλος, σημαντικός είναι ο εμπλουτισμός του dataset με πειράματα διαφορετικών εργαστηριών. Εφόσον τα βίντεο τα οποία τροφοδότησαν τα μοντέλα, προέρχονται από το ίδιο εργαστήριο, δεν μπορεί να θεωρηθεί ότι το μοντέλο έχει γενικευτεί στο πρόβλημα, αλλά στο συγκεκριμένο περιβάλλοντα χώρο του εργαστηρίου και στη συγκεκριμένη ράτσα επιμυών. Η ενίσχυση των δειγμάτων με νέα πειράματα διαφορετικής προέλευσης θα ενίσχυε αδιαμφισβήτητα τα αποτελέσματα της εκτίμησης.

Παράρτημα Α΄

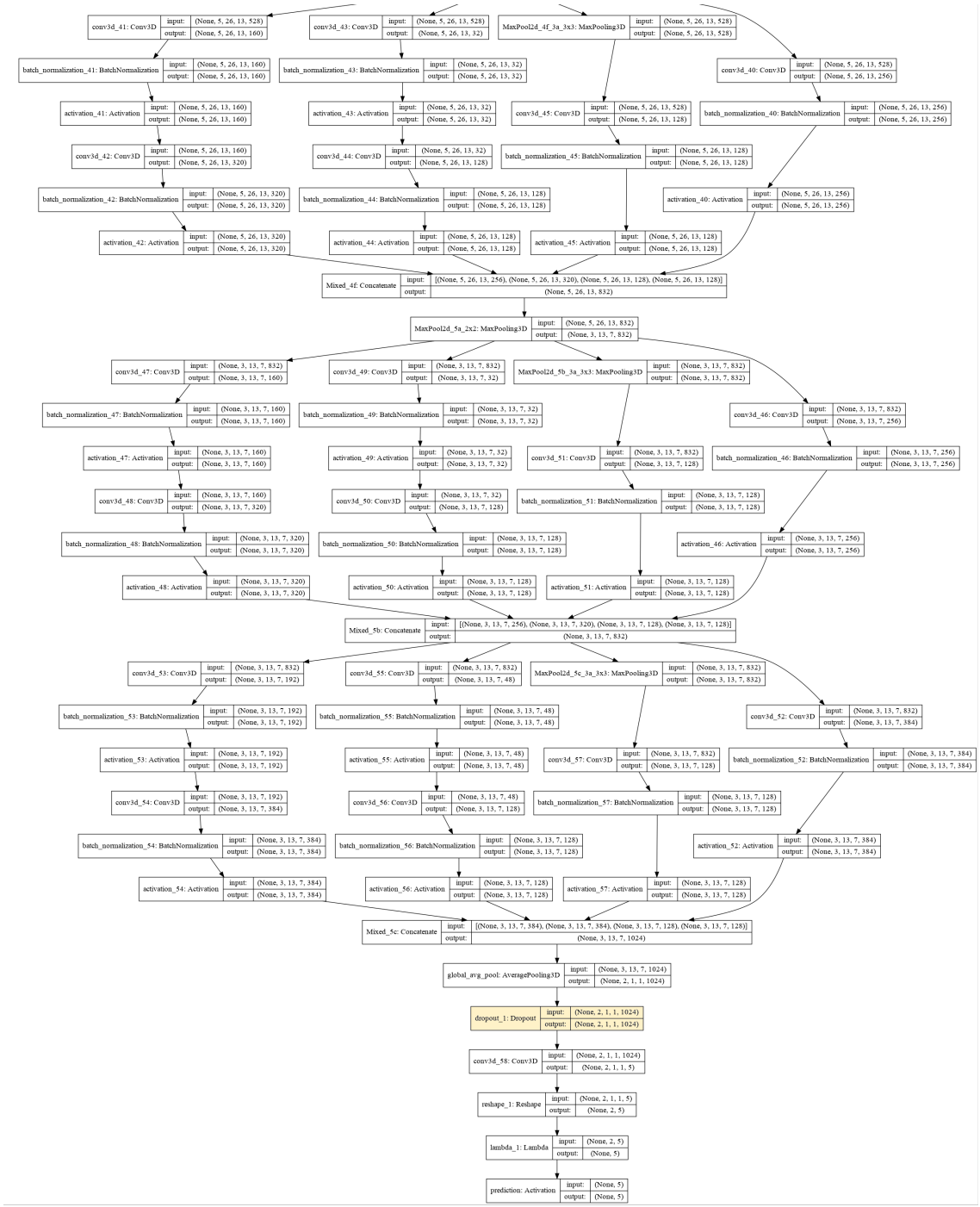
Διαγραμματική Αναπαράσταση του Μοντέλου Inflated 3D



Σχήμα Α΄.1: Διαγραμματική αναπαράσταση του μοντέλου Inflated 3D - Μέρος $1^ο$



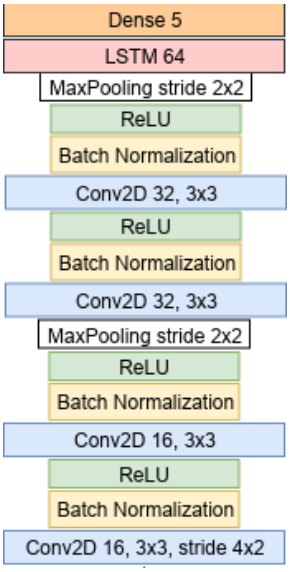
Σχήμα Α΄.2: Διαγραμματική αναπαράσταση του μοντέλου Inflated 3D - Μέρος $2^ο$



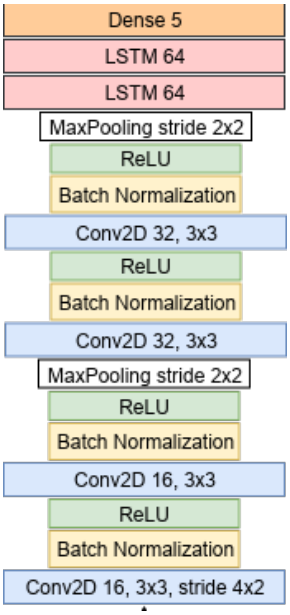
Σχήμα Α΄.3: Διαγραμματική αναπαράσταση του μοντέλου Inflated 3D - Μέρος $3^ο$

Παράρτημα Β΄

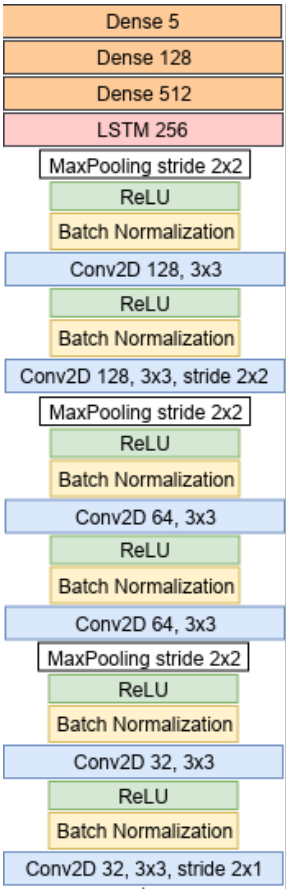
Διαγραμματική Αναπαράσταση των Μοντέλων CNN-LSTM



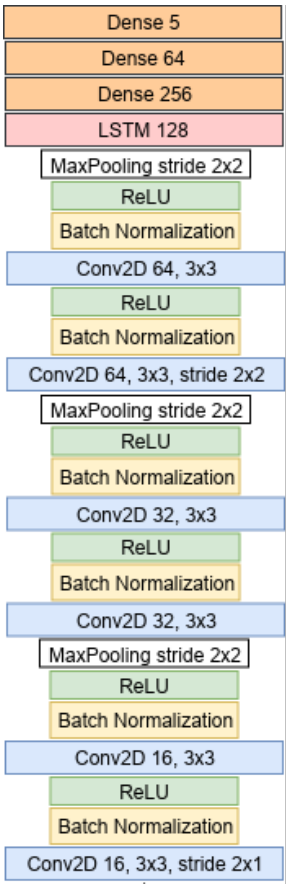
Σχήμα Β΄.1: Αρχιτεκτονική CNN-LSTM: 4conv-1lstm-0dense



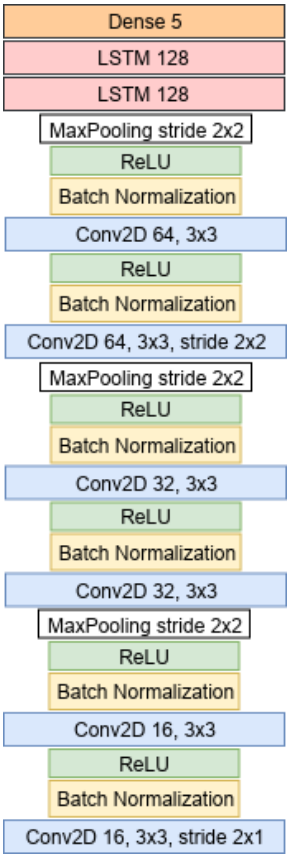
Σχήμα Β΄.2: Αρχιτεκτονική CNN-LSTM: 4conv-2lstm-0dense



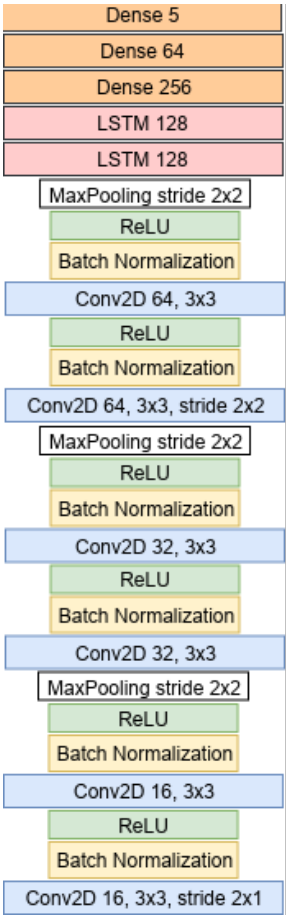
Σχήμα Β΄.3: Αρχιτεκτονική CNN-LSTM: 6conv-1lstm-2dense-wide



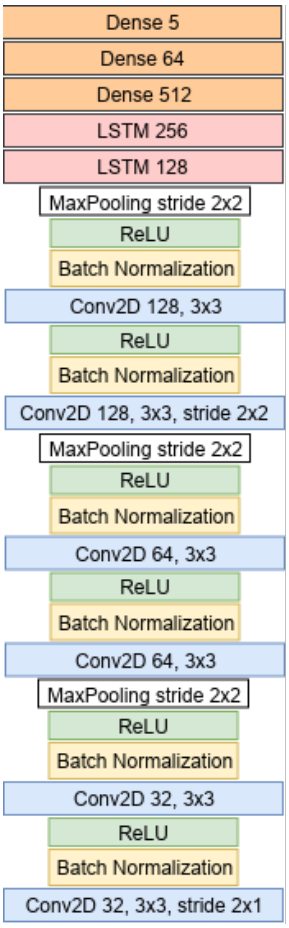
Σχήμα Β΄.4: Αρχιτεκτονική CNN-LSTM: 6conv-1lstm-2dense



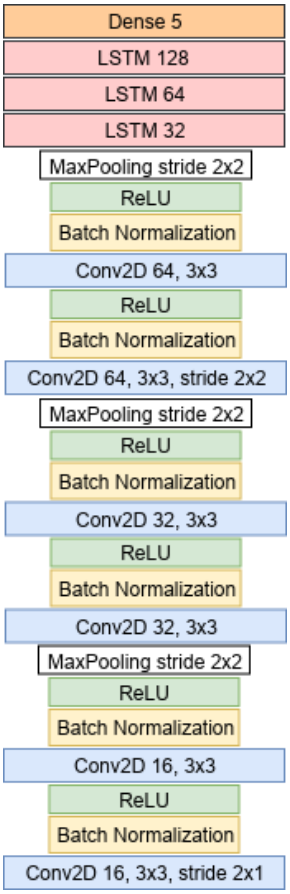
Σχήμα Β΄.5: Αρχιτεκτονική CNN-LSTM: 6conv-2lstm-0dense



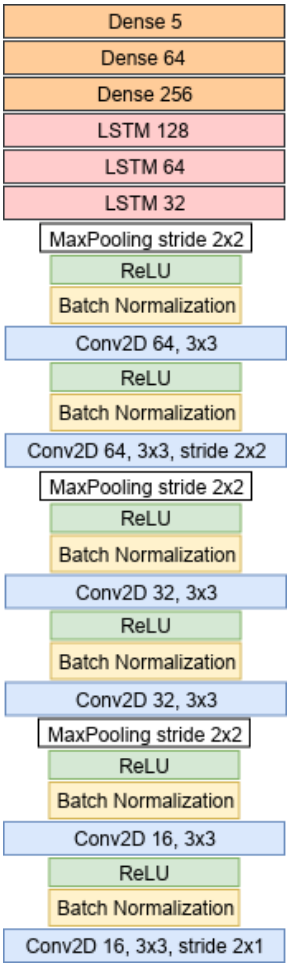
Σχήμα Β΄.6: Αρχιτεκτονική CNN-LSTM: 6conv-2lstm-2dense



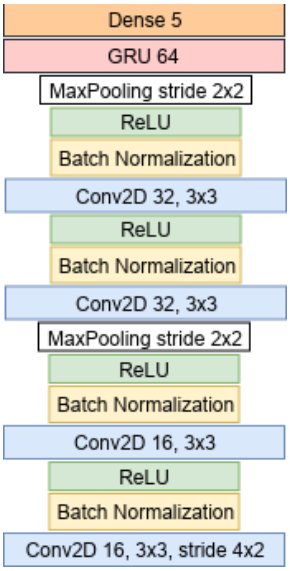
Σχήμα Β΄.7: Αρχιτεκτονική CNN-LSTM: 6conv-2lstm-2dense-wide



Σχήμα Β΄.8: Αρχιτεκτονική CNN-LSTM: 6conv-3lstm-0dense



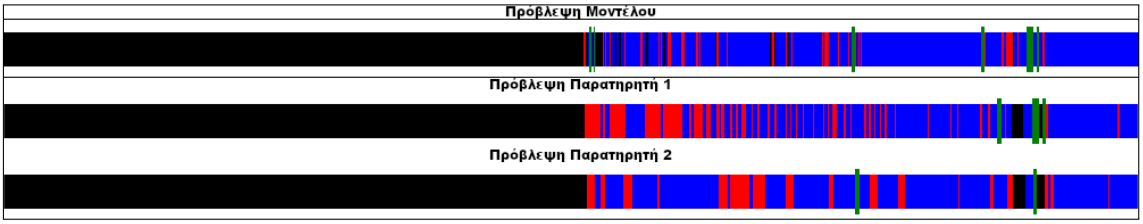
Σχήμα Β΄.9: Αρχιτεκτονική CNN-LSTM: 6conv-3lstm-2dense



Σχήμα Β΄.10: Αρχιτεκτονική CNN-LSTM: 4conv-gru-0dense

Παράρτημα Γ΄

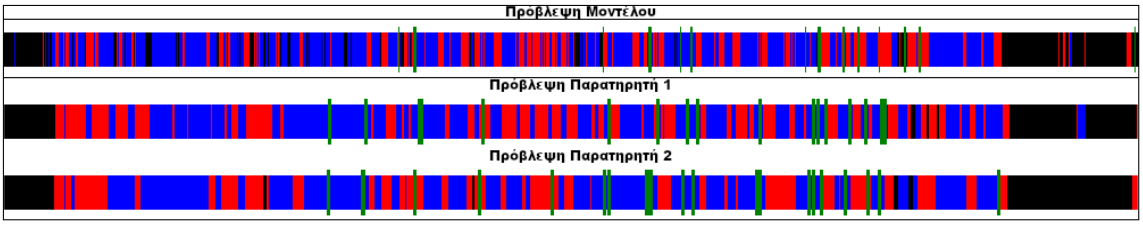
Αποτελέσματα του μοντέλου I3D



Σχήμα Γ΄.1: Αποτελέσματα Πειράματος 4C-2012-F23

Video: 4C-2012-F23	Immobility	Swimming	Climbing	Head Shake	Diving
Μοντέλο I3D	122	14	162	6	0
Παρατηρητής 1	99	39	158	5	0
Παρατηρητής 2	109	29	160	2	0

Πίνακας Γ΄.1: Πίνακας συνολικών δευτερολέπτων ανά κατηγορία πειράματος 4C-2012-F23



Σχήμα Γ΄.2: Αποτελέσματα Πειράματος 4C-2012-F32

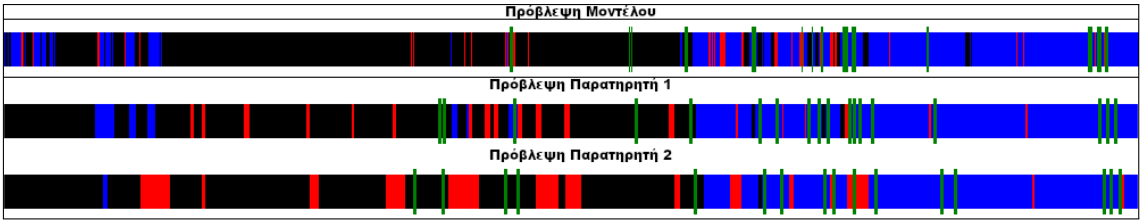
Video: 4C-2012-F32	Immobility	Swimming	Climbing	Head Shake	Diving
Μοντέλο I3D	151	78	66	7	0
Παρατηρητής 1	140	98	47	18	0
Παρατηρητής 2	139	92	51	22	0

Πίνακας Γ΄.2: Πίνακας συνολικών δευτερολέπτων ανά κατηγορία πειράματος 4C-2012-F32

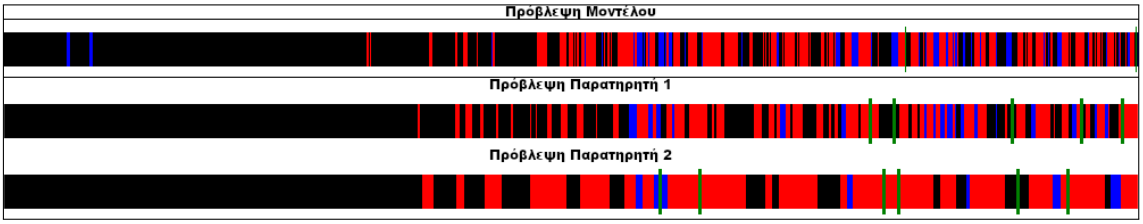
Σχήμα Γ΄.3: Αποτελέσματα Πειράματος 4C-2012-F41

Video: 4C-2012-F41	Immobility	Swimming	Climbing	Head Shake	Diving
Μοντέλο I3D	234	8	60	7	0
Παρατηρητής 1	234	12	54	18	0
Παρατηρητής 2	239	10	50	22	0

Πίνακας Γ΄.3: Πίνακας συνολικών δευτερολέπτων ανά κατηγορία πειράματος 4C-2012-F41



Σχήμα Γ΄.4: Αποτελέσματα Πειράματος 4C-F8



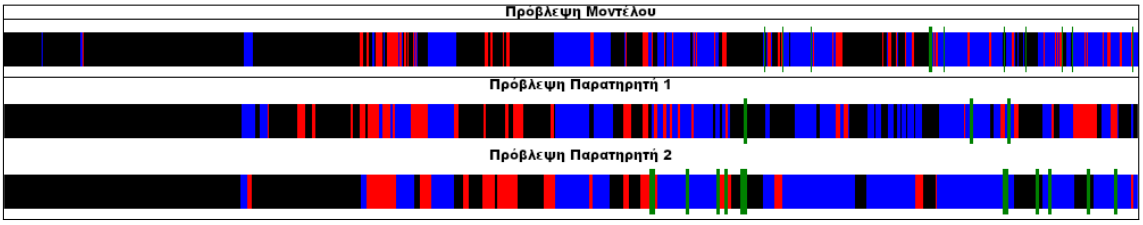
Σχήμα Γ΄.5: Αποτελέσματα Πειράματος 4C-F18

Video: 4C-F18	Immobility	Swimming	Climbing	Head Shake	Diving
Μοντέλο I3D	24	84	192	0	0
Παρατηρητής 1	18	78	198	5	0
Παρατηρητής 2	12	121	161	6	0

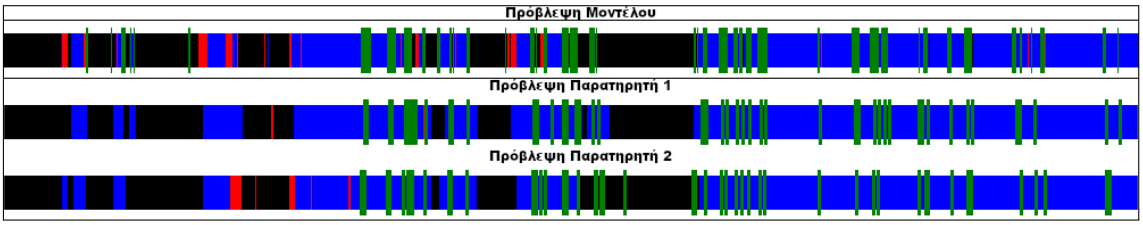
Πίνακας Γ΄.4: Πίνακας συνολικών δευτερολέπτων ανά κατηγορία πειράματος 4C-F18

Video: 4C-F29	Immobility	Swimming	Climbing	Head Shake	Diving
Μοντέλο I3D	97	30	171	4	0
Παρατηρητής 1	95	44	158	3	0
Παρατηρητής 2	114	38	137	13	0

Πίνακας Γ΄.5: Πίνακας συνολικών δευτερολέπτων ανά κατηγορία πειράματος 4C-F29



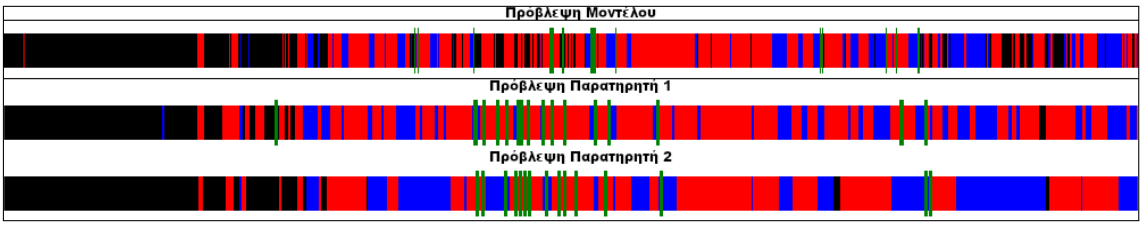
Σχήμα Γ΄.6: Αποτελέσματα Πειράματος 4C-F29



Σχήμα Γ΄.7: Αποτελέσματα Πειράματος 4C-F30

Video: 4C-F30	Immobility	Swimming	Climbing	Head Shake	Diving
Μοντέλο I3D	141	13	99	48	0
Παρατηρητής 1	164	1	93	50	0
Παρατηρητής 2	157	6	96	48	0

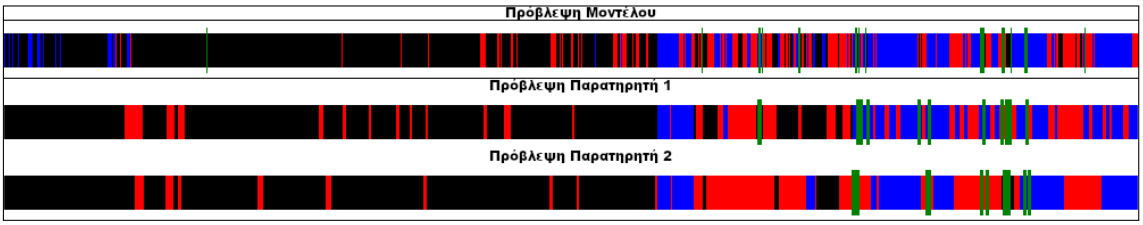
Πίνακας Γ΄.6: Πίνακας συνολικών δευτερολέπτων ανά κατηγορία πειράματος 4C-F30



Σχήμα Γ΄.8: Αποτελέσματα Πειράματος 4C-F31

Video: 4C-F31	Immobility	Swimming	Climbing	Head Shake	Diving
Μοντέλο I3D	55	136	106	15	0
Παρατηρητής 1	64	156	105	22	0
Παρατηρητής 2	87	122	126	22	0

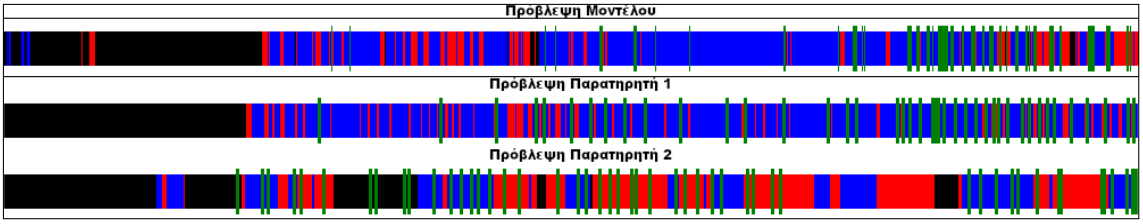
Πίνακας Γ΄.7: Πίνακας συνολικών δευτερολέπτων ανά κατηγορία πειράματος 4C-F31



Σχήμα Γ΄.9: Αποτελέσματα Πειράματος 4C-F37

Video: 4C-F37	Immobility	Swimming	Climbing	Head Shake	Diving
Μοντέλο I3D	70	56	169	6	0
Παρατηρητής 1	51	63	176	12	0
Παρατηρητής 2	49	71	171	11	0

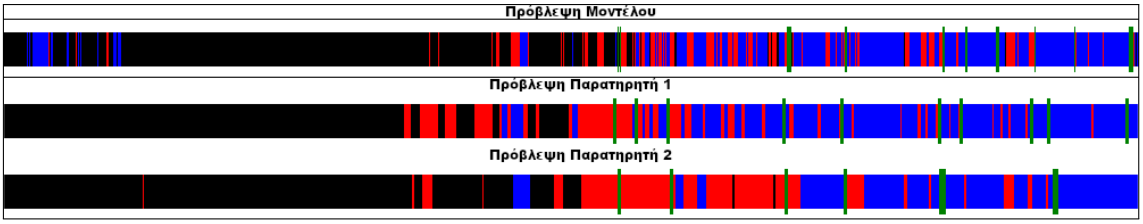
Πίνακας Γ΄.8: Πίνακας συνολικών δευτερολέπτων ανά κατηγορία πειράματος 4C-F37



Σχήμα Γ΄.10: Αποτελέσματα Πειράματος 4C-F39

Video: 4C-F39	Immobility	Swimming	Climbing	Head Shake	Diving
Μοντέλο I3D	159	48	70	26	0
Παρατηρητής 1	161	37	65	44	0
Παρατηρητής 2	77	96	83	52	0

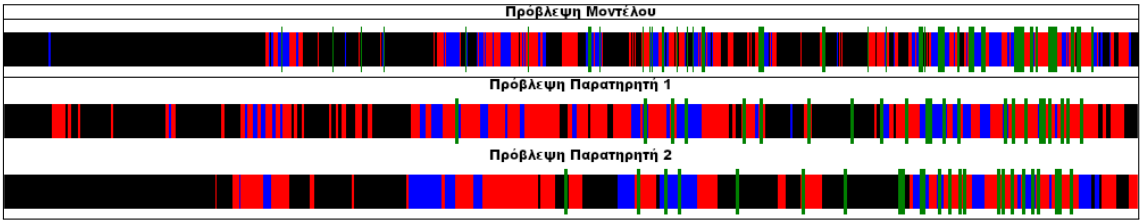
Πίνακας Γ΄.9: Πίνακας συνολικών δευτερολέπτων ανά κατηγορία πειράματος 4C-F39



Σχήμα Γ΄.11: Αποτελέσματα Πειράματος 4C-F45

Video: 4C-F45	Immobility	Swimming	Climbing	Head Shake	Diving
Μοντέλο I3D	105	43	148	6	0
Παρατηρητής 1	109	55	127	11	0
Παρατηρητής 2	83	67	143	8	0

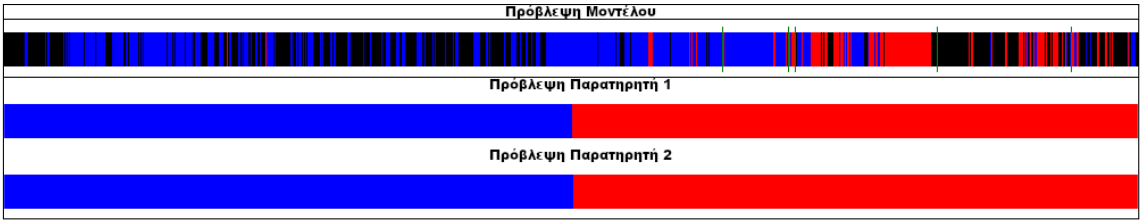
Πίνακας Γ΄.10: Πίνακας συνολικών δευτερολέπτων ανά κατηγορία πειράματος 4C-F45



Σχήμα Γ΄.12: Αποτελέσματα Πειράματος 4C-F46

Video: 4C-F46	Immobility	Swimming	Climbing	Head Shake	Diving
Μοντέλο I3D	42	71	167	25	0
Παρατηρητής 1	40	122	118	24	0
Παρατηρητής 2	44	88	147	25	0

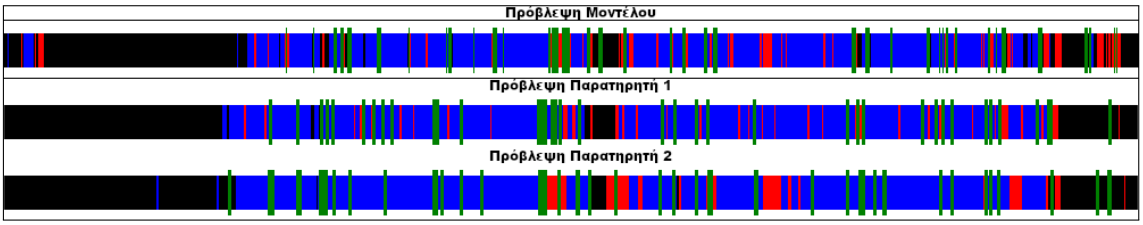
Πίνακας Γ΄.11: Πίνακας συνολικών δευτερολέπτων ανά κατηγορία πειράματος 4C-F46



Σχήμα Γ΄.13: Αποτελέσματα Πειράματος 4C-M7

Video: 4C-M7	Immobility	Swimming	Climbing	Head Shake	Diving
Μοντέλο I3D	133	41	129	1	0
Παρατηρητής 1	150	149	1	0	0
Παρατηρητής 2	150	149	1	0	0

Πίνακας Γ΄.12: Πίνακας συνολικών δευτερολέπτων ανά κατηγορία πειράματος 4C-M7



Σχήμα Γ΄.14: Αποτελέσματα Πειράματος 4C-M15

Video: 4C-M15	Immobility	Swimming	Climbing	Head Shake	Diving
Μοντέλο I3D	156	30	89	30	0
Παρατηρητής 1	158	20	87	40	0
Παρατηρητής 2	154	25	83	45	0

Πίνακας Γ΄.13: Πίνακας συνολικών δευτερολέπτων ανά κατηγορία πειράματος 4C-M15

Κατάλογος σχημάτων

Κατάλογος πινάκων

Βιβλιογραφία

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems 25* (F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds.), pp. 1097–1105, Curran Associates, Inc., 2012.

[2] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv:1409.1556 [cs]*, Sept. 2014.

[3] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (Boston, MA, USA), pp. 1–9, IEEE, June 2015.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *arXiv:1512.03385 [cs]*, Dec. 2015.

[5] I. Laptev, "On Space-Time Interest Points," *International Journal of Computer Vision*, vol. 64, pp. 107–123, Sept. 2005.

[6] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior Recognition via Sparse Spatio-Temporal Features," in *2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, (Beijing, China), pp. 65–72, IEEE, 2005.

[7] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *CVPR 2011*, (Colorado Springs, CO, USA), pp. 3169–3176, IEEE, June 2011.

[8] H. Wang and C. Schmid, "Action Recognition with Improved Trajectories," in *2013 IEEE International Conference on Computer Vision*, (Sydney, Australia), pp. 3551–3558, IEEE, Dec. 2013.

[9] K. Simonyan and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos," *arXiv:1406.2199 [cs]*, June 2014.

[10] J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond Short Snippets: Deep Networks for Video Classification," *arXiv:1503.08909 [cs]*, Mar. 2015.

[11] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-term Recurrent Convolutional Networks for Visual Recognition and Description," *arXiv:1411.4389 [cs]*, Nov. 2014.

[12] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks," *arXiv:1412.0767 [cs]*, Dec. 2014.

[13] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," *arXiv:1705.07750 [cs]*, May 2017.

[14] R. D. Porsolt, A. Bertin, and M. Jalfre, "Behavioral despair in mice: A primary screening test for antidepressants.," *Archives internationales de pharmacodynamie et de therapie*, vol. 229, pp. 327–336, Oct. 1977.

[15] R. D. Porsolt, M. L. Pichon, and M. Jalfre, "Depression: A new animal model sensitive to antidepressant treatments," *Nature*, vol. 266, p. 730, Apr. 1977.

[16] R. D. Porsolt, G. Anton, N. Blavet, and M. Jalfre, "Behavioural despair in rats: A new model sensitive to antidepressant treatments," *European Journal of Pharmacology*, vol. 47, pp. 379–391, Feb. 1978.

[17] M. J. Detke, M. Rickels, and I. Lucki, "Active behaviors in the rat forced swimming test differentially produced by serotonergic and noradrenergic antidepressants," *Psychopharmacology*, vol. 121, pp. 66–72, Sept. 1995.

[18] J. F. Cryan, A. Markou, and I. Lucki, "Assessing antidepressant activity in rodents: Recent developments and future needs," *Trends in Pharmacological Sciences*, vol. 23, pp. 238–245, May 2002.

[19] G. Farnebäck, "Two-Frame Motion Estimation Based on Polynomial Expansion," in *Image Analysis* (G. Goos, J. Hartmanis, J. van Leeuwen, J. Bigun, and T. Gustavsson, eds.), vol. 2749, pp. 363–370, Berlin, Heidelberg: Springer Berlin Heidelberg, 2003.

[20] J. S. Pérez, E. Meinhardt-Llopis, and G. Facciolo, "TV-L1 Optical Flow Estimation," *Image Processing On Line*, vol. 3, pp. 137–150, July 2013.

[21] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, (San Diego, CA, USA), pp. 886–893, IEEE, 2005.

[22] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, (Anchorage, AK, USA), pp. 1–8, IEEE, June 2008.

[23] N. Dalal, B. Triggs, and C. Schmid, "Human Detection Using Oriented Histograms of Flow and Appearance," in *Computer Vision – ECCV 2006* (A. Leonardis, H. Bischof, and A. Pinz, eds.), vol. 3952, pp. 428–441, Berlin, Heidelberg: Springer Berlin Heidelberg, 2006.

[24] D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, F. Perronnin, J. Sánchez, and T. Mensink, "Improving the Fisher Kernel for Large-Scale Image Classification," in *Computer Vision – ECCV 2010* (K. Daniilidis, P. Maragos, and N. Paragios, eds.), vol. 6314, pp. 143–156, Berlin, Heidelberg: Springer Berlin Heidelberg, 2010.

[25] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image Classification with the Fisher Vector: Theory and Practice," *International Journal of Computer Vision*, vol. 105, pp. 222–245, Dec. 2013.

[26] S. Brahmbhatt, "What is a Fisher Vector? - Quora," 2017.

[27] C. Olah, "Understanding LSTM Networks."

[28] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," *arXiv:1406.1078 [cs, stat]*, June 2014.

[29] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," p. 8.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," in *2015 IEEE International Conference on Computer Vision (ICCV)*, (Santiago, Chile), pp. 1026–1034, IEEE, Dec. 2015.

[31] J. Kiefer and J. Wolfowitz, "Stochastic Estimation of the Maximum of a Regression Function," *The Annals of Mathematical Statistics*, vol. 23, pp. 462–466, Sept. 1952.

[32] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv:1412.6980 [cs]*, Dec. 2014.

[33] I. Sutskever, J. Martens, and G. Dahl, "On the importance of initialization and momentum in deep learning," p. 9.

[34] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," p. 30.

[35] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *arXiv:1502.03167 [cs]*, Feb. 2015.

[36] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional Sift Descriptor and Its Application to Action Recognition," in *Proceedings of the 15th ACM International Conference on Multimedia*, MM '07, (New York, NY, USA), pp. 357–360, ACM, 2007.

[37] A. Klaeser, M. Marszalek, and C. Schmid, "A Spatio-Temporal Descriptor Based on 3D-Gradients," in *Procedings of the British Machine Vision Conference 2008*, (Leeds), pp. 99.1–99.10, British Machine Vision Association, 2008.

[38] B. D. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision (IJCAI)," in *[No Source Information Available]*, vol. 81, Apr. 1981.

[39] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild," *arXiv:1212.0402 [cs]*, Dec. 2012.

[40] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," p. 8.

[41] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The Kinetics Human Action Video Dataset," *arXiv:1705.06950 [cs]*, May 2017.

[42] N. Kokras, K. Antoniou, H. G. Mikail, V. Kafetzopoulos, Z. Papadopoulou-Daifoti, and C. Dalla, "Forced swim test: What about females?," *Neuropharmacology*, vol. 99, pp. 408–421, Dec. 2015.

[43] N. Kokras, D. Baltas, F. Theocharis, and C. Dalla, "Kinoscope: An Open-Source Computer Program for Behavioral Pharmacologists," *Frontiers in Behavioral Neuroscience*, vol. 11, May 2017.

[44] anenbergb, "Improved Dense Trajectories." https://github.com/anenbergb/CS221_Project/, 2014.

[45] F. Chollet, "Keras," *GitHub repository*, 2015.

[46] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," p. 19.

[47] dlpbc, "Keras-kinetics-i3d." https://github.com/dlpbc/keras-kinetics-i3d, 2017.