Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων

Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Αναγνώριση ανθρώπινων ενεργειών σε βίντεο με την χρήση Βαθιών Νευρωνικών δικτύων

Διπλωματικη Εργασια

της

ΣΤΑΜΟΥ ΦΙΛΟΜΕΝΑΣ

Επιβλέπων: Ανδρέας Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Εργαστηριο Ευφυων Συστηματων

Αθήνα, Νοέμβριος 2019

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Ευφυών Συστημάτων

Αναγνώριση ανθρώπινων ενεργειών σε βίντεο με την χρήση Βαθιών Νευρωνικών δικτύων

Διπλωματικη Εργασια

της

ΣΤΑΜΟΥ ΦΙΛΟΜΕΝΑΣ

Επιβλέπων: Ανδρέας Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 19η Νοεμβρίου 2019.

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

Αθήνα, Νοέμβριος 2019

(Υπογραφή)

.....

Σταμου Φιλομενα

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

©2019 – All rights reserved Στάμου Φιλομένα, 2019.
Με επιφύλαξη παντός δικαιώματος.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα. Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Ευφυών Συστημάτων

Περίληψη

Το θέμα της παρούσας διπλωματικής εργασίας είναι η *Αναγνώριση Ανθρώπινων Ενεργειών* σε ψηφιακά βίντεο (*Video Action Recognition*) κάνοντας χρήση τεχνικών της *Βαθιάς Μηχανικής Μάθησης* (*Deep Learning*). Το συγκεκριμένο πρόβλημα έχει βρεθεί στο επίκεντρο σημαντικών επιστημονικών και ερευνητικών προσπαθειών κατά τη διάρκεια των τελευταίων χρόνων, χάρη στην εφαρμογή που βρίσκει σε ένα ευρύ φάσμα τομέων.

Καθημερινά προκύπτει ένας τεράστιος όγκος ψηφιακών δεδομένων, με αποτέλεσμα να κρίνεται αναγκαία η βαθύτερη κατανόηση της δομής τους και η ανακάλυψη τρόπων επεξεργασίας και εξαγωγής χρήσιμης γνώσης από αυτά. Η πληροφορία που περιέχεται σε ένα ψηφιακό βίντεο μπορεί να φανεί χρήσιμη σε κλάδους όπως η παρακολούθηση χώρων μέσω κάμερας (*video surveillance*), η αυτόματη οδήγηση (*self-driving cars*) ή η αλληλεπίδραση μεταξύ ανθρώπου-υπολογιστή (*human-computer interaction*).

Προκειμένου να προσεγγίσουμε το περιεχόμενο του *Video Action Recognition*, αρχικά παρουσιάζουμε ένα σύνολο μεθόδων και αρχιτεκτονικών που έχουν χρησιμοποιηθεί για την επίλυση του προβλήματος. Εστιάζουμε την προσοχή μας στις τεχνικές που προέρχονται από τον χώρο της Βαθιάς Μηχανικής Μάθησης και μελετάμε τις επιδόσεις που έχουν σημειώσει.

Στο Κεφάλαιο 5 του εγγράφου υλοποιούμε το δικό μας μοντέλο αναγνώρισης ενεργειών σε βίντεο, το οποίο είναι βασισμένο στα *Συνελικτικά Νευρωνικά Δίκτυα (CNN)* και στα δίκτυα Δύο-Ρευμάτων (*Two-Stream Networks*). Χρησιμοποιούμε τα 13320 δεδομένα βίντεο που περιέχονται στο *dataset UCF-101*, τα επεξεργαζόμαστε και εξάγουμε τα χαρακτηριστικά τους, προκειμένου να καταλήξουμε σε μία πρόβλεψη σχετικά με την αναπαριστούμενη ενέργεια του κάθε βίντεο.

Λέξεις Κλειδιά

Αναγνώριση ενεργειών σε βίντεο, όραση υπολογιστών, βαθιά μηχανική μάθηση, συνελικτικά νευρωνικά δίκτυα, οπτική ροή, δίκτυο δύο ρευμάτων.

Abstract

This diploma thesis deals with *Video Action Recognition* utilizing *Deep Learning* techniques. Video activity recognition, although being an emerging task, has been the subject of important research efforts due to the importance of its everyday applications.

The huge amount of data that are generated on an everyday basis has encouraged the research community to better investigate videos and to develop ways in order to exclude valuable knowledge through data (*Data Mining*). This field is useful to a number of applications, such as *video-surveillance*, *self-driving cars* and *human-computer interaction*.

Activity recognition consists of identifying some actions from a series of observations. As part of the document, we discuss about the main techniques used for activity recognition in computer vision, namely *Video-based Activity Recognition* focusing on the state-of-the-art methods while at the same time mentioning other techniques used for the same task that the research community has known for several years. For each of the analyzed models, its contribution over previous works an the proposed approach performance are examined.

On the Chapter 5 of this paper we try to implement a video action recognition technique that uses *Deep Convolutional Neural Networks (CNN)* and combines both *spatial* and *temporal* information from video frames. We present all the preprocessing that is applied to out data (*dataset UCF-101*) prior to feeding them into our model and the results of our predictions.

Keywords

Video action recognition, computer vision, deep learning, convolutional neural networks, optical flow, two-stream network.

Ευχαριστίες

Η διπλωματική μου εργασία πραγματοποιήθηκε σε συνεργασία με το *Εργαστήριο Ευφυών Συστημάτων* του τομέα *Τεχνολογίας Πληροφορικής και Υπολογιστών* του Εθνικού Μετσόβιου Πολυτεχνείου. Αρχικά, θα ήθελα να ευχαριστήσω τον καθηγητή κ.*Ανδρέα Σταφυλοπάτη*, ο οποίος μου ανέθεσε το θέμα της εν λόγω εργασίας και υπήρξε ο επιβλέπων μου καθόλη τη διάρκεια της έρευνάς μου.

Οφείλω ένα μεγάλω ευχαριστώ στον διδάκτορα κ.*Γεώργιο Σιόλα*, ο οποίος μου έδωσε το κίνητρο να ασχοληθώ με τον τομέα τον Νευρωνικών Δικτύων και με στήριξε σε όλα τα στάδια εκπόνησης της διπλωματικής. Η άψογη συμπεριφορά και η διαρκής διάθεσή του για παροχή βοήθειας υπήρξαν ενισχυτικές και καθοριστικές για το τελικό αποτέλεσμα της εργασίας.

Ακολούθως, θα ήθελα να ευχαριστήσω τους καθηγητές κ.*Γεώργιο Στάμου* για τη μεγάλη εμπιστοσύνη που μου έχει δείξει κατά τη διάρκεια των σπουδών μου και για τη συνέχεια αυτών, καθώς και τον κ.*Παναγιώτη Τσανάκα*, ο οποίος συμμετέχει στην τριμελή επιτροπή έγκρισης της διπλωματικής εργασίας.

Η ενασχόληση με το αντικείμενο της *'Αναγνώρισης Ανθρώπινων Ενεργειών σε βίντεο'* συντελέστηκε σε μία ιδιαίτερα ασταθή περίοδο της ζωής μου και με βοήθησε να ανακαλύψω το ενδιαφέρον μου για τον κλάδο της Μηχανικής Μάθησης. Ευχαριστώ όλους όσους ενδιαφέρθηκαν για την πρόοδο της έρευνάς μου και με ενθάρρυναν να τη συνεχίσω.

Τέλος, ευχαριστώ τον αναγνώστη αυτού του εγγράφου, για τον χρόνο που αφιερώνει και την προσοχή που δείχνει.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

Σύμφωνα με τον Turaga (βλέπε [62]), ως '*ανθρώπινη ενέργεια (activity)*' ορίζουμε απλά μοτίβα κίνησης, τα οποία συνήθως εκτελούνται από ένα ή περισσότερα άτομα και διαρκούν κάποια χρονική διάρκεια. Αυτά τα μοτίβα κίνησης ενδέχεται να περιλαμβάνουν ακολουθίες πράξεων, όπως αλληλεπιδράσεις μεταξύ διαφορετικών ανθρώπων ή μεταξύ ανθρώπων και αντικειμένων του περιβάλλοντός τους. Ο όρος *Activity Recognition* χρησιμοποιείται όταν επιθυμούμε να αναφερθούμε στην οπτική παρατήρηση μιας δραστηριότητας και στην αυτόματη εξαγωγή νοήματος από αυτήν. Ενώ κάτι τέτοιο συμβαίνει αβίαστα στους ανθρώπους μέσω της εμπειρίας τους, τα υπολογιστικά μηχανήματα πρέπει να εκπαιδευτούν προκειμένου να αποκτήσουν την αντίστοιχη ικανότητα.

Στο εισαγωγικό αυτό κεφάλαιο περιγράφουμε το θέμα που μας απασχολεί στην παρούσα εργασία και παρουσιάζουμε τη γενική δομή που ακολουθούμε στο έγγραφο.

1.1 Αντικείμενο της διπλωματικής εργασίας (Video Action Recognition-VAR)

Σκοπός της εργασίας μας είναι η προσέγγιση του προβλήματος της '*Αναγνώρισης Ανθρώπινων Ενεργειών* (Human Activity Recognition)' σε ψηφιακά δεδομένα βίντεο. Πρόκειται για ένα ανοιχτό ερευνητικό αντικείμενο, παρά την αξιοσημείωτη πρόοδο που έχει συντελεστεί, διότι τα αποτελέσματα που έχουν επιτευχθεί μέχρι στιγμής δε συμβαδίζουν με τις υψηλές προδιαγραφές ακριβείας (accuracy). Ουσιαστικά, η εργασία του Action Recognition περιλαμβάνει τον εντοπισμό διαφορετικών ενεργειών σε video clips, όπου οι ενέργειες ενδέχεται να διαδραματίζονται σε ολόκληρη τη διάρκεια του βίντεο ή σε ένα τμήμα του.

Ουσιαστικά, το '*Video Action Recognition-VAR*' μπορεί να αντιμετωπιστεί ως μία φυσική επέκταση του προβλήματος της κατηγοριοποίησης εικόνων (*Image Classification*) σε πολλαπλά πλαίσια (frames). Συναντάμε δύο κύρια είδη συστημάτων που προορίζονται για τον εντοπισμό και την κατηγοριοποίηση ανθρώπινων ενεργειών:

- Πρώτο είδος συστημάτων VAR: γίνεται χρήση *κινητών αισθητήρων* (βλέπε εργασία [68]) ή *φυσιολογικών δεδομένων* (βλέπε εργασία [18]) και στη συνέχεια εφαρμόζονται

ταξινομητές (classifiers) ώστε να προσδιορίσουν το είδος της παρατηρούμενης ενέργειας. Αυτά τα συστήματα υπόσχονται υψηλά ποσοστά ακριβείας αλλά βρίσκουν εφαρμογή σε ένα περιορισμένο πεδίο εφαρμογών.

- Δεύτερο είδος συστημάτων VAR: εδώ αξιοποιούνται *κάμερες* και μοντέλα *ασύρματων ραδιοσυχνοτήτων*, με σκοπό την εξαγωγή χαρακτηριστικών από την οπτική είσοδο, όπως είναι η *θέση* (position), το *σχήμα* (shape) ή το *χρώμα* (color) των αντικειμένων.

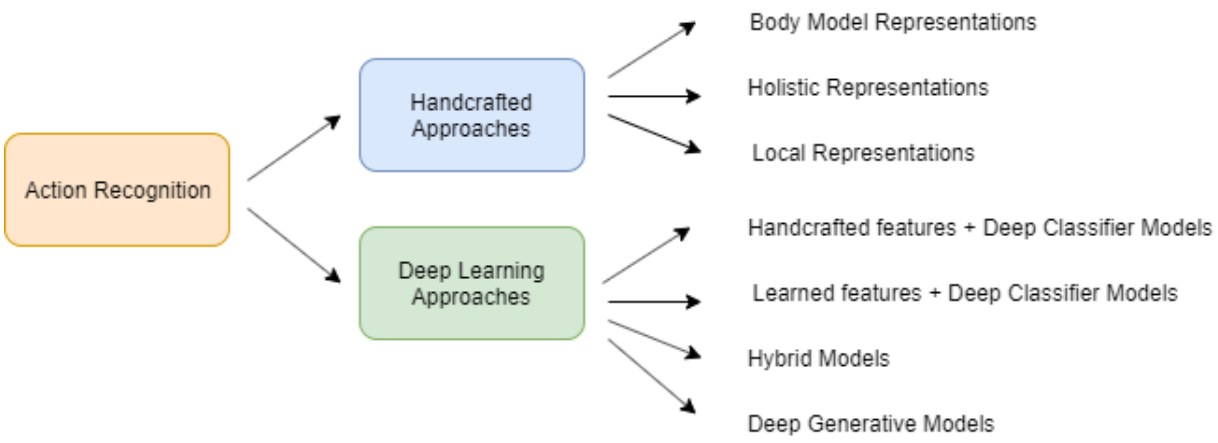
Τα τελευταία χρόνια έχει αυξηθεί η χρήση των συσκευών που καταγράφουν *multi-modal video*, δηλαδή βίντεο που συνδυάζουν κίνηση, ήχο και κείμενο, με αποτέλεσμα να παρέχουν πληροφορία σε μεγαλύτερο βάθος συγκριτικά με την απλή χρωματική πληροφορία που μπορούσαν να καταγράψουν οι κλασικές κάμερες. Αυτά τα συστήματα παρέχουν ακριβείς αναπαραστάσεις του σχήματος του ανθρώπινου σώματος, το οποίο χρησιμοποιείται για τη διαμόρφωση ποικίλων χαρακτηριστικών που αφορούν το σχήμα της ενέργειας (βλέπε εργασία [40]). Μάλιστα, τα συγκεκριμένα συστήματα έχουν χρησιμοποιηθεί από ερευνητές σε προσεγγίσεις βασισμένες σε ιστογράμματα (*histogram-based approaches*) ή στην καταγραφή της ανθρώπινης στάσης σώματος (*human posture*).

Η ανάλυση ενός βίντεο (video analysis) πραγματοποιείται σε διαφορετικά επίπεδα λεπτομέρειας ανάλογα με την πληροφορία που θέλουμε να λάβουμε από αυτό. Για παράδειγμα, ενδέχεται να επεξεργαστούμε κάποιο βίντεο εστιάζοντας την προσοχή μας στα εξής:

- Στις θέσεις ανθρώπων και αντικειμένων.
- Στον τρόπο αλληλεπίδρασης μεταξύ διαφορετικών αντικειμένων.
- Στη στάση του ανθρώπινου σώματος.
- Στη μεταβολή της κίνησης ή της θέσης αντικειμένων και χώρου.

Για αρκετό καιρό, το πρόβλημα της αναγνώρισης ενεργειών σε βίντεο στηριζόταν σε τεχνικές εξαγωγής *handcrafted features* σε συνδυασμό με ταξινομητές (βλέπε εργασία [110]). Ωστόσο, τα τελευταία χρόνια η διαθεσιμότητα μεγάλων ποσοτήτων δεδομένων έχει οδηγήσει στην χρήση Βαθιών Νευρωνικών Δικτύων (*Deep Neural Networks*) για την επίλυση του προβλήματος (βλέπε εργασία [37]). Η επιτυχία της *Βαθιάς Μηχανικής Μάθησης* και συγκεκριμένα των *Συνελικτικών Νευρωνικών Δικτύων* (Convolutional Neural Networks-CNN) στην επίλυση του προβλήματος διαφαίνεται από τα αποτελέσματα που έφεραν όταν εφαρμόστηκαν στο δίκτυο ImageNet (βλέπε εργασία [25]). Στο Σχήμα 1.1 συνοψίζουμε τις διαφορετικές προσεγγίσεις που χρησιμοποιούνται στον κλάδο του Video Action Recognition:

Με την πάροδο του χρόνου, οι προσεγγίσεις που βασίζονται σε handcrafted features συνδυάστηκαν με εφαρμογές πραγματικού χρόνου (real-time applications). Το ενδιαφέρον ενισχύθηκε με την άνθιση των Βαθιών Αρχιτεκτονικών (*Deep Architectures*) το 2012. Μεγάλος αριθμός ερευνητικών προσπαθειών επικεντρώνεται στην εφαρμοσιμότητα αυτών των αρχιτεκτονικών στην Αναγνώριση Ενεργειών και στους πιθανούς συνδυασμούς τους με τις handcrafted προσεγγίσεις.



Σχήμα 1.1: Προσεγγίσεις του Video Action Recognition

1.2 Προκλήσεις ενασχόλησης με τον κλάδο

Κατά τη διαδικασία της *Αναγνώρισης Ενεργειών* συναντάμε αρκετούς περιορισμούς σε σχέση με την εξαγωγή χαρακτηριστικών και την ταξινόμησή τους. Στη συνέχεια, παραθέτουμε ορισμένους από αυτούς:

- Κάθε άνθρωπος εκτελεί την εκάστοτε ενέργεια σύμφωνα με ένα δικό του τρόπο, με αποτέλεσμα σε ορισμένες περιπτώσεις να μην είναι απολύτως προσδιορίσιμη η κατηγορία στην οποία ανήκει κάποια κίνηση.
- Ορισμένα είδη ενεργειών παρουσιάζουν σημαντικές ομοιότητες μεταξύ τους, οπότε δυσχαιρένεται ο διαχωρισμός τους.
- Πιθανές αλλαγές στην οπτική γωνία του παρατηρητή-κάμερας καθορίζουν το πώς μοιάζει η κάθε ενέργεια.
- Η παρουσία σύνθετων αντικειμένων στο παρασκήνιο (background) καθιστούν δύσκολο τον εντοπισμό καθαρών ανρθώπινων μορφών ή σχημάτων
- Η διάρκεια της κάθε ενέργειας συναντά διακυμάνσεις, που σχετίζονται με τον χρόνο που αφιερώνεται για την εκτέλεση ή την καταγραφή της.

1.3 Οργάνωση του εγγράφου

Στην εργασία μας αντιμετωπίζουμε την *Αναγνώριση των Ανθρώπινων Ενεργειών σε βίντεο* ως ένα πρόβλημα που μπορεί να αντιμετωπιστεί αποτελεσματικά αξιοποιώντας τη γνώση των Συνελικτικών Νευρωνικών Δικτύων και των δικτύων Δύο-Ρευμάτων. Αρχικά, στο Κεφάλαιο 2 εισάγουμε τον αναγνώστη στις βασικές αρχές που διέπουν τα Νευρωνικά Δίκτυα και τη Βαθιά Μηχανική Μάθηση, ενώ στο Κεφάλαιο 3 γίνεται μία πρώτη προσέγγιση της ανάλυσης του θέματός μας. Συγκεκριμένα, αναφερόμαστε στις κυριότερες έννοιες που αφορούν την

ανάλυση, επεξεργασία και εξαγωγή γνώσης από τα ψηφιακά δεδομένα βίντεο και παρουσιάζουμε μία χρονική εξέλιξη των τεχνικών που έχουν εφαρμοστεί στον κλάδο του Video Action Recognition.

Στο Κεφάλαιο 4 αναφερόμαστε σε καινοτόμα μοντέλα που έχουν υλοποιηθεί προκειμένου να προσεγγίσουν το πρόβλημα που μελετάμε, ενώ η δική μας υλοποίηση περιγράφεται αναλυτικά στο Κεφάλαιο 5, όπου εξηγούμε το μοντέλο που κατασκευάζουμε, την τεχνική εκπαίδευσης που ακολουθούμε και φυσικά τις προβλέψεις στις οποίες οδηγούμαστε. Τέλος, στο τελευταίο κεφάλαιο (Κεφάλαιο 6) καταλήγουμε στα κυριότερα συμπεράσματα που προκύπτουν από την έρευνα που πραγματοποιήσαμε και παράλληλα προτείνουμε ορισμένες κατευθύνσεις στις οποίες θα μπορούσαμε να κινηθούμε μελλοντικά ώστε να βελτιώσουμε την επίδοση του δικτύου μας.

Κεφάλαιο 2

Θεωρητικό υπόβαθρο της περιοχής των Νευρωνικών Δικτύων

Ο κλάδος της Τεχνητής Νοημοσύνης (*Artificial Intelligence-AI*) ασχολείται με τη σχεδίαση και την υλοποίηση υπολογιστικών συστημάτων (μηχανών) που έχουν την ικανότητα να μιμούνται στοιχεία της ανθρώπινης συμπεριφοράς. Στόχος αυτών των συστημάτων είναι να αναπτύξουν κάποια στοιχεώδη ευφυΐα, δηλαδή να μπορούν να προσαρμοστούν, να εξάγουν συμπεράσματα και να επιλύουν προβλήματα χωρίς την παρέμβαση του ανθρώπινου παράγοντα. Στο πλαίσιο ερευνών της Τεχνητής Νοημοσύνης, γεννήθηκε ο τομέας της Μηχανικής Μάθησης (*Machine Learning-ML*), ο οποίος επικεντρώνεται στην κατασκευή μοντέλων που υλοποιούν συγκεκριμένους αλγορίθμους και χρησιμοποιούν πειραματικά δεδομένα, με σκοπό την εξαγωγή χρήσιμων προβλέψεων ή συμπερασμάτων. Τα μοντέλα που κατεξοχήν βρίσκουν εφαρμογή στις εφαρμογές της Μηχανικής Μάθησης είναι τα Τεχνητά Νευρωνικά Δίκτυα (*Artificial Neural Networks-ANNs*).

Στη συνέχεια, ακολουθεί μία σύντομη περιγραφή των εννοιών, των δομών, του μαθηματικού υποβάθρου και των μοντέλων που θα χρειαστούν για την κατανόηση των επόμενων κεφαλαίων.

2.1 Τεχνητά Νευρωνικά Δίκτυα (ANN)

Τα *Τεχνητά Νευρωνικά Δίκτυα* (Artificial Neural Networks-ANNs, βλέπε εργασίες [36, 74, 33, 71]) αποτελούν ένα αφηρημένο αλγοριθμικό κατασκεύσμα το οποίο εμπίπτει στον κλάδο της Τεχνητής Νοημοσύνης (Artificial Intelligence-AI). Αυτά τα δίκτυα είναι εμπνευσμένα από το Κεντρικό Νευρικό Σύστημα (Central Nervous System-CNS) του ανθρώπου, δηλαδή από τον τρόπο με τον οποίο ο άνθρωπος επεξεργάζεται την πληροφορία και προσπαθούν να προσομοιάσουν τη λειτουργία του. Στόχος τους δηλαδή είναι να συνδυάσουν τον τρόπο σκέψης του ανθρώπινου εγκεφάλου με τον αφηρημένο μαθηματικό τρόπο σκέψης. Χρησιμοποιούνται για την εκτίμηση και προσέγγιση συναρτήσεων οι οποίες δέχονται ένα μεγάλο

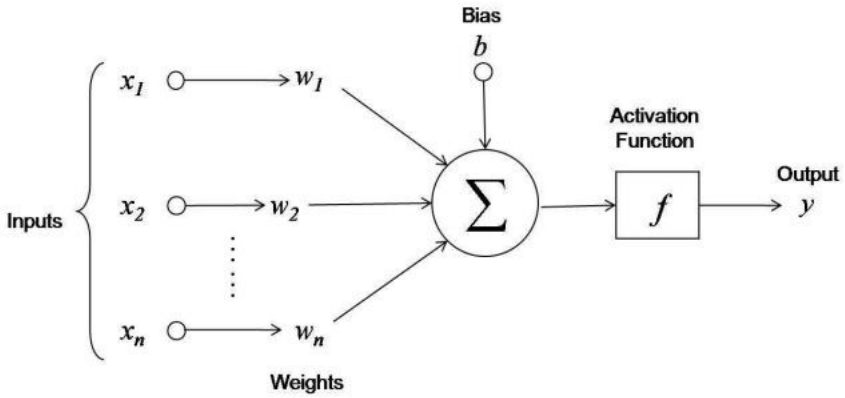
αριθμό δεδομένων εισόδου και έχουν σημειώσει αξιοσημείωτα αποτελέσματα στους τομείς της Αναγνώρισης Εικόνων (*Image Recognition*), της Αναγνώρισης Φωνής (*Speech Recognition*) και της Επεξεργασίας Φυσικής Γλώσσας (*Natural Language Processing-NLP*).

2.1.1 Απλά Νευρωνικά Δίκτυα

Η μοντελοποίηση των Τεχνητών Νευρωνικών Δικτύων γίνεται μέσω ενός συστήματος διασυνδεδεμένων νευρώνων, οι οποίοι ανταλλάσσουν μηνύματα μεταξύ τους. Βάσει του μοντέλου που όρισαν το 1943 οι McCulloch-Pitts (βλέπε εργασία [61]), ο κάθε νευρώνας δέχεται σαν είσοδο ένα διάνυσμα $x = [x_0, x_1, ..., x_n] \in \mathbb{R}^n$ και παράγει μια έξοδο $y \in \mathbb{R}$. Το διάνυσμα εισόδου πολλαπλασιάζεται με ένα διάνυσμα βαρών $W \in \mathbb{R}^n$, οι τιμές του οποίου μεταβάλλονται ανάλογα με την *εμπειρία* του δικτύου. Το αποτέλεσμα αυτού του πολλαπλασιασμού οδηγείται σε μια *μη γραμμική Συνάρτηση Ενεργοποίησης* (Activation Function) από την οποία παράγεται η έξοδος y του δικτύου. Στην Ενότητα 2.2 θα αναλύσουμε τον τρόπο λειτουργίας των πιο συχνά χρησιμοποιούμενων Συναρτήσεων Ενεργοποίησης.

Στο Σχήμα 2.1 παρατηρούμε το γενικό μοντέλο ενός νευρώνα, όπου $x_1, x_2, ..., x_n$ είναι οι έξοδοι των διαφόρων νευρώνων 1,2,...,n, οι οποίες ταυτόχρονα αποτελούν τις εισόδους για άλλους νευρώνες. Τα διάφορα σήματα x_i πολλαπλασιάζονται με ένα συντελεστή βαρύτητας w_i και εντέλει η συνολική έξοδος του νευρώνα j αποτελεί το συνολικό άθροισμα όλων των επιμέρους εισόδων μετά τον πολλαπλασιασμό τους με τους συντελεστές βαρύτητας:

$$S_j = \sum_{i=0}^{n} x_i w_i \tag{2.1}$$



Σχήμα 2.1: Αναπαράσταση Τεχνητού Νευρώνα

Όπως παρατηρούμε στο Σχήμα 2.1, κάθε νευρώνας διαθέτει ένα *συναπτικό βάρος*, το οποίο συμβολίζεται με το γράμμα b και ονομάζεται πόλωση ή κατώφλι (bias,threshold). Η τιμή εισόδου αυτής της σύναψης είναι πάντα η μονάδα και έχει ιδιαίτερη σημασία, διότι καθορίζει την ενεργοποίηση του νευρώνα. Συγκεκριμένα, εάν το συνολικό άθροισμα των υπόλοιπων εισόδων του εκάστοτε νευρώνα είναι μεγαλύτερο από την τιμή της πόλωσης, τότε ο νευρώνας ενεργοποιείται, διαφορετικά παραμένει ανενεργός.

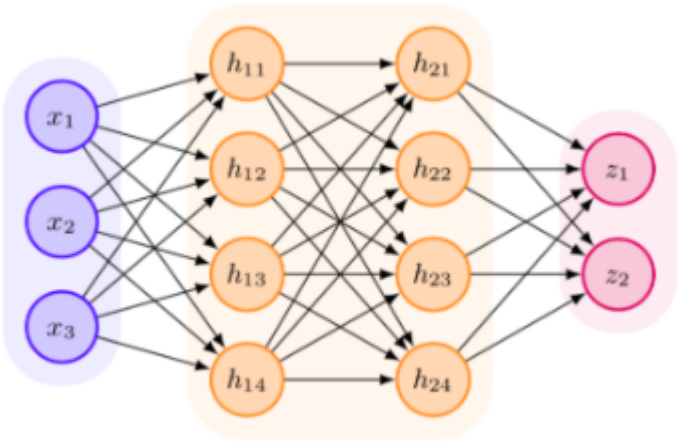
Η συνολική έξοδος του δικτύου υπολογίζεται μέσω της σχέσης:

$$y = f(\sum_{i=0}^{n} x_i w_i + b) \tag{2.2}$$

Τα πιο απλά Νευρωνικά Δίκτυα ονομάζονται *Perceptrons* και προτάθηκαν το 1957 από τον Rosenblatt (βλέπε εργασία [57]). Αποτελούνται από ένα επίπεδο απλών νευρώνων, οι οποίοι λειτουργούν τόσο ως είσοδοι όσο και ως έξοδοι του δικτύου. Κάθε νευρώνας είναι ανεξάρτητος από τους υπόλοιπους, άρα και η εκπαίδευση του κάθε νευρώνα γίνεται ανεξάρτητα από τους υπόλοιπους νευρώνες. Ωστόσο, το 1969 αποδείχθηκε από τους Minsky-Papert (βλέπε εργασία [75]) ότι τα Τεχνητά Νευρωνικά Δίκτυα ενός επιπέδου δεν είναι σε θέση να λύσουν μη γραμμικά προβλήματα, ενώ με έρευνες του 1982 έγινε γνωστό ότι τα Πολυεπίπεδα Νευρωνικά Δίκτυα (Multilayer Perceptrons) μπορούν να προσεγγίσουν οποιαδήποτε συνάρτηση.

2.1.2 Πολυεπίπεδα Νευρωνικά Δίκτυα

Τα Νευρωνικά Δίκτυα που χρησιμοποιούνται στις περισσότερες εφαρμογές βασίζονται σε πολυεπίπεδους νευρώνες Perceptrons και καλούνται *Multi-Layer Perceptrons- MLPs*. Στα MLP οι νευρώνες είναι οργανωμένοι κατά επίπεδα: *επίπεδο εισόδου* (input layer), *κρυφά επίπεδα* (hidden layer) και *επίπεδο εξόδου* (output layer). Στο Σχήμα 2.2 οι νευρώνες εισόδου σημειώνονται με μπλε χρώμα, οι κρυμμένοι νευρώνες με ποπτοκαλί και οι νευρώνες εξόδου με ροζ. Οι κρυμμένοι νευρώνες (hidden neurons) αναλαμβάνουν τον πολλαπλασιασμό της κάθε εισόδου τους με το αντίστοιχο συναπτικό βάρος και την άθροιση όλων των γινομένων. Στο άθροισμα του κάθε νευρώνα προστίθεται ο όρος της πόλωσης (bias) και ακολούθως δίνεται το αποτέλεσμα ως όρισμα στη Συνάρτηση Ενεργοποίησης, η οποία είναι υλοποιημένη εσωτερικά στον κάθε κόμβο και μπορεί να διαφέρει για τον κάθε νευρώνα.



Σχήμα 2.2: Απλό Νευρωνικό Δίκτυο 3 εισόδων και 2 εξόδων

Ανάλογα με τον τρόπο σύνδεσης των νευρώνων μεταξύ τους, διακρίνουμε τις ακόλουθες κατηγορίες Τεχνητών Νευρωνικών Δικτύων:

- **Πλήρως Συνδεδεμένα** (Fully Connected): πρόκειται για δίκτυα στα οποία ο κάθε νευρώνας συνδέεται με όλους τους νευρώνες του επόμενου επιπέδου.
- **Μερικώς Συνδεδεμένα** (Partially Connected): δίκτυα στα οποία υπάρχουν νευρώνες που δε συνδέονται με όλους τους νευρώνες του επόμενου επιπέδου.
- **Εμπρόσθιας Τροφοδότησης** (Feedforward): δίκτυα όπου απουσιάζουν οι συνδέσεις μεταξύ νευρώνων ενός επιπέδου και νευρώνων προηγούμενου επιπέδου.
- **Με Ανατροφοδότηση** (Feedback): όπου σε αντίθεση με την προηγούμενη κατηγορία, οι νευρώνες ενός επιπέδου συνδέονται με νευρώνες προηγούμενου επιπέδου.

2.2 Συναρτήσεις Ενεργοποίησης

Η έξοδος ενός Νευρωνικού Δικτύου μπορεί να είναι ένας οποιοσδήποτε αριθμός, ο οποίος αντιστοιχεί σε κάποια κατηγορία. Δεν τυχαίνει πάντα να προκύπτει αποτέλεσμα εντός του διαστήματος [0,1], αλλά με την εφαρμογή του κατάλληλου πολλαπλασιαστή, δηλαδή της κατάλληλης διαφορίσιμης μη-γραμμικής συνάρτησης, το αποτέλεσμα μεταφέρεται σε οποιοδήποτε διάστημα θελήσουμε, ώστε να καταστεί εύκολη η ερμηνεία του αντίστοιχου αποτελέσματος. Η πρόβλεψη ενός δικτύου δύναται να αντιμετωπιστεί ως πιθανότητα, οπότε μία Συνάρτηση Μεταφοράς Γαυσς θεωρείται Συνάρτηση Πυκνότητας Πιθανότητας.

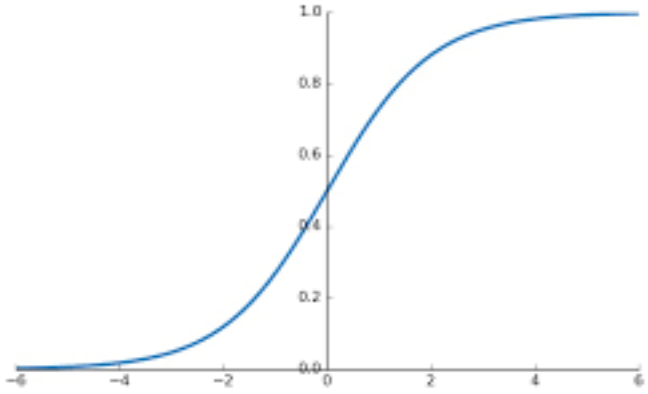
Η Συνάρτηση Ενεργοποίησης μπορεί να είναι *Βηματική* (Step Transfer Function), *Γραμμική* (Linear Transfer Function), *Μη-Γραμμική* (Non-Linear Transfer Function) ή *Στοχαστική* (Stochastic Transfer Function). Οι συναρτήσεις ενεργοποίησης που κατεξοχήν χρησιμοποιούνται στα Νευρωνικά Δίκτυα είναι η *Σιγμοειδής Συνάρτηση* (Sigmoid Function), η *Υπερβολική Εφαπτομένη* (Hyperbolic tangent), η *Μονάδα Γραμμικής Ανόρθωσης* (Rectified Linear Unit-ReLU) και η *Softmax*. Στη συνέχεια, κάνουμε μία σύντομη περιγραφή της καθεμίας από αυτές:

- **Σιγμοειδής Συνάρτηση**: Η εφαρμογή της συγκεκριμένης συνάρτησης ενεργοποίησης εγγυάται ότι οι τιμές της εισόδου αντιστοιχίζονται σε τιμές του διαστήματος [0,1]. Όπως φαίνεται και στο Σχήμα 2.3, οι μικρότερες τιμές αντιστοιχίζονται προσεγγιστικά κοντά στο 0 και οι μεγαλύτερες κοντά στο 1. Αυτό το γεγονός οδηγεί σε απειροελάχιστες τιμές κλίσης, δηλαδή προκαλεί μία εξασθένιση κλίσης (Vanishing Gradients) και έτσι η εκπαίδευση του μοντέλου γίνεται με πολύ αργούς ρυθμούς.

  Η Σιγμοειδής Συνάρτησης υπολογίζεται μέσω του τύπου:

$$f(x) = \sigma(x) = \frac{1}{1 + e^{-x}} \tag{2.3}$$

- **Υπερβολική Εφαπτομένη:** Όπως και η σιγμοειδής συνάρτηση, έτσι και η υπερβολική εφαπτομένη αντιστοιχίζει την είσοδο σε ένα διάστημα [-1,1]. Όπως βλέπουμε στο Σχήμα 2.4, δε μεταβάλλει σημαντικά τις τιμές που βρίσκονται κοντά στο 0, ενώ

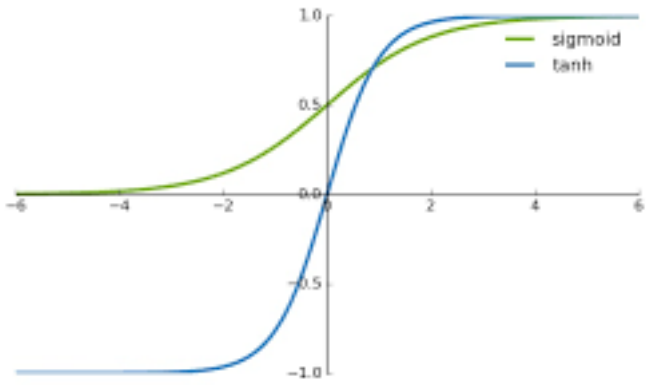


Σχήμα 2.3: Σιγμοειδής Συνάρτηση Ενεργοποίησης

οι μικρές τιμές τείνουν να προσεγγίσουν το -1 και οι μεγάλες το 1. Αυτή η συνάρτηση ενεργοποίησης συναντάει συχνότερα εφαρμογή στα Επαναλαμβανόμενα Νευρωνικά Δίκτυα (Recurrent Neural Network-RNN).

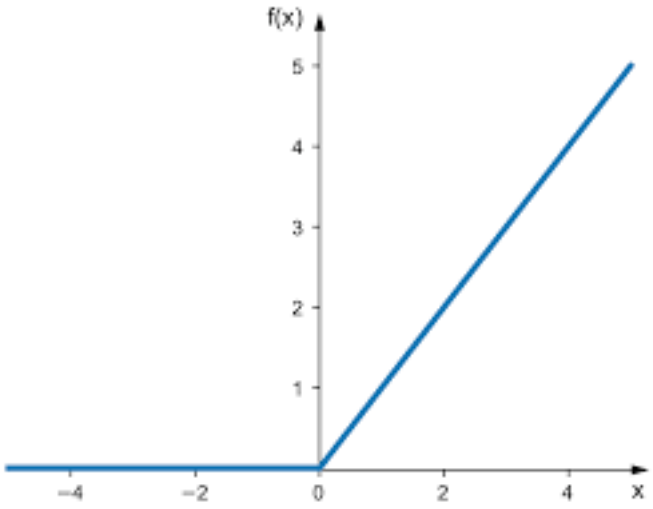
Ο τύπος που δίνει την Υπερβολική Εφαπτομένη είναι ο εξής:

$$f(x) = \tanh(x) = \frac{(e^x - e^{-x})}{(e^x + e^{-x})} \tag{2.4}$$



Σχήμα 2.4: Υπερβολική Εφαπτομένη

- **Μονάδα Γραμμικής Ανόρθωσης**: Πρόκειται για την πιο συχνά χρησιμοποιούμενη συνάρτηση ενεργοποίησης στις σύγχρονες εφαρμογές των Multilayer Perceptrons και των Deep Neural Networks. Όπως φαίνεται στο Σχήμα 2.5, οι αρνητικές τιμές εισόδου μηδενίζονται, με αποτέλεσμα να μη λαμβάνονται καθόλου υπόψη κατά τη διαδικασία εκπαίδευσης. Επίσης, η έξοδος που αντιστοιχεί σε θετικές εισόδους δε διαθέτει κάποιο ανώτατο όριο, γεγονός που καθιστά τη ReLU ιδανική συνάρτηση ενεργοποίησης για προβλήματα στα οποία δεν έχουμε πολύ μεγάλες τιμές εισόδου.



Σχήμα 2.5: Συνάρτηση Ενεργοποίησης ReLU

Η εξίσωση που ορίζει τη συνάρτηση ενεργοποίησης ReLU είναι η ακόλουθη:

$$f(x) = (0, max) \tag{2.5}$$

- **Συνάρτηση Ενεργοποίησης Softmax**: Η συνάρτηση softmax εφαρμόζεται στο στρώμα εξόδου των περισσότερων Τεχνητών Νευρωνικών Δικτύων, ανεξάρτητα από την συνάρτηση ενεργοποίησης που χρησιμοποιείται στους νευρώνες του δικτύου. Ο σκοπός της χρήσης της είναι να κανονικοποιήσει τις τιμές εξόδου ώστε να κατανεμηθούν στο διάστημα [0,1] και να αρθοίζουν στη μονάδα (Σχήμα 2.6). Υπενθυμίζουμε ότι οι νευρώνες του στρώματος εξόδου λαμβάνουν οποιαδήποτε τιμή, η οποία αντιστοιχεί σε κάποια τιμή πιθανότητας.

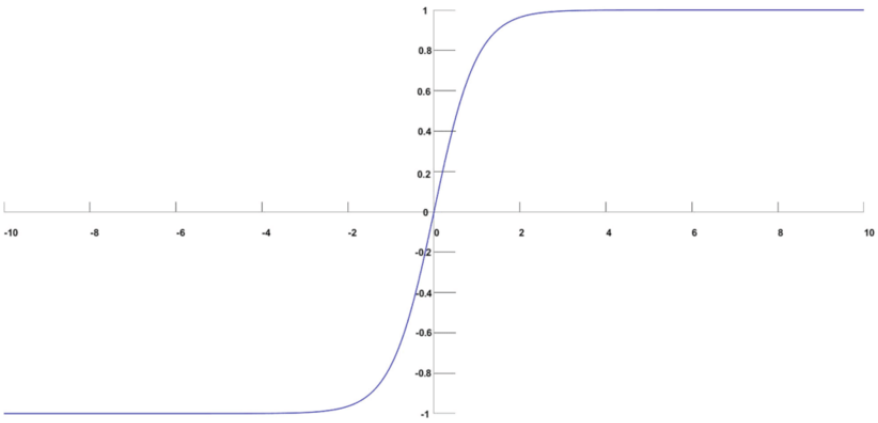
  Η συνάρτηση Softmax δίνεται από τη σχέση:

$$Softmax(y_j) = \frac{e^{y_j}}{\sum_{j=0}^{N} e^{y_j}} \tag{2.6}$$

2.3 Συναρτήσεις Κόστους

Οι Συναρτήσεις Κόστους (Cost Functions) αποτελούν ένα μέτρο της *επίδοσης* των Τεχνητών Νευρωνικών Δικτύων, δεδομένου ενός δείγματος του συνόλου εκπαίδευσης και μίας αναμενόμενης τιμής εξόδου. Μία Συνάρτηση Κόστους δεν είναι διάνυσμα, αλλά μία μεμονωμένη τιμή, επειδή κρίνει τη συμπεριφορά του δικτύου ως ολότητα. Συγκεκριμένα, μία Συνάρτηση Κόστους είναι της μορφής $C(W, B, S^r, E^r)$, όπου W είναι τα βάρη (weights) του δικτύου μας, B είναι οι πολώσεις (biases), S^r είναι η είσοδος ενός δείγματος εκπαίδευσης και E^r είναι η επιθυμητή έξοδος για τη δεδομένη είσοδο.

Μία ευρέως χρησιμοποιούμενη loss function στον κλάδο της Βαθιάς Μηχανικής Μάθησης (Deep Learning) είναι η *Cross-Entropy Loss Function*, την οποία χρησιμοποιούμε και στην



Σχήμα 2.6: Συνάρτηση Ενεργοποίησης Softmax

εκπαίδευση του μοντέλου που κατασκευάζουμε στο πλαίσιο αυτής της εργασίας. Συνεπώς, κρίνουμε αναγαία την ανάλυση του βασικού πλαισίου πίσω από αυτή τη Συνάρτηση Κόστους. Αρχικά, ορίζουμε ως Surprisal-s το βαθμό στον οποίο μένουμε ικανοποιημένοι παίρνοντας ένα συγκεκριμένο αποτέλεσμα από το δίκτυό μας. Όταν η έξοδος του δικτύου έχει χαμηλή πιθανότητα y_i, αυτό σημαίνει ότι αναμένουμε σε μικρό βαθμό τη συγκεκριμένη πρόβλεψη για το δίκτυο, οπότε το s αναμένεται μεγάλο και δίνεται από τη σχέση: $s = log\frac{1}{y_i}$. Γνωρίζοντας την παράμετρο s για ανεξάρτητες εξόδους, θα θέλαμε να γνωρίζουμε την ίδια παράμετρο για ένα ολόκληρο γεγονός (event), οπότε σκεφτόμαστε να πάρουμε το σταθμισμένο μέσο όρο των επιμέρους s. Αυτός ο σταθμισμένος μέσος όρος ονομάζεται Entropy-e και για n εξόδους, υπολογίζεται μέσω της έκφρασης:

$$e = \sum_{0}^{n} y_i log\frac{1}{y_i} \tag{2.7}$$

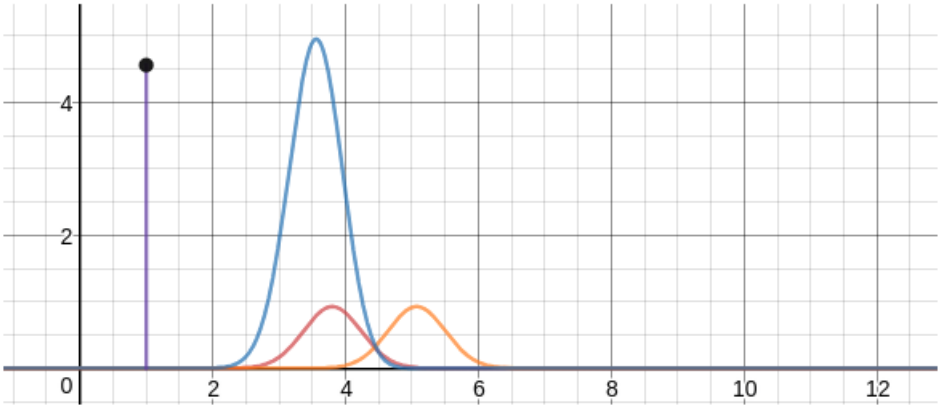
Έπειτα, θεωρώντας ότι η πραγματική πιθανότητα της κάθε εξόδου είναι p_i αλλά η επιθυμητή πιθανότητα είναι q_i, κάθε γεγονός συμβαίνει με πιθανότητα p_i και το surprisal ταυτίζεται με το q_i. Έτσι, καταλήγουμε στον τύπο υπολογισμού του Cross Entropy:

$$c = \sum_{0}^{n} p_i log\frac{1}{q_i} \tag{2.8}$$

Στο Σχήμα 2.7, με πορτοκαλί χρώμα απεικονίζεται η εκτιμούμενη κατανομή πιθανότητας, ενώ με κόκκινο η πραγματική κατανομή πιθανότητας. Παρατηρούμε ότι όσο η εκτιμούμενη κατανομή πιθανότητας απομακρύνεται από την πραγματική κατανομή, το *Cross Entropy* αυξάνεται και αντίστροφα. Άρα, είναι βάσιμο να συμπεράνουμε ότι ελαχιστοποιώντας το *Cross Entropy* μετακινούμαστε πιο κοντά στην πραγματική κατανομή πιθανότητας και αυτός είναι άλλωστε ένας λόγος για τον οποίο προσπαθούμε να μειώσουμε το *Cross Entropy* και η αναμενόμενη κατανομή να μην πλησιάζει στην πραγματική.

2.3.1 Αλγόριθμος Πίσω Διάδοσης Σφάλματος

Ο αλγόριθμος Πίσω Διάδοσης Σφάλματος (*Backpropagation algorithm*) αποτελεί έναν αλγόριθμο βελτιστοποίησης που χρησιμοποιείται ευρέως σε προβλήματα Μηχανικής Μάθη-



Σχήμα 2.7: Συνάρτηση Κόστους Cross-Entropy

σης. Οι αρχές στις οποίες βασίζεται προέρχονται από τη θεωρία ελέγχου που δημοσιεύθηκε το 1960 από τον Henry J. Kelley (βλέπε εργασία [51]). Χρησιμοποιείται ευρέως στα νευρωνικά δίκτυα Εμπρόσθιας Τροφοδότησης (*Feedforward neural networks*) και συγκεκριμένα σε εφαρμογές της Επιβλεπόμενης Μάθησης (*Supervised Learning*). Η μέθοδος της μάθησης με οπισθοδιάδοση λάθους *δε συγκλίνει πάντα* στη βέλτιστη λύση. Ο λόγος που συμβαίνει αυτό σχετίζεται με την αναζήτηση τύπου 'αναρρίχησης λόφου' (hill climbing) που εκτελεί η μέθοδος στον χώρο όλων των συντελεστών βαρύτητας, θεωρώντας ως ευρετική συνάρτηση (heuristic function) την κλίση του συνολικού σφάλματος και προσπαθώντας να βρει το ολικό ελάχιστο της συνάρτησης του συνολικού σφάλματος. Συνεπώς, υπάρχει περίπτωση η μέθοδος να παγιδευτεί σε τοπικά ελάχιστα και να μη μπορέσει να βρει τα βέλτιστα βάρη του δικτύου. Για να αντιμετωπιστεί αυτός ο κίνδυνος, μπορεί να πραγματοποιηθεί στοχαστική μεταβολή των βαρών ή να αυξηθεί ο αριθμός των νευρώνων του κάθε επιπέδου του δικτύου.

Ακολούθως, περιγράφουμε τον τρόπο λειτουργίας του αλγορίθμου. Αρχικά, τα παραδείγματα μάθησης παρουσιάζονται στο μη-εκπαιδευμένο δίκτυο, υπολογίζονται οι έξοδοι και για κάθε νευρώνα του επιπέδου εξόδου (output layer) υπολογίζεται το σφάλμα. Βάσει αυτού του σφάλματος συμβαίνει η αντίστοιχη αλλαγή στα βάρη των νευρώνων του επιπέδου εισόδου (input layer). Με κατεύθυνση από το επίπεδο εξόδου προς το επίπεδο εισόδου, για κάθε εσωτερικό νευρώνα υπολογίζεται η συμμετοχή του στα σφάλματα των νευρώνων εξόδου και βάσει αυτής ανανεώνονται τα βάρη στην είσοδό του. Η συμμετοχή ενός νευρώνα στα σφάλματα των νευρώνων του επόμενου επιπέδου είναι ανάλογη της τρέχουσας εισόδου του και των συντελεστών βαρύτητας που τον συνδέουν με τους νευρώνες του επόμενου επιπέδου.

Εάν θεωρήσουμε ένα νευρώνα k του επιπέδου εξόδου, a_k την έξοδο του συγκεκριμένου νευρώνα για ένα δεδομένο παράδειγμα εισόδου και e_k την επιθυμητή έξοδο για το ίδιο παράδειγμα, τότε ορίζουμε ως Σφάλμα του νευρώνα τη διαφορά $a_k - e_k$. Το πραγματικό σφάλμα κάθε νευρώνα πολλαπλασιάζεται με την παράγωγο της Συνάρτησης Ενεργοποίησης βάσει του γενικευμένου Κανόνα Δέλτα και μας δίνει το Προσαρμοσμένο Σφάλμα του νευρώνα k: $d_k = (a_k - e_k)f'(S_k)$. Τα σφάλματα των νευρώνων των κρυφών επιπέδων μπορούν να υπολογιστούν από τα σφάλματα των νευρώνων του ακριβώς επόμενου επιπέδου. Έστω ο νευρώνας

i ενός κρυφού επιπέδου και w_{ik} τα βάρη που συνδέουν αυτόν το νευρώνα με όλους τους νευρώνες k του επόμενου επιπέδου. Τότε, το σφάλμα του νευρώνα i υπολογίζεται μέσω της σχέσης:

$$d_i = f'(S_i) \sum w_{ik} d_k \tag{2.9}$$

Έτσι, με τους δύο παραπάνω τύπους προσδιορίζουμε τα σφάλματα για όλους τους νευρώνες του δικτύου, μέχρι και το επίπεδο εισόδου. Διακρίνουμε δύο διαφορετικούς τρόπους παρουσίασης των παραδειγμάτων και τροποποίησης των βαρών:

- **Αυξητική Εκπαίδευση** (Incremental Training), όπου για κάθε ξεχωριστό παράδειγμα τροποποιούμε τα βάρη και
- **Μαζική Εκπαίδευση** (Batch Training), στην οποία περιμένουμε να παρουσιαστούν όλα τα παραδείγματα από μία φορά, υπολογίζουμε τις αλλαγές στα βάρη για το κάθε παράδειγμα και εφαρμόζουμε αυτές τις αλλαγές ακριβώς μετά την παρουσίαση των παραδειγμάτων.

Ανεξάρτητα από τον τρόπο που επιλέγουμε να τροποποιήσουμε τα βάρη, η παρουσίαση όλων των παραδειγμάτων από μία φορά ονομάζεται *Εποχή Εκπαίδευσης* (epoch). Το συνολικό σφάλμα για όλα τα παραδείγματα ορίζεται ως το άθροισμα των τετραγώνων των σφαλμάτων των νευρώνων εξόδου, ενώ ως συνθήκη τερματισμού ορίζεται είτε η πτώση αυτού του σφάλματος κάτω από ένα όριο είτε η συμπλήρωση ενός προκαθορισμένου αριθμού εποχών εκπαίδευσης.

2.3.2 Συναρτήσεις Βελτιστοποίησης βασισμένες στην Κάθοδο Κλίσης

Ο κλάδος της Αναγνώρισης Προτύπων ασχολείται με την ελαχιστοποίηση κάποιας Συνάρτησης Κόστους (Cost Function), η οποία συνήθως έχει τη μορφή ενός αθροίσματος:

$$Q(w) = lnQ_i(w) \tag{2.10}$$

όπου η τιμή της παραμέτρου w θέλουμε να ελαχιστοποιεί τη Συνάρτηση Κόστους Q(w).

Το άθροισμα σφάλματος Q(w) αποτελείται από επιμέρους σφάλματα Q(i) που αντιστοιχούν σε κάθε δείγμα i του συνόλου εκπαίδευσης (training set) ή του συνόλου ελέγχου (testing set). Για τον υπολογισμό της τιμής του w που ελαχιστοποιεί το Q(w) χρησιμοποιείται η Μέθοδος της Καθόδου Κλίσης (*Gradient Descent*).

Ανάλογα με το πλήθος των δεδομένων που χρησιμοποιούνται για τον υπολογισμό της κλίσης της Συνάρτησης Βελτιστοποίησης, διακρίνουμε τις ακόλουθες αλγοριθμικές προσεγγίσεις:

- **Batch Gradient Descent**: Πρόκειται για έναν αλγόριθμο που υπολογίζει τις κλίσεις της Συνάρτησης Κόστους σε σχέση με τις παραμέτρους w για όλα τα παραδείγματα του συνόλου δεδομένων εκπαίδευσης, βάσει του τύπου:

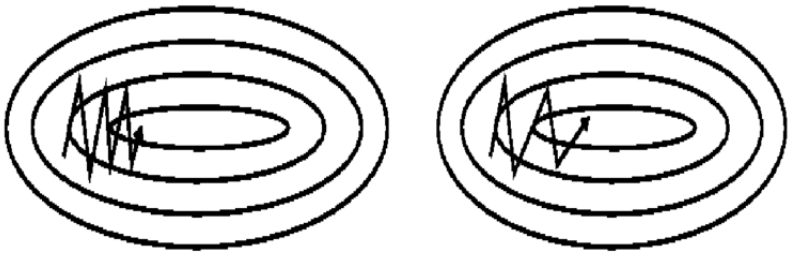
$$w = w - l\nabla_w Q(w) \tag{2.11}$$

Η συγκεκριμένη μέθοδος δεν επιτρέπει την ενημέρωση του μοντέλου με νέα δεδομένα που προκύπτουν από την εκπαίδευση, ενώ ενδέχεται να εκτελεστούν αχρείαστοι υπολογισμοί, διότι συχνά υπολογίζονται οι κλίσεις για όμοια παραδείγματα πριν από κάθε ενημέρωση των παραμέτρων. Είναι σίγουρο ότι εάν χρησιμοποιηθεί σε μία κυρτή επιφάνεια λάθους θα οδηγήσει στην εύρεση ολικού ελαχίστου της Συνάρτησης Κόστους.

- **Stochastic Gradient Descent-SGD**: Σε αντίθεση με τον προηγούμενο αλγόριθμο, εδώ οι παράμετροι ενημερώνονται για κάθε δείγμα εκπαίδευσης x_i με ετικέτα y_i, σύμφωνα με τη σχέση:

$$w = w - l\nabla_w Q(w; x_i, y_i) (2.12)$$

Σε αντίθεση με τον Batch Gradient Descent, εδώ δε συμβαίνουν περιττοί υπολογισμοί και ο υπολογισμός της κλίσης ή η ενημέρωση των βαρών γίνεται με χρήση ενός ή λίγων (mini batch SGD) δειγμάτων εκπαίδευσης. Ωστόσο, οι συχνές ενημερώσεις των παραμέτρων οδηγούν σε μεγάλη διασπορά στις τιμές με αποτέλεσμα να δημιουργούνται διακυμάνσεις στη Συνάρτηση Κόστους. Επίσης, ο αλγόριθμος Στοχαστικής Μείωσης Κλίσης παρουσιάζει προβλήματα πλοήγησης σε περιοχές όπου η επιφάνεια καμπυλώνεται πολύ πιο απότομα σε μία διάσταση από ότι σε κάποια άλλη, οι οποίες όμως είναι παρόμοιες γύρω από το τοπικό ελάχιστο. Σε αυτές τις περιπτώσεις, ο SGD ταλαντώνεται στις πλαγιές αυτής της περιοχής, η οποία συχνά αποκαλείται ως 'χαράδρα', ενώ παράλληλα σημειώνει αργή πρόοδο στην προσπάθεια προσέγγισης του τοπικού ελαχίστου. Υπάρχει μία μέθοδος που βοηθάει στην επιτάχυνση του αλγορίθμου στη σχετική κατεύθυνση και μειώνει τις ταλαντώσεις, η οποία ονομάζεται 'momentum'. Στο ακόλουθο Σχήμα 2.8 παρατηρούμε τη διαφορά της προόδου του SGD χωρίς εφαρμογή του momentum και με την εφαρμογή του:



Σχήμα 2.8: SGD χωρίς momentum (αριστερά) και SGD με momentum (δεξιά)

- **Adagrad**: Ο συγκεκριμένος αλγόριθμος βελτιστοποίησης βασίζεται στην κλίση και χρησιμοποιείται κατά βάση σε αραιά δεδομένα. Προσαρμόζει το ρυθμό μάθησης (learning rate) για κάθε παράμετρο, εφαρμόζοντας μεγάλες ενημερώσεις για σπάνιες παραμέτρους και μικρές ενημερώσεις για παραμέτρους που εμφανίζονται συχνά. Ο αλγόριθμος Adagrad χρησιμοποιεί διαφορετικό ρυθμό μάθησης l για την κάθε παράμετρο w_i τη χρονική στιγμή t. Ο τύπος που ακολουθεί ορίζει την κλίση της Συνάρτησης Κόστους σε σχέση

με την παράμετρο w_i την χρονική στιγμή t:

$$g_{t,i} = \nabla_w J(w_i) \quad (2.13)$$

και η ενημέρωση της κάθε παραμέτρου w_i κάθε χρονική στιγμή t γίνεται ακολουθώντας τη σχέση: $w_{t+1,i} = w_{t,i} - l g_{t,i}$. Επεκτάσεις του αλγορίθμου Adagrad αποτελούν οι αλγόριθμοι Adadelta, RMSProp και ο αλγόριθμος Adam.

2.4 Μέθοδοι Μηχανικής Μάθησης

Το κύριο χαρακτηριστικό των Τεχνητών Νευρωνικών Δικτύων είναι η εγγενής ιδιότητά τους να βελτιώνουν σταδιακά την ικανότητά τους στην επίλυση ενός δοθέντος προβλήματος. Η μάθηση των Νευρωνικών Δικτύων επιτυγχάνεται μέσω της επαναληπτικής εκπαίδευσής τους και της προσαρμογής των παραμέτρων τους, συνήθως των βαρών και της πόλωσης, έως ότου οδηγηθεί το δίκτυο σε μία λειτουργική κατάσταση. Στόχος της εκπαίδευσης των δικτύων είναι να αποκτήσουν την ικανότητα γενίκευσης, δηλαδή να μπορούν να κάνουν ορθές προβλέψεις για καινοφανή δεδομένα εισόδου, τα οποία δεν έχουν ξανασυναντήσει.

Αρχικά, με τον όρο *Μηχανική Μάθηση* (Machine Learning-ML) (βλέπε εργασία [27]) αναφερόμαστε στο τμήμα της επιστήμης υπολογιστών που ασχολείται με την χρήση αλγορίθμων και εργαλείων της στατιστικής με σκοπό την εξαγωγή γνώσης από δεδομένα. Αυτός ο κλάδος έχει οδηγήσει σε σημαντικά αποτελέσματα μέσω μίμησης του τρόπου λειτουργίας του ανθρώπινου εγκεφάλου για την παραγωγή της ανθρώπινης σκέψης. Το Machine Learning περιλαμβάνει μία σειρά βημάτων, με τη διαδικασία να αρχίζει από παρατηρήσεις που έχουν γίνει και έχουν καταγραφεί με τη μορφή εμπειρικών δεδομένων, να συνεχίζει με την αναγνώριση μοτίβων στα παρατηρούμενα δεδομένα και να καταλήγει στην λήψη καλύτερων αποφάσεων για μελλοντικά παραδείγματα. Στόχος της διαδικασίας είναι η αυτοματοποίηση της λήψης αποφάσεων από την πλευρά των υπολογιστών, χωρίς την ανάγκη επέμβασης του ανθρώπινου παράγοντα.

Η μεγάλη άνθιση του κλάδου έχει συντελεστεί τα τελευταία χρόνια και οφείλεται στην πληθώρα των δεδομένων που παράγονται καθημερινά, στην ανακάλυψη αποδοτικών τρόπων αποθήκευσης αυτών των δεδομένων και στην χρήση ταχύτερων υπολογιστικών μηχανημάτων. Ανάλογα με την εφαρμογή που ασχολούμαστε, υπάρχουν διαφορετικοί τρόποι Μηχανικής Μάθησης που ακολουθούνται. Οι πιο ευρέως υιοθετημένες μέθοδοι είναι η *Επιβλεπόμενη Μάθηση* (Supervised Learning), η *Μη Επιβλεπόμενη Μάθηση* (Unsupervised Learning) και η *Ενισχυτική Μάθηση* (Reinforcement Learning-RL).

2.4.1 Επιβλεπόμενη Μάθηση

Η Επιβλεπόμενη Μηχανική Μάθηση ή Μάθηση Με Επίβλεψη (Supervised Learning, βλέπε εργασία [14, 81]) εκπαιδεύει αλγορίθμους που βασίζονται σε δεδομένα εισόδου και εξόδου, τα οποία έχουν λάβει ετικέτες (labels) από τον άνθρωπο. Έτσι, το μοντέλο δέχεται ως είσοδο κατηγοριοποιημένα δεδομένα (*labeled data*) και εκπαιδεύεται πάνω σε αυτά ώστε να γενικεύει

επαρκώς και να πραγματοποιεί σωστές προβλέψεις για άγνωστα δεδομένα εισόδου. Πιο αναλυτικά, αρχικά ορίζονται τυχαίες τιμές στα βάρη των συνδέσεων των νευρώνων του μοντέλου, οι οποίες κατά την εκπαίδευση τροποποιούνται και διορθώνονται βάσει του σφάλματος που παίρνουμε, δηλαδή ανάλογα με το πόσο απέχουμε από τον επιθυμητό στόχο-ετικέτα. Το βήμα τροποποίησης των βαρών καλείται *learning rate* και είναι μία από τις σημαντικότερες παραμέτρους στα Νευρωνικά Δίκτυα., ενώ οι χρησιμοποιούμενες μέθοδοι για την ελαχιστοποίηση του σφάλματος είναι οι ίδιες που εφαρμόζονται και σε άλλες τεχνικές ελαχιστοποίησης.

Τα προβλήματα που αντιμετωπίζει η Επιβλεπόμενη Μάθηση χωρίζονται σε προβλήματα ταξινόμησης (*classification problems*) και σε προβλήματα παλινδρόμησης (*regression problems*). Ένα γνωστό παράδειγμα προβλήματος ταξινόμησης είναι η αναγνώριση και κατηγοριοποίηση εικόνων. Διαθέτουμε δεδομένα ελέγχου (test data) που αποτελούνται από φωτογραφίες, καθεμία από τις οποίες έχει κάποιο label, εκπαιδεύουμε το μοντέλο στα test data και εντέλει το χρησιμοποιούμε για να κατηγοριοποιήσουμε άγνωστες εικόνες δίχως κάποια ετικέτα. Από την άλλη πλευρά, μιλώντας για προβλήματα παλινδρόμησης αναφερόμαστε σε μοντέλα που στην έξοδό τους δίνουν κάποια αριθμητική τιμή και όχι κατηγορία, όπως συνέβαινε στα προβλήματα ταξινόμησης. Για παράδειγμα, στο πρόβλημα της αναγνώρισης εικόνων ή στο πρόβλημα της ανίχνευσης αντικειμένων από εικόνες, θα είχαμε παλινδρόμηση εάν αναζητούσαμε τις συντεταγμένες που ορίζουν ένα ορθογώνιο γύρω από κάποιο αντικείμενο της εικόνας. Επίσης, ένα πρόβλημα παλινδρόμησης θεωρείται και η αξιοποίηση ιστορικών δεδομένων για την πρόβλεψη στοιχείων που αφορούν το μέλλον, όπως η πρόβλεψη μία μελλοντικής τιμής κάποιας χρηματοπιστωτικής μετοχής, βάσει προηγούμενων αλλαγών στις τιμές της.

2.4.2 Μη Επιβλεπόμενη Μάθηση

Αναφερόμαστε σε Μη Επιβλεπόμενη Μάθηση ή Μάθηση Χωρίς Επίβλεψη (Unsupervised Learning, βλέπε εργασία [6, 72]) όταν διαχειριζόμαστε δεδομένα που δεν έχουν κατηγοριοποιηθεί ή αντιστοιχηθεί με κάποια ετικέτα, δηλαδή δεν έχουμε εξάγει καμία εμπειρία ή γνώση από αυτά. Έτσι, εκπαιδεύουμε αλγορίθμους ώστε να ανακαλύπτουν τη δομή στα δεδομένα που δέχονται στην είσοδο, δηλαδή να βρίσκουν τα μοτίβα (patterns) που κρύβονται στα δεδομένα εκπαίδευσης. Προκειμένου να εξάγουμε γνώση, παρέχουμε την πληροφορία στο δίκτυο χωρίς να προβούμε σε κάποιον έλεγχο και το ίδιο το δίκτυο διορθώνει τα σφάλματα στα δεδομένα που δέχεται μέσω του μηχανισμού ανάδρασης (feedback). Όταν το δίκτυο παύει να τροποποιεί τις τιμές των βαρών, θεωρούμε ότι η εκπαίδευσή του έχει ολοκληρωθεί διότι το λάθος στην έξοδο τείνει να μηδενιστεί.

Τα δύο είδη της Μάθησης Χωρίς Επίβλεψη είναι η ομαδοποίηση (*clustering*) και η μείωση των διαστάσεων (*dimensionality reduction*). Το clustering αναφέρεται στην ομαδοποίηση των παρατηρήσεων με τέτοιο τρόπο ώστε τα δεδομένα που ανήκουν στην ίδια ομάδα (group) να παρουσιάζουν παρόμοιες ιδιότητες ή χαρακτηριστικά μεταξύ τους, ενώ συγκρινόμενα με δεδομένα άλλων ομάδων να εμφανίζουν απολύτως διαφορετικές ιδιότητες. Με τον όρο dimensionality reduction εννοούμε την σύμπτυξη των δεδομένων μέσω της αφαίρεσης τυχαίων μεταβλητών χωρίς να χάνεται η δομή και η σημασία του συνόλου των δεδομένων. Αυτή η

σύμπτυξη των δεδομένων οδηγεί σε μεγαλύτερη ευκολία αποθήκευσης των δεδομένων, σε ταχύτερη εκτέλεση υπολογισμών πάνω σε αυτά, σε αμεσότερη οπτική αναπαράστασή τους και σε βελτίωση της επίδοσης του μοντέλου. Χαρακτηριστικά παραδείγματα δικτύων που βασίζουν την εκπαίδευσή τους στη Μη Επιβλεπόμενη Μάθηση είναι τα *Δίκτυα Kohonen* και τα *Self-organizing maps*.

2.4.3 Ενισχυτική Μάθηση

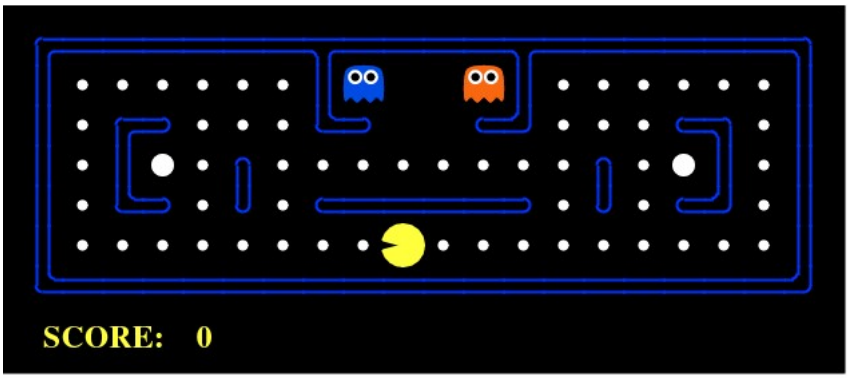
Σε προβλήματα Ενισχυτικής Μάθησης (Reinfrorcement Learning-RL, βλέπε εργασία [94]) ένα πρόγραμμα υπολογιστή αλληλεπιδρά με ένα δυναμικό περιβάλλον, στο οποίο πρέπει να επιτευχθεί κάποιος στόχος χωρίς να ακολουθούνται ρητές οδηγίες. Οι αλγόριθμοι Reinforcement Learning χρησιμοποιούν ένα σύστημα επιβράβευσης (*reward system*) και συνεχών δοκιμών-λαθών προκειμένου να μεγιστοποιηθεί η τελική επιβράβευση reward ενός υποκειμένου-πράκτορα (*agent*). Για να εξηγήσουμε τον τρόπο λειτουργίας των προβλημάτων Ενισχυτικής Μάθησης θα χρησιμοποιήσουμε ένα γνωστό παιχνίδι, το Pac-man, ένα στιγμιότυπο του οποίου φαίνεται στο Σχήμα 2.9. Στη συγκεκριμένη περίπτωση, ως agent θεωρούμε τον Pac, ο οποίος εξερευνά το λαβύρινθο ακολουθώντας τις τελείες (dots) και λαμβάνοντας κάποια επιβράβευση για την κάθε τελεία που διαπερνάει. Έτσι, ο πράκτορας μαθαίνει να μην επιστρέφει σε διαδρομές του λαβυρίνθου από τις οποίες έχει ήδη διέλθει και έχει καταναλώσει τις τελείες που ήταν νωρίτερα τοποθετημένες εκεί. Ωστόσο, ενδέχεται τώρα να έχει εμφανιστεί κάποιο τρόπαιο κατά μήκος μίας διαδρομής που έχουμε ήδη διασχίσει και έτσι να μεγιστοποιείται το reward για εκείνη τη διαδρομή. Συνεπώς, ο αλγόριθμος της Ενισχυτικής Μάθησης θα πρέπει να αποφασίζει εάν συμφέρει τον πράκτορα να συνεχίσει την εξερεύνηση των καταστάσεων εκμεταλλευόμενος και την τρέχουσα ευκαιρία ή να την αγνοήσει.

Προκειμένου οι αλγόριθμοι ενισχυτικής μάθησης να χειριστούν αυτή τη δυσκολία, εισάγεται ένα επίπεδο τυχαιότητας, το οποίο ονομάζεται epsilon-greedy στρατηγική. Ο πράκτορας του προβλήματος, εν προκειμένω ο Pac, όταν βρίσκεται σε ορισμένες καταστάσεις, ακολουθεί κάποια τυχαία διαδρομή και εσκεμμένα αγνοεί τρόπαια. Το ποσοστό αυτών των καταστάσεων το αποκαλούμε epsilon και η διαδικασία υπολογισμού της καινούργιας πιθανότητας για την κάθε κατάσταση είναι γνωστή ως Markov Decision Process-MDP.

Γενικεύοντας το παράδειγμα με το παιχνίδι Pac-Man, οι αλγόριθμοι του Reinforcement Learning αρχίζουν εξερευνητικά και όσο τα συστήματα επιβράβευσης του παιχνιδιού γίνονται περισσότερο κατανοητά, τόσο καλύτερα τείνουν προς την εκμετάλλευση δεδομένων καταστάσεων.

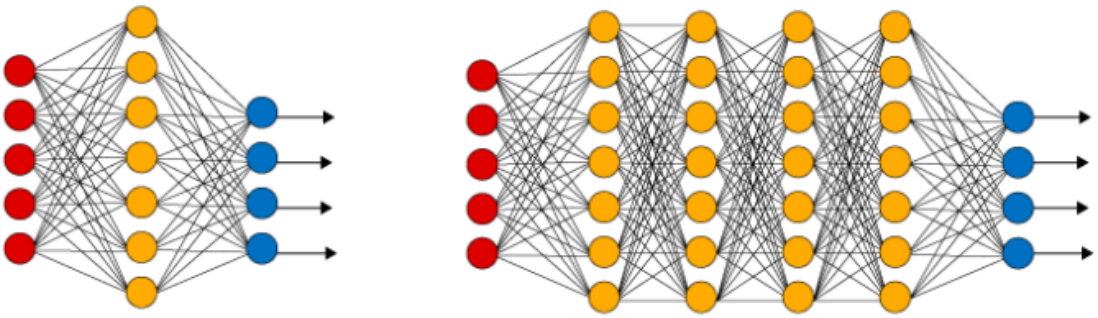
2.5 Συνελικτικά Νευρωνικά Δίκτυα (CNN)

Τα Συνελικτικά Νευρωνικά Δίκτυα (Convolutional Neural Networks-CNNs) αποτελούν ένα είδος Νευρωνικών Δικτύων που έχει αναπτυχθεί στο πλαίσιο της Βαθιάς Μηχανικής Μάθησης (Deep Learning). Ο κλάδος της Βαθιάς Μηχανικής Μάθησης άνθισε τα τελευταία χρόνια χάρη στην πτώση των τιμών του υλικού και στην ανάπτυξη των Μονάδων Επεξερ-



Σχήμα 2.9: Στιγμιότυπο του παιχνιδιού Pac-Man

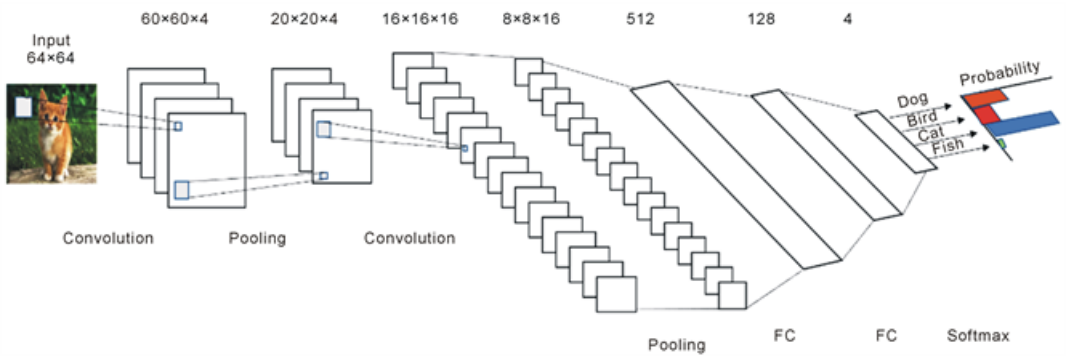
γασίας Γραφικών (Graphics Processing Units-GPUs), φέρνοντας επανάσταση στον επιστημονικό χώρο. Ο όρος 'Deep Learning' χρησιμοποιήθηκε για πρώτη φορά στη Μηχανική Μάθηση από τον Dechter(1986) (βλέπε εργασία [78]) και στην Τεχνητή Νοημοσύνη από τον Aizenberg(2000) (βλέπε εργασία [39]). Οι αλγόριθμοι της Βαθιάς Μηχανικής Μάθησης προσπαθούν να μοντελοποιήσουν αφαιρέσεις υψηλού επιπέδου σε δεδομένα, μέσω πολλαπλών επιπέδων επεξεργασίας και μη γραμμικών μετασχηματισμών.



Σχήμα 2.10: Απλό (αριστερά) και Βαθύ (δεξιά) Νευρωνικό Δίκτυο

Το Σχήμα 2.10 αντιπαραβάλλει ένα απλό νευρωνικό δίκτυο (αριστερή διάταξη) πέντε εισόδων (κόμβοι με κόκκινο χρώμα) και τεσσάρων εξόδων (κόμβοι με μπλε χρώμα) με ενα βαθύ νευρωνικό δίκτυο (δεξιά διάταξη) πέντε εισόδων (κόμβοι με κόκκινο χρώμα), τεσσάρων κρυφών επιπέδων (κόμβοι με κίτρινο χρώμα) και τεσσάρων εξόδων (κόμβοι με μπλε χρώμα).

Οι τεχνικές της Βαθιάς Μηχανικής Μάθησης έγιναν ευρέως γνωστές μετά την ανάπτυξη μιας αρχιτεκτονικής Συνελικτικών Νευρωνικών Δικτύων (CNN) από τον Alex Krizhevsky, η οποία ονομάστηκε AlexNet (βλέπε εργασία [3]) και κέρδισε το διαγωνισμό του ImageNet το 2012. Σε σχέση με τα κλασικά δίκτυα εμπρόσθιας τροφοδότησης (feedforward networks), τα CNN διαθέτουν μικρότερο αριθμό συνδέσεων και παραμέτρων, γεγονός που καθιστά ευκολότερη την εκπαίδευσή τους, ενώ η θεωρητικά βέλτιστη επίδοσή τους είναι πιθανό να είναι ελάχιστα χειρότερη από αυτήν που θα παρουσίαζε ένα feedforward network.



Σχήμα 2.11: Συνελικτικό Νευρωνικό Δίκτυο

2.5.1 Τρόπος λειτουργίας των CNN

Τα Συνελικτικά Νευρωνικά Δίκτυα σχεδιάζονται ώστε να αναγνωρίζουν δισδιάστατα σχήματα με υψηλό βαθμό ανεκτικότητας στη μετατόπιση, την κλιμάκωση, τη στρέβλωση και σε άλλες μορφές παραμόρφωσης. Για να επιτευχθεί αυτός ο σκοπός, το Νευρωνικό Δίκτυο εκπαιδεύεται με *Επιβλεπόμενο τρόπο* και τα βήματα που ακολουθούνται κατά την εκπαίδευσή του είναι τα ακόλουθα:

- **Εξαγωγή Χαρακτηριστικών** (Feature Extraction): κάθε νευρώνας δέχεται τις συναπτικές εισόδους του από ένα προηγούμενο επίπεδο και το υποχρεώνει να εξάγει τοπικά χαρακτηριστικά. Μετά την εξαγωγή ενός χαρακτηριστικού, η θέση του γίνεται λιγότερο σημαντική και διατηρείται η πληροφορία για τη σχετική θέση του ως προς άλλα χαρακτηριστικά που δεν έχουν ακόμα εξαχθεί.

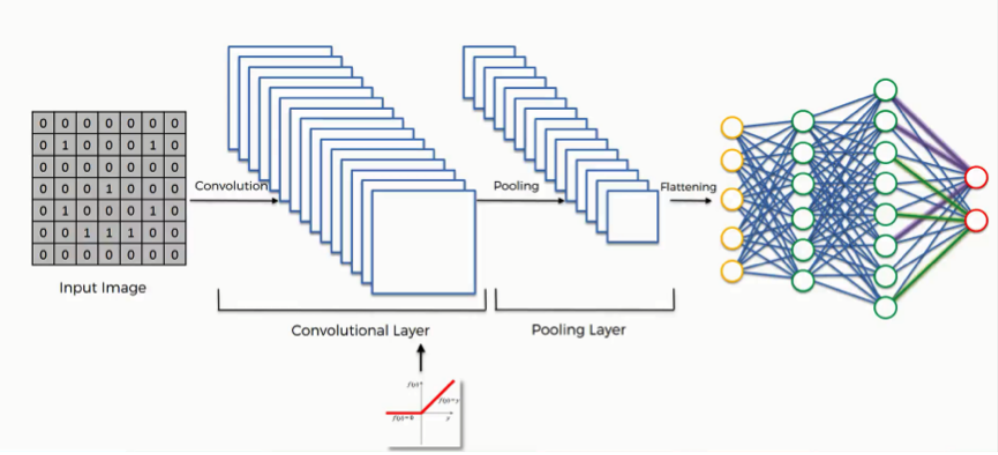
- **Αντιστοίχιση Χαρακτηριστικών** (Feature Mapping): κάθε υπολογιστικό επίπεδο του δικτύου διαθέτει πολλαπλούς χάρτες χαρακτηριστικών (feature maps), καθένας από τους οποίους είναι σε μορφή ενός επιπέδου μέσα στο οποίο οι μεμονωμένοι νευρώνες ελέγχονται ώστε να μοιράζονται το ίδιο σύνολο συναπτικών βαρών. Αυτό έχει ως αποτέλεσμα το δίκτυο να καθίσταται ανεξάρτητο από τη μετατόπιση, η οποία επιβάλλεται στη λειτουργία ενός feature map μέσω της χρήσης μίας συνέλιξης με ένα πυρήνα (kernel) μικρού μεγέθους. Μετά τη συνέλιξη, εφαρμόζεται μία Συνάρτηση Ενεργοποίησης, όπως είναι η sigmoid ή η ReLU. Μέσω του διαμοιρασμού των βαρών, μειώνεται το πλήθος των ελεύθερων παραμέτρων.

- **Υποδειγματοληψία** (Downsampling): μετά από κάθε Συνελικτικό Επίπεδο (Convolutional Layer) βρίσκεται ένα Υπολογιστικό Επίπεδο (Pooling Layer), το οποίο είναι υπεύθυνο για την εκτέλεση υποδειγματοληψίας,όπως είναι το average pooling. Με αυτό τον τρόπο μειώνεται η ανάλυση του feature map και η έξοδός του γίνεται λιγότερο ευαίσθητη σε παραμορφώσεις.

- **Εξαγωγή προβλέψεων** (Prediction Mapping): αφού ολοκληρωθούν οι υπολογισμοί των προηγούμενων επιπέδων, στο τέλος της αρχιτεκτονικής του Συνελικτικού Νευ-

ρωνικού Δικτύου προστίθεται ο κατάλληλος αριθμός από Πλήρως Συνδεδεμένα Επίπεδα (Fully Connected Layer), προκειμένου να εξαχθεί η τελική έξοδος του δικτύου.

Η λειτουργία ενός Συνελικτικού Νευρωνικού Δικτύου περιλαμβάνει τη φάση του Προωθητικού Περάσματος (*Forward Pass*), κατά το οποίο μία είσοδος, που συνήθως πρόκειται για μία εικόνα, περνάει από διαδοχικά επίπεδα επεξεργασίας έως ότου εξαχθεί η τελική πρόβλεψη του δικτύου. Κάθε επίπεδο επεξεργασίας δέχεται μία είσοδο, τη μετασχηματίζει και εξάγει μία έξοδο, η οποία αποτελεί την είσοδο του επόμενου επιπέδου. Η διαδικασία που ακολουθείται είναι σειριακή και όταν ολοκληρωθεί, ακολουθεί η διαδικασία στην οποία βασίζεται η εκπαίδευση του δικτύου και η οποία είναι γνωστή ως οπισθοδιάδοση σφάλματος (*Backward Error Propagation*). Παρά τις ποικίλες τροποποιήσεις στην αρχιτεκτονική των CNN, η γενική μορφή της αρχιτεκτονικής τους είναι αυτή που απεικονίζεται στο Σχήμα 2.12:



Σχήμα 2.12: Γενική μορφή αρχιτεκτονικής ενός CNN

2.5.2 Επίπεδα Επεξεργασίας των CNN

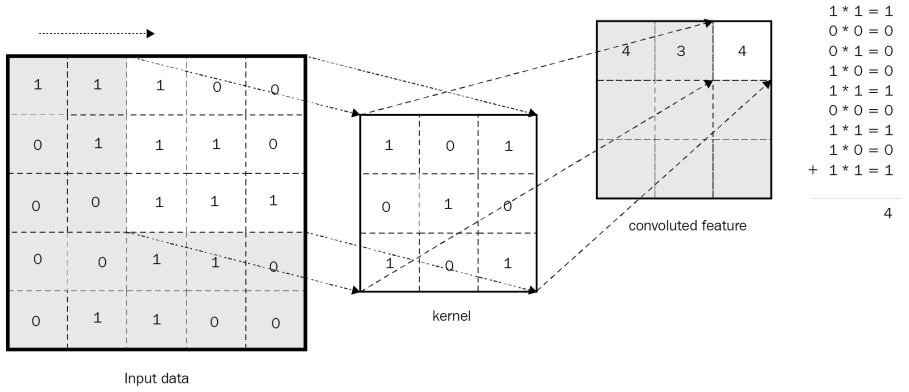
Σε επίπεδο υλοποίησης, ένα μοντέλο Βαθιάς Μηχανικής Μάθησης και συγκεκριμένα ένα μοντέλο Συνελικτικών Νευρωνικών Δικτύων χρησιμοποιεί Νευρωνικά Δίκτυα, τα οποία περιέχουν διασυνδεδεμένους νευρώνες, οι οποίοι είναι ομαδοποιημένοι σε επίπεδα (layers). Στη συνέχεια, εξηγούμε τον τρόπο κατασκευής και λειτουργίας της κάθε κατηγορίας επιπέδου:

2.5.2.1 Επίπεδο Εισόδου (Input Layer)

Πρόκειται για το αρχικό επίπεδο του δικτύου, το οποίο αναλαμβάνει να φορτώσει τα ακατέργαστα δεδομένα (raw data). Οι διαστάσεις του επιπέδου προσδιορίζονται από τις διαστάσεις των δεδομένων εισόδου. Για παράδειγμα, εάν το δίκτυο τροφοδοτείται με ψηφιακές εικόνες αναπαράστασης RGB, τότε το Input Layer θα έχει μήκος ίσο με το μήκος των εικόνων, ύψος ίσο με το ύψος τους και βάθος ίσο με τον αριθμό των καναλιών R,G και B της κάθε εικόνας.

2.5.2.2 Συνελικτικό Επίπεδο (Convolutional Layer)

Το Συνελικτικό Επίπεδο αποτελεί το βασικότερο στοιχείο στην αρχιτεκτονική των Συνελικτικών Νευρωνικών Δικτύων. Αναλαμβάνει να δεχτεί τα δεδομένα εισόδου και μέσω ενός συνόλου συνδεδεμένων νευρώνων του προηγούμενου επιπέδου να τα μετασχηματίσει. Μέσω αυτών των μετασχηματισμών επιτυγχάνεται η εξαγωγή χαρακτηριστικών (feature extraction) της εικόνας εισόδου και η δημιουργία διαφορετικών *feature maps*, όπως φαίνεται και στο Σχήμα 2.13:

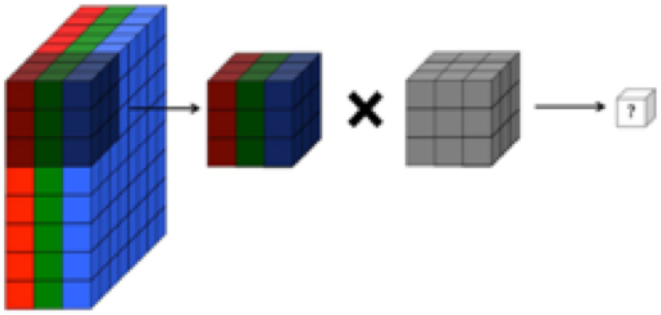


Σχήμα 2.13: Επεξεργασία εισόδου για την παραγωγή feature maps

Στη συνέχεια, διευκρινίζουμε ορισμένες έννοιες που σχετίζονται άμεσα με τον τρόπο λειτουργίας των Συνελικτικών Επιπέδων και που θα βοηθήσουν στην καλύτερη κατανόηση της διαδικασίας εξαγωγής των χαρακτηριστικών των δεδομένων εισόδου.

- **Πράξης Συνέλιξης** (Convolution): Το μαθηματικό υπόβαθρο στο οποίο στηρίζεται η λειτουργία του Convolutional Layer είναι η πράξη της συνέλιξης (convolution), η οποία περιγράφει έναν κανόνα για τη συνένωση πληροφορίας προερχόμενης από δύο διαφορετικά μέρη. Η είσοδος της συνέλιξης είναι είτε ακατέργαστα δεδομένα είτε κάποιος χάρτης χαρακτηριστικών (*feature map*), ο οποίος έχει σχηματιστεί από προηγούμενα επίπεδα του CNN. Ένας πυρήνας (*kernel*) ή αλλιώς φίλτρο (*filter*) ολισθαίνει στα δεδομένα εισόδου και παράγει τα συνελικτικά χαρακτηριστικά (*convoluted features*). Σε κάθε βήμα, τα στοιχεία του φίλτρου πολλαπλασιάζονται ένα-προς-ένα (element-wise) με τα αντίστοιχα στοιχεία του πίνακα εισόδου και η έξοδος της διαδικασίας έχει μεγαλύτερη τιμή αν το χαρακτηριστικό που αναζητείται ανιχνεύεται στην είσοδο. Ο πίνακας εξόδου που προκύπτει ονομάζεται χάρτης χαρακτηριστικών (feature map) ή χάρτης ενεργοποίησης (activation map). Εάν τα δεδομένα εισόδου είναι εικόνες, εφαρμόζοντας πολλά διαφορετικά φίλτρα εξάγουμε ποικίλα χαρακτηριστικά τους, με αποτέλεσμα η συνένωση των διαφορετικών feature maps να οδηγεί σε τρισδιάστατη (3D) έξοδο.

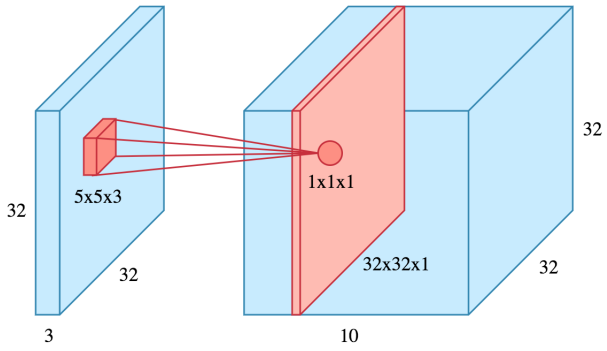
- **Φίλτρα** (Filters): Η εφαρμογή των φίλτρων στα δεδομένα εισόδου γίνεται με τρόπο που θυμίζει κινούμενο παράθυρο. Συγκεκριμένα, πραγματοποιείται ένα-προς-ένα (element-wise) πολλαπλασιαμός των στοιχείων του κάθε φίλτρου και της περιοχής του πίνακα



Σχήμα 2.14: Εφαρμογή συνέλιξης στα δεδομένα εισόδου

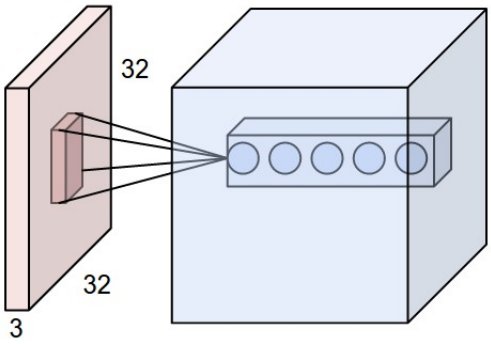
εισόδου. Ο αριθμός των φίλτρων που θα χρησιμοποιηθούν καθορίζεται από τις παραμέτρους του συνελικτικού επιπέδου. Τα φίλτρα είναι υπεύθυνα για την ενεργοποίηση συγκεκριμένων μοτίβων ή χαρακτηριστικών εφόσον αυτά τα μοτίβα εμφανίζονται στις αντίστοιχες θέσεις των δεδομένων εκπαίδευσης. Καθώς προχωράμε σε πιο βαθιά επίπεδα του CNN, τα φίλτρα τείνουν να αναγνωρίζουν πιο αφηρημένους συνδυασμούς χαρακτηριστικών.

- **Πίνακες Χαρακτηριστικών** (Feature Maps): Όπως αναφέραμε και παραπάνω, κατά την εφαρμογή διαφορετικών φίλτρων στα δεδομένα εισόδου παράγονται έξοδοι δύο διαστάσεων. Αυτές οι δισδιάστατες έξοδοι του κάθε φίλτρου ονομάζονται feature maps και στοιβάζονται δημιουργώντας μία τρισδιάστατη αναπαράσταση της εξόδου, όπως μας δείχνει και το Σχήμα 2.15:



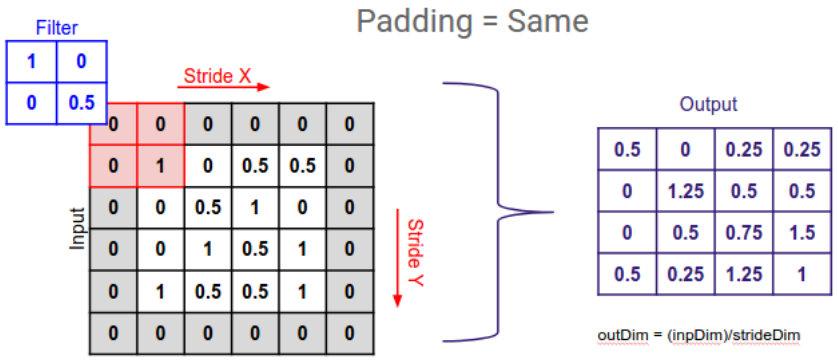
Σχήμα 2.15: Συστάδα από feature maps

- **Τοπική σύνδεση μεταξύ των νευρώνων**: Αντί να συνδεθούν όλοι οι νευρώνες με αυτούς του προηγούμενου επιπέδου, επιλέγουμε να συνδέσουμε τον καθένα με ένα ορισμένο σύνολο νευρώνων. Οι συνδέσεις είναι τοπικές στον χώρο, ως προς το ύψος και το πλάτος, όμως πάντα λαμβάνεται υπόψη το βάθος του όγκου εισόδου. Εάν θεωρήσουμε έναν όγκο εισόδου με διαστάσεις 32x32x3 και φίλτρο διαστάσεων 5x5, τότε κάθε νευρώνας του συνελικτικού επιπέδου θα έχει βάρη σε μία περιοχή του όγκου εισόδου διαστάσεων 5x5x3. Άρα, οι συνολικές τιμές των βαρών του νευρώνα θα είναι 5x5x3.



Σχήμα 2.16: Τοπική σύνδεση των νευρώνων του Συνελικτικού Επιπέδου

- **Υπερπαράμετροι**: Η διάταξη της εξόδου που θα προκύψει από ένα Συνελικτικό Επίπεδο καθορίζεται αρχικά από το μέγεθος του φίλτρου που θα χρησιμοποιήσουμε και από το πλήθος των φίλτρων που θα εφαρμοστούν για τον εντοπισμό των διαφορετικών χαρακτηριστικών (*features*) της εισόδου. Μία σημαντική παράμετρος κατά την χρήση φίλτρων αποτελεί το βήμα (*stride*), το οποίο ρυθμίζει κατά πόσο θα ολισθαίνει κάθε φορά το φίλτρο καθώς διασχίζει το σύνολο των δεδομένων εισόδου. Για παράδειγμα, εάν το βήμα έχει τιμή 1, αυτό σημαίνει ότι το φίλτρο θα μετακινείται στην εικόνα κατά ένα pixel τη φορά, ενώ αν έχει μεγαλύτερη τιμή, τότε θα μειώνονται οι διαστάσεις της εξόδου. Συχνά κρίνεται αναγκαίο να συμπληρώνουμε τις ακραίες τιμές της εισόδου χρησιμοποιώντας μηδενικές τιμές (*zero-padding*), όπως παρατηρούμε να συμβαίνει στην είσοδο διαστάσεων 4x4 που παρουσιάζεται στο Σχήμα 2.17, προκειμένου να μπορέσει να εφαρμοστεί το φίλτρο διαστάσεων 2x2, με βήμα ολίσθησης (stride) ίσο με 1.



Σχήμα 2.17: Zero-padding σε δεδομένα εισόδου 4x4, με φίλτρο 2x2 και βήμα 1

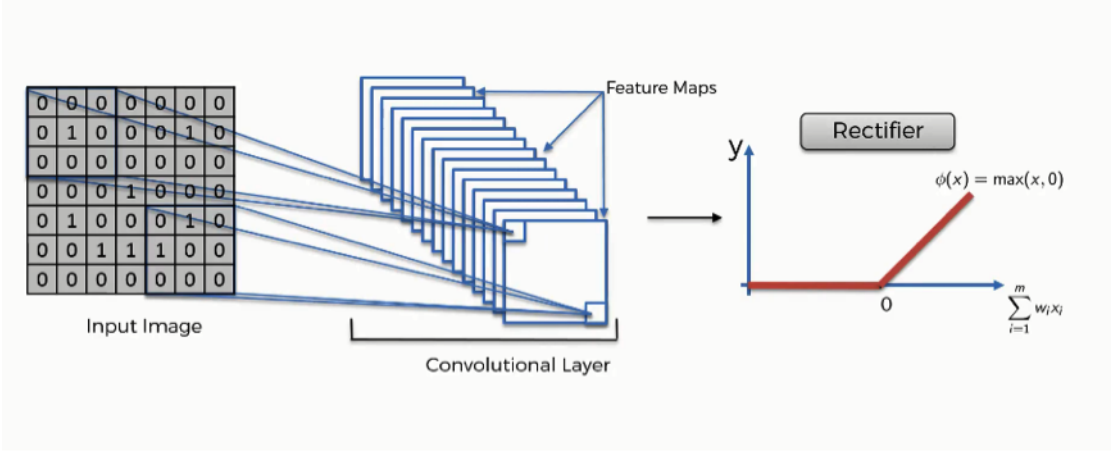
Ο συνολικός αριθμός των νευρώνων που προκύπτουν μετά το Convolutional Layer προσδιορίζεται από τη σχέση $\frac{I-F+2Z}{S}+1$, όπου με I συμβολίζουμε το μέγεθος του όγκου εισόδου, με F το μέγεθος του φίλτρου, με S το βήμα που εφαρμόζεται και με Z το πλήθος των μηδενικών που χρησιμοποιήθηκαν κατά το zero-padding.

2.5.2.3 Επίπεδο Γραμμικής Ανόρθωσης (Rectified Linear Unit)

Το επίπεδο επεξεργασίας Rectified Linear Unit-ReLU δεν επηρεάζει το μέγεθος της εισόδου, με αποτέλεσμα η είσοδος και η έξοδός του να έχουν το ίδιο ακριβώς μέγεθος. Αυτό που κάνει η ReLU είναι να αποκόπτει συγκεκριμένα στοιχεία της εισόδου, βάσει της σχέσης:

$$f(x) = (0, max) \tag{2.14}$$

Το συγκεκριμένο επίπεδο *δε διαθέτει παραμέτρους* και δε συμμετέχει στην εκπαίδευση του δικτύου. Ο σκοπός ύπαρξής του είναι η αύξηση της *μη-γραμμικότητας* του Συνελικτικού Νευρωνικού Δικτύου, δεδομένου ότι τα περισσότερα δεδομένα που επεξεργαζόμαστε είναι μη-γραμμικά. Συγκεκριμένα, η συνάρτηση ReLU μετατρέπει κάθε αρνητική τιμή εισόδου σε μηδενική ώστε να αναδείξει μη-γραμμικές συσχετίσεις στα δεδομένα εισόδου. Για παράδειγμα, σε περίπτωση που η είσοδος έχει θετικές τιμές σε μία περιοχή της εικόνας που παρουσιάζεται ένα μοτίβο, τότε ενδέχεται να έχει αρνητικές ή μηδενικές τιμές σε άλλες περιοχές που δεν εμφανίζουν το συγκεκριμένο μοτίβο. Όπως βλέπουμε στο Σχήμα 2.18, εφαρμόζοντας τη συνάρτηση ReLU όλες οι αρνητικές τιμές γίνονται μηδενικές και ενεργοποιείται η έξοδος μόνο στις περιοχές με το επιθυμητό μοτίβο. Ο λόγος που συχνά προτιμάται η συκεκριμένη Συνάρτηση Ενεργοποίησης αντί για κάποια άλλη όπως είναι η Σιγμοειδής (sigmoid), έγκειται στο ότι η κλίση της δεν αυξάνεται υπερβολικά, αλλά παραμένει μηδενική ή σταθερή.



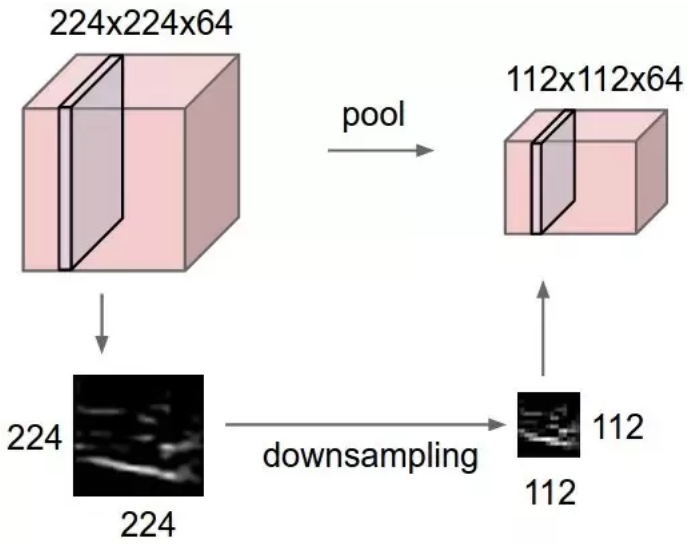
Σχήμα 2.18: Εφαρμογή της Συνάρτησης Ενεργοποίησης ReLU

2.5.2.4 Συγκεντρωτικό Επίπεδο (Pooling Layer)

Pooling Layers ονομάζουμε τα επίπεδα του CNN που συνήθως παρεμβάλλονται μεταξύ διαδοχικών Συνελικτικών Επιπέδων (Convolutional Layer) και στοχεύουν στη μείωση του χώρου που δεσμεύεται λόγω της αναπαράστασης των δεδομένων και των παραμέτρων του δικτύου. Έτσι, μειώνονται οι διαστάσεις σε κάθε χάρτη χαρακτηριστικών (*feature map*), δηλαδή πραγματοποιείται υποδειγματοληψία (*downsampling*), χωρίς να επηρεάζεται η διάσταση του βάθους των δεδομένων ή να έχουμε απώλεια σημαντικών πληροφοριών. Ως αποτέλε-

σμα, επιτυγχάνεται ο έλεγχος της υπερπροσαρμογής (*overfitting*) του μοντέλου στα δεδομένα εκπαίδευσης.

Προκειμένου να υλοποιήσουμε το downsampling, ορίζουμε ένα παράθυρο διαστάσεων 2x2 που ολισθαίνει πάνω στο εκάστοτε feature map και έχει την ιδιότητα να διατηρεί το μεγαλύτερο στοιχείο της εκάστοτε περιοχής (*max pooling*) ή να υπολογίζει το μέσο όρο των στοιχείων της περιοχής (*average pooling*) ή να βρίσκει το άθροισμά τους (*sum pooling*). Κατά το max pooling έχει παρατηρηθεί ότι εξάγεται πιο έντονη πληροφορία σε σχέση με το sum pooling. Στο Σχήμα 2.19 απεικονίζεται ένα παράδειγμα εφαρμογής του max pooling σε δεδομένα εισόδου διαστάσεων 224x224x64. Εφαρμόζονται φίλτρα διαστάσεων 2x2x64 και βήμα (*stride*) ίσο με 2, οπότε παρατηρείται μία μείωση των διαστάσεων (πλάτος, ύψος) των δεδομένων εισόδου κατά 50%, ενώ απορρίπτεται το 75% των συνολικών ενεργοποιήσεων.

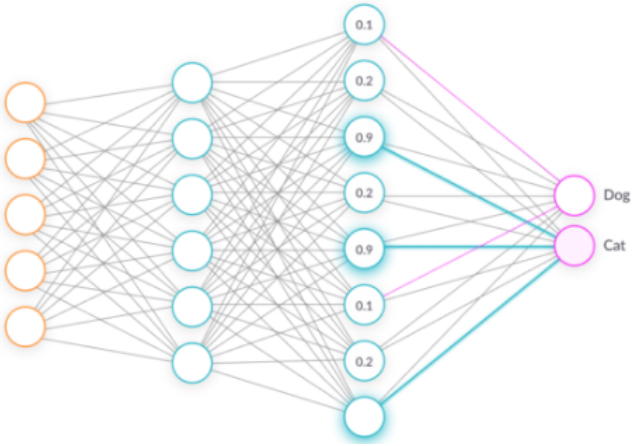


Σχήμα 2.19: Εφαρμογή του Max Pooling

2.5.2.5 Πλήρως-Συνδεδεμένο Επίπεδο (Fully-Connected Layer)

Τα Fully-Connected Layers αποτελούν μία αρχιτεκτονική πολλών επιπέδων με νευρώνες, η οποία χρησιμοποιεί μία συνάρτηση ενεργοποίησης στην έξοδό της. Σε ένα Πλήρως-Συνδεδεμένο επίπεδο, κάθε νευρώνες συνδέεται με όλους τους νευρώνες του προηγούμενου επιπέδου. Σε ένα συνηθισμένο πρόβλημα ταξινόμησης (*classification*) εικόνων, ο ρόλος του Πλήρως-Συνδεδεμένου Επιπέδου έγκειται στην αξιοποίηση των χαρακτηριστικών που έχουν προέλθει από τα *Convolutional* και *Pooling Layers*, με σκοπό την ταξινόμηση της εικόνας εισόδου σε διάφορες κλάσεις, οι οποίες χρησιμοποιήθηκαν και κατά την εκπαίδευση του μοντέλου. Πέρα από την χρήση του Fully-Connected Layer στην ταξινόμηση, πρόκειται για έναν άμεσο τρόπο εκμάθησης μη-γραμμικών συνδυασμών των χαρακτηριστικών, οι οποίοι ενδέχεται να είναι σημαντικότεροι για την ταξινόμηση. Οι έξοδοι ενός Fully Connected Layer αθροίζουν στη μονάδα, διότι έχουμε χρησιμοποιήσει τη Softmax ως Συνάρτηση Ενεργοποίησης του επιπέδου εξόδου του, με αποτέλεσμα οι έξοδοι να έχουν αντιστοιχηθεί σε τιμές ενός διανύσματος

τιμών [0,1]. Πλήρως-Συνδεδεμένα Επίπεδα εμφανίζονται και σε προβλήματα Παλινδρόμησης (*regression*), όπου στο τελευταίο επίπεδο υπάρχει ένας μοναδικός νευρώνας, του οποίου η έξοδος αποτελεί την τελική έξοδο του μοντέλου.

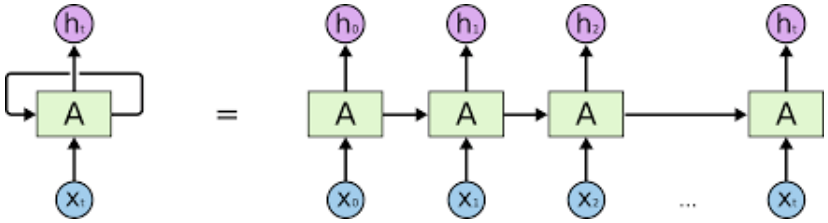


Σχήμα 2.20: Πλήρως-Συνδεδεμένο Επίπεδο

2.6 Αναδρομικά Νευρωνικά Δίκτυα (RNN)

Συνήθως, στα Νευρωνικά Δίκτυα θεωρούμε ότι όλες οι είσοδοι και οι έξοδοί τους είναι ανεξάρτητες μεταξύ τους. Ωστόσο, σε ένα υποθετικό πρόβλημα όπου θέλουμε να προβλέψουμε την επόμενη λέξη σε μία πρόταση, χρειάζεται να γνωρίζουμε τη λέξη που προηγήθηκε. Τα Αναδρομικά Νευρωνικά Δίκτυα (*Recurrent Neural Networks-RNNs*) ονομάζονται έτσι επειδή εκτελούν την ίδια ενέργεια για κάθε στοιχείο μιας ακολουθίας, με το αποτέλεσμα να παραμένει ανεξάρτητο από προηγούμενους υπολογισμούς. Ένας διαφορετικός τρόπος για να σκεφτούμε τα RNN είναι ότι διαθέτουν *μνήμη*, η οποία αποθηκεύει πληροφορία για ό,τι έχει υπολογιστεί μέχρι στιγμής. Ένα Αναδρομικό Νευρωνικό Δίκτυο περιέχει βρόχους (loops) που επιτρέπουν την πληροφορία να διαπερνάει τους νευρώνες κατά την ανάγνωση μίας εισόδου. Στο Σχήμα 2.21, ως x_t συμβολίζουμε την είσοδο, ως A ένα τμήμα του Δικτύου και ως h_t την έξοδό του. Θα μπορούσαμε να τροφοδοτήσουμε το σύστημα με τις λέξεις μίας πρότασης ή με τους χαρακτήρες μίας συμβολοσειράς.

Ορισμένα είδη Αναδρομικών Νευρωνικών Δικτύων είναι τα *LSTM, GRU* και *bi-directional RNN*. Κατεξοχήν, RNN χρησιμοποιούνται στους κλάδους της *Επεξεργασίας Φυσικής Γλώσσας* (Natural Language Processing-NLP), της *Όρασης Υπολογιστών* (Computer Vision), της *Ανάλυσης Βίντεο* (Video Analysis) και της *Δημιουργίας Εικόνων* (Image Generation). Το μεγαλύτερο πλεονέκτημα που προσφέρει η χρήση τους έγκειται στο γεγονός ότι οποιοσδήποτε αριθμός εισόδων και εξόδων μπορεί μέσω ενός δικτύου RNN να μετατραπεί σε ένα μοντέλο μίας εισόδου-μίας εξόδου ή πολλών εισόδων-πολλών εξόδων.

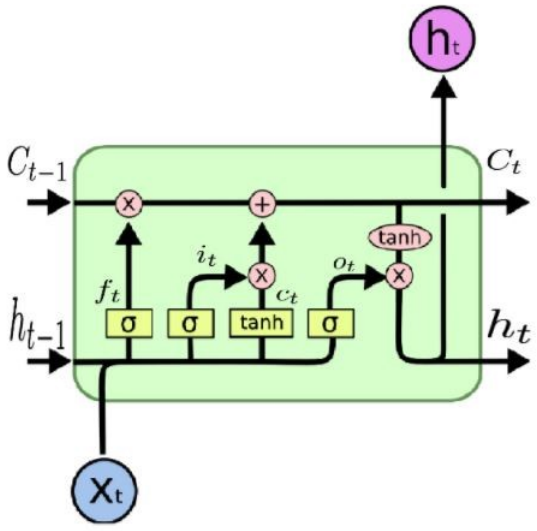


Σχήμα 2.21: Αναδρομικό Νευρωνικό Δίκτυο

2.7 Δίκτυα Μακράς Βραχυπρόθεσμης Μνήμης (LSTM)

Τα νευρωνικά δίκτυα Μακράς Βραχυπρόθεσμης Μνήμης (Long Short-Term Memory-LSTM) αποτελούν μία υποκατηγορία των Αναδρομικών Νευρωνικών Δικτύων (RNN) και μπορούν να εκπαιδεύονται στην εκμάθηση μακροσκελών ακολουθιών. Σχεδιάζονται με τρόπο ώστε να αποφεύγεται η εξάρτηση που μπορεί να διαθέτουν μεταξύ τους και έχουν την ικανότητα να απομνημονεύουν πληροφορία που βρίσκεται σε πολύ μακρινές χρονικές στιγμές. Η αρχιτεκτονική των LSTM δικτύων προτάθηκε το 1997 από τους Hochreiter και Schmidhuber, ενώ πλέον χρησιμοποιείται σε μία πληθώρα σύγχρονων προβλημάτων. Οι δομικές μονάδες ενός δικτύου LSTM μπορούν να γίνουν αντιληπτές ως κύτταρα (*cells*), τα οποία αποτελούνται από μία πύλη εισόδου (*input gate*), μία πύλη εξόδου (*output gate*) και μία πύλη λήθης (*forget gate*). Το κύτταρο απομνημονεύει τιμές ανά τυχαία χρονικά διαστήματα και οι τρεις πύλες αναλαμβάνουν τη διανομή της πληροφορίας εντός και εκτός του κυττάρου.

Τα δίκτυα που βασίζονται σε LSTM αρχιτεκτονική είναι κατάλληλα για προβλήματα ταξινόμησης, επεξεργασίας και εξαγωγής προβλέψεων βάσει χρονικών ακολουθιών δεδομένων. Βρίσκουν εφαρμογή στους τομείς της αναγνώρισης γραφικού χαρακτήρα (*handwriting gesture recognition*) ή της αναγνώρισης φωνής και βίντεο (*speech/video recognition*).



Σχήμα 2.22: Δίκτυο LSTM

Κεφάλαιο 3

Προσέγγιση του θέματος (Video Action Recognition-VAR)

Η κατανόηση της ψηφιακής πληροφορίας από τον άνθρωπο αποτελεί μία διαδικασία που συμβαίνει αβίαστα μέσω της όρασής μας. Μπορούμε με ευκολία να διακρίνουμε χαρακτηριστικά, αντικείμενα, γεγονότα ή ενέργειες που απεικονίζονται σε μία ψηφιακή εικόνα ή σε ένα ψηφιακό βίντεο. Επίσης, έχουμε την ικανότητα να αποκτάμε άμεσα μία αντίληψη για την ποιότητα της ανάλυσης του ψηφιακού πολυμέσου, να ανακαλούμε στη μνήμη μας κάποια σχετική εμπειρία που μπορεί να έχουμε βιώσει στο παρελθόν και να εξάγουμε χρήσιμη γνώση από αυτό που βλέπουμε. Ωστόσο, οι υπολογιστές δεν είναι σε θέση να εκτελέσουν αυτόματα τις αντίστοιχες λειτουργίες που πραγματοποιεί ο ανθρώπινος εγκέφαλος. Η *Υπολογιστική Όραση* (Computer Vision) αποτελεί ένα επιστημονικό πεδίο που εντάσσεται στον ευρύτερο κλάδο της Τεχνητής Νοημοσύνης (Artificial Intelligence-AI) και προσπαθεί να αναπαράγει μέσω αλγορίθμων την αίσθηση της όρασης, ώστε αυτή να χρησιμοποιηθεί από ηλεκτρονικούς υπολογιστές ή ρομποτικές εφαρμογές. Τα περισσότερα συστήματα Μηχανικής Όρασης συνδυάζουν την *Ψηφιακή Επεξεργασία Σήματος* (Digital Signal Processing) με τη *Μηχανική Μάθηση* (Machine Learning), ώστε να αναπτύξουν εύρωστα συστήματα ταξινόμησης εικόνων και κατ΄ επέκταση βίντεο.

Στο παρόν κεφάλαιο, επικεντρωνόμαστε στο ρόλο που διαδραματίζουν οι σύγχρονες τεχνικές *Βαθιάς Μηχανικής Μάθησης* (Deep Learning) στην αναγνώριση των ενεργειών που διαδραματίζονται κατά τη διάρκεια ενός ψηφιακού βίντεο. Εξετάζουμε μεθόδους και μοντέλα, τα οποία επεξεργάζονται ακολουθίες από πλαίσια (frames) ενός βίντεο και προβλέπουν την ενέργεια που απεικονίζεται σε αυτό.

3.1 Χρήσιμοι Ορισμοί

3.1.1 Όραση Υπολογιστών

Ο κλάδος της 'Υπολογιστικής Όρασης' (*Computer Vision-CV*, βλέπε εργασία [100]) ασχολείται με την αλγοριθμική αναπαράσταση τρισδιάστατων (3D) σκηνών, μέσω της επεξερ-

γασίας δισδιάστατων (2D) εικόνων που λαμβάνονται από οπτικούς αισθητήρες, όπως είναι οι κάμερες. Στη συνέχεια, στόχος είναι η εξαγωγή μίας πρόβλεψης ή απόφασης σχετικά με το αναπαριστούμενο αντικείμενο της αρχικής σκηνής. Τα συστήματα που βασίζονται στη Μηχανική Όραση δε διαθέτουν εμπειρία ούτε κάποιον ενσωματωμένο τρόπο αναγνώρισης προτύπων στις εικόνες, ενώ αδυνατούν να κατανοήσουν μικροαλλαγές στα απεικονιζόμενα αντικείμενα ή στην τοποθέτηση της κάμερας. Η μόνη ικανότητα που διαθέτουν είναι να διαχειρίζονται σύνολα ή διανύσματα τιμών, όπου η κάθε τιμή διαθέτει και ένα επιπλέον στοιχείο θορύβου, με αποτέλεσμα να παρέχονται ελλειπείς πληροφορίες για τα στοιχεία της εικόνας. Οι εφαρμογές της Όρασης Υπολογιστών προσπαθούν να μετατρέψουν αυτό το θορυβώδες σύνολο τιμών στην αντίληψη της πραγματικής εικόνας και των χαρακτηριστικών που διαθέτει. Ο συγκεκριμένος τομέας εφαρμογών σχετίζεται άμεσα με την ψηφιακή επεξεργασία εικόνων.

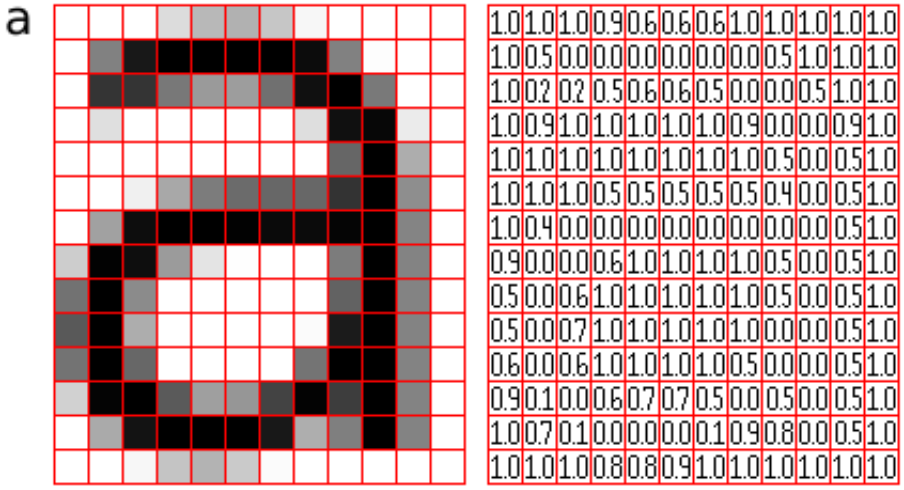
Υπάρχουν *αλγόριθμοι χαμηλού επιπέδου*, οι οποίοι λαμβάνουν μία εικόνα ως είσοδο και επιστρέφουν μία τροποποιημένη εικόνα ως έξοδο. Ακόμη, υπάρχουν *αλγόριθμοι ενδιάμεσου επιπέδου*, οι οποίοι παράγουν ως έξοδο ορισμένα χαρακτηριστικά ανωτέρου επιπέδου, όπως είναι οι ακμές μιας εικόνας. Τέλος, οι *αλγόριθμοι υψηλού επιπέδου* επικεντρώνονται στην επεξεργασία των εικόνων εισόδου για την εξαγωγή χρήσιμης πληροφορίας, δηλαδή για τον εντοπισμό συκεκριμένων χαρακτηριστικών και αντικειμένων.

Ως χαρακτηριστικά (features) θεωρούμε ορισμένα κομμάτια πληροφορίας, που στην περίπτωση των εικόνων μπορεί να πρόκειται για συγκεκριμένες δομές όπως σημεία, για απότομες αλλαγές στη φωτεινότητα ή για αντικείμενα. Αυτά τα features αποτελούν *Οπτικούς Περιγραφητές* (Visual Descriptors, βλέπε εργασία [107]) της εικόνας, οι οποίοι άλλοτε περιέχουν πληροφορίες χαμηλού επιπέδου όπως χρώμα, υφή, σχήμα και κίνηση και άλλες φορές περιέχουν πληροφορίες υψηλού επιπέδου σχετικά με αντικείμενα ή γεγονότα που παρουσιάζονται στην εικόνα. Η περιγραφή αυτών των χαρακτηριστικών χρησιμοποιείται συχνά ως τεχνική στους τομείς της *Μηχανικής Μάθησης* (Machine Learning-ML, βλέπε εργασία [90]), της *Αναγνώρισης Προτύπων* (Pattern Recognition) και της *Επεξεργασίας Εικόνας* (Image Processing).

3.1.2 Ψηφιακή Εικόνα

Από μαθηματική σκοπιά, ως 'Ψηφιακή Εικόνα' (Digital Image) ορίζεται μια δισδιάστατη συνάρτηση f(x,y), όπου x,y αποτελούν τις διαστάσεις της εικόνας ως προς τον χώρο, ενώ η τιμή της συνάρτησης υποδηλώνει την ένταση του χρώματος στο συγκεκριμένο σημείο (x,y). Το σημείο (x,y) προσδιορίζει ένα εικονοστοιχείο (*pixel*), η απόχρωση του οποίου μπορεί να είναι είτε ασπρόμαυρη είτε έγχρωμη. Μία εικόνα αποτελείται από πολλά κανάλια και το Ιστόγραμμα Χρώματος (Color Histogram) είναι μία αναπαράσταση της κατανομής των χρωμάτων στο κάθε κανάλι. Διακρίνουμε τα κανάλια *RGB (Red,Green,Blue)*, τα οποία είναι επιρρεπή σε αλλαγές της φωτεινότητας, και τα κανάλια *HSV (Hue,Saturation,Lightness)*.

Συχνά διαχωρίζουμε τις ψηφιακές εικόνες σε επιμέρους τμήματα (*segments*) αναθέτοντας μία ετικέτα στο κάθε εικονοστοιχείο (pixel). Pixels με όμοια οπτικά χαρακτηριστικά, όπως χρώμα, ένταση ή υφή, αντιστοιχίζονται στην ίδια κατηγορία και αποκτούν κοινή ετικέτα. Αυ-



Σχήμα 3.1: Αναπαράσταση των pixel του ψηφιακού γράμματος 'a'

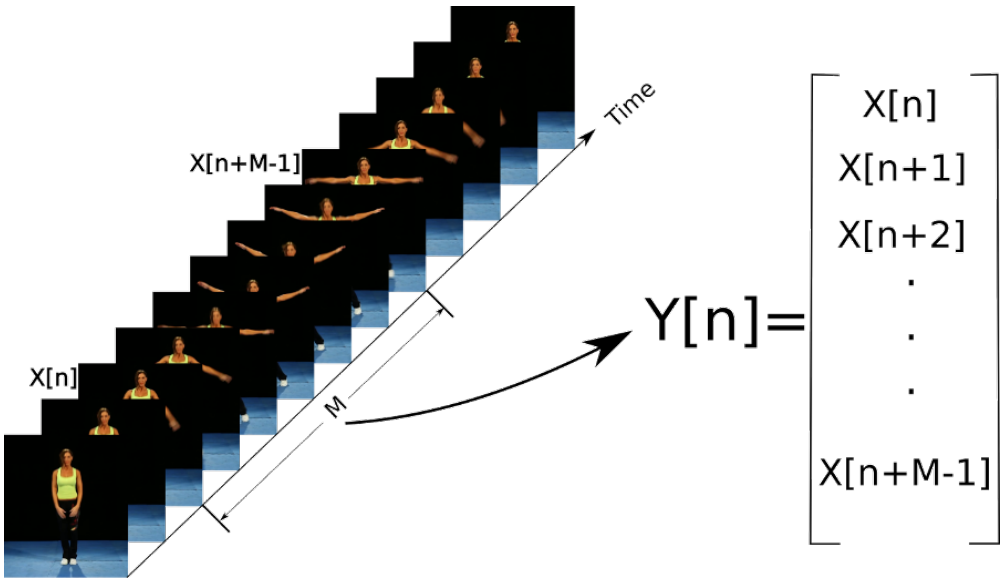
τή η διαδικασία είναι γνωστή ως *Image Segmentation* και χρησιμοποιείται κατεξοχήν για την ανίχνευση κινήσεων ή τον εντοπισμό διαφορών μεταξύ παρόμοιων εικόνων. Επιπλέον, μία ψηφιακή εικόνα χωρίζεται σε διαφορετικά επίπεδα (layers), τα οποία αφορούν το προσκήνιο (foreground) και το παρασκήνιο (background). Το κάθε επίπεδο διαθέτει τα δικά του χαρακτηριστικά και αντιμετωπίζεται ως ξεχωριστή οντότητα. Για παράδειγμα, μέσω του διαχωρισμού της εικόνας σε επίπεδα, είναι δυνατή η αφαίρεση του φόντου (*background subtraction*) προκειμένου να εντοπιστούν αντικείμενα που βρίσκονται στο προσκήνιο της εικόνας.

Σε μία εικόνα εμφανίζεται ένας αριθμός από στοιχεία ενδιαφέροντος, δηλαδή χαρακτηριστικά (features). Μία επιθυμητή ιδιότητα ενός ανιχνευτή χαρακτηριστικών είναι η επανάληψη, δηλαδή η ανίχνευση του ίδιου χαρακτηριστικού σε δύο ή περισσότερα διαφορετικές εικόνες ή frames. Η ανίχνευση χαρακτηριστικών (*feature detection*) αποτελεί μία λειτουργία επεξεργασίας εικόνας χαμηλού επιπέδου και ουσιαστικά εξετάζει κάθε εικονοστοιχείο ώστε να δει αν υπάρχει μοτίβο που να αποτελεί χαρακτηριστικό. Για τη σωστή και ολοκληρωμένη εξαγωγή χαρακτηριστικών (feature extraction), απαιτείται η χρήση του κατάλληλου αλγορίθμου, παράλληλα με την αξιοποίηση φίλτρων για την επεξεργασία της εικόνας, εφόσον έχει συνδυαστεί με τον κατάλληλο περιγραφέα. Μετά το feature extraction, προκειμένου να ανιχνεύσουμε κάποιο αντικείμενο στην εικόνα απαιτείται να εκπαιδεύσουμε τον κατάλληλο αλγόριθμο ταξινόμησης ώστε να μάθει τις διαφορές μεταξύ των διαφορετικών κατηγοριών εικόνων. Οι αλγόριθμοι έχουν τη δυνατότητα να ανιχνεύουν μόνο στοιχεία στα οποία έχουν εκπαιδευτεί, μέσω της κατάλληλης εξαγωγής χαρακτηριστικών και της χρήσης περιγραφέων εικόνας.

Οι σύγχρονες εφαρμογές αναγνώρισης εικόνων κάνουν χρήση τεχνικών της Βαθιάς Μηχανικής Μάθησης και συγκεκριμένα των Συνελικτικών Νευρωνικών Δικτύων (CNN). Το πλεονέκτημα αυτών των μεθόδων έγκειται στο ότι καθιστούν περιττή την προεπεξεργασία των δεδομένων εισόδου, δηλαδή δεν είναι αναγκαία η χρήση κάποιου περιγραφητή για την εξαγωγή των χαρακτηριστικών της εικόνας.

3.1.3 Ψηφιακό Βίντεο

Σε ένα 'Ψηφιακό Βίντεο' (Digital Video, βλέπε εργασία [95]) αναπαρίστανται κινούμενες οπτικές εικόνες με τη μορφή κωδικοποιημένων ψηφιακών δεδομένων (digital data), σε αντίθεση με τα αναλογικά σήματα που χρησιμοποιούνται για την αναπαράσταση ενός Αναλογικού Βίντεο (Analog Video). Ένα βίντεο αποτελείται από μία ακολουθία ψηφιακών εικόνων, γνωστών ως πλαίσια (*frames*), τα οποία διαδέχονται το ένα το άλλο με ένα συγκεκριμένο ρυθμό (*frames per second-fps*). Το κάθε πλαίσιο είναι μία Ψηφιακή Εικόνα, δηλαδή ένας ορθογώνιος χάρτης αποτελούμενος από ένα σύνολο εικονοστοιχείων (*pixels*). Η μοναδική ιδιότητα που διαθέτουν τα pixels είναι το χρώμα τους, που δεν είναι τίποτε άλλο παρά ένας συγκεκριμένος αριθμός δυαδικών ψηφίων (bits). Όσο μεγαλύτερο είναι το βάθος χρώματος (color depth) ενός βίντεο, δηλαδή όσο περισσότερα bits χρησιμοποιούνται για την χρωματική αναπαράσταση των pixels, τόσο αυξάνονται οι χρωματικές διακυμάνσεις που μπορούν να αποτυπωθούν. Συχνά, στις εφαρμογές του '*Video Action Recognition-VAR*' χρησιμοποιούμε μεμονωμένα πλαίσια του βίντεο, τα οποία αποκαλούνται *still frames* και είναι στατικές εικόνες που έχουν ληφθεί από το μία ακολουθία 'κινούμενων' πλαισίων.



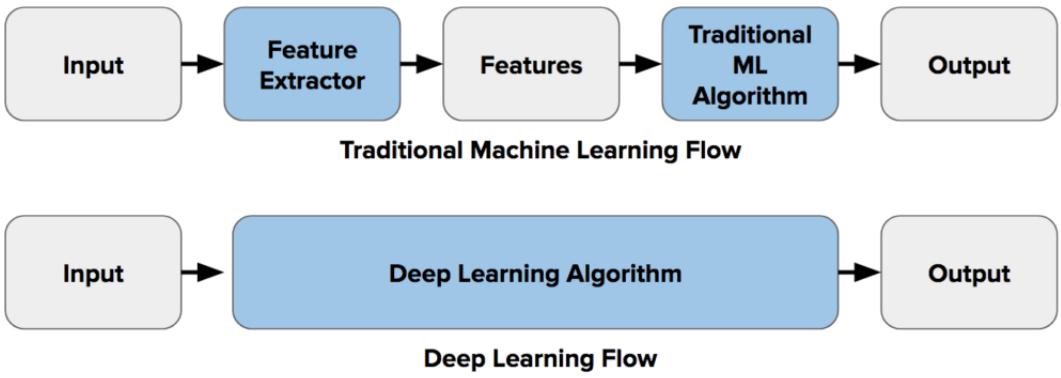
Σχήμα 3.2: Αναπαράσταση διαδοχικών πλαισίων ενός βίντεο

Ένα μέτρο που χρησιμοποιείται για τον προσδιορισμό του ρυθμού με τον οποίο λαμβάνουμε πληροφορία από μία ακολουθία ψηφιακού βίντεο είναι το *bit rate*. Σε περίπτωση που μιλάμε για βίντεο που δεν έχει συμπιεστεί, το bit rate αναφέρεται απευθείας στην ποιότητα του βίντεο. Το μέγεθος ενός βίντεο (video size) εξαρτάται από το bit rate και από τη διάρκειά του. Μέσω της συμπίεσης ενός βίντεο (video compression), μειώνεται το bit rate, με αποτέλεσμα να επηρεάζει σε μικρότερο βαθμό την ποιότητα του βίντεο.

3.1.4 Διαφορά μεταξύ handcrafted features και learned features

Η '*Αναγνώριση Ανθρώπινων Ενεργειών*' βασίζεται στον εντοπισμό χαρακτηριστικών στο κάθε πλαίσιο (frame) του βίντεο και στην παρατήρηση του τρόπου μεταβολής τους κατά μήκος διαδοχικών πλαισίων. Ορισμένα από αυτά τα χαρακτηριστικά αποκαλούνται '*handcrafted features*' διότι προκύπτουν άμεσα από τις τιμές των εικονοστοιχείων (pixels) του κάθε πλαισίου, ενώ τα '*learned features*' εξάγονται μετά από επεξεργασία.

Ειδικότερα, με τον όρο *handcrafted features* αναφερόμαστε στις εγγενείς ιδιότητες των εικόνων, δηλαδή στην πληροφορία που περιέχουν και την οποία μπορούμε να λάβουμε μέσω ποικίλων αλγοριθμικών τεχνικών. Δύο παραδείγματα handcrafted χαρακτηριστικών είναι οι *ακμές* (edges) και οι *γωνίες* (corners) των εικόνων. Ένας αλγόριθμος που στοχεύει στην ανίχνευση ακμών (*edge detector*) βασίζεται στις ξαφνικές αλλαγές στην ένταση (intensity) κάποιων pixel της εικόνας. Σε περίπτωση που αναφερόμαστε σε gray-scale εικόνες και χρησιμοποιούμε 8-bit για την αναπαράσταση της πληροφορίας του χρώματος, το κάθε pixel μπορεί να λάβει μία από τις 2^8 $(= 256)$ πιθανές τιμές. Η τιμή '0' θεωρούμε ότι αντιστοιχεί στο απόλυτο μαύρο χρώμα, ενώ η τιμή '255' στο απόλυτο άσπρο. Περιοχές της εικόνας όπου παρατηρούνται σημαντικές αλλαγές στις τιμές των pixel αποτελούν ενδείξεις ύπαρξης κάποιας ακμής. Ο αλγόριθμος που μόλις περιγράψαμε αποτελεί την πιο απλή και βασική μορφή αλγορίθμου ανίχνευσης ακμών, ενώ πιο δομημένες προτάσεις αποτελούν οι *Harris corners detectors* και *Hogg detectors*.

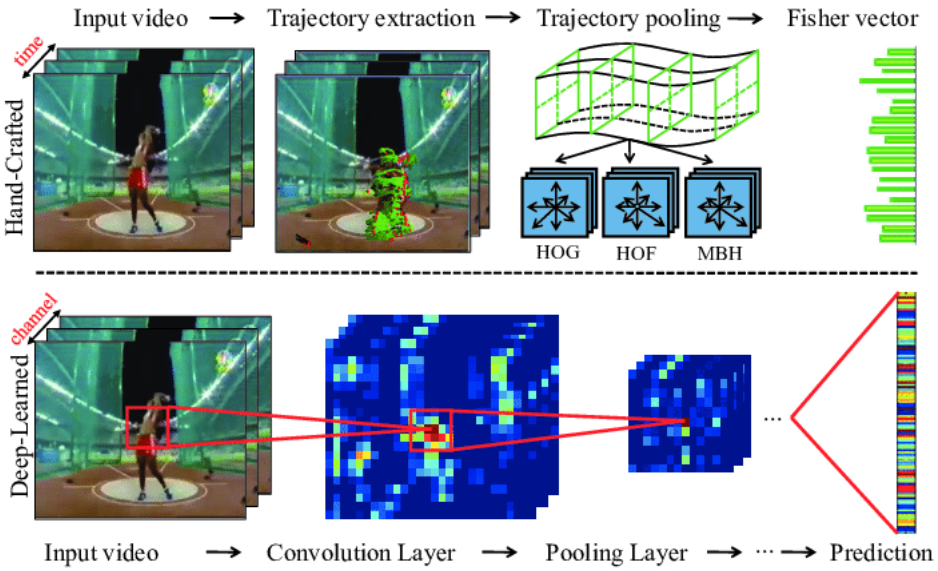


Σχήμα 3.3: Διαφορετικές προσεγγίσεις αξιοποίησης των χαρακτηριστικών των πλαισίων

Τα *handcrafted features* έχουν χρησιμοποιηθεί σε συνδυασμό με τις παραδοσιακές προσεγγίσεις της Μηχανικής Μάθησης για την Αναγνώριση Αντικειμένων (Object Recognition) και την Όραση Υπολογιστών (Computer Vision), όπως είναι οι *Μηχανές Διανυσμάτων Κατάστασης* (Support Vector Machines-SVM). Ωστόσο, οι πιο πρόσφατες προσεγγίσεις είναι δομημένες με τέτοιον τρόπο ώστε τα χαρακτηριστικά των πλαισίων να εξάγονται χωρίς να βασίζονται στα handcrafted features. Για παράδειγμα, οι αρχιτεκτονικές των *CNN* εκπαιδεύονται κατάλληλα και μέσω της εφαρμογής ειδικών φίλτρων (kernels/filters) πετυχαίνουν την εξαγωγή των *learned features* των πλαισίων.

3.2 Επισκόπηση των μεθόδων που έχουν χρησιμοποιηθεί στην αναγνώριση ενεργειών

Οι πρώτες προσπάθειες που έχουν καταγραφεί στην '*Αναγνώριση ανθρώπινων ενεργειών σε βίντεο*' (Video Action Recognition-VAR) στοχεύουν στον εντοπισμό της βασικής ανθρώπινης φιγούρας που απεικονίζεται στο κάθε βίντεο και στο διαχωρισμό της από τον περιβάλλων χώρο. Στη συνέχεια, η έρευνα επικεντρώνεται στην ανίχνευση κινήσεων και ενεργειών που πραγματοποιούνται από τους πρωταγωνιστές του βίντεο. Προς αυτή την κατεύθυνση, διαδοχικά frames υποβάλλονται σε επεξεργασία ώστε να εξαχθούν τα χαρακτηριστικά του καθενός, ενώ πλέον χρησιμοποιούνται τεχνικές *Μηχανικής Μάθησης* για την εκπαίδευση κατάλληλων μοντέλων τα οποία θα μπορούν να εντοπίζουν συγκεκριμένα χαρακτηριστικά σε μεγάλες συλλογές βίντεο.



Σχήμα 3.4: Εντοπισμός χωροχρονικών σημείων ενδιαφέροντος

Το VAR αποτελεί ένα πρόβλημα ταξινόμησης (classification) και προσεγγίζεται μέσω *Επιβλεπόμενης Μηχανικής Μάθησης* (Supervised Learning), όπου το σύστημά μας εκπαιδεύεται χρησιμοποιώντας δεδομένα της μορφής: $\{(x_1, y_1), (x_2, y_2), (x_3, y_3), ..., (x_n, y_n)\}$. Ως $\{x_1, x_2, x_3, ..., x_n\}$ συμβολίζουμε τα τρισδιάστατα (R^3) βίντεο που τροφοδοτούνται στην είσοδο του μοντέλου, ενώ τα $\{y_1, y_2, y_3, ..., y_n\}$ ($y_i \in \Omega$) αποτελούν το είδος της ενέργειας (action label) που αντιστοιχεί στην κάθε είσοδο. Στόχος ενός προβλήματος αναγνώρισης ενεργειών είναι η προσέγγιση μίας συνάρτησης ταξινόμησης της μορφής: $F(x) : R^3 \rightarrow \Omega$.

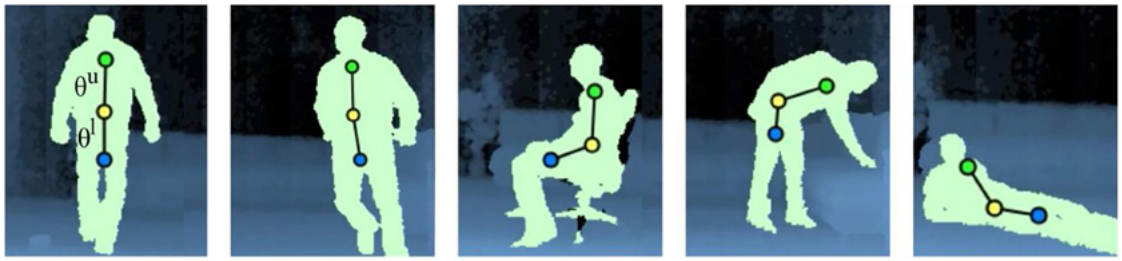
Στη συνέχεια, παρουσιάζουμε μία σφαιρική εικόνα των ερευνών που έχουν κατά καιρούς διεξαχθεί στο πλαίσιο του VAR. Οι σχετικές εργασίες κάνουν χρήση συνόλων δεδομένων (dataset) για την εκπαίδευση και αξιολόγηση των μοντέλων τους.

3.2.1 Προσεγγίσεις βασισμένες σε handcrafted features

Πριν την άνθιση του κλάδου της *Βαθιάς Μηχανικής Μάθησης* (Deep Learning), το πρόβλημα της αναγνώρισης ενεργειών σε βίντεο προσεγγιζόταν μέσω αλγορίθμων του Computer Vision. Συνήθως εντοπίζονταν τοπικά χαρακτηριστικών υψηλής διαστατικότητας από τα πλαίσια του βίντεο, τα οποία περιέγραφαν μία συγκεκριμένη περιοχή και εξάγονταν είτε σε πυκνά (βλέπε εργασία [102]) είτε σε αραιά σύνολα σημείων ενδιαφέροντος (interest points, βλέπε εργασία [38]). Η κωδικοποίηση των χαρακτηριστικών γινόταν μέσω της μεθόδου '*Bag of Visual Words*' και η τελική απόφαση του αλγορίθμου προέκυπτε μέσω κάποιας '*Μηχανής Διανυσμάτων Υποστήριξης*' (Support Vector Machine-SVM). Ακολουθεί ανάλυση των κυριότερων τεχνικών που βασίζονται στα handcrafted features των πλαισίων.

3.2.1.1 Μοντέλα περιγραφής του ανθρώπινου σώματος (Body Models)

Στο πλαίσιο των πρώτων προσπαθειών που είχαν γίνει για τον εντοπισμό κινήσεων σε βίντεο, είχε προταθεί η χρήση μιας απλοποιημένης αναπαράστασης του ανθρώπινου σώματος μέσω φωτεινών πηγών (βλέπε εργασία [46]). Στις αρθρώσεις του σώματος είχαν τοποθετηθεί *φωτεινές πηγές*, οι οποίες άλλαζαν θέση κατά τη μετακίνηση του ανθρώπου (Moving Light Displays-MLD) και βάσει του μοτίβου της κίνησης που προέκυπτε, προσδιοριζόταν η συντελούμενη ενέργεια.



Σχήμα 3.5: Εντοπισμός ενεργειών μέσω Κινούμενων Φωτεινών Πηγών

Η μέθοδος των '*Κινούμενων Φωτεινών Πηγών*' οδήγησε στην αναπαράσταση της κίνησης (motion) μέσω μίας ακολουθίας δισδιάστατων εικόνων, η καθεμία από τις οποίες αντιστοιχίζεται σε μία ενέργεια. Λόγω μικροαλλαγών που ενδέχεται να παρουσιάζονται στη στάση του σώματος διαφορετικών ανθρώπων ή λόγω διαφοροποιήσεων στο περιβάλλον όπου διαδραματίζεται κάποια ενέργεια, οι επιστήμονες αναγκάστηκαν να κατασκευάσουν τρισδιάστατα μοντέλα προκειμένου να αναπαραστήσουν το ανθρώπινο σώμα (βλέπε εργασία [64]). Έτσι, η αναγνώριση των ανθρώπινων δραστηριοτήτων έπαψε να εξαρτάται από αυτές τις συνθήκες και παρουσίασε ποσοστό επιτυχίας 90%.

3.2.1.2 Ολιστικές Αναπαραστάσεις (Holistic Representations)

Σε αντίθεση με πριν, οι *Ολιστικές Αναπαραστάσεις* δεν απαιτούν τον εντοπισμό και την επισήμανση των διαφορετικών μερών του ανθρώπινου σώματος. Στηρίζονται στην προεπεξερ-

γασία των εικόνων εκτελώντας συγκεκριμένες διαδικασίες, όπως είναι η αφαίρεση του παρασκηνίου (*background subtraction*) και η εξαγωγή των χαρακτηριστικών (*feature extraction*), ενώ παράλληλα χρησιμοποιούν περιγράμματα (*contours*) ή σκιαγραφήσεις (*silhouettes*) του ανθρώπινου σώματος (βλέπε εργασία [12]).

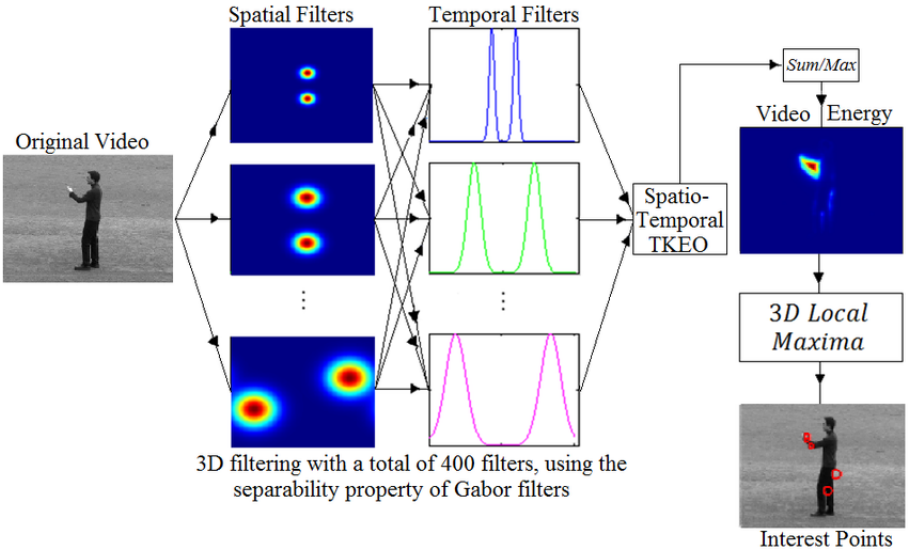
Αρχικά, η έρευνα βασίστηκε σε εικόνες με μαύρο παρασκήνιο (background) και επικεντρώθηκε στη δημιουργία ενός μοντέλου όπου οι εικόνες διαδοχικών χειρονομιών συνδέονται μεταξύ τους ώστε να προσδιορίσουν κάποια ενέργεια. Αυτή η μέθοδος τροποποιήθηκε (βλέπε εργασία [106]) ώστε να μετατρέψει τα διαδοχικά frames του βίντεο σε ένα διάνυσμα ενοποιημένων χαρακτηριστικών της εικόνας, όπου θα γίνεται αποκλειστική χρήση σκιαγραφήσεων (silhouettes). Στη συνέχεια, αυτό το διάνυσμα χαρακτηριστικών (*feature vector*) αξιολογείται μέσω του 'Κρυφού Μαρκοβιανού Μοντέλου' (*Hidden Markov Model-HMM*). Ακολούθησε το έργο των Davis και Bobick (βλέπε εργασία [9]), οι οποίοι δημιούργησαν τις '*Εικόνες Ιστορίας της Κίνησης*' (Motion History Images-MHI) και τις '*Εικόνες Ενέργειας της Κίνησης*' (Motion Energy Images-MEI) λαμβάνοντας τις σκιαγραφήσεις στο πεδίο του χρόνου, προκειμένου να αξιοποιήσουν την πληροφορία του κάθε πλαισίου.

Στην πορεία, η επιστημονική κοινότητα εστίασε την έρευνά της στην χρήση '*Πυκνών Οπτικών Ροών*' (Dense Optical Flows, βλέπε εργασία [20]). Τα διανυσματικά πεδία των Οπτικών Ροών δεν απαιτούν την αφαίρεση του παρασκηνίου (background subtraction), αλλά παραμένουν ευαίσθητα σε αλλαγές του φωτεισμού ή των επιφανειών των αντικειμένων. Για την εξαγωγή των χαρακτηριστικών του κάθε πλαισίου (feature extraction) και τον εντοπισμό αντικειμένων (object detection), γίνεται χρήση κλίσεων (gradients) και συγκεκριμένα χρησιμοποιούνται '*Ιστογράμματα Προσανατολισμένων Κλίσεων*' (Histograms of Oriented Gradients-HOG, βλέπε εργασία [21, 96]). Παρ' όλα αυτά, η αδυναμία των Ολιστικών Αναπαραστάσεων στη διαχείριση διακυμάνσεων της οπτικής γωνίας του παρατηρητή-κάμερας (*viewpoint variations*), οδήγησε στην εκτεταμένη χρήση των *Τοπικών Αναπαραστάσεων*.

3.2.1.3 Τοπικές Αναπαραστάσεις (Local Representations)

- **Εντοπισμός Σημείων Ενδιαφέροντος** (Interest Points Detection): η έρευνα που πραγματοποίηθηκε πάνω στα χωροχρονικά σημεία ενδιαφέροντος (space-time interest points) ενός βίντεο, έδωσε το έναυσμα για την χρήση *Τοπικών Αναπαραστάσεων* στην εξαγωγή χαρακτηριστικών (feature extraction) από μία εικόνα. Έγινε χρήση του ανιχνευτή γωνιών '*Harris corner detector*', ο οποίος βασίζεται στην ιδέα ενός μικρού κινούμενου παραθύρου, μέσα στο οποίο παρατηρούνται αλλαγές στη φωτεινότητα των pixel. Έτσι, ο αλγόριθμος εντοπίζει αλλαγές κατά μήκος σημείων που παρουσιάζουν μεγάλη αστάθεια στην κίνησή τους, όπως φαίνεται στο Σχήμα 3.6.

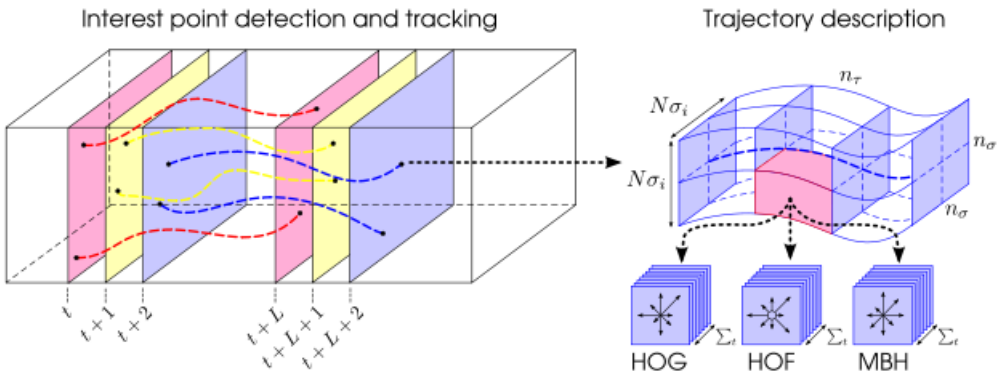
- **Τοπικοί Περιγραφητές** (Local Descriptors): Οι τοπικοί περιγραφητές διακρίνονται σε *περιγραφητές ακμών* (edge detectors) και *περιγραφητές κίνησης* (motion detectors). Η παλαιότερη χρήση κυβοειδών μοντέλων για την αναπαράσταση του ανθρώπινου σώματος δεν αποδείχθηκε αρκετά αποτελεσματική, οπότε κρίθηκε αναγκαία η χρήση διαφορετικών μεθόδων για τη μελέτη της παρατηρούμενης κίνησης κατά μήκος των πλαισίων



Σχήμα 3.6: Χρήση τρισδιάστατων φίλτρων Gabor για την ανίχνευση χωροχρονικών σημείων ενδιαφέροντος

ενός βίντεο. Τα '*Ιστογράμματα Προσανατολισμένων Κλίσεων*' (Histogram of Oriented Gradients-HOG) χρησιμοποιήθηκαν στον εντοπισμό της κίνησης (βλέπε εργασία [54]) και επεκτάθηκαν στο χωροχρονικό πεδίο. Μάλιστα, η ίδια ιδέα εφαρμόστηκε και για τα πεδία οπτικών ροών (optical flow fields), δεδομένου ότι κωδικοποιούν την παρατηρούμενη κίνηση στα βίντεο σε επίπεδο-pixel. Έτσι, δημιουργήθηκαν τα '*Ιστογράμματα Οπτικών Ροών*' (Histogram of optical Flow-HoF) και τα '*Ιστογράμματα Περιθωρίων της Κίνησης*' (Motion Boundary Histogram-MBH, βλέπε εργασία [23]). Ωστόσο, ο υπολογισμός των πεδίων οπτικών ροών έχει υψηλές υπολογιστικές απαιτήσεις, με αποτέλεσμα να γίνεται χρήση τεχνικών αποσυμπίεσης (decompression techniques).

- **Προσεγγίσεις βασισμένες στις Τροχιές** (Trajectory-based Approaches): το σημαντικότερο μειονέκτημα της χρήσης κυβοειδών αναπαραστάσεων έγκειται στην πιθανή μετατόπιση των εντοπισμένων σημείων ενδιαφέροντος εντός των χρονικών περιθωρίων του κύβου. Γι' αυτό κρίθηκε αναγκαία η αξιοποίηση της *τροχιάς* που διαγράφεται κατά μήκος των διαδοχικών πλαισίων ενός βίντεο, καθώς εκτυλίσσεται η εκάστοτε ενέργεια που απεικονίζεται. Ως τροχιά μιας ενέργειας ('*Action Trajectory*') εννοούμε τη μεταβολή ενός feature του βίντεο στο πεδίο του χρόνου. Οι αναπαραστάσεις που βασίστηκαν σε action trajectories (βλέπε εργασίες [67, 66, 102]) ενσωμάτωσαν τα ιστογράμματα MBH, HoG και HoF για να δημιουργήσουν πλήρεις αναπαραστάσεις των χαρακτηριστικών των πλαισίων, ενώ ο υπολογισμός των trajectories έγινε μέσω χρήσης οπτικών ροών. Μάλιστα, σε μεταγενέστερες εργασίες χρησιμοποιήθηκαν *περιγραφητές SURF* και *πυκνές οπτικές ροές* (dense optical flows) προκειμένου να μειωθούν οι επιδράσεις των κινήσεων της κάμερας καταγραφής. Τέλος, στην εργασία [77] προστέθηκε η χρήση πολλαπλώς επιπέδων στοιβάδων, καθεμία από τις οποίες αποτελούταν από Fisher Vectors.



Σχήμα 3.7: Εντοπισμός χωροχρονικών σημείων ενδιαφέροντος

Οι προσεγγίσεις που βασίζονται στα handcrafted features των εικόνων είναι σύνθετες στην κατασκευή τους και τροποποιούνται με δυσκολία, με αποτέλεσμα να μην παρέχουν μία ενοποιημένη καθολική λύση στο πρόβλημα του VAR. Αυτές οι τεχνικές αντικαταστάθηκαν από την χρήση μεθόδων της Βαθιάς Μηχανικής Μάθησης (Deep Learning).

3.2.2 Προσεγγίσεις βασισμένες σε learned features

Χάρη στις τεχνικές της *Βαθιάς Μηχανικής Μάθησης* (Deep Learning), η εκμάθηση των χαρακτηριστικών των εικόνων συμβαίνει παράλληλα με την ταξινόμησή τους. Τα *Συνελικτικά Νευρωνικά Δίκτυα* (CNN) έφεραν επανάσταση στον χώρο του image classification και του image recognition, ενώ αρκετές είναι οι εφαρμογές των *Αναδρομικών Νευρωνικών Δικτύων* (RNN) και των *Δικτύων Μακράς Βραχυπρόθεσμης Μνήμης* (LSTM).

- **Συνελικτικά Νευρωνικά Δίκτυα** (CNN) [92, 91]: το συγκεκριμένο είδος δικτύων αποτελείται από έναν αριθμό συνελικτικών επιπέδων, καθένα από τα οποία είναι υπεύθυνο για την εξαγωγή διαφορετικών χαρακτηριστικών από τα πλαίσια του βίντεο. Τα επίπεδα χαμηλότερου επιπέδου αναλαμβάνουν την εξαγωγή απλών χαρακτηριστικών των εικόνων, ενώ τα υψηλότερα επίπεδα χρησιμοποιούν φίλτρα για να εξάγουν πιο σύνθετα χαρακτηριστικά. Ο τρόπος σχεδιασμού των φίλτρων γίνεται βάσει της αρχής του *διαμοιρασμού βαρών* (weight sharing), η οποία επιτρέπει τη μείωση του αριθμού των παραμέτρων που προορίζονται για εκπαίδευση. Το κάθε επίπεδο αυξάνει το βάθος και την πολυπλοκότητα του δικτύου, ενώ την ίδια στιγμή αυξάνει τις διαστάσεις των *convoluted features*. Τα CNN χρησιμοποιούνται εκτενώς διότι αποτελούν έναν αποτελεσματικό τρόπο εκμάθησης χαρακτηριστικών και μπορούν να χρησιμοποιηθούν ως end-to-end μοντέλα ταξινόμησης.

- **Αναδρομικά Νευρωνικά Δίκτυα** (RNN) [60]: τα δίκτυα RNN έχουν τη δυνατότητα να μοντελοποιούν ακολουθιακές συμπεριφορές χάρη στις συνδέσεις ανάδρασεις που παρουσιάζει η αρχιτεκτονική τους. Βρήκαν μεγάλη εφαρμογή στους τομείς της *αναγνώρισης γραφικού χαρακτήρα* (handwriting recognition, βλέπε εργασία [19]) και της

αναγνώρισης φωνής (speech recognition, βλέπε εργασία [53]), με αποτέλεσμα να επεκταθεί η χρήση τους στη μοντελοποίηση των χρονικών συσχετίσεων μεταξύ πλαισίων κάποιου βίντεο. Συγκεκριμένα, κάθε δίκτυο RNN που προορίζεται για το *Video Action Recognition* βασίζεται στις ανανεώσεις του διανύσματος τρέχουσας μνήμης του βάσει του τρέχοντος frame, του προηγούμενου διανύσματος μνήμης και της προηγούμενης θέσης ενός αντικειμένου που απεικονίζεται στο βίντεο.

- **Δίκτυα Μακράς Βραχυπρόθεσμης Μνήμης** (LSTM) [105]: συνήθως αυτά τα δίκτυα χρησιμοποιούνται σε συνδυασμό με τα CNN ή τα RNN, προκειμένου να αντιμετωπίσουν το πρόβλημα της κλίσης *(gradient problem)*.

3.2.2.1 Συνδυασμός handcrafted features και deep classifiers

Αρκετές μέθοδοι που προσπαθούν να επιλύσουν το πρόβλημα του VAR βασίζονται σε handcrafted features και προσπαθούν να ενσωματώσουν τη διάσταση του χρόνου από διαδοχικά πλαίσια βίντεο, ώστε η εκπαίδευση του *Βαθιού Νευρωνικού Δικτύου* να αρχίσει αφού θα έχουν ήδη εξαχθεί ορισμένα χαρακτηριστικά των πλαισίων. Στην εργασία [52] προτάθηκε η χρήση μίας τροποποιημένης μορφής CNN, στα οποία η πληροφορία χαμηλού επιπέδου αναπαρίσταται από handcrafted features. Η ακολουθία ενεργειών ενός οποιουδήποτε προσώπου σε ένα βίντεο παράγει έναν τρισδιάστατο όγκο, ο οποίος εξάγεται μέσω τρισδιάστατων *φίλτρων Gabor* (βλέπε Σχήμα 3.6 και εργασία [47]). Τα συγκεκριμένα φίλτρα χρησιμοποιούν μεμονωμένα πλαίσια ώστε να εξάγουν την πληροφορία που αφορά το εξωτερικό περίγραμμα του πρωταγωνιστή ενός βίντεο ή βασίζονται σε πολλαπλά frames με σκοπό την παραγωγή ενός χωροχρονικού όγκου. Η αξία αυτών των όγκων έγκειται στο ότι καθιστούν τις ενέργειες ανεξάρτητες της οπτικής γωνίας της κάμερας καταγραφής (viewpoint variations). Ένα *3D CNN* εφαρμόζεται ξεχωριστά στο κάθε spatio-temporal volume και πραγματοποιείται το feature extraction. Στη συνέχεια, τα χαρακτηριστικά που λαμβάνουμε αντιστοιχίζονται σε συγκεκριμένες κατηγορίες βάσει ενός μοντέλου ταξινόμησης (βλέπε εργασία [52]). Μετά από έρευνα των Jhuang et al. (βλέπε εργασία [41]), δημιουργήθηκε ένα *δίκτυο εμπρόσθιας τροφοδότησης* (feed-forward network), το οποίο εντοπίζει χωροχρονικά χαρακτηριστικά αυξανόμενης πολυπλοκότητας προκειμένου να προσδιοριστούν οι μονάδες που παρουσιάζουν ευαισθησία στην κίνηση.

3.2.2.2 Συνδυασμός learned features και deep classifiers

Τα *3D CNN* στοχεύουν στην εξαγωγή χαρακτηριστικών που αφορούν τη θέση των αντικειμένων και των προσώπων ενός βίντεο και κάνουν χρήση δισδιάστατων μετασχηματισμών, αξιοποιώντας την τρίτη διάσταση για την εξαγωγή της χρονικής πληροφορίας. Τα 3D CNN, όπως παρουσιάστηκαν στην εργασία [42] των Ji et al., εφαρμόζουν ένα 3D φίλτρο (filter/kernel), το οποίο έχει προκύψει από την εφαρμογή του ίδιου 2D φίλτρου σε μία συγκεκριμένη θέση πολλαπλών πλαισίων του βίντεο. Ως αποτέλεσμα, τα χαρακτηριστικά που προκύπτουν από τις 3D συνελίξεις παραμένουν ανεξάρτητα από τις μεταβολές στον χώρο κατά το πέρας του χρόνου. Έχει αποδειχθεί ότι τα 3D CNN οδηγούν σε καλύτερα αποτελέσματα

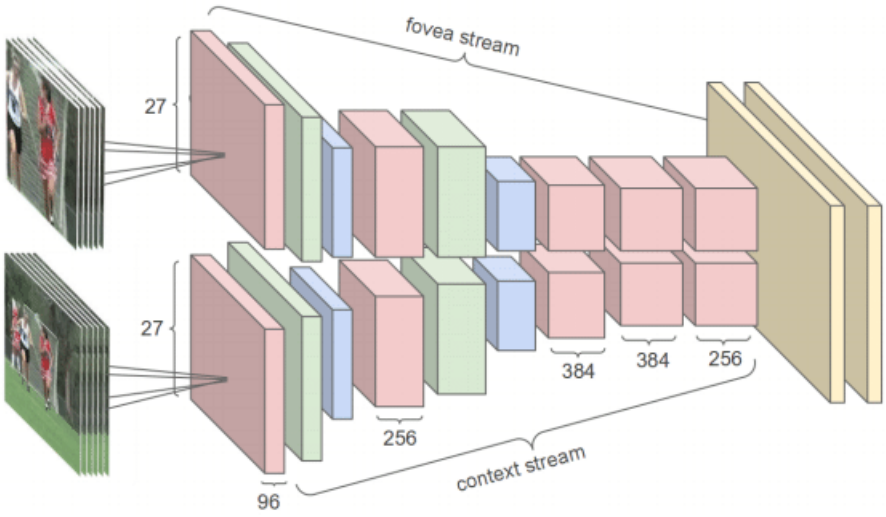
σε σχέση με τις παραδοσιακές εφαρμογές των 2D CNN (βλέπε εργασία [43]). Ωστόσο, οι περισσότερες αρχιτεκτονικές των 3D CNN που έχουν κατασκευαστεί με αυτόν τον τρόπο παρουσιάζουν ένα άνω όριο στον αριθμό των πλαισίων που χρησιμοποιούν για την εξαγωγή της χρονικής πληροφορίας, ενώ απαιτούν αυξημένη υπολογιστική ισχύ για τη λειτουργία τους και ένα μεγάλο αριθμό εκπαιδευμένων δεδομένων. Αυξάνοντας το χρονικό βάθος των δικτύων, δηλαδή χρησιμοποιώντας μεγαλύτερα χρονικά διαστήματα για την εκτέλεση των τρισδιάστατων συνελίξεων (βλέπε εργασία [98]), η επίδοση του 3D CNN βελτιώνεται σημαντικά.

Έκτοτε, η επιστημονική έρευνα επικεντρώθηκε στον τρόπο με τον οποίο μπορούν τα Βαθιά Νευρωνικά Δίκτυα να ενσωματώσουν με επιτυχία τη διάσταση του χρόνου. Οι Karpathy et al. στην εργασία τους [48] δοκίμασαν τις επιδόσεις διαφορετικών μοντέλων που συνδυάζουν την χωρική και χρονική πληροφορία μέσω *early fusion*, *late fusion* και *slow fusion*. Ορισμένες προσεγγίσεις χρησιμοποιούν μόνο ένα πλαίσιο από το κάθε βίντεο (*single-frame approach*) για να τροφοδοτήσουν κάποια αρχιτεκτονική βαθιού νευρωνικού δικτύου, χωρίς να λαμβάνουν υπόψη τους την πληροφορία που αφορά τον χρόνο.

Κατά το *late fusion*, δύο εικόνες που απέχουν ένα συγκεκριμένο αριθμό πλαισίων τροφοδοτούνται σε δύο ανεξάρτητα δίκτυα και στη συνέχεια οδηγούν τα αποτελέσματά τους σε δύο επίσης ανεξάρτητα πλήρως-συνδεδεμένα επίπεδα (fully-connected layers). Η τελική πρόβλεψη του συστήματος προκύπτει συγχωνεύοντας τα αποτελέσματα των δύο fully-connected επιπέδων. Όταν πραγματοποιείται *early fusion*, η συγχώνευση των πλαισίων συμβαίνει σε επίπεδο-pixel προτού αυτά δοθούν ως είσοδος του συστήματος. Τέλος, το *slow fusion* αποτελεί ένα συνδυασμό του late fusion και του early fusion. Η διαδικασία απαιτεί τα συνελικτικά επίπεδα να είναι συνδεδεμένα κατά μήκος διαδοχικών πλαισίων, ώστε να παρέχουν το πλεονέκτημα της χρονικής συνέλιξης πέρα από τη συνέλιξη στον χώρο. Σε σχέση με τα τρία είδη, το slow fusion σημειώνει καλύτερη επίδοση από τις υπόλοιπες χάρη στην χρήση των 3D kernels κατά μήκος των πολλαπλών επιπέδων του δικτύου.

Παράλληλα, οι Karpathy et al. (βλέπε εργασία [49]) πειραματίστηκαν με τα *multi-resolution* μοντέλα δημιουργώντας ένα δίκτυο δύο-ρευμάτων (*two-stream*). Όπως παρατηρούμε και στο Σχήμα 3.8, το ένα ρεύμα καλείται '*context stream*' και αναλαμβάνει την επεξεργασία μίας εικόνας χαμηλής ανάλυσης, ενώ το ρεύμα '*fovea stream*' επεξεργάζεται το κέντρο της εικόνας, το οποίο είναι χαμηλής ανάλυσης. Τα αποτελέσματα των συνελίξεων στα δύο διαφορετικά streams συνδυάζονται στα πλήρως-συνδεδεμένα επίπεδά τους (fully-connected layers) και οδηγούν στο τελικό αποτέλεσμα της ταξινόμησης. Το γεγονός ότι χρησιμοποιούμε βίντεο διαφορετικής ανάλυσης σε χωριστά, αλλά ίδιας αρχιτεκτονικής δίκτυα, οδηγεί σε αξιοσημείωτη μείωση των παραμέτρων που πρέπει να εκπαιδευτούν και σε βελτίωση του συνολικού accuracy.

Τέλος, ορισμένες μέθοδοι κάνουν χρήση 3D CNN και ενός μικρού συνελικτικού πυρήνα (convolutional kernel) διαστάσεων 3x3x3 κατά μήκος του δικτύου και έχουν αποδείξει ότι η διατήρηση σταθερού βάθους σε κάθε επίπεδο οδηγεί σε καλύτερες επιδόσεις σε σχέση με την χρήση μεταβλητού βάθους στον χρόνο για κάθε επίπεδο. Αυτό το δίκτυο ονομάστηκε *C3D* και αποτέλεσε την έμπνευση για ένα γενικό περιγραφητή, ο οποίος υπολογίζει τους μέσους όρους των εξόδων από τα fully-connected επίπεδα, με στόχο την εκμάθηση γενικών χαρακτη-



Σχήμα 3.8: Multi-resolution CNN

ριστικών (generic features) από το κάθε βίντεο, προκειμένου το δίκτυο να μην χρειάζεται να προσαρμόζεται στο εκάστοτε πρόβλημα.

Κάνοντας χρήση τρισδιάστατων φίλτρων αυξάνεται ο αριθμός των παραμέτρων του δικτύου και κατά συνέπεια το κόστος και η πολυπλοκότητά του. Γιάυτό έγιναν προσπάθειες να αντικατασταθεί η χρήση των τρισδιάστατων φίλτρων από 2D και 1D φίλτρα. Μάλιστα, σύμφωνα με έρευνες των Baccouche (βλέπε εργασία [4]) και Donahue (βλέπε εργασία [5]) μπορεί να χρησιμοποιηθεί μία αλληλουχία από CNN και LSTM μονάδες, σχηματίζοντας ένα δίκτυο που ονομάστηκε *Long-term Recurrent Convolutional Network-LRCN* και χρησιμοποιήθηκε για το end-to-end training του δικτύου.

3.2.2.3 Υβριδικά μοντέλα (Hybrid Models)

Τα μοντέλα πολλαπλών ρευμάτων (multi-stream models) έχουν βασιστεί στην ιδέα του διαχωρισμού της χωρικής και χρονικής πληροφορίας. Στην εργασία [86] των Simonyan και Zisserman εφαρμόστηκε για πρώτη φορά η ιδέα των πολλαπλών ρευμάτων, όπου χρησιμοποιείται ένα CNN για την εξαγωγή της χωρικής πληροφορίας των video frames και ένα δεύτερο CNN για την καταγραφή της χρονικής πληροφορίας μέσω *Οπτικών Ροών* (Optical Flows). Τα δύο δίκτυα παρουσιάζουν παρόμοια αρχιτεκτονική και εκπαιδεύονται ανεξάρτητα. Όταν ολοκληρωθεί η εκπαίδευσή τους, υπολογίζεται η έξοδος του καθενός μέσω *softmax* και τα αποτελέσματά τους συνδυάζονται μέσω '*fusion*'.

Τα πειράματα πραγματοποιούνται κάνοντας χρήση επιπέδων πυκνών οπτικών ροών (*dense optical flows*) από διαδοχικά πλαίσια του βίντεο, τροχιών που περιγράφουν την κίνηση (*motion trajectories*) και αμφίδρομων οπτικών ροών (*bi-directional optical flows*). Επίσης, τα δεδομένα της εκπαίδευσης προέρχονται από δύο διαφορετικά dataset, το *UCF-101* και το *HMDB-51*, στην προσπάθεια να αντιμετωπιστεί το πρόβλημα του περιορισμένου αριθμού των training set.

Στον Πίνακα 3.1 που ακολουθεί, συγκεντρώνουμε 10 κομβικές εργασίες που έχουν ασχοληθεί με το VAR χρησιμοποιώντας δεδομένα του dataset UCF-101 και παραθέτουμε τα ποσοστά ακριβείας (accuracy) που έχουν σημειώσει:

Συγγραφέας	Χρονιά δημοσίευσης	Χρησιμοποιούμενη τεχνική	Accuracy
Jiang et al.	2012	Trajectories [45]	78.5%
Simonyan,Zisserman	2014	Two-stream CNN [86]	88.0%
Tran et al.	2015	C3D Generic Descriptor [97]	90.4%
Sun et al.	2015	Factorized spatiotemporal CNNs [93]	88.1%
Wang et al.	2015	Two-stream [103]	89.3%
Wang et al.	2015	Trajectory pooling,Fisher vector [103]	91.5%
Lev et al.	2016	RNN Fisher vector [58]	94.08%
Feichtenhofer et al.	2016	ResNet [30]	93.5%
Li et al.	2016	VLAD [59]	92.2%
Varol et al.	2017	Long-term temporal convolutions [99]	91.7%

Πίνακας 3.1: Σύγκριση των τεχνικών που έχουν εφαρμοστεί στο dataset UCF-101

3.3 Σύνολα δεδομένων προορισμένα για το VAR

Το πλήθος των συνόλων δεδομένων που προορίζονται για την *Αναγνώριση ανθρώπινων ενεργειών σε βίντεο* αυξάνεται διαρκώς τα τελευταία χρόνια, με τη δημιουργία τουλάχιστον ενός dataset κάθε χρόνο μετά το 2005 (βλέπε εργασία [18]). Αυτά τα σύνολα διαφοροποιούνται σημαντικά στα χαρακτηριστικά τους. Περιέχουν διαφορετικό αριθμό κατηγοριών, ενώ η μέση διάρκεια και η ανάλυση των βίντεο παρουσιάζουν επίσης σημαντικές διαφορές. Μάλιστα, σε ορισμένες περιπτώσεις τα βίντεο έχουν κοπεί αναλόγως ώστε να περιλαμβάνουν μόνο την ενέργεια που μας ενδιαφέρει, ενώ σε άλλες περιπτώσεις ενδέχεται να απεικονίζονται διαφορετικές ενέργειες σε τυχαία σημεία του βίντεο. Τα πρώτα dataset που είχαν χρησιμοποιηθεί ήταν τα *KTH* (βλέπε εργασία [84]) και *Weizmann* (βλέπε εργασία [8]), τα οποία διέθεταν έναν περιορισμένο αριθμό κατηγοριών. Τα σύνολα που χρησιμοποιούμε σήμερα (βλέπε εργασία [15]) περιλαμβάνουν περισσότερες ενέργειες, σύνθετα παρασκήνια (backgrounds), πολλούς δράστες (actors), επικαλύψεις αντικειμένων (occlusions) και διακυμάνσεις στην οπτική γωνία καταγραφής (viewpoint variations). Στη συνέχεια, ακολουθεί μία σύντομη περιγραφή του κάθε συνόλου, ενώ στους Πίνακες 3.2 και 3.3 παρουσιάζουμε τα συγκριτικά χαρακτηριστικά τους.

- **UCF-101** [89]: πρόκειται για το dataset που θα χρησιμοποιήσουμε στα πειράματά μας (βλέπε Κεφάλαιο 5), το οποίο περιλαμβάνει μία συλλογή από 13320 διαφορετικά βίντεο διαχωρισμένα σε 101 είδη ενεργειών. Το συγκεκριμένο σύνολο δεδομένων χρησιμοποιήθηκε για πρώτη φορά στο διαγωνισμό '*THUMOS'13*' (βλέπε εργασία [44]) και η

επίδοση των προεινόμενων λύσεων μετρήθηκε βάσει του *accuracy* (mAcc) σε τρία διαφορετικά train/test splits των δεδομένων.

- **HMDB-51** [56]: περιέχει 6766 βίντεο κατανεμημένα σε 51 διακριτές κατηγορίες ενεργειών. Η κάθε κατηγορία περιλαμβάνει τουλάχιστον 100 δείγματα βίντεο, τα οποία παρουσιάζουν σημαντικές διαφορές μεταξύ τους. Οι επιδόσεις των μεθόδων VAR σε αυτό το dataset αξιολογούνται υπολογίζοντας τη μέση τιμή της μετρικής *accuracy* (mAcc) σε καθένα από τα τρία train/test splits στα οποία είναι χωρισμένα τα δεδομένα.

- **JHMDB** [17]: αποτελεί ένα υποσύνολο του HMDB-51, στο οποίο επισημειώνονται οι αρθρώσεις του σώματος των ανθρώπων που πρωταγωνιστούν στο κάθε βίντεο, ώστε να είναι αμεσότερη η παρακολούθηση της κίνησης που εκτελείται. Περιέχονται 928 βίντεο, τα οποία προέρχονται από το HMDB-51 και είναι χωρισμένα σε 21 είδη ενεργειών. Για την αξιολόγηση του JHMDB χρησιμοποιούνται τα ίδια train/test splits με το HMDB-51.

- **Hollywood2** [83]: αποτελείται από 1707 βίντεο, τα οποία προέρχονται από 69 ταινίες του Hollywood και ταξινομούνται σε 12 επικαλυπτόμενες κλάσεις. Συνεπώς, πρόκειται για ένα *multi-label dataset*, στο οποίο η απεικονιζόμενη ενέργεια ενός βίντεο ενδέχεται να εντάσσεται σε παραπάνω από μία κατηγορίες, όπως συμβαίνει με τις ενέργειες HandShake και HugPerson. Το συγκεκριμένο σύνολο δεδομένων συνοδεύεται από ένα train/test split, βάσει του οποίου υπολογίζεται η μέση τιμής της μετρικής precision (mAP).

- **Olympic Sports** [73]: πρόκειται για ένα σύνολο δεδομένων που περιέχει 783 βίντεο αθλητών, οι οποίοι εκτελούν 16 διαφορετικές αθλητικές δραστηριότητες, με 50 ακολουθίες στην κάθε κατηγορία. Κάποιες δραστηριότητες περιλαμβάνουν αλληλεπιδράσεις των αθλητών με αντικείμενα, όπως το Bowling και το Weightlifting. Προκείμενου να αξιολογήσουμε τις επιδόσεις των δικτύων πάνω σε αυτό το dataset χρησιμοποιούμε τα train/test splits και είτε τη μετρική accuracy είτε την precision.

- **ActivityNet** [11]: το συγκεκριμένο dataset αποτελεί ένα από τα πολυπληθέστερα, καθώς περιλαμβάνει 19994 βίντεο χωρίς περικοπές στη διάρκειά τους, τα οποία είναι κατανεμημένα σε 200 κατηγορίες ενεργειών. Στην κάθε κατηγορία ανήκουν τουλάχιστον 100 βίντεο, ενώ ο μέσος όρος στιγμιοτύπων στο κάθε βίντεο υπολογίζεται στην τιμή 1.54. Παρότι το *testing set*, δηλαδή τα δεδομένα που προορίζονται για την αξιολόγηση του dataset, δεν είναι δημοσίως διαθέσιμο, η επίδοση του ActivityNet υπολογίζεται βάσει της μετρικής precision στο *validation set*.

- **High-Five** [76]: περιλαμβάνει 300 βίντεο από 23 διαφορετικές εκπομπές της τηλεόρασης, τα οποία είναι διαχωρισμένα σε 5 διαφορετικές κατηγορίες. Οι 4 από αυτές τις κατηγορίες περιέχουν αλληλεπιδράσεις μεταξύ ανθρώπων, ενώ στην πέμπτη ανήκουν όσα βίντεο δεν απεικονίζουν κάποια αλληλεπίδραση.

Dataset	Actions	Video clips	Training videos	Testing videos	Frames/clip
UCF-101	101	13320	9537	3783	186.50
HMDB-51	51	6766	3570	1530	94.488
JHMDB	21	929	660	268	41.112
Hollywood2	12	1707	901	972	285.62
Olympic Sports	16	783	649	134	180.18
ActivityNet	200	19746	9902	4856	3236.17
High-Five	5	300	150	150	94.276

Πίνακας 3.2: Κατανομή των δεδομένων των dataset

Dataset	Width	Height	Frames/second
UCF-101	240.99 [320-400]	320.02 [226-240]	25.90 [25.00-29.97]
HMDB-51	366.81 [176-592]	240.00 [240-240]	30.00 [30.00-30.00]
JHMDB	320.00 [320-320]	240.00 [240-240]	30.00 [30.00-30.00]
Hollywood2	609.24 [480-720]	338.31 [224-576]	24.75 [23.98-29.97]
Olympic Sports	509.38 [192-1280]	361.70 [144-720]	-
ActivityNet	845.28 [128-1280]	516.44 [96-720]	27.68 [6.00-30.00]
High-Five	607.36 [400-720]	356.37 [288-576]	24.10 [23.98-25.00]

Πίνακας 3.3: Χαρακτηριστικά των βίντεο των dataset

Πέρα από τα σύνολα δεδομένων που αναφέραμε πιο πάνω, έχουν δημιουργηθεί ακόμα τρία dataset τα οποία διαθέτουν μεγάλο πλήθος δεδομένων και προορίζονται για την '*Αναγνώριση Ανθρώπινων Ενεργειών σε βίντεο*':

- **Sports-1M** [48]: πρόκειται για ένα dataset μεγάλης κλίμακας, στο οποίο κάποια ενέργεια μπορεί να ανήκει σε παραπάνω από μία κατηγορίες. Περιέχονται 1.13 εκατομμύρια βίντεο κατανεμημένα σε 487 αθλητικές κατηγορίες, ενώ στην καθεμία ανήκουν από 1000 έως 3000 βίντεο. Το *Sports-1M* αποτέλεσε ένα από τα πρώτα σύνολα δεδομένων που χρησιμοποιούσε συνδέσμους από το *YouTube*, με αποτέλεσμα η διαθεσιμότητα των βίντεο να μην είναι σταθερή.

- **YouTube-8M** [1]: εδώ χρησιμοποιούνται 7 εκατομμύρια συνδέσμων URL, που αντιστοιχούν σε βίντεο συνολικής διάρκειας περίπου 450000 ωρών. Επίσης, διακρίνονται 4716 είδη ενεργειών, ενώ όπως ήταν αναμενόμενο, το κάθε βίντεο απεικονίζει κατά μέσο όρο 3.4 διαφορετικές ενέργειες. Μάλιστα, το *YouTube-8M* συνοδεύεται από 3.2 δισεκατομμύρια οπτικοακουστικά χαρακτηριστικά των βίντεο, τα οποία έχουν εξαχθεί χρησιμοποιώντας *Βαθιά Νευρωνικά Δίκτυα* και συγκεκριμένα τα μοντέλα '*Inception-V3*' και '*VGG-acoustic*' αντίστοιχα.

- **Kinetics** [50]: ένα από τα πιο πρόσφατα dataset, το οποίο περιέχει πάνω από 240000 δεδομένα εκπαίδευσης χωρισμένα σε 400 κατηγορίες ενεργειών, με αποτέλεσμα η κάθε κατηγορία να περιλαμβάνει τουλάχιστον 400 βίντεο. Το συγκεκριμένο σύνολο δεδομένων είναι διαθέσιμο και μέσω συνδέσμων του YouTube.

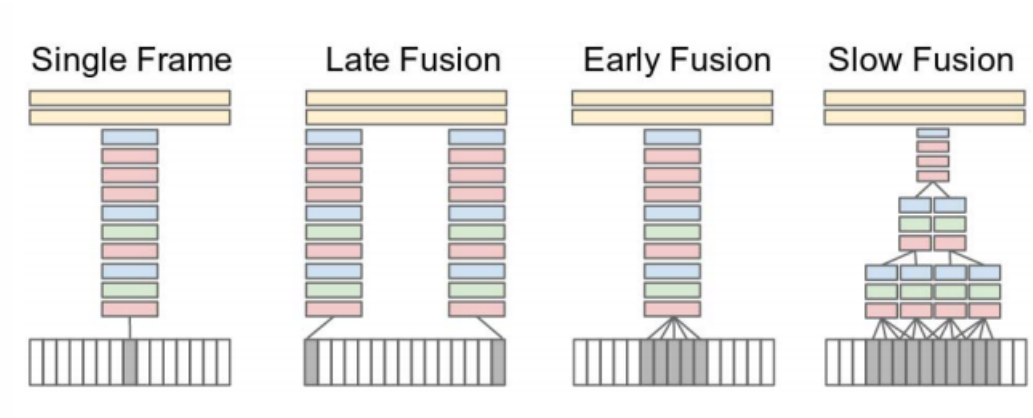
Όπως διαφαίνεται από τα παραπάνω, ορισμένα dataset περιλαμβάνουν video προερχόμενα από το *YouTube*, κάποια άλλα από *εκπομπές της τηλεόρασης*, από δραστηριότητες σε *εσωτερικούς ή εξωτερικούς χώρους* ή από *ταινίες*. Τις μεγαλύτερες δυσκολίες ως προς την επεξεργασία τους παρουσιάζουν εκείνα στα οποία το παρασκήνιο (*background*) μεταβάλλεται ή η οπτική γωνία (*viewpoint*) της κάμερας αλλάζει, αλλά και όσα περιλαμβάνουν video χαμηλής ανάλυσης (*resolution*) και σταθερότητας (*stabilization*) στην εικόνα. Το μοντέλο VAR που κατασκευάζουμε στα πλαίσια της διπλωματικής μας εργασίας κάνει χρήση του dataset *UCF-101*, οπότε κρίνουμε αναγκαία την επιπλέον ανάλυση του συγκεκριμένου στην Ενότητα 5.3.

Κεφάλαιο 4

Υλοποιήσεις VAR βασισμένες σε Βαθιά Νευρωνικά Δίκτυα

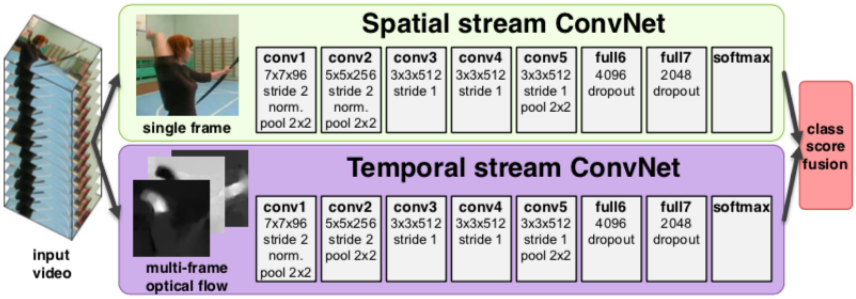
Σε αυτό το κεφάλαιο αναλύουμε τον τρόπο λειτουργίας και τη συμβολή 9 μοντέλων που έχουν χρησιμοποιηθεί στην επίλυση του προβλήματος της '*Αναγνώρισης Ανθρώπινων Ενεργειών σε βίντεο*'. Τα μοντέλα αυτά έχουν προέλθει από την εξέλιξη δύο βασικών τύπων δικτύων, η βασική διαφοροποίηση των οποίων έγκειται στην επιλογή σχεδιασμού για το συνδυασμό της χωροχρονικής πληροφορίας:

- Δίκτυο Μονού-Ρεύματος (**Single-Stream Network**): διαδοχικά πλαίσια του βίντεο αποτελούν την είσοδο του δικτύου, ενώ η συγχώνευση της χρονικής πληροφορίας από αυτά γίνεται μέσω δισδιάστατων προεκπαιδευμένων συνελίξεων και διακρίνεται σε *single frame fusion*, *late fusion*, *early fusion* και σε *slow fusion*. Για την εξαγωγή των τελικών προβλέψεων του δικτύου υπολογίζεται ο μέσος όρος των αποτελεσμάτων που έχουν προκύψει από τα διαφορετικά βίντεο. Παρά τις εκτεταμένες έρευνες που πραγματοποιήθηκαν, διαπιστώθηκε ότι η επίδοση του Single-Stream Network ήταν αρκετά χειρότερη από αυτή των αλγορίθμων που βασίζονταν σε handcrafted features των πλαισίων.



Σχήμα 4.1: Διαφορερικές τεχνικές ενσωμάτωσης της χρονικής πληροφορίας

- Δίκτυο Δύο-Ρευμάτων (**Two-Stream Network**): δεδομένης της αδυναμίας των *Βαθιών Νευρωνικών Δικτύων* να μαθαίνουν τα χαρακτηριστικά της κίνησης (*motion features*), άρχισαν να χρησιμοποιούνται στοιβάδες διανυσμάτων οπτικών ροών (*stacked optical flow vectors*) προκειμένου να μοντελοποιηθεί η κίνηση που εντοπίζεται κατά τη διάρκεια ενός βίντεο. Έτσι, η αρχιτεκτονική των Two-Stream Network) συνδυάζει δύο ανεξάρτητα δίκτυα, από τα οποία το ένα είναι υπεύθυνο για την εξαγωγή της χωρικής πληροφορίας και προεκπαιδεύεται σε ένα μεγάλο σύνολο δεδομένων, το οποίο συνήθως προορίζεται για εφαρμογές του *Image Recognition*. Η είσοδος του *spatial stream* αποτελείται από ένα μεμονωμένο πλαίσιο του βίντεο, ενώ ως είσοδος του *temporal stream* χρησιμοποιούνται στοιβάδες από αμφίδρομες οπτικές ροές (*bi-directional optical flows*), οι οποίες προκύπτουν από 10 διαδοχικά frames. Τα δύο δίκτυα εκπαιδεύονται ξεχωριστά και συνδυάζονται μέσω ενός γραμμικού ταξινομητή *Support Vector Machine-SVM*. Η τελική πρόβλεψη του συστήματος προκύπτει από το μέσο όρο των πλαισίων που έχουν χρησιμοποιηθεί στην είσοδο. Το μοντέλο Two-Stream Network βελτίωσε σημαντικά την επίδοση των Single-Stream Network που περιγράψαμε παραπάνω.



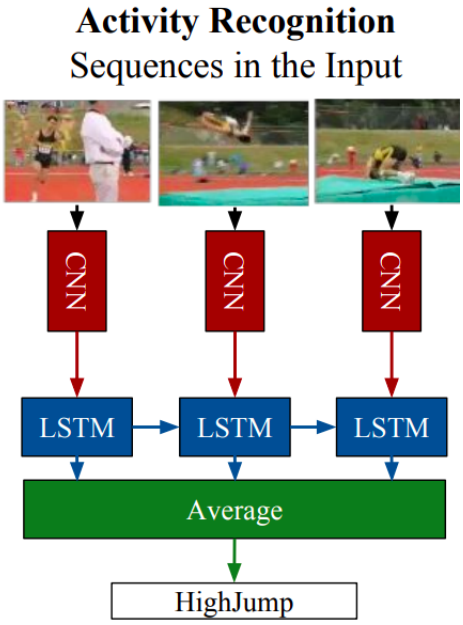
Σχήμα 4.2: Αρχιτεκτονική του μοντέλου Two-Stream Network

4.1 Long-term Recurrent Convolutional Networks - LRCNs

Τα δίκτυα LRCN προορίζονται για την αναγνώριση και την περιγραφή των οπτικών χαρακτηριστικών (visual features) ενός βίντεο. Το συγκεκριμένο είδος αρχιτεκτονικής (βλέπε Σχήμα 4.3) συνδυάζει συνελικτικά επίπεδα και μεγάλου εύρους χρονικές αναδρομές, ενώ υφίσταται end-to-end training. Το ερευνητικό έργο που κρύβεται πίσω από τα μοντέλα LRCN αποτελεί μία προσπάθεια επέκτασης της χρήσης των δικτύων CNN σε μεταβλητού-χρόνου εισόδους και εξόδους. Επίσης, χάρη στην παρουσία των μονάδων LSTM επιτρέπεται η εκμάθηση του συστήματος σε μεγάλου εύρους χρονικές εξαρτήσεις της εισόδου. Τα βάρη τόσο των CNN όσο και των LSTM δικτύων διαμοιράζονται κατά μήκος του χρόνου, οδηγώντας σε μία αναπαράσταση που μπορεί να κλιμακωθεί για αυθαίρετα μεγάλες ακολουθίες.

Αρχικά, είχε προταθεί η χρήση ενός δικτύου LSTM για κάθε χάρτη χαρακτηριστικών (*feature map*) που προκύπτει από την εκπαίδευση του αντίστοιχου CNN, με σκοπό την εξαγωγή της πληροφορίας από τα video clips στο πεδίο του χρόνου. Ωστόσο, η χρονική συγκέντρωση

(temporal pooling) των χαρακτηριστικών που εξάγονται μέσω των συνελικτικών επιπέδων (*convoluted features*) έχει αποδειχθεί πιο αποδοτική σε σχέση με την ύπαρξη διαδοχικών δικτύων CNN και LSTM. Κατά την εκπαίδευση χρησιμοποιούμε *16 τυχαία πλαίσια* από κάθε βίντεο και διαμορφώνουμε τα δύο είδη εισόδων με τα οποία θα τροφοδοτήσουμε το σύστημα. Σχηματίζεται μία είσοδος *RGB εικόνων* και μία άλλη είσοδος αποτελούμενη από *οπτικές ροές* (optical flows).



Σχήμα 4.3: Πρόβλεψη της κατηγορίας 'HighJump' από το μοντέλο LRCN

Παράλληλα, στην εργασία [28] που δημοσιεύθηκε το 2014 από τους Donahue et al., προτείνεται ένα εκπαιδευόμενο μοντέλο αντιστοίχισης εικόνων σε προτάσεις. Χρησιμοποιούνται δίκτυα CNN ως *encoders* για την κωδικοποίηση των οπτικών χαρακτηριστικών των πλαισίων σε ένα διάνυσμα κατάστασης (state vector) και μονάδες δικτύων LSTM που αναλαμβάνουν τον ρόλο των *decoders* προκειμένου να αποκωδικοποιήσουν αυτό το διάνυσμα σε μία συμβολοσειρά φυσικής γλώσσας. Το σύστημα που προκύπτει μπορεί να εκπαιδευτεί σε μεγάλης κλίμακας dataset για εικόνες και κείμενα.

Η τελική πρόβλεψη για το κάθε clip αποτελεί το μέσο όρο όλων των προβλέψεων σε κάθε χρονικό διάστημα (time step). Σε επίπεδο ολόκληρου του βίντεο, το τελικό σκορ αποτελεί το μέσο όρο των προβλέψεων από το κάθε clip. Στον Πίνακα 4.1 καταγράφουμε την επίδοση του μοντέλου LRCN στο σύνολο δεδομένων *UCF-101*. Τα μειονεκτήματα αυτής της μεθόδου οφείλονται στη λανθασμένη ανάθεση ετικετών (*false label assignment*) λόγω της κατάτμησης του κάθε βίντεο σε επιμέρους clip segments. Ακόμη, το μοντέλο δεν έχει την ικανότητα να εξάγει χρονική πληροφορία μεγάλου εύρους, ενώ υπάρχει η ανάγκη προεπεξεργασίας των χαρακτηριστικών των πλαισίων ώστε να μπορέσουν να χρησιμοποιηθούν στο σχηματισμό των οπτικών ροών (optical flows).

Accuracy	Λεπτομέρειες υπολογισμού
82.92%	Σταθμισμένο σκορ από RGB εικόνες και οπτικές ροές
71.1%	Σκορ από RGB εικόνες

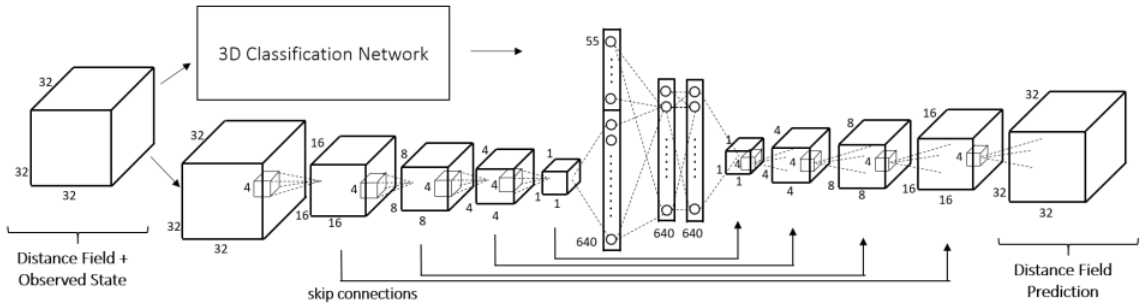
Πίνακας 4.1: Επίδοση του μοντέλου LRCN στο dataset UCF-101

4.2 3D Convolutional Neural Networks - C3D

Η έμπνευση για τα 3D CNN προήλθε από την εργασία [49] των Karpathy et al., ενώ η βασική ιδέα της υλοποίησής τους (βλέπε εργασία [97]) περιλαμβάνει την εκπαίδευση αυτών των μεγάλων δικτύων στο σύνολο δεδομένων *Sports-1M* και στη συνέχεια την χρήση τους ως ανιχνευτές χαρακτηριστικών (*feature extractors*) σε διαφορετικά σύνολα δεδομένων. Ειδικότερα, τα 3D CNN χρησιμοποιούν έναν απλό γραμμικό ταξινομητή (*Support Vector Machine-SVM*) ώστε να ταξινομήσουν τα εξαγόμενα χαρακτηριστικά (*extracted features*) των πλαισίων.

Το μοντέλο παρουσιάζει βελτιωμένες επιδόσεις εφόσον συνδυαστεί με την χρήση *hand-crafted features*. Όπως φαίνεται στο Σχήμα 4.4, μπορούν να χρησιμοποιούνται και αποσυνελικτικά επίπεδα (deconvolutional layer) για την ερμηνεία των προβλέψεων του συνολικού δικτύου. Ερμηνεύοντας το αποτέλεσμα των deconvolutional layers, καταλήγουμε ότι το δίκτυο επικεντρώνεται στην χωρική πληροφορία των πρώτων πλαισίων, ενώ καταγράφει την πληροφορία της κίνησης από τα μεταγενέστερα πλαίσια.

Κατά τη διάρκεια της εκπαίδευσης του μοντέλου, εξάγουμε 5 τυχαία clips διάρκειας 2 δευτερολέπτων από το κάθε βίντεο και θεωρούμε ότι η απεικονιζόμενη ενέργεια ταυτίζεται με την ενέργεια που αναπαρίσταται σε ολόκληρο το βίντεο. Κατά τη διάρκεια ελέγχου του μοντέλου (*testing*), διαλέγουμε τυχαία 10 clips και υπολογίζουμε το μέσο όρο των αποτελεσμάτων τους για την εξαγωγή της τελικής πρόβλεψης.



Σχήμα 4.4: Αρχιτεκτονική του μοντέλου 3D Convolutional Neural Network

Στον Πίνακα 4.2 καταγράφουμε την επίδοση του δικτύου στο σύνολο δεδομένων *UCF-101*:

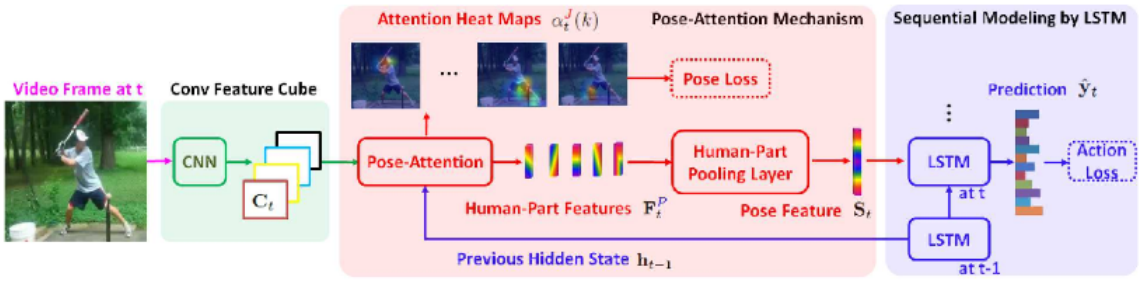
Accuracy	Λεπτομέρειες υπολογισμού
82.3%	Συνδυασμός 3D CNN και SVM
85.2%	Συνδυασμός 3D CNN και SVM
90.4%	Συνδυασμός 3D CNN,iDT και SVM

Πίνακας 4.2: Επίδοση του μοντέλου C3D στο dataset UCF-101

4.3 3D CNNs - Attention Mechanism

Παρότι η ερευνητική εργασία [108] των Yao et al. δεν επικεντρώνεται στην αναγνώριση ενεργειών, αποτελεί μία εργασία-ορόσημο αναφορικά με τις αναπαραστάσεις βίντεο. Συγκεκριμένα, γίνεται χρήση μίας αρχιτεκτονικής *encoder-decoder* που συνδυάζει *3D CNN* και *RNN* δίκτυα για την καταγραφή της τοπικής χωροχρονικής πληροφορίας από το κάθε βίντεο. Μάλιστα, προκειμένου να βελτιωθεί η επίδοση του συστήματος γίνεται χρήση ενός προεκπαιδευμένου δικτύου 3D CNN και ενός μηχανισμού προσοχής (*attention mechanism*).

Η αρχιτεκτονική του συστήματος θυμίζει αυτή που συναντήσαμε στην ενότητα 4.2, με τη διαφορά ότι αντί να οδηγήσουμε τα χαρακτηριστικά (features) που έχουν εξαχθεί από το 3D CNN απευθείας στο LSTM δίκτυο, οι χάρτες χαρακτηριστικών (feature maps) του κάθε clip συγχωνεύονται με μία στοιβάδα 2D feature maps για το ίδιο σύνολο πλαισίων. Έτσι, η αναπαράσταση $\{v_1, v_2, ..., v_n\}$ του κάθε frame i γίνεται πιο περιγραφική. Επίσης, δεν υπολογίζεται ο μέσος όρος των διανυσμάτων που αφορούν τον χρόνο (temporal vectors) για όλα τα frames, αλλά ένας σταθμισμένος μέσος όρος που συνδυάζει τα χρονικά χαρακτηριστικά (temporal features). Τα βάρη προσοχής (*attention weights*) επιλέγονται βάσει της εξόδου του δικτύου LSTM σε κάθε χρονική στιγμή (time step).



Σχήμα 4.5: Χρήση Attention Mechanism στην αναγνώριση ενεργειών

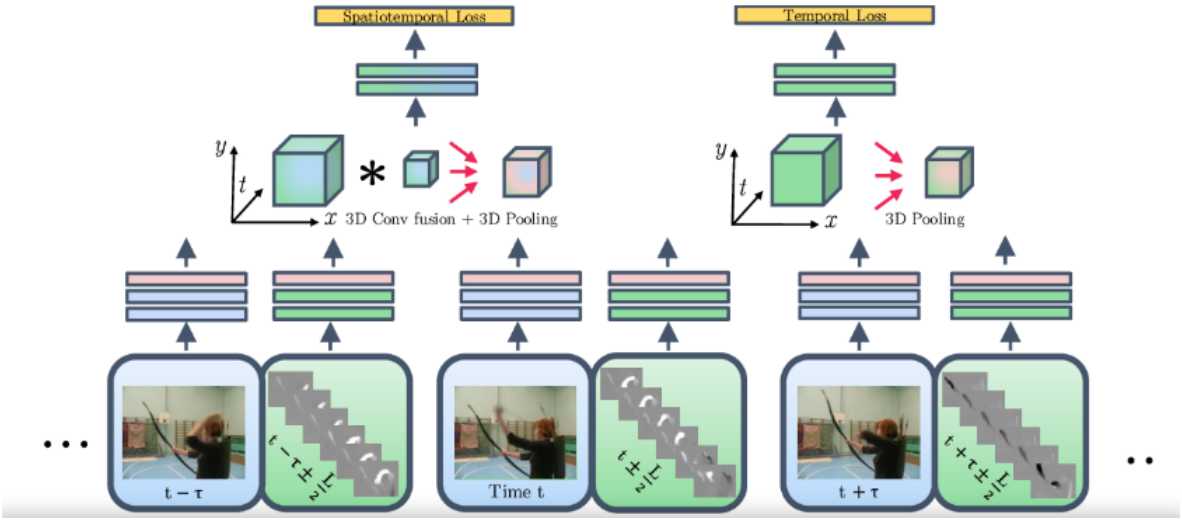
4.4 Two-Stream CNNs

Στην εργασία [29] που δημοσιεύθηκε το 2016 από τους Feichtenhofer, Zisserman και Pinz, γίνεται χρήση μίας καινοτόμας αρχιτεκτονικής που στοχεύει στον εντοπισμό των χωρικών και χρονικών χαρακτηριστικών από τα πλαίσια ενός βίντεο. Συγκεκριμένα, υλοποιούνται δύο δίκτυα CNN, το ένα αποκαλείται '*Spatial Stream CNN*' και αναλαμβάνει την ανίχνευση

των spatial features από τα frames, ενώ το '*Temporal Stream CNN*' αποκωδικοποιεί την πληροφορία της κίνησης που παρατηρείται στο βίντεο. Τα δύο διαφορετικά ρεύματα (streams) συγχωνεύονται μέσω fusion και μας δίνουν την τελική πρόβλεψη του συστήματος.

Προκειμένου το συνολικό σύστημα να μπορέσει να διακρίνει παρόμοιες ενέργειες, όπως το 'Brushing Hair' και το 'Brushing Teeth', το *Spatial Stream Network* συλλέγει τις χωρικές εξαρτήσεις του βίντεο, δηλαδή αναγνωρίζει εάν πρόκειται για 'hair' ή για 'teeth', ενώ το *Temporal Stream Network* εντοπίζει την παρουσία περιοδικής κίνησης για κάθε χωρικό στοιχείο του βίντεο.

Ως εκ τούτου, είναι αναγκαίο να αντιστοιχίσουμε κάθε spatial feature map με ένα temporal feature map που αφορά τη συγκεκριμένη τοποθεσία όπου εντοπίστηκε το χαρακτηριστικό του πλαισίου. Έτσι, μοντελοποιείται η χρονική εξάρτηση που υπάρχει μεταξύ πολλαπλών πλαισίων του βίντεο.



Σχήμα 4.6: Spatiotemporal fusion στα Two-Stream CNNs

Περισσότερες λεπτομέρειες του συγκεκριμένου μοντέλου *Αναγνώρισης Ενεργειών σε βίντεο* παραθέτουμε στο Κεφάλαιο 5, όπου εξηγούμε την υλοποίηση που πραγματοποιήσαμε στο πλαίσιο της παρούσας εργασίας. Στον Πίνακα 4.3 καταγράφουμε την επίδοση της μεθόδου στο σύνολο δεδομένων *UCF-101*:

Accuracy	Λεπτομέρειες υπολογισμού
92.5%	Χρήση Two-Stream fusion
94.2%	Χρήση Two-Stream fusion και iDT

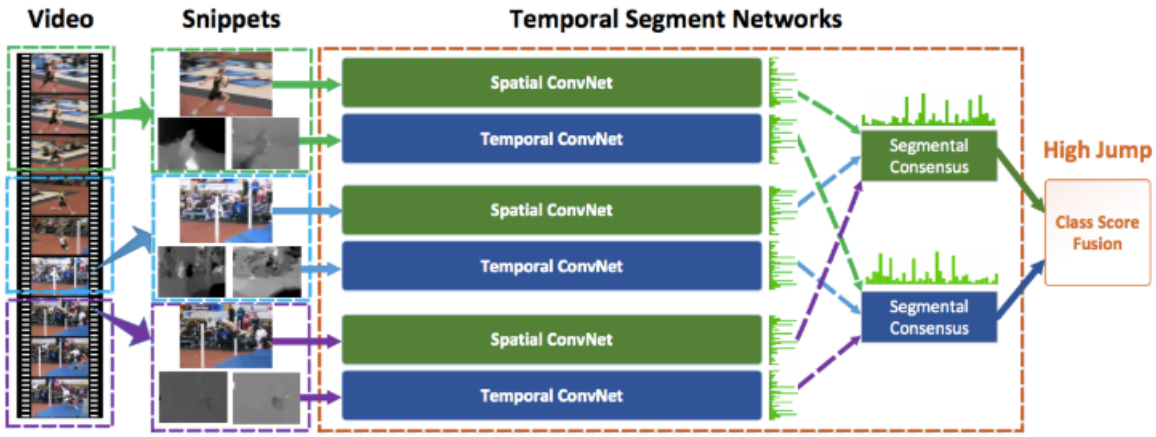
Πίνακας 4.3: Επίδοση του μοντέλου Two-Stream CNN στο dataset UCF-101

4.5 Temporal Segment Networks - TSNs

Στη δημοσιευμένη εργασία [104] του 2016, οι Wang et al. βασίστηκαν στην αρχιτεκτονική Two-Stream CNNs που παρουσιάσαμε στην Ενότητα 4.4 και προσπάθησαν να μοντελοποιήσουν τη μεγάλου εύρους χρονική πληροφορία. Αποδείχθηκε ότι η χρήση των τεχνικών *batch normalisation* και *dropout*, αλλά και η προεκπαίδευση του δικτύου αποτελούν καλές πρακτικές για τη μείωση της υπερπροσαρμογής (overfitting) του συστήματος στα δεδομένα εκπαίδευσης.

Ως είσοδος του μοντέλου, εξετάζεται η χρήση οπτικών ροών (*optical flows*) και δύο καινοτόμες μορφές εισόδου, οι οποίες ονομάζονται *warped optical flows* και *RGB difference*. Για την εξαγωγή των τελικών προβλέψεων σε επίπεδο-βίντεο έχουν εξεταστεί ποικίλες στρατηγικές. Η καλύτερη από αυτές ήταν ο συνδυασμός των προβλέψεων από τα temporal και spatial streams μέσω υπολογισμού του μέσου όρου τους, συγχώνευσης των τελικών spatial και temporal scores και εφαρογής της συνάρτησης softmax σε καθεμία από τις κατηγορίες του συνόλου δεδομένων που χρησιμοποιείται.

Κατά τη διάρκεια της εκπαίδευσης και αξιολόγησης του δικτύου, το κάθε βίντεο χωρίζεται σε k μέρη (segments) ίσης χρονικής διάρκειας. Στη συνέχεια, γίνεται τυχαία δειγματοληψία των snippets από κάθε k-segment. Στο Σχήμα 4.5 απεικονίζεται η αρχιτεκτονική ενός δικτύου που βασίζεται στη λογική των Temporal Segments και Two-Streams.



Σχήμα 4.7: Αρχιτεκτονική ενός Temporal Segment Network

Η επίδοση που σημειώθηκε από την χρήση των μοντέλων *Temporal Segment Network* καταγράφεται στον Πίνακα 4.5:

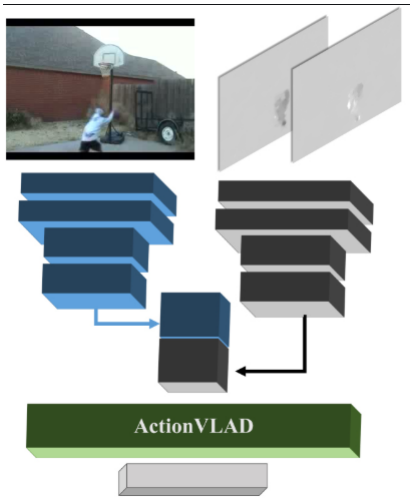
Accuracy	Λεπτομέρειες υπολογισμού
94.0%	TSN με είσοδο RGB frames και optical flows
94.2%	TSN με είδοσο RGB frames, optical flows και warped flows

Πίνακας 4.4: Επίδοση του μοντέλου TSN στο dataset UCF-101

4.6 Action VLAD

Η συμβολή της εργασίας [35] των Girdhar et al. έγκειται στην χρήση εκπαιδευόμενων συνόλων χαρακτηριστικών (VLAD), τα οποία παίρνουν τη θέση των τεχνικών *max-pooling* και *average-pooling*. Η μέθοδος συγχώνευσης που ακολουθείται, έχει κοινά στοιχεία με αυτή του Σάκου Εικονικών Λέξεων (*Bag of Visual Words*). Χρησιμοποιούνται εκπαιδευμένα λεξιλόγια (*vocabularies*), τα οποία αποτελούν αναπαραστάσεις μίας k-ενέργειας ή υπο-ενέργειας, η οποία συσχετίζεται με τα χωροχρονικά χαρακτηριστικά (spatiotemporal features). Η έξοδος του κάθε δικτύου σε μία αρχιτεκτονική two-stream κωδικοποιείται ως 'action words' k-διαστάσεων.

Το *max-pooling* και το *average-pooling* αναπαριστούν ολόκληρη την κατανομή των σημείων ως ένας απλός περιγραφητής (descriptor), ο οποίος ενδέχεται να μην αποδειχθεί βέλτιστος για την αναπαράσταση ενός ολόκληρου βίντεο που αποτελείται από πολλαπλές επιμέρους ενέργειες. Σε αντίθεση, η προτεινόμενη μέθοδος του *video aggregation* αναπαριστά μία ολόκληρη κατανομή από descriptors με πολλαπλές ενέργειες, κατανέμοντας τον χώρο του περιγραφέα (descriptor space) σε k-κελιά (cells) και εφαρμόζοντας pooling σε καθένα από αυτά. Αυτό σημαίνει ότι οι τεχνικές max-pooling και average-pooling ενδείκνυνται όταν πρόκειται για πανομοιότυπα χαρακτηριστικά (features), αλλά δεν έχουν την ικανότητα να καταγράψουν με επαρκή τρόπο ολόκληρη την κατανομή των χαρακτηριστικών. Το δίκτυο *ActionVLAD* δημιουργεί ομάδες (clusters) από τα χαρακτηριστικά εμφάνισης και κίνησης, ενώ παράλληλα υπολογίζει τις αποστάσεις τους από τα πλησιέστερα cluster centers.



Σχήμα 4.8: Αρχιτεκτονική ενός δικτύου Action VLAD

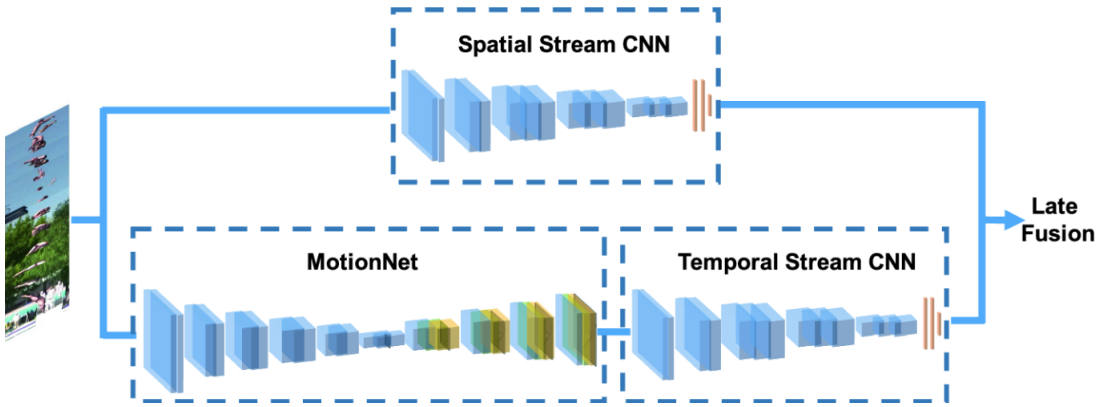
Η χρήση του VLAD ως έναν αποτελεσματικό τρόπο υλοποίησης του pooling και η επέκτασή του σε ένα end-to-end framework το οποίο έχει τη δυνατότητα να εκπαιδευτεί, οδήγησε σε state-of-the-art αποτελέσματα για το πρόβλημα του Video Action Recognition:

Accuracy	Λεπτομέρειες υπολογισμού
92.7%	ActionVLAD
93.6%	Συνδυασμός ActionVlad και iDT

Πίνακας 4.5: Επίδοση του μοντέλου ActionVLAD στο dataset UCF-101

4.7 Hidden Two-Stream Network

Η εργασία [111] των Zhu et al., η οποία δημοσιεύθηκε το 2017, παρουσίασε μία καινοτόμα αρχιτεκτονική για την παραγωγή μίας εισόδου optical flow μέσω ενός ξεχωριστού δικτύου. Η χρήση της μεθόδου των optical flow στην two-stream αρχιτεκτονική, κατέστησε επιτακτική την ανάγκη του σχηματισμού των οπτικών ροών για το κάθε επιλεγμένο πλαίσιο, ώστε να μειωθεί ο αποθηκευτικός χώρος που απαιτούν και η ταχύτητα επεξεργασίας τους. Εδώ, υποστηρίζεται η χρήση μίας μη-επιβλεπόμενης αρχιτεκτονικής για τη δημιουργία optical flow από μία στοιβάδα πλαισίων. Το optical flow μπορεί να αντιμετωπιστεί ως ένα πρόβλημα ανακατασκευής κάποιας εικόνας. Δεδομένου ενός ζεύγους διαδοχικών πλαισίων I_1 και I_2, το CNN παράγει ένα πεδίο ροής V. Στη συνέχεια, χρησιμοποιώντας το προβλέψιμο πεδίο ροής V και το πλαίσιο I_2, είναι δυνατή η ανακατασκευή του πλαισίου I_1 μέσω *inverse warping*, ώστε η διαφορά μεταξύ του I_1 και της ανακατασκευασμένης μορφής του να ελαχιστοποιείται.



Σχήμα 4.9: Αρχιτεκτονική ενός Hidden Two-Stream Network

Εξετάστηκαν διαφορετικές στρατηγικές και αρχιτεκτονικές παραγωγής των optical flow, με μεγαλύτερο ρυθμό frames-per-second (fps) και λιγότερες παραμέτρους, χωρίς να επηρεάζεται σημαντικά το accuracy του δικτύου. Η αρχιτεκτονική του μοντέλου μοιάζει αρκετά σε αυτήν του Two-Stream Network, αλλά πλέον το Temporal Stream αποτελείται από το *MotionNet*, δηλαδή από ένα optical flow generation network και η είσοδός του τροφοδοτείται με διαδοχικά πλαίσια αντί οπτικών ροών. Επίσης, γίνεται χρήση ενός πρόσθετου multi-level loss για τη μη-επιβλεπόμενη εκπαίδευση του MotionNet.

Η επίδοση του συστήματος βελτιώθηκε περαιτέρω αξιοποιώντας την τεχνική fusion, η οποία στηρίζεται σε Temporal Segment Networks - TSNs. Χάρη στα *Hidden Two-Stream Networks* βελτιώθηκε η ταχύτητα και το επαγόμενο κόστος των προβλέψεων του δικτύου,

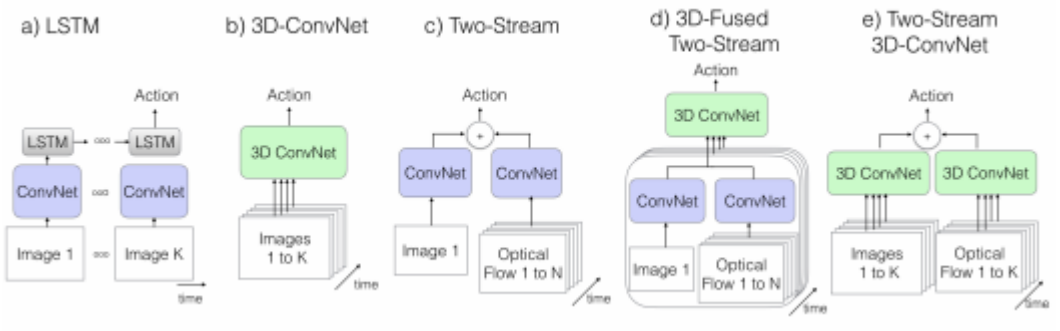
ενώ μέσω της αυτόματης δημιουργίας των οπτικών ροών καταγράφηκαν τα αποτελέσματα του Πίνακα 4.6:

Accuracy	Λεπτομέρειες υπολογισμού
89.8%	Hidden Two-Stream
92.5%	Συνδυασμός Hidden Two-Stream και TSN

Πίνακας 4.6: Επίδοση του μοντέλου Hidden Two-Stream Network στο dataset UCF-101

4.8 Inflated 3D CNNs - I3D

Η έρευνα πάνω στην χρήση των δικτύων 3D CNN για την αναγνώριση ενεργειών σε βίντεο (βλέπε Ενότητα 4.2) συνεχίστηκε στα πλαίσια της εργασίας [13]. Αντί για την χρήση ενός μονού 3D δικτύου, οι Zisserman et al. προτείνουν την αξιοποίηση δύο διαφορετικών 3D δικτύων και στα δύο ρεύματα της two-stream αρχιτεκτονικής που περιγράψαμε νωρίτερα. Επίσης, αξιοποιούνται τα προεκπαιδευμένα δισδιάστατα μοντέλα (2D CNN) προκειμένου να εφαρμοστούν σε μία τρίτη διάσταση. Έτσι, το Spatial Stream Network τροφοδοτείται με πλαίσια του βίντεο, τα οποία έχουν γίνει stacked ως προς τη διάσταση του χρόνου. Οι διαφορετικές αρχιτεκτονικές που δοκιμάζονται, απεικονίζονται στο Σχήμα 4.10.



Σχήμα 4.10: Διαφορετικά αρχιτεκτονικά μοντέλα

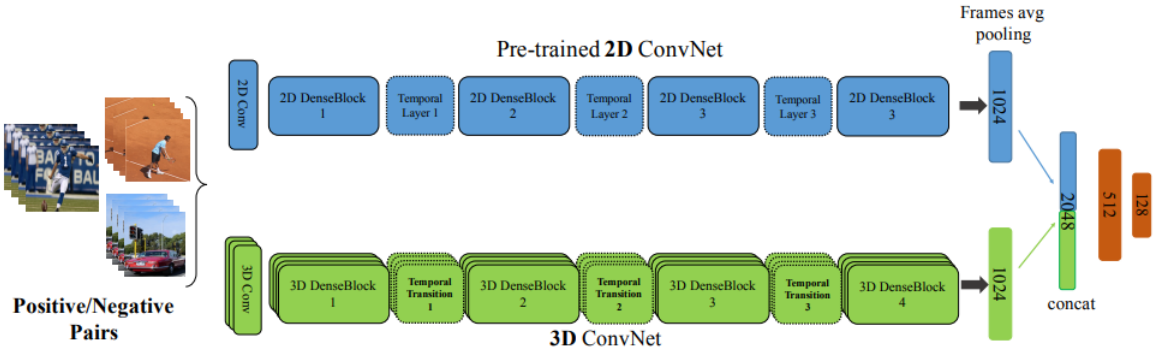
Τα αποτελέσματα των Inflated 3D CNNs στην Two-Stream αρχιτεκτονική συνοψίζονται στον Πίνακα 4.7:

Accuracy	Λεπτομέρειες υπολογισμού
93.4%	Two-Stream I3D
98.0%	Συνδυασμός ImageNet και προεκπαιδευμένου Kinetics

Πίνακας 4.7: Επίδοση του μοντέλου I3D στο dataset UCF-101

4.9 Temporal 3D CNNs - T3D

Η τελευταία εργασία [26] που εξετάζουμε δημοσιεύθηκε το 2017 και αποτελεί μία επέκταση της εργασίας που είχε πραγματοποιηθεί μέχρι στιγμής στα I3D Networks. Συγκεκριμένα, προτείνεται η χρήση ενός Single Stream, η αρχιτεκτονική του οποίου βασίζεται σε αυτήν του δικτύου *3D DenseNet*, με την προσθήκη ενός multi-depth χρονικού επιπέδου συγχώνευσης (*Temporal Transition Layer*). Αυτό το επίπεδο τοποθετείται μετά από τα dense blocks προκειμένου να εντοπιστούν διαφορετικά χρονικά βάθη. Το *multi-depth pooling* πραγματοποιείται μέσω pooling με kernels μεταβλητού χρονικού βάθους.



Σχήμα 4.11: Αρχιτεκτονική ενός Temporal 3D CNN

Την ίδια στιγμή, οι Diba et al. εξέτασαν μία καινούργια τεχνική επιβλεπόμενης μεταφοράς γνώσης (*Supervised Transfer Learning*) μεταξύ των προεκπαιδευμένων 2D CNNs και των T3D Networks. Η προτεινόμενη αρχιτεκτονική εκπαιδεύεται κατάλληλα και το σφάλμα (error) της τελικής πρόβλεψης οδηγείται με οπισθοδιάδοση λάθους (*error backpropagation*) μέσα από το T3D δίκτυο, ώστε να υλοποιηθεί το Transfer Learning.

Η προτεινόμενη αρχιτεκτονική συνδυάζει την χρονική πληροφορία κατά μήκος μεταβλητού βάθους, αλλά δεν οδηγεί σε καλύτερα αποτελέσματα σε σχέση με αυτά των I3D δικτύων. Ωστόσο, η συμβολή της εντοπίζεται στη μελέτη της τεχνικής του *Supervised Transfer Learning* μεταξύ των 2D και 3D δικτύων.

Accuracy	Λεπτομέρειες υπολογισμού
90.3%	T3D
91.7%	Συνδυασμός T3D και Transfer Learning
93.2%	Συνδυασμός T3D και TSN

Πίνακας 4.8: Επίδοση του μοντέλου T3D στο dataset UCF-101

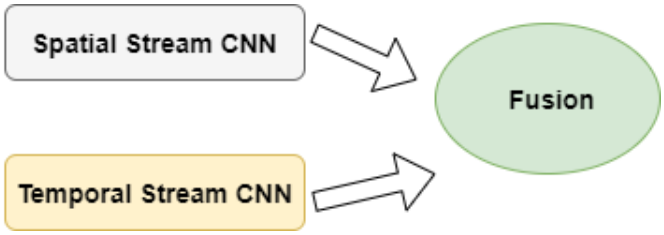
Κεφάλαιο 5

Πειραματική διαδικασία και αποτελέσματα

Σε αυτό το κεφάλαιο αναφερόμαστε στο σύστημα που κατασκευάσαμε προκειμένου να προσεγγίσουμε την επίλυση του προβλήματος '*Video Action Recognition-VAR*'. Ειδικότερα, επεξηγούμε τις λεπτομέρειες υλοποίησης του μοντέλου μας και παραθέτουμε τα πειραματικά αποτελέσματα που λάβαμε. Τα δεδομένα βίντεο που χρησιμοποιούμε προέρχονται από το dataset *UCF-101* και αναπαριστούν 101 διαφορετικές ανθρώπινες ενέργειες, τις οποίες καλείται το σύστημά μας να ταξινομήσει. Η ανάλυση που ακολουθούμε έχει βασιστεί στην ερευνητική εργασία [87] που δημοσιεύθηκε το *2014* από τους *Karen Simonyan* και *Andrew Zisserman*.

5.1 Προσέγγιση της αναγνώρισης ενεργειών μέσω Two-Stream Network

Κάθε βίντεο λόγω της χωροχρονικής του φύσης μπορεί να αποσυντεθεί στις χωρικές (*spatial*) και χρονικές (*temporal*) συνιστώσες του. Το χωρικό τμήμα του μεταφέρει την πληροφορία που αφορά τις σκηνές και τα αντικείμενα που απεικονίζονται στα πλαίσια του βίντεο, ενώ το χρονικό μέρος εκφράζει την κίνηση μεταξύ του παρατηρητή-κάμερας και των αντικειμένων. Χωρίζουμε το σύστημά μας σε δύο ρεύματα (streams), όπως απεικονίζεται στο Σχήμα 5.1. Το κάθε ρεύμα υλοποιείται από ένα *Συνελικτικό Νευρωνικό Δίκτυο* (CNN).



Σχήμα 5.1: Two-Stream Architecture

- **Spatial Stream CNN**: αυτό το δίκτυο δέχεται ως είσοδο πλαίσια (frames) του υπό εξέταση βίντεο, τα οποία δεν είναι παρά RGB εικόνες (χρωματικά κανάλια R,G,B). Από τις στατικές εικόνες μπορούμε να εξάγουμε χρήσιμη πληροφορία και να αναγνωρίσουμε ορισμένες ενέργειες που είναι στενά συνδεδεμένες με συγκεκριμένα αντικείμενα που απεικονίζονται στο βίντεο. Συνεπώς, αυτό το δίκτυο αποτελεί ουσιαστικά ένα δίκτυο αναγνώρισης εικόνων και γι΄ αυτό μπορούμε να το προεκπαιδεύσουμε σε ένα μεγάλο dataset που προορίζεται για εργασίες image recognition.

- **Temporal Stream CNN**: η είσοδος αυτού του δικτύου σχηματίζεται από μία στοιβάδα πεδίων μετατόπισης (*displacement fields*) των οπτικών ροών μεταξύ διαδοχικών πλαισίων. Η είσοδος αυτής της μορφής περιγράφει κατά αποκλειστικό τρόπο την κίνηση μεταξύ των video frames, γεγονός που καθιστά ευκολότερη την αναγνώριση της ζητούμενης ενέργειας. Ο τρόπος σχηματισμού του *optical flow stack* περιγράφεται στην Ενότητα 5.2.

5.2 Αναπαράσταση βίντεο μέσω Οπτικών Ροών (Optical Flow Representations)

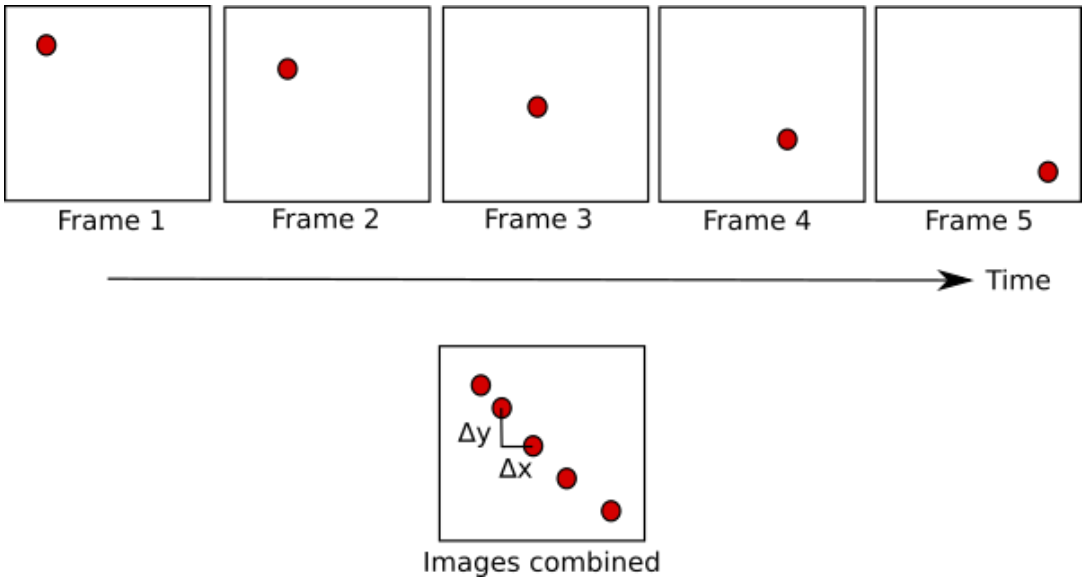
Προτού προχωρήσουμε στην περιγραφή της αρχιτεκτονικής του *Temporal Stream CNN*, κρίνουμε απαραίτητη την επεξήγηση της έννοιας του *Optical Flow*. Η μέθοδος της αναπαράστασης ενός ψηφιακού βίντεο μέσω optical flow χρησιμοποιείται σε μία πληθώρα εφαρμογών, όπως η ανίχνευση κίνησης, η ανίχνευση κινούμενων αντικειμένων ή εμποδίων. Ως Optical Flow (βλέπε εργασία [7]) εννούμε το μοτίβο της φαινομενικής κίνησης των αντικειμένων, επιφανειών και ακμών μιας ψηφιακής εικόνας μεταξύ δύο διαδοχικών καρέ (frames), που οφείλεται στη σχετική κίνηση μεταξύ ενός παρατηρητή (κάμερα) και μιας σκηνής. Πρόκειται για μία κατανομή μεταβολών της φωτεινότητας (intensity) σε μία εικόνα ή frame. Αυτό σημαίνει ότι το Optical Flow μπορεί να αντιμετωπιστεί ως ένα διανυσματικό πεδίο δύο διαστάσεων, όπου κάθε φορέας μετατόπισης αντιστοιχεί στη μετατόπιση των σημείων μεταξύ των δύο frames.

Θεωρώντας ένα pixel με συντεταγμένες (x,y,t), η έντασή του (intensity-I) προσδιορίζεται από μία συνάρτηση I(x,y,t), η οποία μεταβάλλεται εφόσον το εικονοστοιχείο μετακινηθεί. Κατά τη μελέτη των optical flow πραγματοποιούμε δύο υποθέσεις. Αρχικά, υποθέτουμε πως η ένταση ενός pixel κάποιου frame δεν αλλάζει μεταξύ διαδοχικών πλαισίων (*Brightness Constancy Assumption*), δηλαδή ότι ισχύει η σχέση:

$$I(x + \delta x, y + \delta y, t + \delta t) = I(x, y, t) \tag{5.1}$$

Επίσης, θεωρούμε ότι γειτονικά pixels σε μία εικόνα είναι πιθανό να ανήκουν στο ίδιο αντικείμενο ή επιφάνεια, δηλαδή για κάθε pixel του *pixel grid* οι γειτονικοί κόμβοι μπορεί να ανήκουν στην ίδια επιφάνεια και τα optical flows αυτών των pixel να είναι παρόμοια. Αυτή η θεώρηση είναι γνωστή ως *'Spatial Smoothness'*. Οι μέθοδοι που βασίζονται στην αναπαράσταση των βίντεο μέσω οπτικών ροών προσπαθούν να υπολογίσουν την κίνηση μεταξύ δύο πλαισίων εικόνας, τα οποία έχουν ληφθεί τις χρονικές στιγμές t και t+Δt. Αυτές

οι μέθοδοι ονομάζονται διαφορικές διότι στηρίζονται σε προσεγγίσεις των σειρών Taylor και χρησιμοποιούν μερικές παραγώγους ως προς τις χωρικές και χρονικές συνιστώσες.



Σχήμα 5.2: Δημιουργία Optical flow

Συγκεκριμένα, ένα pixel που μετακινείται κατά δx, δy και δt μεταξύ δύο πλαισίων, αποκτά μία ένταση $I(x+\delta x, y+\delta y, t+\delta t)$.Υποθέτουμε ότι η μετακίνηση του pixel είναι μικρή, οπότε μέσω Σειρών Taylor παίρνουμε το ανάπτυγμα:

$$I(x+\delta x, y+\delta y, t+\delta t) = I(x,y,t) + \frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t \tag{5.2}$$

Από τις Σχέσεις 5.1, 5.2 συμπεραίνουμε τα εξής:

$$\frac{\partial I}{\partial x}\Delta x+\frac{\partial I}{\partial y}\Delta y+\frac{\partial I}{\partial t}\Delta t = 0 \implies \frac{\partial I}{\partial x}\frac{\Delta x}{\Delta t}+\frac{\partial I}{\partial y}\frac{\Delta y}{\Delta t}+\frac{\partial I}{\partial t}\frac{\Delta t}{\Delta t} = 0 \implies \frac{\partial I}{\partial x}V_x+\frac{\partial I}{\partial y}V_y+\frac{\partial I}{\partial t} = 0 \tag{5.3}$$

Στην τελευταία Σχέση, τα V_x, V_y αποτελούν τις x,y συνιστώσες της ταχύτητας (*velocity*) του *optical flow* της έντασης I(x,y,t) και τα $\frac{\partial I}{\partial x}$, $\frac{\partial I}{\partial y}$ και $\frac{\partial I}{\partial t}$ αντιστοιχούν στις μερικές παραγώγους της έντασης του εικονοστοιχείου (x,y,t). Χρησιμοποιώντας τη γραφή I_x, I_y και I_t για τις μερικές παραγώγους στη σχέση που καταλήξαμε, οδηγούμαστε στην εξίσωση:

$$I_x V_x + I_y V_y = -I_t \tag{5.4}$$

Αυτή η εξίσωση περιέχει δύο αγνώστους και δεν επιδέχεται κάποια λύση. Το συγκεκριμένο πρόβλημα των αλγορίθμων οπτικής ροής ονομάζεται '*aperture problem*' και για να επιλυθεί χρειάζεται ένα επιπλέον σύνολο εξισώσεων. Όλες οι μέθοδοι που βασίζονται στα Optical Flow Representations επιδέχονται επιπλέον περιορισμούς ώστε να εκτιμήσουν την πραγματική ροή της έντασης των pixel.

5.3 Σύνολο δεδομένων UCF-101

Το μοντέλο που κατασκευάζουμε χρησιμοποιεί δεδομένα που ανήκουν στο dataset *UCF-101* (βλέπε εργασία [89]), το οποίο αποτελεί επέκταση του dataset *UCF-50* (βλέπε εργασία [79]). Πρόκειται για το πιο διαδεδομένο dataset στον κλάδο του *Video Action Recognition*. Συγκεκριμένα, αποτελείται από *13320 βίντεο*, τα οποία ανήκουν σε 101 διαφορετικές κατηγορίες και παράγουν *27 ώρες* δεδομένων. Τα βίντεο που ανήκουν στην κάθε action class χωρίζονται σε *25 ομάδες* (groups), καθεμία από τις οποίες περιλαμβάνει *4-7 βίντεο*. Τα βίντεο που ανήκουν στην ίδια ομάδα διαθέτουν κάποια κοινά χαρακτηριστικά, όπως είναι το background ή τα πρόσωπα που συμμετέχουν. Επίσης, η μέση διάρκεια του κάθε clip είναι *7.21sec*, ενώ το frame rate σε όλα είναι *25fps* και η ανάλυσή τους *320x240*. Τα κυριότερα χαρακτηριστικά των δεδομένων συνοψίζονται στον Πίνακα 5.1.

Η βάση δεδομένων περιλαμβάνει ρεαλιστικά βίντεο που έχουν δημοσιευτεί στο YouTube από τυχαίους χρήστες, ενώ δε διαθέτουν κάποια δομημένη μορφή, η κάμερα βιντεοσκόπησης ενδέχεται να κινείται, το background να μη διακρίνεται με μεγάλη ευκρίνεια, οι συνθήκες φωτισμού να ποικίλλουν, η ποιότητα των frames να είναι χαμηλή και να υπάρχει μερική επικάλυψη διαφορετικών αντικειμένων. Το διακριτικό πλεονέκτημα της προκειμένης βάσης δεδομένων από τα υπόλοιπα dataset που παριστάνουν ανθρώπινες ενέργειες έγκειται στο μεγάλο πλήθος κλάσεων που διαθέτει και στο γεγονός ότι τα ιδεο έχουν βιντεοσκοπηθεί σε μη ελεγχόμενα περιβάλλοντα.

Πλήθος ενεργειών	101
Πλήθος βίντεο	13320
Πλήθος ομάδων ανά ενέργεια	25
Πλήθος βίντεο ανά ομάδα	4-7
Μέση χρονική διάρκεια βίντεο	7.21sec
Συνολική χρονική διάρκεια βίντεο	1600mins
Ελάχιστη χρονική διάρκεια βίντεο	1.06sec
Μέγιστη χρονική διάρκεια βίντεο	71.04sec
Ρυθμός πλαισίων	25fps
Ανάλυση βίντεο	320x240

Πίνακας 5.1: Σύνοψη των χαρακτηριστικών του dataset UCF-101

Η κάθε ενέργεια ανήκει σε κάποια από τις εξής πέντε ομάδες: '*Αλληλεπίδραση Ανθρώπου με Αντικείμενο*' (Human-Object Interaction), '*Αλληλεπίδραση Ανθρώπου με Άνθρωπο*' (Human-Human Interaction), '*Αποκλειστική Κίνηση του Σώματος*', '*Μουσικά Όργανα*' και '*Αθλήματα*'. Συνολικά, οι 101 διαφορετικές ενέργειες που παρουσιάζονται στο UCF 101 είναι οι ακόλουθες: Apply Eye Makeup, Apply Lipstick, Archery, Baby Crawling, Balance Beam, Band Marching, Baseball Pitch, Basketball, Basketball Dunk, Bench Press, Biking, Billiards, Blow Dry Hair, Blowing Candles, Body Weight Squats, Bowling, Boxing-

Punching Bag, Boxing-Speed Bag, Breast Stroke, BrushingTeeth, Clean And Jerk, Cliff Diving, Cricket Bowling, Cricket Shot, Cutting In Kitchen, Diving, Drumming, Fenching, Field Hockey Penalty, Floor Gymnastics, Frisbee Catch, Front Crawl, Golf Swing, Haircut, Hammering, Hammer Throw, Handstand Pushups, Handstand Walking, HeadMassage, High Jump, Horse Race, Horse Riding, Hula Hoop, Ice Dancing, Javelin Throw, Juggling Balls, Jumping Jack, Jump Rope, Kayaking, Knitting, Long Jump, Lunges, Military Parade, Mixing, Mopping Floor, Nunchucks, Parallel Bars, Pizza Tossing, Playing Cello, Playing Daf, Playing Dhol, Playing Flute, Playing Quitar, Playing Piano, Playing Sitar, Playing Tabla, Playing Violin, Pole Vault, Pommel Horse, Pull Ups, Rafting, Rock Climbing Indoor, Rope Climbing, Rowing, Salsa Spin, Shaving Beard, Shotput, Skate Boarding, Skiing, Skiijet, Sky Diving, Soccer Juggling, Soccer Penalty, Still Rings, Sumo Wrestling, Surfing, Swing, Table Tennis Shot, Tai Chi, Tennis Swing, Throw Discus, Trampoline Jumping, Typing, Uneven Bars, Volleyball Spiking, Walking With Dog, Wall Pushups, Writing On Board, YoYo.

Στο Σχήμα 5.3 που ακολουθεί, παρουσιάζεται ένα δείγμα για καθεμία από τις 101 διαφορετικές κατηγορίες ενεργειών του dataset UCF-101:

Σχήμα 5.3: Σύνολο δεδομένων UCF-101

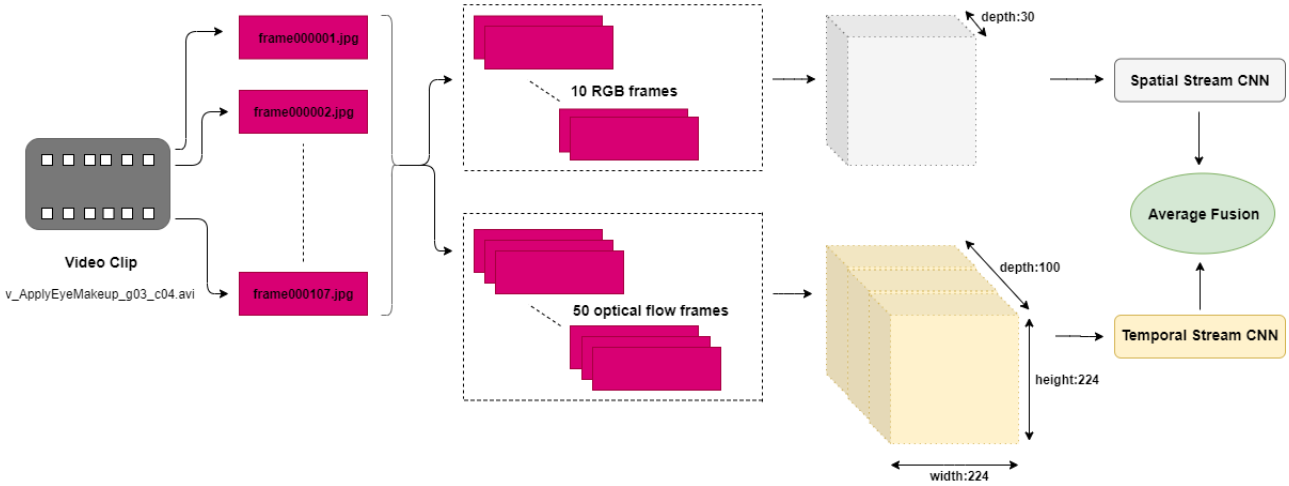
Επιπλέον, αξίζει να αναφερθούμε στη σύμβαση που ακολουθείται για την ονομασία του κάθε βίντεο που περιλαμβάνεται στο dataset. Ο φάκελος των δεδομένων περιλαμβάνει 101 διαφορετικούς υποφακέλους, ο καθένας από τους οποίους περιλαμβάνει τα βίντεο μιας κατηγορίας ενέργειας. Η μορφή του ονόματος του κάθε βίντεο είναι η εξής: $v\_X\_gY\_cZ.avi$, όπου τα X,Y και Z αναπαριστούν το label μιας συγκεκριμένης κατηγορίας, τον αριθμό της ομάδας (group) στην οποία ανήκει το δεδομένο βίντεο (clip) και τον αριθμό του clip κατά αντιστοιχία. Για παράδειγμα, το βίντεο v_ApplyEyeMakeup_g03_c04.avi αντιστοιχεί στο clip 4 του group 3 της action class "ApplyEyeMakeup".

5.4 Λεπτομέρειες υλοποίησης του συστήματος

Ο τρόπος με τον οποίο υλοποιούμε το *Two-Stream Network*, τόσο από την πλευρά των αρχιτεκτονικών επιλογών που κάνουμε όσο και από την πλευρά της εκπαιδευτικής διαδικασίας που ακολουθούμε, αποτελεί καθοριστικό παράγοντα για την επίδοση οποιουδήποτε συνδυασμού *Συνελικτικών Νευρωνικών Δικτύων*.

5.4.1 Αρχιτεκτονική του μοντέλου

Όπως αναφέραμε και στην ενότητα 5.1, το σύστημα που κατασκευάζουμε αποτελείται από δύο Συνελικτικά Νευρωνικά Δίκτυα, τα οποία εκπαιδεύονται ξεχωριστά και παρουσιάζουν κοινή δομή στην αρχιτεκτονική τους. Στα πλαίσια της παρούσας εργασίας θα αναφερόμαστε στο πρώτο CNN ως '*Spatial Stream CNN*', διότι ο ρόλος του συνοψίζεται στον εντοπισμό της χωρικής πληροφορίας που περιέχεται στα frames του εξεταζομένου βίντεο. Το δεύτερο CNN θα αποκαλείται '*Temporal Stream CNN*' και είναι υπεύθυνο για την καταγραφή της χρονικής πληροφορίας, δηλαδή της κίνησης που παρουσιάζεται μεταξύ των αντικειμένων των πλαισίων κατά τη διάρκεια του βίντεο. Ο αρχιτεκτονικός σχεδιασμός των δύο δικτύων βασίζεται σε αυτόν που έχει ακολουθηθεί από τους Zeiler και Fergus στην έρευνά τους (βλέπε εργασία [109]) επί των τεχνικών οπτικοποιήσης της λειτουργίας των ενδιάμεσων συνελικτικών επιπέδων ενός CNN και του ρόλου των ταξινομητών. Το αρχιτεκτονικό μοντέλο που κατασκευάσουμε απεικονίζεται στο Σχήμα 5.4.

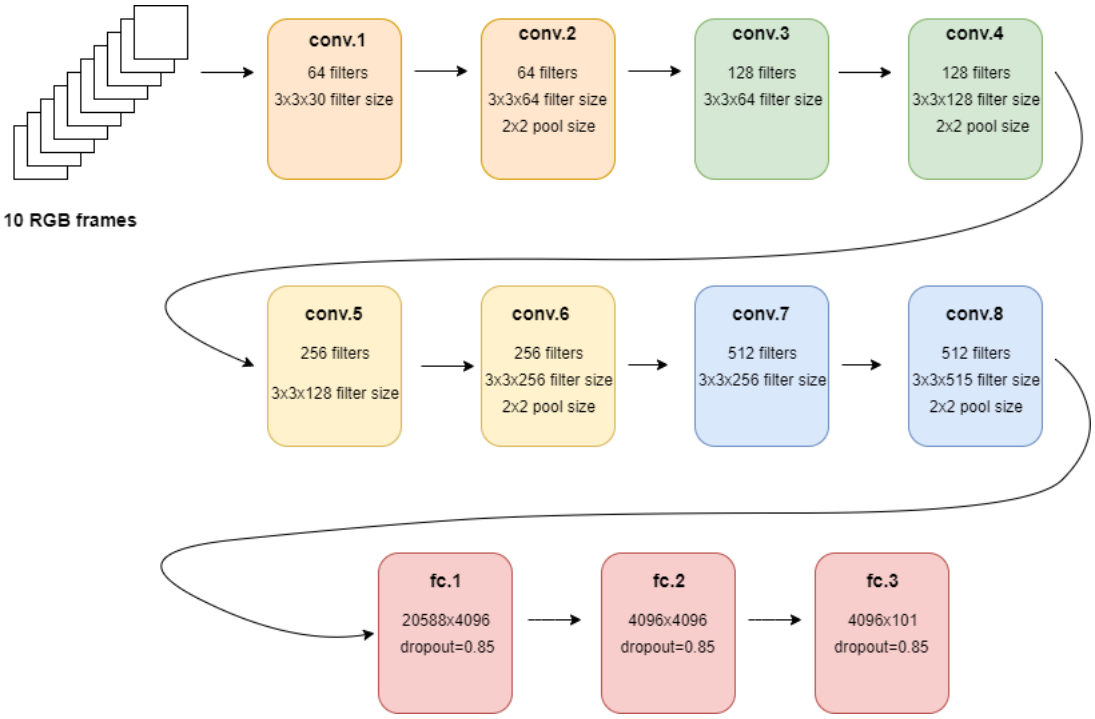


Σχήμα 5.4: Αρχιτεκτονική του χρησιμοποιούμενου μοντέλου

Κάθε CNN που στοχεύει στην αναγνώριση αντικειμένων σε εικόνες, δέχεται ως είσοδο μία δισδιάστατη (*2D*) εικόνα x_i που διαθέτει τρία χρωματικά κανάλια (R,G,B) και μέσω μίας ακολουθίας επιπέδων την αντιστοιχίζει με ένα διάνυσμα πιθανότητας (*probability vector*) $\hat{y}_i$, το οποίο έχει τόσες διαστάσεις όσες είναι οι κατηγορίες ταξινόμησης των εικόνων. Στην υλοποίησή μας οι κατηγορίες που διαθέτουμε είναι 101 και για τον εντοπισμό τους χρησιμοποιούμε *11 επίπεδα* (layers), από τα οποία τα πρώτα 7 είναι συνελικτικά (*convolutional*) και τα

τελευταία 3 είναι πλήρως-συνδεδεμένα (*fully-connected*). Μετά από αρκετούς πειραματισμούς, καταλήξαμε στη συγκεκριμένη οργάνωση της αρχιτεκτονικής διότι αυτή οδηγεί το μοντέλο μας σε βέλτιστες επιδόσεις, ενώ διαπιστώσαμε ότι *αφαιρώντας κάποιο από τα ενδιάμεσα συνελικτικά επίπεδα το μοντέλο λειτουργεί με μικρότερη ακρίβεια.*

- **Spatial Stream CNN**: Κάθε *convolutional layer* εκτελεί την πράξη της συνέλιξης (convolution) μεταξύ της εξόδου του προηγούμενου επιπέδου (ή της εικόνας εισόδου, εάν πρόκειται για το πρώτο συνελικτικό επίπεδο του δικτύου) και ενός συνόλου εκπαιδευομένων φίλτρων (*filters*), προκειμένου να παραγάγει χάρτες με τα χαρακτηριστικά (*feature maps*) της εικόνας. Κάθε φίλτρο ή feature map που προκύπτει έχει τετραγωνικό σχήμα. Στη συνέχεια, τα αποτελέσματα των συνελίξεων διέρχονται μέσα από μία γραμμική συνάρτηση ανόρθωσης (*Rectified Linear Unit-ReLU*), ενώ σε ορισμένες περιπτώσεις ενδέχεται να συγχωνευθούν με τα γειτονικά τους στοιχεία μέσω *max pooling* ή να υποστούν κάποιου είδους κανονικοποίηση (*normalisation*).



Σχήμα 5.5: Αρχιτεκτονική του Spatial Stream CNN

Ειδικότερα, το πρώτο συνελικτικό επίπεδο (conv.1) του Spatial Stream CNN δέχεται ως είσοδο μία στοιβάδα 10 RGB εικόνων διαστάσεων 224x224x3, καθεμία από τις οποίες διαθέτει τρία κανάλια χρώματος (R,G,B), οπότε ο συνολικός όγκος της εισόδου έχει διαστάσεις 224x224x30. Χρησιμοποιούνται 64 διαφορετικά φίλτρα (filters) μεγέθους 3x3x30, ενώ η έξοδος του conv.1 είναι ένας χάρτης χαρακτηριστικών (feature map).

Το δεύτερο συνελικτικό επίπεδο (conv.2) παίρνει απευθείας ως είσοδο την έξοδο του πρώτου συνελικτικού επιπέδου και τη φιλτράρει χρησιμοποιώντας 64 kernels διαστάσεων 3x3x64, ενώ στη συνέχεια εφαρμόζει pooling διαστάσεων 2x2 στην έξοδό του. Ακο-

λούθως, το τρίτο συνελικτικό επίπεδο (conv.3) χρησιμοποιεί 128 φίλτρα διαστάσεων 3x3x64, όπως και το τέταρτο, με τη διαφορά ότι το τελευταίο εφαρμόζει pooling διαστάσεων 2x2. Ακολουθεί αντίστοιχο μοτίβο και στα υπόλοιπα τέσσερα συνελικτικά επίπεδα, όπως φαίνεται και στο Σχήμα 5.5.

Αξίζει να σημειωθεί ότι *όσο μικρότερες είναι οι διαστάσεις των φίλτρων που χρησιμοποιούμε, τόσο πιο έντονα είναι τα χαρακτηριστικά που εξάγουμε.*

- **Temporal Stream CNN**: Ο ρόλος των συνελικτικών επιπέδων του δικτύου είναι η ανάλυση της κάθε εικόνας εισόδου στα χαρακτηριστικά της (features) και η παραγωγή των αντίστοιχων feature maps. Στη συνέχεια, η έξοδος του τελευταίου συνελικτικού επιπέδου (conv.8) οδηγείται στα *fully-connected layers*, τα οποία αναλαμβάνουν την τελική πρόβλεψη του δικτύου, δηλαδή την αντιστοίχιση της εικόνας με κάποια ετικέτα (label).

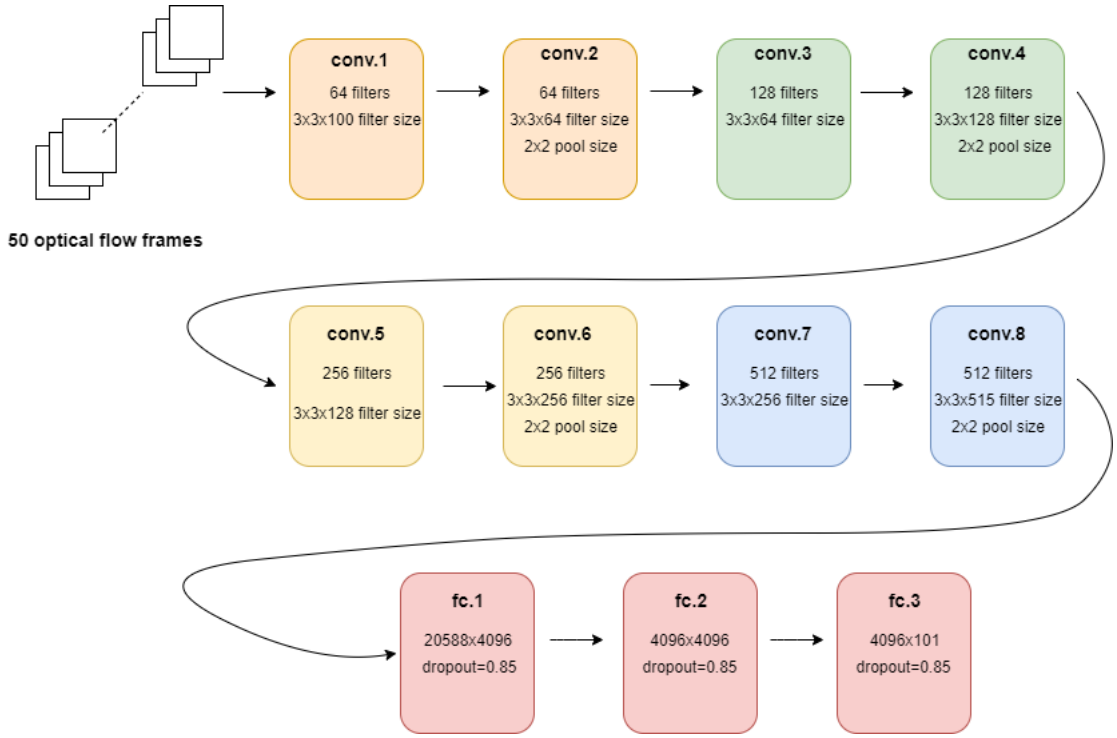
  Για να συμβεί αυτό, μετά τις πράξεις των convolution, activating, pooling και normalising, ακολουθεί το flattening, δηλαδή η μετατροπή της εξόδου των συνελικτικών επιπέδων σε ένα διάνυσμα τιμών, που η καθεμία αναπαριστά την πιθανότητα ένα συγκεκριμένο χαρακτηριστικό της εικόνας να ανήκει σε μία ετικέτα. Συγκεκριμένα, το πρώτο πλήρως-συνδεδεμένο επίπεδο (fc.1) παίρνει τα χαρακτηριστικά του top-convolutional layer σε μορφή διανύσματος διαστάσεων $3 \cdot 3 \cdot 515$ και χρησιμοποιεί 4096 units. Το δεύτερο πλήρως-συνδεδεμένο επίπεδο (fc.2) αποτελείται και αυτό από 4096 units, ενώ το τελευταίο από 101. Καθεμία από τις 101 τελικές εξόδους αντιστοιχεί και σε μία κλάση-ενέργεια του προβλήματος, δηλαδή σε ένα από τα 101 βίντεο του dataset UCF-101.

  Η αρχιτεκτονική του νευρωνικού μας δικτύου διαθέτει έναν τεράστιο αριθμό παραμέτρων, γεγονός που συνεπάγεται ανεπάρκεια του CNN να εκπαιδευτεί και να μάθει τόσο μεγάλο πλήθος παραμέτρων χωρίς να παρουσιαστεί *overfitting* στα δεδομένα εισόδου. Προκειμένου να μειώσουμε το overfitting του μοντέλου μας, υλοποιούμε τη μέθοδο κανονικοποίησης *dropout* στα fully-connected layers, η οποία αποδείχθηκε εξαιρετικά αποτελεσματική.

  Μέσω αυτού του τρόπου, δηλαδή θέτοντας την τιμή της παραμέτρου dropout ίση με 0.85, μηδενίζουμε την έξοδο κάθε κρυμμένου νευρώνα (hidden neuron) με πιθανότητα 0.85. Με αυτό τον τρόπο, όσοι νευρώνες 'απορριφθούν' δε θα λάβουν μέρος στο forward pass, ούτε θα συμμετέχουν στο backpropagation. Αυτό σημαίνει ότι κάθε φορά που το δίκτυό μας εκπαιδεύεται, προσαρμόζει την αρχιτεκτονική, αλλά μοιράζεται πάντα τις ίδιες τιμές βαρών. Στόχος είναι να καταστήσουμε τον κάθε νευρώνα ανεξάρτητο από την παρουσία άλλων νευρώνων. Κατά το testing, χρησιμοποιούμε όλους τους νευρώνες αλλά πολλαπλασιάζουμε τις εξόδους τους με έναν παράγοντα 0.85, που αποτελεί μία λογική προσέγγιση για να πάρουμε το γεωμετρικό μέσο όρο των κατανομών προβλέψεων που προέκυψαν από τα dropout δίκτυα.

  Παράλληλα, ένας άλλος τρόπος μείωσης του overfitting είναι το *data augmentation*,

δηλαδή η τεχνητή προσαύξηση του χρησιμοποιούμενου dataset χωρίς οι τροποποιήσεις που επιβάλλουμε στα δεδομένα να επηρεάζουν τις ετικέτες τους. Αυτός ο τρόπος αναλύεται στο πλαίσιο της περιγραφής της εκπαίδευσης του μοντέλου μας, στην ενότητα 5.4.2.



Σχήμα 5.6: Αρχιτεκτονική του Spatial Stream CNN

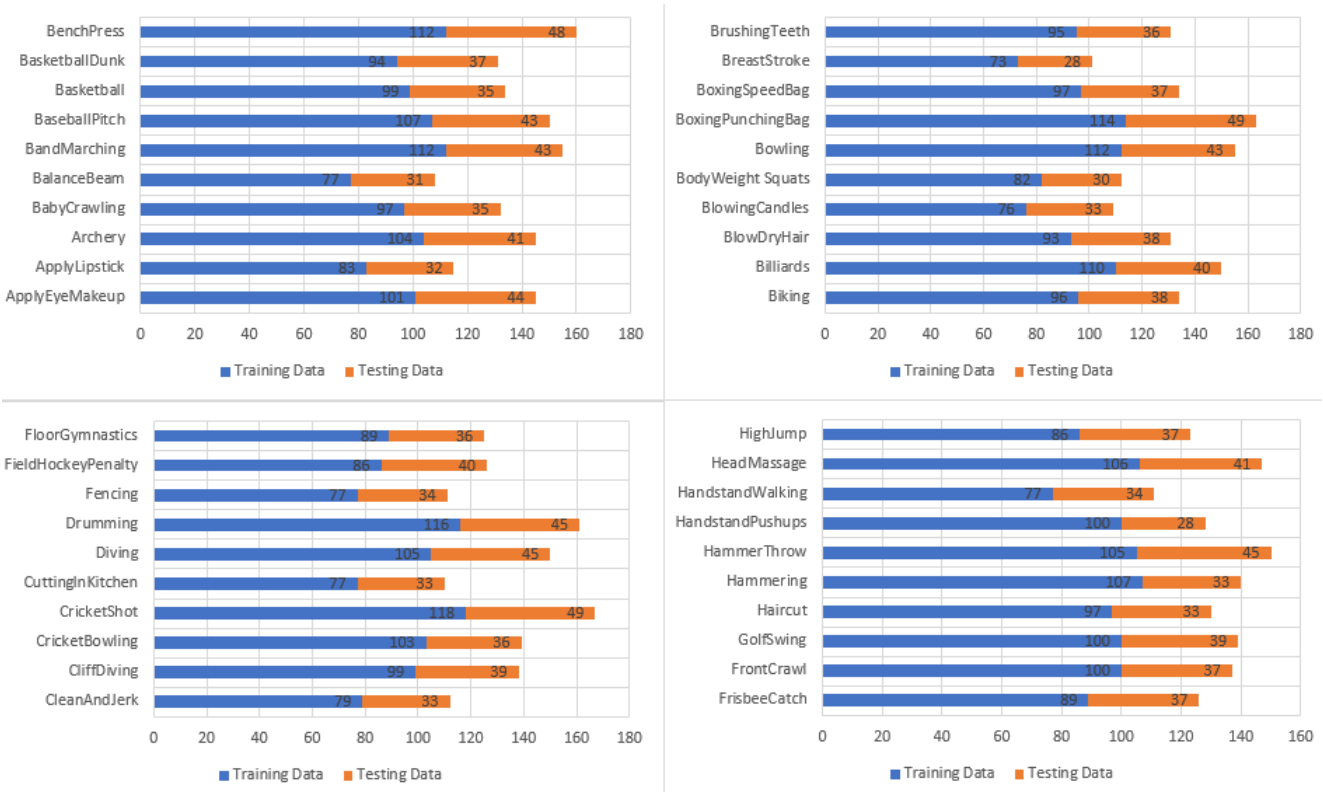
Όπως παρατηρούμε στα σχήματα 5.5 και 5.6, τα χαρακτηριστικά της αρχιτεκτονικής των δύο δικτύων Spatial Stream CNN και Temporal Stream CNN είναι σχδόν ίδια, με τη μοναδική διαφορά να παρατηρείται στο πρώτο συνελικτικό επίπεδο (conv.1), όπου στην πρώτη περίπτωση χρησιμοποιούμε φίλτρο βάθους 30 διαστάσεων, ενώ στη δεύτερη βάθους 100.

5.4.2 Εκπαίδευση του μοντέλου

Αφού ορίσουμε την αρχιτεκτονική του δικτύου, δηλαδή επιλέξουμε τον αριθμό των επιπέδων, το πλήθος των νευρώνων, το τρόπο σύνδεσης αυτών και τις συναρτήσεις ενεργοποίησης σε κάθε επίπεδο, καλούμαστε να το εκπαιδεύσουμε χρησιμοποιώντας τον αλγόριθμο πίσω διάδοσης σφάλματος (*backpropagation*) και της μεθόδου καθόδου κλίσης (*gradient descent*). Η εκπαίδευση του νευρωνικού δικτύου έχει ως στόχο την κατάλληλη επιλογή των συναπτικών βαρών και πολώσεων για όλους του νευρώνες ώστε να παράγονται οι σωστές έξοδοι για κάθε είσοδο.

Σε γενικές γραμμές, η εκπαίδευση του μοντέλου μας γίνεται χρησιμοποιώντας ένα μεγάλο σύνολο από κατηγοριοποιημένες εικόνες x, y, δηλαδή εικόνες που διαθέτουν κάποια ετικέτα (label) y. Χρησιμοποιούμε τα ψηφιακά δεδομένα βίντεο που μας παρέχονται από το dataset

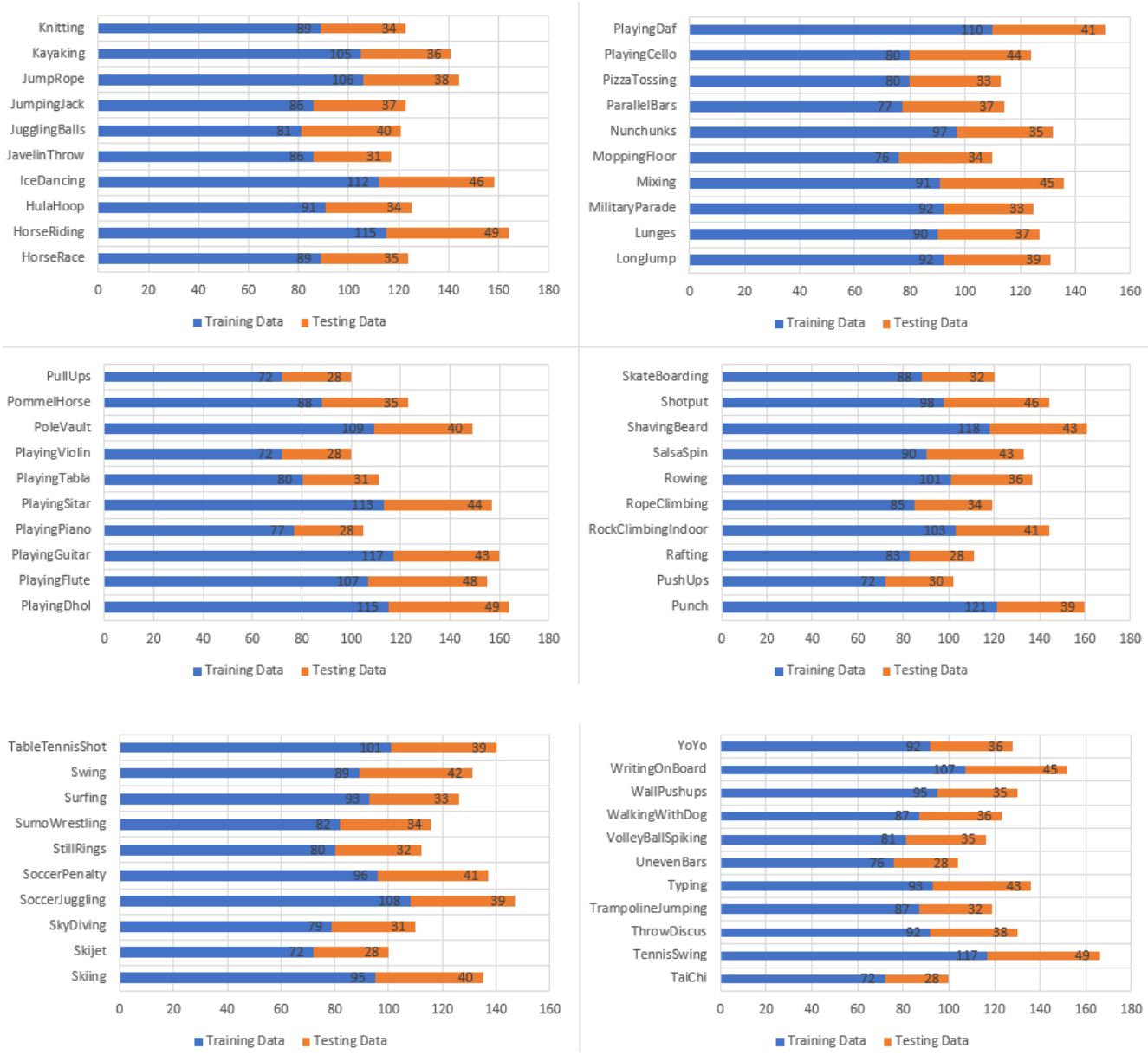
UCF-101, τα οποία είναι κατηγοριοποιημένα σε 101 κατηγορίες ενεργειών, έχουν αναλυθεί σε optical flows και έχουν χωριστεί σε *training data* και *testing data*. Τα δεδομένα είναι χωρισμένα σε τρία splits και εμείς στα πειράματά μας χρησιμοποιούμε το πρώτο. Στις ακόλουθες εικόνες έχουμε απεικονίσει την κατανομή των βίντεο για καθεμία από τις 101 κατηγορίες ενεργειών.



Προτού εισάγουμε τα δεδομένα μας στο σύστημα, τα υποβάλλουμε σε ορισμένες τροποποιήσεις (*transformations*). Συγκεκριμένα, αλλάζουμε το μέγεθος του κάθε πλαισίου (frame) ώστε η μικρότερη διάστασή του να ισούται με 256, στη συνέχεια κρατάμε μόνο την κεντρική περιοχή (center) διαστάσεων 256x256. Ακολούθως, από αυτή την περιοχή επιλέγουμε τυχαία ένα κομμάτι της, το οποίο έχει διαστάσεις 224x224. Στη συνέχεια, αυτό το τμήμα θα υποστεί RGB jitering, δηλαδή οι οριζόντιες γραμμές του θα μετακινηθούν, και horizontal flipping, που σημαίνει ότι θα περιστραφεί ως προς τον x-άξονα.

Οι αρχικές τιμές παραμέτρων των δύο CNN που υλοποιούμε, προέρχονται από το μοντέλο VGG-16, το οποίο έχει προεκπαιδευτεί στο ILSVRC-2012. Το συγκεκριμένο dataset έχει σχηματιστεί στα πλαίσια του διαγωνισμού 'ImageNet Large-Scale Visual Recognition Challenge' και διαθέτει ένα υποσύνολο του ευρέως διαδεδομένου συνόλου δεδομένων 'ImageNet' (βλέπε εργασία [25]), με περίπου 1.000 εικόνες στην καθεμία από τις 1.000 κατηγορίες εικόνων. Αθροιστικά, χρησιμοποιούνται περίπου *1.2 εκατομμύρια εικόνων στο training* του δικτύου, *50.000 εικόνες στο validation* και *150.000 εικόνες στο testing*.

Η ανανέωση των βαρών στο κάθε CNN γίνεται βάσει του αλγορίθμου στοχαστικής καθόδου κλίσης (*Stochastic Gradient Descent-SGD*) έχοντας επιλέξει το μέγεθος του *mini-*



Σχήμα 5.7: Κατανομή των βίντεο στο UCF-101-split01

batch να ισούται με *128*, την αρχική τιμή του συντελεστή εκπαίδευσης (*learning rate*) του δικτύου να είναι $10^{-}3$ και το *momentum* να είναι ορισμένο στην τιμή *0,9*. Ο ρυθμός (*step*) με τον οποίο μειώνουμε το learning rate παραμένει σταθερό στην τιμή *10*.

Προκειμένου να ελέγξουμε πόσο καλά έχει εκπαιδευτεί το μοντέλο μας στα δεδομένα εισόδου, το τροφοδοτούμε με ένα τυχαίο βίντεο, από το οποίο διαλέγουμε *25 πλαίσια*, τα οποία απέχουν ίση χρονική απόσταση μεταξύ τους. Από το κάθε πλαίσιο, παράγουμε μέσω μετασχηματισμών, όπως *cropping* και *flipping*, 10 διαφορετικές εικόνες-πλαίσια, τα οποία σχηματίζουν το *stack των 10 RGB frames* που αποτελούν τις εισόδους του Spatial Stream CNN. Δεδομένου ότι η κάθε RGB εικόνα διαθέτει τρία χρωματικά κανάλια R,G,B, η είσοδός του θα έχει βάθος *30*.

Επίσης, από τα 25 πλαίσια σχηματίζεται το *stack των 50 optical flow frames*, τα οποία τροφοδοτούν το Temporal Stream CNN. Το κάθε optical flow frame έχει αναλυθεί σε δύο άλλα frames, που αντιστοιχούν στις οριζόντιες και κατακόρυφες μετακινήσεις των διαδοχικών πλαισίων, με αποτέλεσμα να δημιουργείται μία είσοδος *βάθους 100* για το Temporal Stream CNN.

Στη συνέχεια, η πρόβλεψη του συστήματός μας για το αρχικό βίντεο προκύπτει από τον υπολογισμό του μέσου όρου των προβλέψεων για τα 25 πλαίσια. Αξίζει να σημειώσουμε ότι κατά το pre-training του Spatial Stream CNN, χρησιμοποιούμε τις ίδιες τεχνικές *data augmentation* τόσο στα δεδομένα εκπαίδευσης όσο και σε αυτά που προορίζονται για το testing του μοντέλου. Οι τεχνικές αυτές δεν επηρεάζουν αρνητικά την αναγνώριση της εκάστοτε κατηγορίας αναπαριστούμενης ενέργειας από την κάθε εικόνα, παρά μόνο δημιουργούν επιπλέον παραδείγματα για τη συγκεκριμένη κατηγορία.

5.4.3 Αξιολόγηση του μοντέλου

Αφού έχουμε εκπαιδεύσει το μοντέλο μας για *25 εποχές* (epochs), επιθυμούμε να αξιολογήσουμε τις εξόδους που μας έδωσε, δηλαδή την πιθανότητα ή την κλάση που προέβλεψε. Προκειμένου να εξετάσουμε την αποτελεσματικότητα του συστήματος, χρησιμοποιούμε κάποια μετρική επί των δεδομένων ελέγχου (*testing data*).

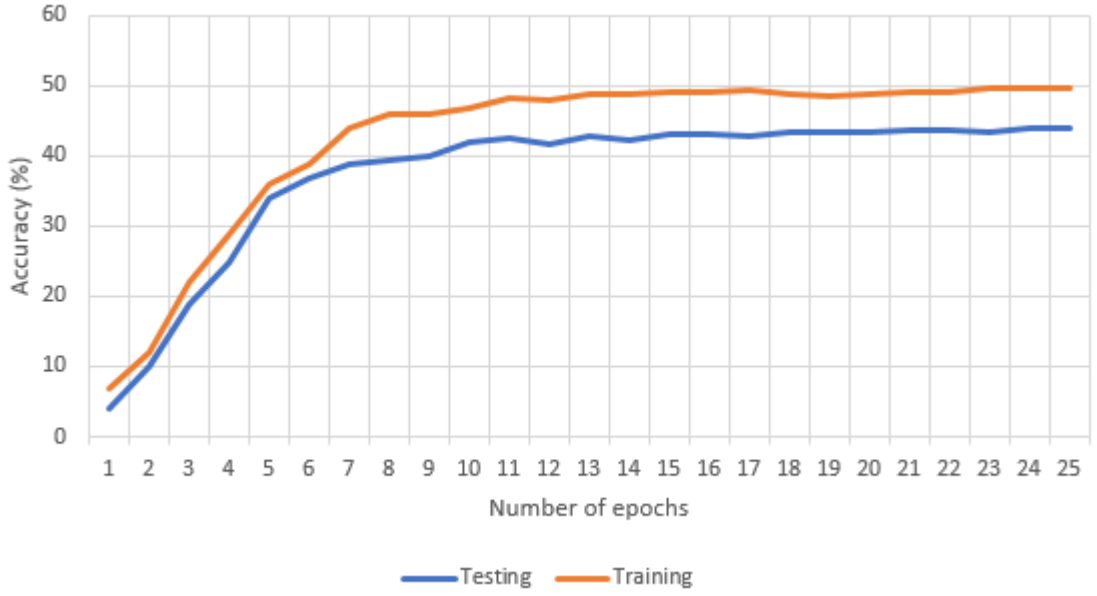
Διαφορετικές μετρικές (metrics) επίδοσης χρησιμοποιούνται για την αξιολόγηση διαφορετικών αλγορίθμων της Μηχανικής Μάθησης. Συγκεκριμένα, μετρικές που προορίζονται για τα προβλήματα ταξινόμησης, όπως είναι αυτό της αναγνώρισης ανθρώπινων ενεργειών σε βίντεο, είναι τα *Loss,Accuracy* και *Confusion Matrix*, ενώ τα *Recall* και *Precision* χρησιμοποιούνται κυρίως σε αλγορίθμους ταξινόμησης, που εφαρμόζονται κατά κόρον από τις μηχανές αναζήτησης (search engines).

Τα πειράματά μας αξιολογούνται βάσει της μετρικής *testing accuracy*, η οποία δείχνει τον αριθμό των σωστών προβλέψεων του μοντέλου ως προς το σύνολο των προψλέψεων που πραγματοποίησε. Η συγκεκριμένη τεχνική αξιολόγησης δεν αποδεικνύεται ιδιαίτερα αποτελεσματική σε περιπτώσεις όπου οι κατηγορίες ταξινόμησης δεν είναι ισορροπημένες.

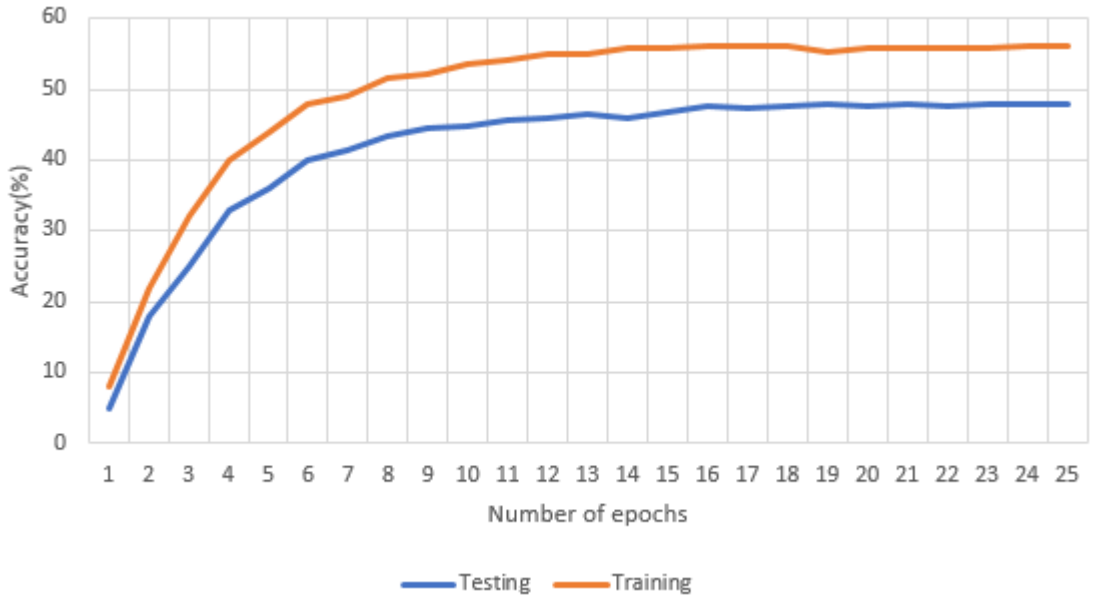
Στη συνέχεια, απεικονίζουμε τη βελτίωση των *training accuracy* και *testing accuracy* κατά την εξέλιξη των 25 εποχών εκπαίδευσης του μοντέλου μας. Παρατηρούμε ότι σε όλες τις εποχές το training accuracy δίνει καλύτερες τιμές σε σχέση με το testing accuracy. Η τελική τιμή του testing accuracy παρουσιάζεται ξεχωριστά για το κάθε δίκτυο στον Πίνακα 5.2:

Δίκτυο	Accuracy
Spatial Stream CNN	44.97%
Temporal Stream CNN	48.17%
Two-Stream Network	56.23%

Πίνακας 5.2: Επίδοση του μοντέλου

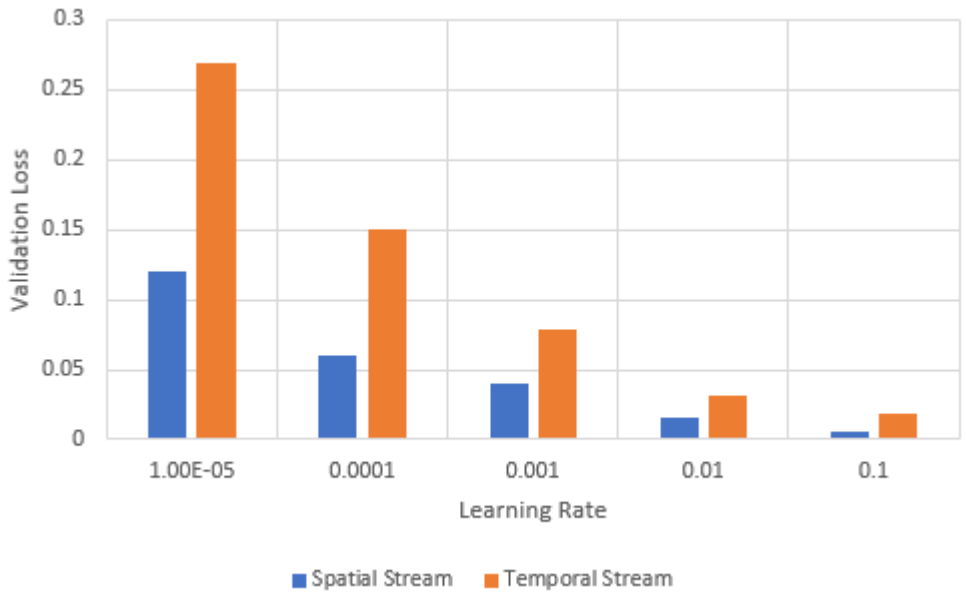


Σχήμα 5.8: Accuracy του Spatial Stream CNN



Σχήμα 5.9: Accuracy του Temporal Stream CNN

Στο Διάγραμμα 5.10 παρακολουθούμε τον τρόπο με τον οποίο επηρεάζεται το accuracy των δύο stream από αλλαγές στο ρυθμό εκμάθησης του δικτύου. *Το learning rate προσδιορίζει το step size σε κάθε επανάληψη*, καθώς το δίκτυο προσπαθεί να ελαχιστοποιήσει τη loss function.



Σχήμα 5.10: Loss συναρτήσει learning rate

Κεφάλαιο 6

Συμπεράσματα και προτάσεις

Στα προγούμενα κεφάλαια προσεγγίσαμε το πρόβλημα της '*Αναγνώρισης Ανθρώπινων Ενεργειών σε βίντεο*' εξετάζοντας τις διαφορερικές τεχνικές που έχουν κατά καιρούς χρησιμοποιηθεί. Παρουσιάσαμε μεθόδους που βασίζονται στα *handcrafted features* των πλαισίων του βίντεο, ενώ δώσαμε μεγαλύτερη έμφαση σε εκείνες τις μεθόδους που έχουν υιοθετήσει τη λογική και τις αρχές της *Βαθιάς Μηχανικής Μάθησης*. Υλοποιήσαμε το δικό μας μοντέλο *Two-Stream Network* και υπολογίσαμε την επίδοσή του χρησιμοποιώντας δεδομένα του ευρέως διαδεδομένου dataset *UCF-101*. Σε αυτό το κεφάλαιο συνοψίζουμε το έργο που διατελέσαμε στο πλαίσιο της παρούσας διπλωματικής μας εργασίας και καταγράφουμε τα συμπεράσματα στα οποία καταλήξαμε, δίνοντας κατευθύνσεις για μελλονική έρευνα πάνω στον κλάδο της αναγνώρισης ενεργειών.

6.1 Συμπεράσματα της διπλωματικής εργασίας

Αρχικά, προσπαθώντας να προσεγγίσουμε το θέμα μας και να κατανοήσουμε τις δυσκολίες που παρουσιάζει, μελετήσαμε αρκετές δημοσιευμένες εργασίες της τελευταίας δεκαετίας. Ανακαλύψαμε ότι ορισμένες υλοποιήσεις αξιοποιούν τα handcrafted χαρακτηριστικά των πλαισίων, ενώ άλλες βασίζονται στην ικανότητα των *Βαθιών Νευρωνικών Δικτύων* να εξάγουν μόνα τους τα χαρακτηριστικά από τις εικόνες-πλαίσια που δέχονται ως είσοδο. Εντέλει, οι μέθοδοι που παρουσιάζουν την καλύτερη επίδοση είναι αυτές που προσπαθούν να συνδυάσουν τις δύο αυτές προσεγγίσεις σε μία *υβριδική αρχιτεκτονική*. Είδαμε να χρησιμοποιούνται imporved Dense Trajectories για την εξαγωγή των handcrafted features, deep layers για την εκμάθηση χαρακτηριστικών, Fisher Vectors για την αναπαράσταση των βίντεο και τεχνικές μείωσης των διαστάσεών τους (*dimensionality reduction*).

Επιχειρώντας να ερμηνεύσουμε τις εντυπωσιακές επιδόσεις ορισμένων συστημάτων που βασίζονται αποκλειστικά στα *handcrafted features* των frames, καταλήγουμε ότι αυτό οφείλεται στην πρότερη γνώση που είναι ενσωματωμένη σε αυτού του είδους τα χαρακτηριστικά. Εξ΄ ορισμού, τα handcrafted features έχουν δημιουργηθεί βάσει της ανθρώπινης εμπειρίας και εξειδίκευσης, οπότε είναι λογικό να παρακάμπτουν ένα μεγάλο αριθμό ενδιάμεσων βημάτων, τα οποία 'διέσχισαν' οι άνθρωποι ώστε να καταλήξουν σε συμπεράσματα αναφορικά με τα

δεδομένα που διαχειρίζονται. Αυτό σημαίνει ότι η διαδικασία σχηματισμού αυτών των χαρακτηριστικών μπορεί να αντιμετωπιστεί ως μία τεχνική ενσωμάτωσης παλαιότερης γνώσης απευθείας στα μοντέλα της μηχανικής μάθησης.

Σε μία προσπάθεια αξιολόγησης του μοντέλου που υλοποιήσαμε στο Κεφάλαιο 5, εξάγουμε το συμπέρασμα ότι η λογική του *Two-Stream Network*, που περιλαμβάνει το συνδυασμό δύο ανεξάρτητων δικτύων CNN μέσω fusion, οδηγεί σε συνολικά μεγαλύτερο accuracy του συστήματος σε σχέση με τις τεχνικές *single-stream*. Η αρχιτεκτονική wo-Stream ανταγωνίζεται τις επιδόσεις άλλων σύγχρονων μεθόδων αναπαράστασης βίντεο και επεξεργασίας τους, αλλά αδυνατεί να προσδιορίσει τη θέση (localization) της ενέργειας ως προς το πεδίο του χώρου ή του χρόνου. Επίσης, η εκπαίδευση του *Temporal Stream CNN* πάνω σε Οπτικές Ροές (*Optical Flows*) μειώνει σημαντικά την ανάγκη εκτενούς προεπεξεργασίας των δεδομένων. Παράλληλα, στο πλαίσιο της υλοποίησης που κάναμε διαπιστώσαμε τη σημασία του βάθους στα CNN. Το κάθε επίπεδο (layer) του δικτύου προσπαθεί να μάθει την καλύτερη δυνατή αναπαράσταση των δεδομένων του προηγούμενου επιπέδου.

6.2 Προτάσεις για μελλοντική έρευνα

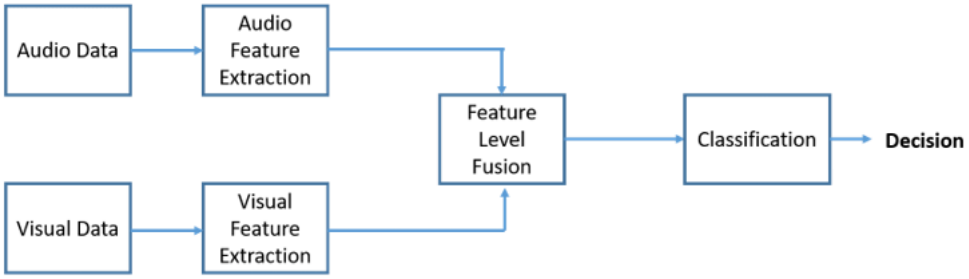
Είναι γεγονός ότι το περιορισμένο μέγεθος των dataset που προορίζονται για τον τομέα του *Video Action Recognition*, μειώνει την επίδοση των συστημάτων. Ενώ διατίθεται μία μεγάλη συλλογή δεδομένων εικόνων που είναι κατάλληλη για εφαρμογές του *Image Recognition*, αυτά τα δεδομένα δε μπορούν να χρησιμοποιηθούν και για το VAR διότι δεν ενσωματώνουν την πληροφορία της κίνησης, την οποία χρειαζόμαστε για την αναγνώριση μιας ενέργειας. Μία πιθανή λύση στο πρόβλημα θα αποτελούσε η *τεχνητή σύνθεση δεδομένων* (data synthesis), με σκοπό την ενσωμάτωση της ανθρώπινης γνώσης και εμπειρίας στα training data, αντί να ενσωματώσουμε την αντίστοιχη εξειδίκευση στο μοντέλο της μηχανικής μάθησης. Η γνώση που θα δώσουμε στα τεχνητά δεδομένα ενδέχεται να αφορά τα αντικείμενα που απεικονίζονται στα πλαίσια του βίντεο, τις σκηνές, το φωτεισμό, τους ανθρώπους ή τις κινήσεις τους. Έτσι, είναι δυνατή η δημιουργία ενός συνόλου δεδομένων που θα περιέχει ποικίλα και ρεαλιστικά αποσπάσματα βίντεο, τα οποία θα απεικονίζουν ανθρώπινες ενέργειες.

Φυσικά, τα τεχνητά δεδομένα θα μπορούσαν να συγχωνευθούν με τα πραγματικά, ώστε να οδηγήσουν σε ακόμα πιο πλήρη σύνολα δεδομένων. Ακόμη, θα μπορούσε να εφαρμοστεί *Multi-Task Learning* στο δίκτυο που αναλαμβάνει τον εντοπισμό της κίνησης κατά μήκος των διαδοχικών πλαισίων του βίντεο, συνδυάζοντας δύο ή περισσότερα διαφορετικά dataset. Η εφαρμογή του Multi-Task Learning μπορεί να βελτιώσει το συνολικό accuracy του συστήματος δεδομένου ότι κατά τη διαδικασία της εκπαίδευσης εκμεταλλευόμαστε όλα τα διαθέσιμα training data, ενώ οδηγούμαστε σε καλύτερες αναπαραστάσεις των κατηγοριών που διαθέτουν λιγότερα δείγματα σε σχέση με άλλες.

Παράλληλα, η έρευνα θα μπορούσε να εστιάσει στις *υβριδικές αρχιτεκτονικές* και στις μεθόδους που έχουν τη δυνατότητα να αξιοποιούν το δυναμικό χαρακτήρα της χρονικής πληροφορίας σε ένα βίντεο, όπως συμβαίνει με τα *Long Short Term Memory recurrent network (LSTM)*. Επίσης, στις αρχιτεκτονικές των *Deep Neural Network*, οι τιμές των υπερπαρα-

μέτρων που επιλέγονται πρέπει να εξετάζονται ενδελεχώς.

Τέλος, ένας πιθανός τρόπος βελτίωσης της επίδοσης ενός συστήματος που στοχεύει στην αναγνώριση ενεργειών, θα ήταν η χρήση ενός *Multi-Modal System* (βλέπε εργασία [63]), δηλαδή ενός συστήματος που δεν εκμεταλλεύεται μόνο τα οπτικά χαρακτηριστικά των πλαισίων, αλλά και άλλου είδους χαρακτηριστικά που σχετίζονται με την ίδια τη φύση του βίντεο. Μία πιθανή υλοποίηση θα μπορούσε να αποτελεί αυτή του Σχήματος 6.1, όπου συνδυάζεται η χρήση ενός CNN που εξάγει *visual features* με ένα άλλο CNN που εξάγει *audio features* από τα πλαίσια του βίντεο.



Σχήμα 6.1: Multi-modal σύστημα για την αναγνώριση ενεργειών σε βίντεο

Βιβλιογραφία

[1] Sami Abu-El-Haija et al. "Youtube-8m: A large-scale video classification benchmark". In: *arXiv preprint arXiv:1609.08675* (2016).

[2] Jake K Aggarwal and Michael S Ryoo. "Human activity analysis: A review". In: *ACM Computing Surveys (CSUR)* 43.3 (2011), p. 16.

[3] Md. Zahangir Alom et al. "The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches". In: *CoRR* abs/1803.01164 (2018). arXiv: 1803.01164. URL: http://arxiv.org/abs/1803.01164.

[4] M Baccouche et al. "International Workshop on Human Behavior Understanding". In: (2011).

[5] M Baccouche et al. "International Workshop on Human Behavior Understanding". In: (2011).

[6] H.B. Barlow. "Unsupervised Learning". In: *Neural Computation* 1.3 (1989), pp. 295–311. DOI: 10.1162/neco.1989.1.3.295. eprint: https://doi.org/10.1162/neco.1989.1.3.295. URL: https://doi.org/10.1162/neco.1989.1.3.295.

[7] Brian G.Schunck Berthold K.P.Horn. "Determining optical flow". In: (). URL: http://image.diku.dk/imagecanon/material/HornSchunckOptical_Flow.pdf.

[8] Moshe Blank et al. "Actions as space-time shapes". In: *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*. Vol. 2. IEEE. 2005, pp. 1395–1402.

[9] Aaron F Bobick and James W Davis. "The recognition of human movement using temporal templates". In: *IEEE Transactions on Pattern Analysis & Machine Intelligence* 3 (2001), pp. 257–267.

[10] "Introduction". In: *Handbook of Image and Video Processing (Second Edition)*. Ed. by AL BOVIK. Second Edition. Communications, Networking and Multimedia. Burlington: Academic Press, 2005, p. 1. ISBN: 978-0-12-119792-6. DOI: https://doi.org/10.1016/B978-0-12-119792-6.50141-8. URL: http://www.sciencedirect.com/science/article/pii/B9780121197926501418.

[11] Fabian Caba Heilbron et al. "Activitynet: A large-scale video benchmark for human activity understanding". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 961–970.

[12] Stefan Carlsson and Josephine Sullivan. "Action recognition by shape matching to key frames". In: *Workshop on models versus exemplars in computer vision*. Vol. 1. 18. Citeseer. 2001.

[13] Joao Carreira and Andrew Zisserman. "Quo vadis, action recognition? a new model and the kinetics dataset". In: *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 6299–6308.

[14] Rich Caruana and Alexandru Niculescu-Mizil. "An Empirical Comparison of Supervised Learning Algorithms". In: *Proceedings of the 23rd International Conference on Machine Learning*. ICML '06. Pittsburgh, Pennsylvania, USA: ACM, 2006, pp. 161–168. ISBN: 1-59593-383-2. DOI: 10.1145/1143844.1143865. URL: http://doi.acm.org/10.1145/1143844.1143865.

[15] Jose M Chaquet, Enrique J Carmona, and Antonio Fernández-Caballero. "A survey of video datasets for human action and activity recognition". In: *Computer Vision and Image Understanding* 117.6 (2013), pp. 633–659.

[16] Guangchun Cheng et al. "Advances in Human Action Recognition: A Survey". In: *CoRR* abs/1501.05964 (2015). arXiv: 1501.05964. URL: http://arxiv.org/abs/1501.05964.

[17] Guilhem Chéron, Ivan Laptev, and Cordelia Schmid. "P-cnn: Pose-based cnn features for action recognition". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 3218–3226.

[18] Kaelbling L. Chieu H. Lee W. "Activity Recognition from Physiological Data Using Conditional Random Fields." In: Workshop at ICML, 2018. URL: https://dspace.mit.edu/handle/1721.1/30197.

[19] Dan Claudiu Ciresan et al. "Convolutional neural network committees for handwritten character classification". In: *2011 International Conference on Document Analysis and Recognition*. IEEE. 2011, pp. 1135–1139.

[20] Ross Cutler and Matthew Turk. "View-based interpretation of real-time optical flow for gesture recognition". In: *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE. 1998, pp. 416–421.

[21] Navneet Dalal and Bill Triggs. "Histograms of oriented gradients for human detection". In: 2005.

[22] Navneet Dalal and Bill Triggs. "Histograms of oriented gradients for human detection". In: 2005.

[23] Navneet Dalal, Bill Triggs, and Cordelia Schmid. "Human detection using oriented histograms of flow and appearance". In: *European conference on computer vision*. Springer. 2006, pp. 428–441.

[24] Trevor Darrell and Alex Pentland. "Space-time gestures". In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 1993, pp. 335–340.

[25] Jia Deng et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[26] Ali Diba et al. "Temporal 3d convnets: New architecture and transfer learning for video classification". In: *arXiv preprint arXiv:1711.08200* (2017).

[27] Pedro Domingos. "A Few Useful Things to Know About Machine Learning". In: *Commun. ACM* 55.10 (Oct. 2012), pp. 78–87. ISSN: 0001-0782. DOI: 10.1145/2347736.2347755. URL: http://doi.acm.org/10.1145/2347736.2347755.

[28] Jeff Donahue et al. *Long-term Recurrent Convolutional Networks for Visual Recognition and Description*. 2014. arXiv: 1411.4389 [cs.CV].

[29] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. "Convolutional two-stream network fusion for video action recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 1933–1941.

[30] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. "Convolutional two-stream network fusion for video action recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 1933–1941.

[31] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. "Convolutional two-stream network fusion for video action recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 1933–1941.

[32] Crowley J. Caviar Fisher R. Santos-Victor J. "Context Aware Vision Using Image-Based Active Recognition". In: 2005. URL: https://homepages.inf.ed.ac.uk/rbf/CAVIAR/.

[33] Carlos Gershenson. "Artificial Neural Networks for Beginners". In: *CoRR* cs.NE/0308031 (2003). URL: http://arxiv.org/abs/cs.NE/0308031.

[34] J.J. (1950) Gibson. "The Perception of the Visual World". In: Oxford, England: Houghton Mifflin. URL: https://psycnet.apa.org/record/1951-04286-000.

[35] Rohit Girdhar et al. "Actionvlad: Learning spatio-temporal aggregation for action classification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 971–980.

[36] Simon Haykin. "1 FEEDFORWARD NEURAL NETWORKS : AN INTRODUCTION". In: 2004.

[37] Samitha Herath, Mehrtash Harandi, and Fatih Porikli. "Going deeper into action recognition: A survey". In: *Image and vision computing* 60 (2017), pp. 4–21.

[38] Laptev I. "On Space-Time Interest Points". In: Campus Beaulieu, 35042 Rennes Cedex, France, 2005. URL: http://www.irisa.fr/vista/Papers/2005_ijcv_laptev.pdf.

[39] J.P.L.Vanderwalle I.Aizenberg N.N.Aizenberg. *Valued and Universal Binary Neurons: Theory,Learning and Applications*. Springer Science Business Media, 2000. URL: https://www.springer.com/gp/book/9780792378242.

[40] Ahmad Jalal, Shaharyar Kamal, and Daijin Kim. "Shape and motion features approach for activity tracking and recognition from kinect video camera". In: *2015 IEEE 29th International Conference on Advanced Information Networking and Applications Workshops*. IEEE. 2015, pp. 445–450.

[41] Hueihan Jhuang et al. "A biologically inspired system for action recognition". In: *2007 IEEE 11th International Conference on Computer Vision*. Ieee. 2007, pp. 1–8.

[42] Shuiwang Ji et al. "3D convolutional neural networks for human action recognition". In: *IEEE transactions on pattern analysis and machine intelligence* 35.1 (2012), pp. 221–231.

[43] Shuiwang Ji et al. "3D convolutional neural networks for human action recognition". In: *IEEE transactions on pattern analysis and machine intelligence* 35.1 (2012), pp. 221–231.

[44] Yu-Gang Jiang et al. *THUMOS challenge: Action recognition with a large number of classes*. 2014.

[45] Yu-Gang Jiang et al. "Trajectory-based modeling of human actions with motion reference points". In: *European Conference on Computer Vision*. Springer. 2012, pp. 425–438.

[46] Gunnar Johansson. "Visual perception of biological motion and a model for its analysis". In: *Perception & psychophysics* 14.2 (1973), pp. 201–211.

[47] Judson P Jones and Larry A Palmer. "An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex". In: *Journal of neurophysiology* 58.6 (1987), pp. 1233–1258.

[48] Andrej Karpathy et al. "Large-scale video classification with convolutional neural networks". In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2014, pp. 1725–1732.

[49] Andrej Karpathy et al. "Large-scale video classification with convolutional neural networks". In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2014, pp. 1725–1732.

[50] Will Kay et al. "The kinetics human action video dataset". In: *arXiv preprint arXiv:1705.06950* (2017).

[51] Henry J. Kelley. "Gradient Theory of Optimal Flight Paths". In: *Grumman Aircraft Engineering Corp.* (Oct. 1960). DOI: 10.2514/8.5282. URL: https://doi.org/10.2514/8.5282.

[52] K Kim. "Intelligent immigration control system by using international symposium on neural networks". In: *Proc. The International Symposium on Neural Networks*. 2005, pp. 147–156.

[53] Suyoun Kim, Takaaki Hori, and Shinji Watanabe. "Joint CTC-attention based end-to-end speech recognition using multi-task learning". In: *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2017, pp. 4835–4839.

[54] Alexander Klaser, Marcin Marszałek, and Cordelia Schmid. "A spatio-temporal descriptor based on 3d-gradients". In: 2008.

[55] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Advances in neural information processing systems". In: *Neural Information Processing Systems Foundation* 1269 (2012).

[56] Hilde Kuehne et al. "HMDB51: A Large Video Database for Human Motion Recognition". In: Nov. 2011, pp. 2556–2563. DOI: 10.1109/ICCV.2011.6126543.

[57] E. Kussul et al. "Rosenblatt perceptrons for handwritten digit recognition". In: *IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No.01CH37222)*. Vol. 2. July 2001, 1516–1520 vol.2. DOI: 10.1109/IJCNN.2001.939589.

[58] Guy Lev et al. *Rnn fisher vectors for action recognition and image annotation*. Springer, 2016.

[59] Yingwei Li et al. "Vlad3: Encoding dynamics of deep features for action recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 1951–1960.

[60] Ming Liang and Xiaolin Hu. "Recurrent convolutional neural network for object recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3367–3375.

[61] Ling Zhang and Bo Zhang. "A geometrical representation of McCulloch-Pitts neural model and its applications". In: *IEEE Transactions on Neural Networks* 10.4 (July 1999), pp. 925–929. DOI: 10.1109/72.774263.

[62] "Machine recognition of human activities:A survey". In: *Proceedings of the 18th ACM International Conference on Multimedia*. IEEE Trans.Circuits Syst.Vid.Technol, 2008. DOI: 10.1109/TCSVT.2008.2005594. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.319.4930&rep=rep1&type=pdf.

[63] Petros Maragos, Alexandros Potamianos, and Patrick Gros. *Multimodal Processing and Interaction - Audio, Video, Text*. Aug. 2008. DOI: 10.1007/978-0-387-76316-3.

[64] David Marr and Herbert Keith Nishihara. "Representation and recognition of the spatial organization of three-dimensional shapes". In: *Proceedings of the Royal Society of London. Series B. Biological Sciences* 200.1140 (1978), pp. 269–294.

[65] Marcin Marszałek, Ivan Laptev, and Cordelia Schmid. "Actions in context". In: *CVPR 2009-IEEE Conference on Computer Vision & Pattern Recognition*. IEEE Computer Society. 2009, pp. 2929–2936.

[66] Pyry Matikainen, Martial Hebert, and Rahul Sukthankar. "Trajectons: Action recognition through the motion analysis of tracked features". In: *2009 IEEE 12th international conference on computer vision workshops, ICCV workshops*. IEEE. 2009, pp. 514–521.

[67] Ross Messing, Chris Pal, and Henry Kautz. "Activity recognition using the velocity histories of tracked keypoints". In: *2009 IEEE 12th international conference on computer vision*. IEEE. 2009, pp. 104–111.

[68] Napoletano P. Unimib Shar Micucci D. Mobilio M. "A dataset for human activity recognition using acceleration data from smartphones." In: Appl.Sci, 2017. DOI: 10.3390/app7101101. URL: https://www.mdpi.com/2076-3417/7/10/1101.

[69] Thomas B Moeslund, Adrian Hilton, and Volker Krüger. "A survey of advances in vision-based human motion capture and analysis". In: *Computer vision and image understanding* 104.2-3 (2006), pp. 90–126.

[70] Wolfgang E Nagel, Michael M Resch, and Dietmar B Kroner. *High Performance Computing in Science and Engineering'10*. Springer, 2010.

[71] M.M. Nelson and W.T. Illingworth. "A practical guide to neural nets". In: (1991).

[72] Yuval Netzer et al. "Reading Digits in Natural Images with Unsupervised Feature Learning". In: *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*. 2011. URL: http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf.

[73] Juan Carlos Niebles, Chih-Wei Chen, and Li Fei-Fei. "Modeling temporal structure of decomposable motion segments for activity classification". In: *European conference on computer vision*. Springer. 2010, pp. 392–405.

[74] Simon O.Haykin. *Neural Nwtworks and Learning Machines, 3rd Edition*. McMaster University, Hamilton, Ontario, Canada: Pearson, 2009.

[75] Mikel Olazaran. "A Sociological Study of the Official History of the Perceptrons Controversy". In: *Social Studies of Science* 26.3 (1996), pp. 611–659. DOI: 10.1177/030631296026003005. eprint: https://doi.org/10.1177/030631296026003005. URL: https://doi.org/10.1177/030631296026003005.

[76] Alonso Patron-Perez et al. "High Five: Recognising human interactions in TV shows." In: *BMVC*. Vol. 1. Citeseer. 2010, p. 2.

[77] Xiaojiang Peng et al. "Action recognition with stacked fisher vectors". In: *European Conference on Computer Vision*. Springer. 2014, pp. 581–595.

[78] R.Dechter. "Learning while searching in contrast-satisfaction-problems". In: Univerity of California, Computer Science Department, Cognitive Systems Laboratory, 1986.

[79] Kishore K. Reddy and Mubarak Shah. "Recognizing 50 Human Action Categories of Web Videos". In: *Mach. Vision Appl.* 24.5 (July 2013), pp. 971–981. ISSN: 0932-8092. DOI: 10.1007/s00138-012-0450-4. URL: http://dx.doi.org/10.1007/s00138-012-0450-4.

[80] Kishore K Reddy and Mubarak Shah. "Recognizing 50 human action categories of web videos". In: *Machine Vision and Applications* 24.5 (2013), pp. 971–981.

[81] Martin Riedmiller. "Advanced supervised learning in multi-layer perceptrons-From backpropagation to adaptive learning algorithms". In: *Computer Standards Interfaces* 16.3 (1994), pp. 265–278. DOI: 10.1016/0920-5489(94)90017-5. URL: https://doi.org/10.1016/0920-5489(94)90017-5.

[82] Mikel D Rodriguez, Javed Ahmed, and Mubarak Shah. "Action MACH a spatio-temporal Maximum Average Correlation Height filter for action recognition." In: *CVPR*. Vol. 1. 1. 2008, p. 6.

[83] Anna Rohrbach et al. "A dataset for movie description". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3202–3212.

[84] Christian Schuldt, Ivan Laptev, and Barbara Caputo. "Recognizing human actions: a local SVM approach". In: *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.* Vol. 3. IEEE. 2004, pp. 32–36.

[85] Phil Simon. "too Big to Ignore". In: *The business case for big data*. Wiley and SAS Business Series. 2013. ISBN: 978-1-118-63817-0.

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

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

[88] Ray J. Solomonoff. "A formal theory of inductive inference. Part ||". In: *Information and control* (), pp. 224–254.

[89] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild". In: *CoRR* abs/1212.0402 (2012). arXiv: 1212.0402. URL: http://arxiv.org/abs/1212.0402.

[90] D. F. Specht. "A general regression neural network". In: *IEEE Transactions on Neural Networks* 2.6 (Nov. 1991), pp. 568–576. DOI: 10.1109/72.97934.

[91] Suraj Srinivas et al. "A taxonomy of deep convolutional neural nets for computer vision". In: *Frontiers in Robotics and AI* 2 (2016), p. 36.

[92] Suraj Srinivas et al. "A taxonomy of deep convolutional neural nets for computer vision". In: *Frontiers in Robotics and AI* 2 (2016), p. 36.

[93] Lin Sun et al. "Human action recognition using factorized spatio-temporal convolutional networks". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 4597–4605.

[94] Richard S. Sutton and Andrew G. Barto. *Introduction to Reinforcement Learning*. 1st. Cambridge, MA, USA: MIT Press, 1998. ISBN: 0262193981.

[95] A.Murat Tekalp. *Digital Video Processing*. Prentice Hall. URL: http://ptgmedia.pearsoncmg.com/images/9780133991000/samplepages/9780133991000.pdf.

[96] Christian Thurau and Václav Hlavác. "Pose primitive based human action recognition in videos or still images". In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2008, pp. 1–8.

[97] Du Tran et al. "Learning spatiotemporal features with 3d convolutional networks". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 4489–4497.

[98] Gül Varol, Ivan Laptev, and Cordelia Schmid. "Long-term temporal convolutions for action recognition". In: *IEEE transactions on pattern analysis and machine intelligence* 40.6 (2017), pp. 1510–1517.

[99] Gül Varol, Ivan Laptev, and Cordelia Schmid. "Long-term temporal convolutions for action recognition". In: *IEEE transactions on pattern analysis and machine intelligence* 40.6 (2017), pp. 1510–1517.

[100] Andrea Vedaldi and Brian Fulkerson. "Vlfeat: An Open and Portable Library of Computer Vision Algorithms". In: *Proceedings of the 18th ACM International Conference on Multimedia*. MM '10. Firenze, Italy: ACM, 2010, pp. 1469–1472. ISBN: 978-1-60558-933-6. DOI: 10.1145/1873951.1874249. URL: http://doi.acm.org/10.1145/1873951.1874249.

[101] Schmid C. Wang H. "Action Recognition with Improved Trajectories". In: LEAR, INRIA, France, 2013. URL: https://www.cv-foundation.org/openaccess/content_iccv_2013/papers/Wang_Action_Recognition_with_2013_ICCV_paper.pdf.

[102] H Wang et al. "Action recognition by Dense Trajectories Computer Vision and Pattern Recognition (CVPR)". In: *2011 IEEE Conference on, IEEE*, pp. 3169–3176.

[103] Limin Wang, Yu Qiao, and Xiaoou Tang. "Action recognition with trajectory-pooled deep-convolutional descriptors". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 4305–4314.

[104] Limin Wang et al. "Temporal segment networks: Towards good practices for deep action recognition". In: *European conference on computer vision*. Springer. 2016, pp. 20–36.

[105] S Waterhouse, D MacKay, and T Robinson. *Advances in Neural Information Processing Systems*. 1996.

[106] Junji Yamato, Jun Ohya, and Kenichiro Ishii. "Recognizing human action in time-sequential images using hidden markov model". In: *Proceedings 1992 IEEE Computer Society conference on computer vision and pattern recognition*. IEEE. 1992, pp. 379–385.

[107] Yan Ke and R. Sukthankar. "PCA-SIFT: a more distinctive representation for local image descriptors". In: *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.* Vol. 2. June 2004, pp. II–II. DOI: 10.1109/CVPR.2004.1315206.

[108] Li Yao et al. "Describing videos by exploiting temporal structure". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 4507–4515.

[109] Matthew D Zeiler and Rob Fergus. "Visualizing and understanding convolutional networks". In: *European conference on computer vision*. Springer. 2014, pp. 818–833.

[110] Fan Zhu et al. "From handcrafted to learned representations for human action recognition: A survey". In: *Image and Vision Computing* 55 (2016), pp. 42–52.

[111] Yi Zhu et al. "Hidden two-stream convolutional networks for action recognition". In: *Asian Conference on Computer Vision*. Springer. 2018, pp. 363–378.

[112] Andrew Zisserman. "Return of the Devil in the Details: Delving Deep into Convolutional Nets". In: Visual Geometry Group, Department of Engineering Science, University of Oxford. URL: https://arxiv.org/pdf/1405.3531.pdf.

6.2.0.0.0.1