Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων

Τομεας Σηματων, Ελεγχου και Ρομποτικης

Αυτόματη Αναγνώριση Ανθρώπινων Δράσεων χρησιμοποιώντας Βαθιά Συνελικτικά Νευρωνικά Δίκτυα

Διπλωματικη Εργασια

του

ευαγγελου α. νικολουδακη

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Εργαστηριο Ορασης Υπολογιστων, Επικοινωνιας Λογου και Επεξεργασιας Σηματων

Αθήνα, Ιούλιος 2017

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής
Εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων

Αυτόματη Αναγνώριση Ανθρώπινων Δράσεων χρησιμοποιώντας Βαθιά Συνελικτικά Νευρωνικά Δίκτυα

Διπλωματικη Εργασια

του

ευαγγελου α. νικολουδακη

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 18η Ιουλίου 2017.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Πέτρος Μαραγκός	Κωνσταντίνος Τζαφέστας	Γεράσιμος Ποταμιάνος
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Αναπληρωτής Καθηγητής Παν/μίου Θεσσαλίας

Αθήνα, Ιούλιος 2017

(Υπογραφή)

.....

Ευαγγελος Νικολουδακης

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

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής
Εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων

Ευχαριστίες

Θα ήθελα καταρχήν να ευχαριστήσω τον καθηγητή κ. Πέτρο Μαραγκό για την επίβλεψη της παρούσας διπλωματικής εργασίας και για την ευκαιρία που μου έδωσε να την εκπονήσω στο εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων. Ο χρόνος του και η καθοδήγησή του στα πλαίσια της παρούσας εργασίας αλλά και κατά τη διδασκαλία των μαθημάτων της Όρασης Υπολογιστών και της Αναγνώρισης Προτύπων μου έδωσαν την ευκαιρία να γνωρίσω αυτούς τους τομείς της επιστήμης. Επίσης, ευχαριστώ θερμά τον Δρ. Βασίλη Πιτσικάλη για την καθοδήγησή του και τη συμβολή του στην πρόοδο της έρευνας. Ιδιαίτερη αναφορά θα ήθελα να κάνω στον Πέτρο Κούτρα για τις συμβουλές, το ενδιαφέρον του και τη γόνιμη συνεργασία που είχαμε. Θα ήθελα ακόμη να ευχαριστήσω τη Νάνσυ Ζλατίντση και τον Ισίδωρο Ροδομαγουλάκη για τη βοήθεια τους σε οτιδήποτε χρειάστηκε σε αυτή την όμορφη εμπειρία. Τέλος, θα ήθελα να ευχαριστήσω τους γονείς μου για την καθοδήγηση και την ηθική συμπαράσταση που μου προσέφεραν όλα αυτά τα χρόνια.

Περίληψη

Η παρούσα Διπλωματική Εργασία πραγματεύεται το πρόβλημα της αυτόματης αναγνώρισης ανθρώπινων δράσεων, στο πλαίσιο της αλληλεπίδρασης ανθρώπου-μηχανής. Για την εξαγωγή χαρακτηριστικών χρησιμοποιούνται τόσο hand-crafted τεχνικές όσο και τεχνικές βαθιάς μάθησης. Συγκεκριμένα, δίνεται έμφαση στην εφαρμογή των βελτιωμένων πυκνών τροχιών με Bag-of-Words κωδικοποίηση καθώς και στην χρήση Τρισδιάστατων Συνελικτικών Νευρωνικών Δικτύων (3D ConvNets) και Συνελικτικών Νευρωνικών Δικτύων Διπλής Ροής (Two-Stream ConvNets), από τα οποία εξάγουμε χαρακτηριστικά βαθιάς μάθησης. Πειραματιζόμαστε χρησιμοποιώντας τις διαθέσιμες αναπαραστάσεις βίντεο, αξιολογώντας την επίδοση τους σε μία σειρά από δημοφιλείς βάσεις δεδομένων, καθώς και στη βάση ανθρώπινων δράσεων Cognimuse, η οποία δημιουργήθηκε στο ερευνητικό πλαίσιο της παρούσας εργασίας και παρουσιάζει ιδιαίτερες προκλήσεις. Για την ταξινόμηση των δράσεων χρησιμοποιούνται μη γραμμικές Μηχανές Διανυσμάτων Υποστήριξης με πυρήνα x^2, στο επίπεδο των οποίων εφαρμόζεται σύμμειξη των παραγόμενων αναπαράστασεων βίντεο, επιτυγχάνοντας state-of-the-art επίδοση στη βάση δεδομένων HMDB51 και πολύ υψηλή ακρίβεια αναγνώρισης στη βάση δεδομένων Hollywood2.

Λέξεις Κλειδιά

Αναγνώριση Ανθρώπινων Δράσεων, Αναγνώριση Ανθρώπινων Χειρονομιών, Αναπαράσταση Βίντεο, Πυκνές Τροχιές, 3Δ Συνελικτικά Νευρωνικά Δίκτυα, Συνελικτικά Νευρωνικά Δίκτυα Διπλής Ροής, C3D Χαρακτηριστικά, TSN Χαρακτηριστικά, Μηχανές Διανυσμάτων Υποστήριξης

Abstract

This thesis focuses on the problem of automatic human action recognition, within the scope of human-computer interaction. Features are extracted by using both hand-crafted and deep-learning techniques. Specifically, we use improved dense trajectories with Bag-of-Words encoding, alongside with deep-learned features extracted from 3D and Two-Stream Convolutional Neural Networks. We experiment with the produced video representations, which are evaluated on widely-used action databases and the challenging Cognimuse dataset, that constructed for the purpose of our research. A multi-class x^2 SVM is used for action classification, in which we apply a kernel fusion scheme for the available extracted features. Our approach outperforms the state-of-the-art performance on HMDB51 dataset and achieves very high recognition accuracy on Hollywood2 dataset.

Keywords

Human Action Recognition, Human Gesture Recognition, Video Representation, Dense Trajectories, Bag of Words, 3D Convolutional Neural Networks, Two-Stream Convolutional Neural Networks, C3D Features, TSN Features, Support Vector Machines

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Γενικά για την Όραση Υπολογιστών

Η Όραση Υπολογιστών αποτελεί έναν πολύ σημαντικό κλάδο της επιστήμης και της τεχνολογίας. Αναπτύσσεται ραγδαία τα τελευταία χρόνια με εφαρμογές που γίνονται ολοένα και περισσότερο αντιληπτές στην καθημερινή ζωή. Σκοπός της Όρασης Υπολογιστών είναι να αναπαράγει αλγοριθμικά την αίσθηση της όρασης σε μία μηχανή, συνήθως έναν ηλεκτρονικό υπολογιστή ή ένα ρομπότ. Η όραση, ως μία βασική αίσθηση, για τον άνθρωπο είναι μία διαδικασία εγγενής και αυτόματη, η οποία εξελίσσεται κατά την πάροδο της ζωής του, επιτρέποντας του, για παράδειγμα, να αναγνωρίζει συνεχώς περισσότερα αντικείμενα ή πρόσωπα. Αντλώντας έμπνευση από αυτή την ανθρώπινη διαδικασία, η Όραση Υπολογιστών προσπαθεί να βοηθήσει τις μηχανές να αντιληφθούν και να ερμηνεύσουν τον κόσμο μέσω οπτικών ερεθισμάτων, με παρόμοιο τρόπο όπως ο άνθρωπος.

Πιο συγκεκριμένα, η Όραση Υπολογιστών σχετίζεται με τη θεωρία και τις μεθόδους που αφορούν στην εξαγωγή πληροφοριών μέσω της ανάλυσης οπτικών δεδομένων. Τα δεδομένα αυτά μπορούν να έχουν τη μορφή μιας απλής, έγχρωμης ή μη, ψηφιακής εικόνας, μίας ακολουθίας εικόνων (βίντεο), μίας εικόνας βάθους, εικόνων που περιέχουν πολλαπλές όψεις του ίδιου αντικειμένου, ενός πολυδιάστατου βιοϊατρικού σήματος κ.α. Για τη σχεδίαση και την ανάπτυξη αλγορίθμων και συστημάτων που λαμβάνουν και αναλύουν την οπτική πληροφορία, η Όραση Υπολογιστών δανείζεται και συνδυάζει έννοιες και μεθόδους από διαφορετικούς τομείς της επιστήμης όπως η επεξεργασία σημάτων, η μηχανική μάθηση, τα εφαρμοσμένα μαθηματικά, η φυσική, η νευροβιολογία και ο αυτόματος έλεγχος. Μερικές πρακτικές εφαρμογές της Όρασης Υπολογιστών είναι:

- Αλληλεπίδραση Ανθρώπου-Μηχανής, μέσω συστημάτων αναγνώρισης δράσεων ή χειρονομιών.

- Ταξινόμηση και αναγνώριση λέξεων, χαρακτήρων, συμβόλων σε έγγραφα.
- Αναγνώριση αντικειμένων σε βίντεο από κάμερες παρακολούθησης.
- Βιοϊατρικές εφαρμογές, όπως η βελτίωση των MRI εικόνων και η ανάπτυξη συστημάτων αυτόματης διάγνωσης από παρόμοια βιοϊατρικά δεδομένα.

1.2 Το πρόβλημα της αναγνώρισης ανθρώπινων δράσεων και χειρονομιών

Το πρόβλημα της Αναγνώρισης Ανθρώπινων Δράσεων (Human Action Recognition) αποτελεί ένα πολύ σημαντικό ερευνητικό πεδίο, το οποίο γνωρίζει αλματώδη πρόοδο τα τελευταία χρόνια. Σκοπός του προβλήματος είναι η ανάπτυξη αλγορίθμων για την αυτόματη επεξεργασία της οπτικής πληροφορίας και την αναγνώριση των δράσεων που πραγματοποιούνται από έναν ή περισσότερους ανθρώπους, με τελικό αποτέλεσμα την «κατανόηση» από τη μηχανή του τι είναι αυτό που «βλέπει», την «αντίληψη» του τι πραγματοποιείται και τι σημαίνει. Οι λέξεις «κατανόηση» και «αντίληψη» υποδεικνύουν ότι στόχος του προβλήματος είναι η προσομοίωση από τη μηχανή μέρους των δυνατοτήτων του ανθρώπινου εγκέφαλου, με στόχο την επικοινωνία ανθρώπου-μηχανής.

Μία δράση μπορεί να οριστεί με διάφορους τρόπους ανάλογα με το πως πραγματοποιείται, που αναφέρεται και πως αλληλεπιδρά, ενδεχομένως, με το περιβάλλον. Στη βιβλιογραφία, η πιο δημοφιλής προσέγγιση [3] κατηγοριοποιεί τις ανθρώπινες δράσεις σε μια ιεραρχία τριών επιπέδων: α) τις θεμελειώδεις δράσεις (action primitives), β) τις δράσεις (actions) και γ) τις δραστηριότητες (activities). Ως *θεμελιώδης δράση* ορίζεται μία απλή, μεμονωμένη κίνηση που πραγματοποιείται συνήθως από κάποιο μέλος του ανθρώπινου σώματος, όπως π.χ. η κίνηση ενός χεριού προς μία κατεύθυνση. Μία *δράση* είναι μια ακολουθία από διαδοχικές θεμελειώδεις δράσεις, όπως π.χ. το βούρτσισμα των μαλλιών που περιλαμβάνει διαδοχικές κινήσεις των χεριών. Μία *δραστηριότητα* περιλαμβάνει πολλαπλές δράσεις και είναι ένα γενικότερο γεγονός το οποίο σχετίζεται με το περιβάλλον, τα αντικείμενα και τους ανθρώπους που αλληλεπιδρούν στις δράσεις αυτές. Ένα παράδειγμα είναι η δραστηριότητα «παίζω τένις» που αποτελείται από δράσεις όπως «τρέχω», «χτυπάω το μπαλάκι», «κάνω σερβίς» κ.α. Η πλειοψηφία των μεθόδων που έχουν αναπτυχθεί καθώς και η εν εξελίξει ερευνητική δραστηριότητα επικεντρώνεται κυρίως στις δράσεις του δεύτερου επιπέδου, όπως τις καταγράψαμε παραπάνω, καθώς αφ΄ ενός δεν είναι τόσο πολύπλοκες όσο μία γενικού χαρακτήρα *δραστηριότητα* και αφ΄ ετέρου περιγράφουν μία πιο ολοκληρωμένη και ξεχωριστή σημασιολογικά κίνηση σε σχέση με μία απλή *θεμελιώδη δράση*.

Οι χειρονομίες αποτελούν ένα υποσύνολο των *δράσεων* που περιγράψαμε παραπάνω και, αν τις ορίσουμε αυστηρά, αφορούν αποκλειστικά συγκεκριμένες κινήσεις των

χεριών, που σκοπό έχουν να μεταδώσουν ένα νόημα. Παρ΄ όλα αυτά, στην πράξη, οι χειρονομίες μπορούν να περιλαμβάνουν ταυτόχρονα κινήσεις του κεφαλιού, μορφασμούς του προσώπου ή ακόμα και μετατοπίσεις του υπόλοιπου σώματος [4]. Οι χειρονομίες έχουν ιδιαίτερο ενδιαφέρον ερευνητικά, καθώς τις χρησιμοποιούμε καθημερινά στην επικοινωνία μας είτε για να συνοδέψουμε το λόγο μας είτε για να μεταδώσουμε ένα μήνυμα αποκλειστικά μέσω αυτών. Κατ΄ επέκταση, είναι ένας ελπιδοφόρος τρόπος για να επικοινωνήσουμε με μία μηχανή που εκμεταλλεύεται την οπτική πληροφορία. Χαρακτηριστικό παράδειγμα χειρονομιών είναι η νοηματική γλώσσα, η οποία περιέχει καθορισμένες κινήσεις των χεριών, με συγκεκριμένη αρχή και τέλος, όπου η καθεμία έχει και ένα διαφορετικό νόημα.

1.2.1 Περιγραφή και προκλήσεις του προβλήματος της αναγνώρισης ανθρώπινων δράσεων

Όπως περιγράψαμε παραπάνω, το πρόβλημα της αναγνώρισης ανθρώπινων δράσεων απαιτεί την επεξεργασία οπτικής πληροφορίας, η οποία είναι ουσιαστικά η είσοδος στο σύστημα μας. Η πληροφορία αυτή μπορεί να είναι ένα αποθηκευμένο βίντεο που περιλαμβάνει μία δράση, όπως π.χ. μία απομονωμένη σκηνή από μία ταινία, ή συνεχόμενα καρέ που λαμβάνονται σε πραγματικό χρόνο, για παράδειγμα από μία κάμερα παρακολούθησης. Τα δεδομένα αυτά μπορεί ενδεχομένως να συνοδεύονται από κάποια επιπλέον πληροφορία όπως τον «χάρτη βάθους» (depth map) ή τις συντεταγμένες του χωροχρονικού εντοπισμού μίας δράσης (action localization).

Στη βιβλιογραφία, η αναγνώριση ανθρώπινων δράσεων ταυτίζεται, συνήθως, με την *ταξινόμηση* ενός βίντεο που περιέχει μία δράση σε μία προκαθορισμένη κατηγορία (action classification). Παραδείγματος χάριν, αν σε ένα βίντεο υπάρχει ένας άνθρωπος ο οποίος τρέχει, ο αλγόριθμος πρέπει να ταξινομήσει το βίντεο αυτό στην κατηγορία «τρέχω». Τυπικά, ωστόσο, το πρόβλημα της αναγνώρισης δράσεων (action recognition) συνδυάζει αρχικά τον χωροχρονικό εντοπισμό μίας δράσης σε ένα βίντεο (action localization) και εν συνεχεία την ταξινόμηση της δράσης αυτής σε μία κλάση (action classification). Στο προηγούμενο παράδειγμα, δηλαδή, θα θέλαμε ο αλγόριθμος, επιπλέον, να εντοπίζει το που αρχίζει και που τελειώνει το τρέξιμο, για να θεωρήσουμε ότι γίνεται πραγματική «κατανόηση» της δράσης. Παρόλο που ο χωροχρονικός εντοπισμός μίας δράσης αποτελεί ένα συναφές και εξίσου σημαντικό πρόβλημα, απαιτεί ξεχωριστή έρευνα και επιβαρύνει με επιπλέον πολυπλοκότητα ένα ήδη δύσκολο πρόβλημα. Επομένως, στην παρούσα διπλωματική θα ακολουθήσουμε τη σύμβαση που επιτάσσει η βιβλιογραφία, η οποία συγχέει την *αναγνώριση* με την *ταξινόμηση* δράσεων.

Η πρόοδος που έχει σημειωθεί τα τελευταία χρόνια πάνω στην αυτόματη αναγνώριση ανθρώπινων δράσεων είναι ραγδαία. Ωστόσο, η πολυπλοκότητα των κινήσεων και

οι διαφορετικές συνθήκες που μπορεί να πραγματοποιούνται οι ανθρώπινες δράσεις καθιστούν το πρόβλημα αυτό εξαιρετικά δυσεπίλυτο, με δυσκολίες που εξακολουθούν να παραμένουν ανυπέρβλητα εμπόδια. Συνεπώς, η έρευνα για την ανάπτυξη ενός συστήματος αναγνώρισης ανθρώπινων δράσεων, έρχεται αντιμέτωπη με πολλές προκλήσεις [5], οι σημαντικότερες εκ των οποίων είναι οι ακόλουθες:

- **Μεταβολή της γωνίας λήψης**: Η γωνία λήψης είναι μάλλον η σημαντικότερη παράμετρος για ένα βίντεο. Η μορφή που έχει μία δράση και κατ΄ επέκταση η δυνατότητα αναγνώρισης της επηρεάζονται σε μεγάλο βαθμό από τη γωνία την οποία καταγράφεται. Σε διαφορετικές γωνίες λήψης η θέση και η πόζα των ανθρώπων που συμμετέχουν σε μία δράση διαφέρει και σημεία του σώματος τους μπορεί να είναι ή να μην είναι ορατά. Έτσι, τα διάφορα μοτίβα κινήσεων για μία δράση εμφανίζονται με διαφορετικό τρόπο ακόμη και σε μικρές αποκλίσεις της γωνίας λήψης, καθιστώντας την αναγνώριση της πολύ δύσκολη.

- **Πιθανές επικαλύψεις (occlusions)**: Σε ένα ιδανικό βίντεο μία δράση θα ήταν πλήρως ορατή και διαχωρίσιμη από το υπόλοιπο περιβάλλον. Στην πράξη, όμως, σε ένα δυναμικό, ρεαλιστικό περιβάλλον, όπως, για παράδειγμα, αυτό που εμφανίζεται στη σκηνή μίας ταινίας ή σε μία ακολουθία καρέ από μία κάμερα παρακολούθησης, είναι πολύ πιθανόν να παρατηρούνται επικαλύψεις ενός ανθρώπου με άλλους ανθρώπους ή με ένα αντικείμενο, με αποτέλεσμα να κρύβονται μέρη του ανθρώπινου σώματος που είναι σημαντικά για να αναγνωριστεί μία δράση.

- **Μεταβλητότητα εκτέλεσης μίας δράσης**: Ο τρόπος εκτέλεσης μίας δράσης μπορεί να διαφέρει σημαντικά από άτομο σε άτομο. Ο κάθε άνθρωπος διαθέτει ξεχωριστά μορφολογικά χαρακτηριστικά, όπως το μέγεθος του κορμού του ή οι αναλογίες των άκρων του, και μπορεί να πραγματοποιεί μία κίνηση με διαφορετική ταχύτητα ή με ελαφρώς διαφοροποιημένο μοτίβο από έναν άλλον άνθρωπο ή ακόμα και σε διαφορετικές επαναλήψεις της ίδιας κίνησης από τον ίδιο.

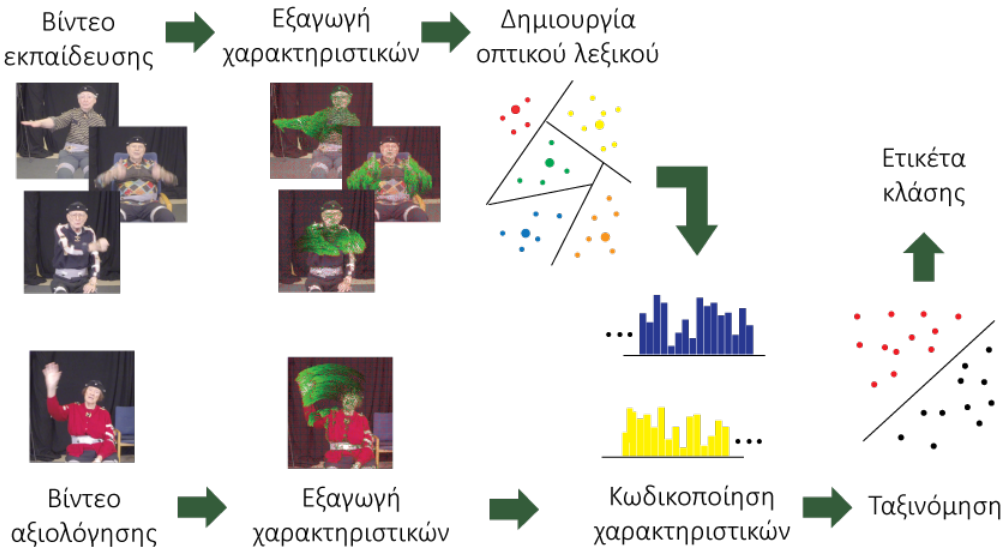
- **Κίνηση της κάμερας**: Σε ένα βίντεο είναι αρκετά συνηθισμένο να μεταβάλλονται τόσο η γωνία λήψης όσο και η απόσταση της κάμερας κατά τη διάρκεια π.χ. περιστροφή της κάμερας, ζουμ κ.α. Έτσι, συχνά εμφανίζονται κάποια παραπλανητικά μοτίβα κινήσεων τα οποία οδηγούν σε μία περίπλοκη συνισταμένη κίνηση από την οποία είναι δύσκολο να διαχωριστεί η δράση που αντιστοιχεί στον άνθρωπο.

- **«Ακατάστατο» (cluttered) φόντο**: Το δυναμικό, «ακατάστατο» φόντο είναι μία μορφή περισπασμού από την δράση που πραγματοποιείται καθώς προσθέτει ασαφή πληροφορία. Η εξαγωγή χαρακτηριστικών οπτικής ροής «ανιχνεύει» την

ανεπιθύμητη κίνηση του φόντου και την συγχέει με την κίνηση που πραγματοποιείται στο προσκήνιο με αποτέλεσμα να καθιστά πολύ δύσκολη την απομόνωση της δράσης που πραγματοποιείται.

1.2.2 Περιγραφή ενός συστήματος αναγνώρισης ανθρώπινων δράσεων

Τα συστήματα αναγνώρισης ανθρώπινων δράσεων, όπως συναντώνται στη διεθνή βιβλιογραφία, αποτελούνται, στην πλειονότητα τους, από κάποια κοινά επιμέρους υποσυστήματα, όπως φαίνεται στο Σχήμα 1.1. Τα υπό εξέταση δεδομένα χωρίζονται σε δύο υποσύνολα, ένα υποσύνολο εκπαίδευσης και ένα υποσύνολο αξιολόγησης. Κατά κανόνα, το υποσύνολο εκπαίδευσης πρέπει να είναι σημαντικά μεγαλύτερο και συνήθως αποτελείται από περίπου το 70-80% των δεδομένων, ενώ το υπόλοιπο 20-30% διαμορφώνει το υποσύνολο αξιολόγησης. Παρακάτω, περιγράφουμε συνοπτικά το ρόλο καθενός από τους επιμέρους κόμβους του συστήματος.



Σχήμα 1.1: Τα επιμέρους στάδια ενός συστήματος ταξινόμησης ανθρώπινων δράσεων. Σχήμα από [1]

Εξαγωγή χαρακτηριστικών

Η εξαγωγή χαρακτηριστικών αποτελεί το πρώτο και σημαντικότερο στάδιο ενός συστήματος αναγνώρισης ανθρώπινων δράσεων. Πραγματοποιείται από μία μαθηματική μέθοδο που επεξεργάζεται την 3Δ πληροφορία ενός βίντεο, την οποία «μετατρέπει» σε μία άλλη μαθηματική οντότητα, μικρότερης διάστασης, π.χ. διάνυσμα ή 2Δ πίνακας, η

οποία ιδανικά συμπυκνώνει το σημασιολογικό περιεχόμενο και περιγράφει με μοναδικό τρόπο την αρχική 3Δ ακολουθία εικόνων. Με αυτόν τον τρόπο απορρίπτεται η περιττή πληροφορία και διευκολύνονται τα επόμενα στάδια του συστήματος αφού επιτυγχάνεται μείωση της διάστασης του αρχικού πολύπλοκου σήματος του βίντεο. Ένα παράδειγμα χαρακτηριστικών, ευρέως διαδεδομένων, είναι τα χωροχρονικά σημεία ενδιαφέροντος (STIPs) (βλ. Ενότητα 3.1.1).

Τις περισσότερες φορές, η εξαγωγή χαρακτηριστικών απαιτεί την εν συνεχεία «περιγραφή» τους, η οποία γίνεται με διάφορους τρόπους, όπως π.χ. η κατασκευή ιστογραμμάτων με τις τιμές των χαρακτηριστικών. Οι «περιγραφητές» (descriptors), όπως ονομάζονται, υπολογίζονται συνήθως σε μία γειτονιά χαρακτηριστικών, με ενδεικτικά παραδείγματα τους HoG, HoF και MBH. Στη βιβλιογραφία, ο όρος εξαγωγή χαρακτηριστικών ταυτίζεται τόσο με τον υπολογισμό των χαρακτηριστικών όσο και με την «περιγραφή» τους. Στην παρούσα Διπλωματική Εργασία θα θεωρούμε αποτέλεσμα της εξαγωγής χαρακτηριστικών τους «περιγραφητές», ακολουθώντας τις επιταγές της βιβλιογραφίας, ενώ όταν απαιτείται θα διαχωρίζουμε τις δύο έννοιες.

Στην πλειονότητα των περιπτώσεων, οι «περιγραφητές» που εξάγονται έχουν μεγάλη διάσταση και συσχέτιση μεταξύ τους, οπότε ένα επιπλέον βήμα απαιτείται για τη μείωση της διάστασης και τη διαχωρισιμότητα τους προκειμένου να αφαιρεθεί η περιττή πληροφορία και να διευκολυνθεί η περαιτέρω επεξεργασία τους στα επόμενα στάδια. Συνηθισμένες μέθοδοι για την πραγματοποίηση αυτού του βήματος είναι η Ανάλυση σε Πρωτεύουσες Συνιστώσες (Principal Component Analysis, PCA), η Γραμμική Διακριτική Ανάλυση (Linear Discriminant Analysis, LDA) και η Ανάλυση Ανεξάρτητων Συνιστωσών (Independent Component Analysis, ICA).

Δημιουργία Οπτικού Λεξικού

Σε αυτό το στάδιο πραγματοποιείται ο χωρισμός του χώρου των χαρακτηριστικών σε ομάδες που αντιπροσωπεύουν κοινή πληροφορία από το βίντεο. Οι ομάδες αυτές αποτελούν το «λεξικό» σύμφωνα με το οποίο θα γίνει μετέπειτα η «μετάφραση» των χαρακτηριστικών στο στάδιο της κωδικοποίησης. Για τον υπολογισμό του λεξικού χρησιμοποιείται ο αλγόριθμος K-means ή η μοντελοποίηση μέσω της εκπαίδευσης ενός μίγματος Γκαουσιανών Κατανομών (Gaussian Mixture Model).

Κωδικοποίηση Χαρακτηριστικών

Έπειτα από την εξαγωγή τους στο αρχικό τμήμα του συστήματος, τα χαρακτηριστικά δεν είναι συνήθως σε αξιοποιήσιμη μορφή για την ταξινόμηση του βίντεο, καθώς έχουν συνήθως διαφορετικές αναπαραστάσεις (π.χ. διαφορετική διάσταση) που δεν είναι συγκρίσιμες μεταξύ τους. Για αυτό το λόγο, στο στάδιο της κωδικοποίησης τους υπολογίζεται μία ενιαία αναπαράσταση των χαρακτηριστικών, η οποία μπορεί να

αξιοποιηθεί από κάποιον ταξινομητή και έχει συνήθως τη μορφή ενός ιστογράμματος στατιστικών συχνοτήτων με βάση το οπτικό λεξικό, όπως π.χ. στην περίπτωση της Bag-of-Words κωδικοποίησης (βλ. Ενότητα 3.1.4)

Ταξινόμηση

Στο τελικό στάδιο γίνεται η ταξινόμηση του βίντεο εισόδου σε μία από τις προκαθορισμένες κλάσεις. Χρησιμοποιείται ένας ταξινομητής ο οποίος εκπαιδεύεται πάνω στα δεδομένα του υποσυνόλου εκπαίδευσης. Συγκεκριμένα, με πρότερη γνώση των κατηγοριών που ανήκουν τα δεδομένα εκπαίδευσης, ακολουθώντας τα παραπάνω βήματα, εξάγονται τα κωδικοποιημένα χαρακτηριστικά τους και με βάση αυτά υπολογίζεται ένα μοντέλο για κάθε μία από τις κλάσεις. Εν συνεχεία, γίνεται μία εκτίμηση της αποτελεσματικότητας του ταξινομητή πάνω στα δεδομένα του υποσυνόλου αξιολόγησης, τα οποία είναι «άγνωστα» για αυτόν. Ειδικότερα, χρησιμοποιώντας τα εκπαιδευμένα μοντέλα ταξινόμησης, εξετάζεται κάθε «άγνωστο» βίντεο αξιολόγησης και κατατάσσεται αυτόματα σε μία κατηγορία. Το πόσο καλή είναι αυτή η ταξινόμηση εξαρτάται από όλα τα επιμέρους στάδια του συστήματος αναγνώρισης ανθρώπινων δράσεων, δηλαδή από την «ποιότητα» των δεδομένων, τις μεθόδους εξαγωγής και κωδικοποίησης χαρακτηριστικών αλλά και το είδος του ταξινομητή που χρησιμοποιείται. Χαρακτηριστικό παράδειγμα ταξινομητών είναι οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines, SVMs) (βλ. Ενότητα 3.3).

1.3 Διαθέσιμες βάσεις δεδομένων

Η ραγδαία εξέλιξη του πεδίου της αυτόματης αναγνώρισης ανθρώπινων δράσεων ανέδειξε την ανάγκη για πειραματισμούς σε μεγάλους όγκους δεδομένων και την επαλήθευση των αλγόριθμων και των διαφόρων μεθόδων που αναπτύσσονται, με βάση ένα κοινό μέτρο σύγκρισης. Για την κάλυψη της ανάγκης αυτής, δημιουργήθηκαν τα τελευταία χρόνια βάσεις δεδομένων οι οποίες αποτελούν κοινά σημεία αναφοράς για τις περισσότερες εργασίες και καλύπτουν ένα ευρύ φάσμα δυσκολιών και προκλήσεων. Παρακάτω περιγράφουμε αναλυτικά τις σημαντικότερες από αυτές, καθώς και μία επιπλέον που κατασκευάστηκε για τους σκοπούς της παρούσας Διπλωματικής Εργασίας.

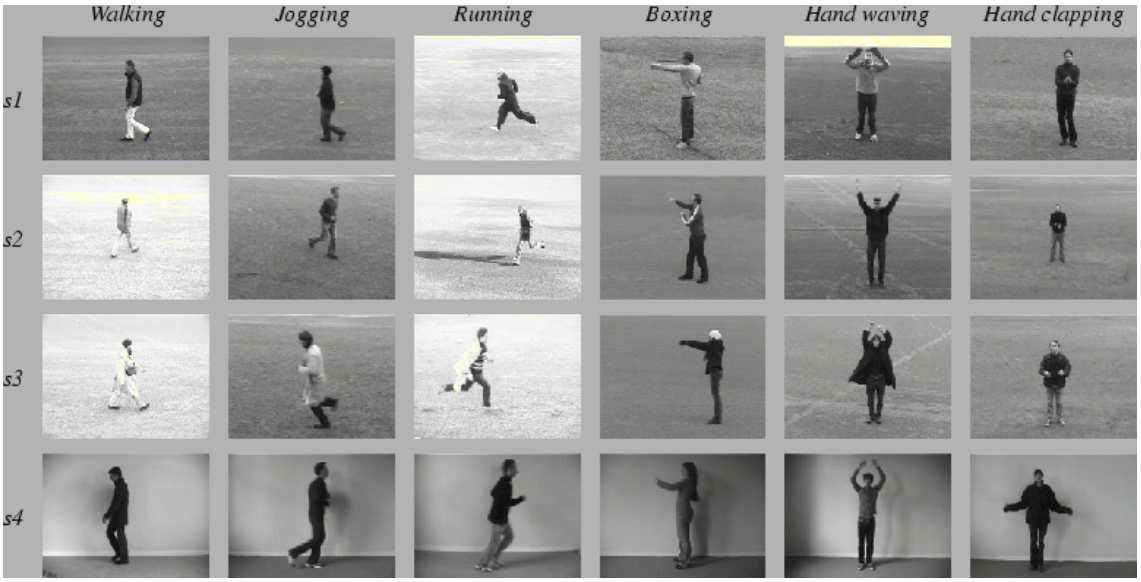
1.3.1 Η βάση δεδομένων KTH

Η βάση ανθρώπινων δράσεων KTH [6] είναι μία από τις πρώτες που κατασκευάστηκαν και παραμένει ακόμη και σήμερα αρκετά δημοφιλής. Τα βίντεο που περιλαμβάνει έχουν μαγνητοσκοπηθεί σε ομοιογενές περιβάλλον, ενώ η εκτέλεση των δράσεων γίνεται ξεκάθαρα, χωρίς επικαλύψεις ή μεγάλες μεταβολές σε διαφορετικές επαναλήψεις της

ίδιας δράσης. Λόγω της απλότητας της και της εύκολης επεξεργασίας των δεδομένων της, παραμένει μέχρι σήμερα σημείο αναφοράς για την αξιολόγηση των διαφόρων συστημάτων αναγνώρισης ανθρώπινων δράσεων και είναι συνήθως η πρώτη βάση δεδομένων που δοκιμάζεται όταν υπάρχει η ανάγκη να βγουν κάποια πρωταρχικά συμπεράσματα σε πρώιμο στάδιο.

Περιέχει 6 κατηγορίες ανθρώπινων δράσεων: *walking*, *jogging*, *running*, *boxing*, *hand waving* και *hand clapping*, οι οποίες πραγματοποιούνται αρκετές φορές από 25 διαφορετικά άτομα σε 4 διαφορετικά σενάρια: σε εξωτερικό χώρο *s1*, σε εξωτερικό χώρο με μεταβολή κλίμακας (ζουμ) *s2*, σε εξωτερικό χώρο με διαφορετικά ρούχα *s3* και σε εσωτερικό χώρο *s4*, όπως φαίνεται στο Σχήμα 1.2. Ο συνολικός αριθμός των βίντεο της βάσης είναι 2391. Οι λήψεις έχουν γίνει με στατική κάμερα και με ρυθμό 25 καρέ ανά δευτερόλεπτο (25 fps). Έπειτα έχει εφαρμοστεί μία υποδειγματοληψία ώστε όλα τα βίντεο να έχουν χωρική ανάλυση 160×120, ενώ η χρονική τους διάρκεια είναι, κατά μέσο όρο, περίπου 4 δευτερόλεπτα.

Όπως προτείνεται από τους δημιουργούς της βάσης, τα δεδομένα χωρίζονται σε σύνολο εκπαίδευσης και σύνολο αξιολόγησης. Το δεύτερο περιλαμβάνει τα βίντεο 9 ατόμων (2,3,5,6,7,8,9,10,22) ενώ το πρώτο τα βίντεο των υπόλοιπων 16 ατόμων. Οι ταξινομητές των διαφόρων μεθόδων εκπαιδεύονται πάνω στα δεδομένα των 16 ατόμων του συνόλου εκπαίδευσης και η ακρίβεια ταξινόμησης τους (classification accuracy) υπολογίζεται ως ο λόγος των σωστών ταξινομήσεων προς το σύνολο όλων των βίντεο των 9 ατόμων του συνόλου αξιολόγησης.



Σχήμα 1.2: Ενδεικτικά καρέ από τη βάση KTH στα 4 διαφορετικά σενάρια

1.3.2 Η βάση δεδομένων UCF101

Η βάση ανθρώπινων δράσεων UCF101 [7] αποτελείται από 13320 ρεαλιστικά βίντεο, που έχουν συλλεχθεί από το Youtube, καθένα από τα οποία ανήκει σε μία από 101 κατηγορίες δράσεων. Αποτελεί επέκταση της βάσης UCF50 [8] η οποία διαθέτει 50 κατηγορίες δράσεων. Ο ρεαλιστικός χαρακτήρας των δεδομένων αποτυπώνεται από τις μεγάλες μεταβολές στην κίνηση της κάμερας, στην εμφάνιση, την πόζα καθώς και στην κλίμακα των αντικειμένων, τις αλλαγές στη γωνία λήψης, το «ακατάστατο» φόντο, τις διαφορετικές συνθήκες φωτισμού κλπ. Ο μεγάλος όγκος και η φύση των δεδομένων σε συνδυασμό με την πολύ μεγάλη ποικιλία κινήσεων καθιστούν την UCF101 μία από τις πιο δύσκολες και απαιτητικές βάσεις ανθρώπινων δράσεων.

Τα βίντεο χωρίζονται σε 25 ομάδες, όπου καθεμία διαθέτει 4-7 βίντεο από κάθε κλάση δράσεων. Τα βίντεο της ίδιας ομάδας μπορεί να έχουν κοινά χαρακτηριστικά, όπως παρόμοιο φόντο, παραπλήσια γωνία λήψης κ.α. Οι κλάσεις ανθρώπινων δράσεων της UCF101, οι οποίες απεικονίζονται στο Σχήμα 1.3, προέρχονται από 5 ευρύτερες κατηγορίες:

1. Αλληλεπίδραση Ανθρώπου με Αντικείμενα: *Apply Eye MakeUp, Apply Lipstick, Blow Dry Hair, Brushing Teeth, Cutting in Kitchen, Hammering, Hula Hoop, Juggling Balls, Jump Rope, Knitting, Mixing Batter, Mopping Floor, Nun Chucks, Pizza Tossing, Shaving Beard, Skate Boarding, Soccer Juggling, Typing, Writing on Board, Yo Yo*

2. Γενικές Κινήσεις Σώματος: *Baby Crawling, Blowing Candles, Body Weight Squats, Handstand Pushups, Handstand Walking, Jumping Jack, Lunges, Pull Ups, Push Ups, Rock Climbing Indoor, Rope Climbing, Swing, Tai Chi, Trampoline Jumping, Walking with a Dog, Wall Push Ups*

3. Αλληλεπίδραση Ανθρώπου με άλλους Ανθρώπους: *Band Marching, Haircut, Head Massage, Military Parade, Salsa Spin*

4. Παίξιμο Μουσικού Οργάνου: *Drumming, Playing Cello, Playing Daf, Playing Dhol, Playing Flute, Playing Guitar, Playing Piano, Playing Sitar, Playing Tabla, Playing Violin*

5. Αθλητισμός: *Archery, Balance Beam, Baseball Pitch, Basketball, Basketball Dunk, Bench Press, Biking, Billiard, Bowling, Boxing-Punching Bag, Breaststroke, Clean and Jerk, Cliff Diving, Cricket Bowling, Cricket Shot, Diving, Fencing, Field Hockey Penalty, Floor Gymnastics, Frisbee Catch, Front Crawl, Golf Swing, Hammer Throw, High Jump, Horse Race, Horse Riding, Ice Dancing, Javelin Throw, Kayaking, Long Jump, Parallel Bars, Pole Vault,*

Pommel Horse, Punch, Rafting, Rowing, Shotput, Skiing, Jetski, Sky Diving, Soccer Penalty, Still Rings, Sumo Wrestling, Surfing, Table Tennis Shot, Tennis Swing, Throw Discus, Uneven Bars, Volleyball Spiking

Σχήμα 1.3: Ενδεικτικά καρέ από τις 101 κλάσεις της UCF101. Το διαφορετικό χρώμα στο περίγραμμα κάθε εικόνας υποδεικνύει την ευρύτερη κατηγορία στην οποία ανήκει

Όσον αφορά το πειραματικό πλαίσιο, οι δημιουργοί της UCF101, προτείνουν τρεις διαφορετικές διαμερίσεις (splits) των δεδομένων σε σύνολα εκπαίδευσης και αξιολόγησης. Κάθε split περιέχει περίπου 80-110 βίντεο εκπαίδευσης και 30-45 βίντεο αξιολόγησης από κάθε κλάση. Παρόλο που για κάθε split, ο συνολικός αριθμός των βίντεο εκπαίδευσης - και αξιολόγησης αντίστοιχα - διαφέρει ελαφρώς, παρατηρείται μια κατανομή των δεδομένων $\sim 72\%$ για εκπαίδευση και $\sim 28\%$ για αξιολόγηση. Κάθε

ταξινομητής μπορεί να εκπαιδευτεί και να αξιολογηθεί ξεχωριστά για κάθε split. Η συνολική ακρίβεια ταξινόμησης για τη βάση υπολογίζεται ως ο μέσος όρος των επιμέρους ακριβειών ταξινόμησης για κάθε split.

1.3.3 Η βάση δεδομένων Hollywood2

Η βάση ανθρώπινων δράσεων Hollywood2 [9] αποτελείται από 1707 βίντεο, προερχόμενα από 69 ταινίες του Hollywood, καθένα από τα οποία ανήκει σε μία από τις εξής 12 κατηγορίες δράσεων: *AnswerPhone, DriveCar, Eat, FightPerson, GetOutCar, HandShake, HugPerson, Kiss, Run, SitDown, SitUp, StandUp*. Στο Σχήμα 1.4 απεικονίζονται αντιπροσωπευτικά καρέ από κάθε κλάση.

Σχήμα 1.4: Ενδεικτικά καρέ από τις 12 κλάσεις της Hollywood2. Από αριστερά προς τα δεξιά, πρώτη σειρά: *GetOutCar, Run, SitUp, Drive Car, Eat, Kiss*, δεύτερη σειρά: *FightPerson, AnswerPhone, StandUp, SitDown, HandShake, HugPerson*

Για την κατασκευή της συγκεκριμένης βάσης οι δημιουργοί της συνέλεξαν τα βίντεο με αυτόματο τρόπο, αντιστοιχίζοντας το κείμενο του σεναρίου με την οπτικοακουστική πληροφορία κάθε ταινίας, αφού είχε προηγηθεί μία ταξινόμηση των δράσεων βασισμένη αποκλειστικά στο σενάριο (text-based script classification) [10]. Σε μία ταινία, η γωνία λήψης και ο φωτισμός μεταβάλλονται συνεχώς, ενώ έντονες είναι και οι εναλλαγές στο φόντο και στις κινήσεις της κάμερας προκαλώντας συνεχώς επικαλύψεις. Συνεπώς, οι δράσεις που περιλαμβάνουν τα βίντεο της Hollywood2 παρουσιάζουν μεγάλη μεταβλητότητα ως προς την εκτέλεση αλλά και το περιβάλλον τους, ενώ, λόγω του τρόπου εξαγωγής τους, σε πολλές περιπτώσεις τα κλιπ είναι αρκετά μεγαλύτερα σε διάρκεια από την δράση αυτή καθ΄ αυτή που περιέχουν. Όπως γίνεται αντιληπτό, η φύση των δεδομένων της Hollywood2 την καθιστά μία από τις πιο δύσκολες και πολύπλοκες βάσεις ανθρώπινων δράσεων που υπάρχουν στη Βιβλιογραφία.

Για τη διεξαγωγή πειραμάτων στη συγκεκριμένη βάση, προτείνεται από τους δημιουργούς της ο διαχωρισμός σε δύο σχεδόν ισοσκελισμένα σύνολα εκπαίδευσης και αξιολόγησης, τα κλιπ των οποίων προέρχονται από διαφορετικές ταινίες. Ο διαχωρισμός αυτός είναι σπάνιος καθώς το σύνολο αξιολόγησης (884 βίντεο) είναι ελαφρώς μεγαλύτερο από το σύνολο εκπαίδευσης (823 βίντεο). Για τη μέτρηση της συνολικής

ακρίβειας ταξινόμησης, ακολουθείται μία διαφορετική διαδικασία, αφού ορισμένα βίντεο περιλαμβάνουν παραπάνω από μία ανθρώπινες δράσεις. Συγκεκριμένα, αρχικά υπολογίζεται η μέση ακρίβεια ταξινόμησης για κάθε κλάση δράσεων (Average Precision - AP) και στη συνέχεια ο μέσος όρος των AP όλων των κλάσεων (mean Average Precision - mAP).

1.3.4 Η βάση δεδομένων HMDB51

Η βάση ανθρώπινων δράσεων HMDB51 [11] περιέχει συνολικά 6766 βίντεο, καθένα από τα οποία ανήκει σε μία από 51 διαφορετικές κατηγορίες δράσεων. Σε κάθε κατηγορία αντιστοιχούν τουλάχιστον 101 βίντεο. Η συλλογή τους έχει γίνει κυρίως από ταινίες και ένα μικρό ποσοστό από διαδικτυακές βάσεις δεδομένων όπως το Perlinger, το Youtube και τα Google videos. Ο ρυθμός δειγματοληψίας τους είναι ίσος με 30 καρέ ανά δευτερόλεπτο, ενώ κάθε βίντεο έχει υποστεί μία προ-επεξεργασία ώστε το ύψος των καρέ να είναι ίσο με 240 pixels με σταθερή, όμως, την αναλογία των διαστάσεων του (aspect ratio). Οι κλάσεις δράσεων της HMDB51, οι οποίες απεικονίζονται στο Σχήμα 1.5, προέρχονται από 5 ευρύτερες κατηγορίες:

1. Γενικές Κινήσεις Προσώπου: *smile, laugh, chew, talk*

2. Κινήσεις Προσώπου Αλληλεπιδρούμενες με Αντικείμενα: *smoke, eat, drink*

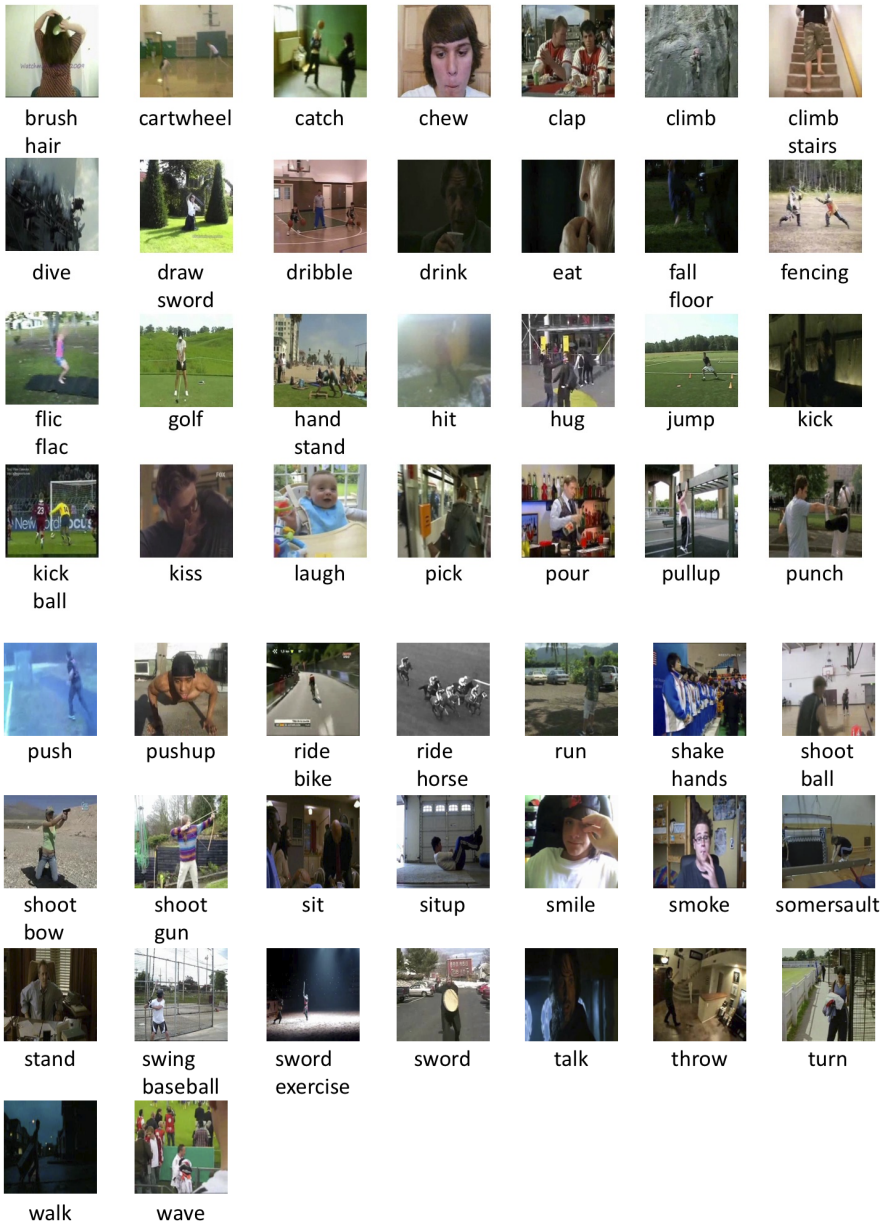
3. Γενικές Κινήσεις Σώματος: *cartwheel, clap hands, climb, climb stairs, dive, fall on the floor, backhand flip, handstand, jump, pull up, push up, run, sit down, sit up, somersault, stand up, turn, walk, wave*

4. Κινήσεις Σώματος Αλληλεπιδρούμενες με Αντικείμενα: *brush hair, catch, draw sword, dribble, golf, hit something, kick ball, pick, pour, push something, ride bike, ride horse, shoot ball, shoot bow, shoot gun, swing baseball bat, sword exercise, throw*

5. Κινήσεις Σώματος Αλληλεπιδρούμενες με Ανθρώπους: *fencing, hug, kick someone, kiss, punch, shake hands, sword fight*

Η HMDB51 ανήκει στην κατηγορία των δύσκολων και πολύπλοκων βάσεων ανθρώπινων δράσεων, όπως η UCF101 και η Hollywood2 που αναλύσαμε παραπάνω. Στα δεδομένα της παρουσιάζονται συχνά επικαλύψεις, απότομες κινήσεις της κάμερας και «ακατάστατο» φόντο. Ο μεγάλος αριθμός των κλάσεων της προσθέτει επιπλέον πολυπλοκότητα και την καθιστά ιδιαίτερα απαιτητική.

Όσον αφορά την αξιολόγηση της, ακολουθείται ένα πειραματικό μοτίβο όμοιο με αυτό της UCF101. Σύμφωνα με αυτό, τα δεδομένα χωρίζονται σε τρία splits, όπου



Σχήμα 1.5: Ενδεικτικά καρέ από τις 51 κλάσεις της HMDB51

καθένα περιέχει 70 βίντεο εκπαίδευσης και 30 βίντεο αξιολόγησης από κάθε κλάση. Ένας ταξινομητής μπορεί να εκπαιδευτεί και να αξιολογηθεί ξεχωριστά για κάθε split. Η συνολική ακρίβεια ταξινόμησης για τη βάση υπολογίζεται ως ο μέσος όρος των επιμέρους ακριβειών ταξινόμησης για κάθε split.

1.3.5 Η βάση δεδομένων Cognimuse

Η βάση ταινιών Cognimuse [12] αποτελείται από τριαντάλεπτα αποσπάσματα επτά ταινιών, στα οποία έχει πραγματοποιηθεί, από τα μέλη του εργαστηρίου, εντοπισμός και επισημείωση (annotation) των ακουστικών και οπτικών γεγονότων. Οι επτά αυτές ταινίες είναι οι εξής:

- *Beautiful Mind (BMI)*: Βιογραφία, Δράμα – 2001
- *Chicago (CHI)*: Μιούζικαλ – 2002
- *Crash (CRA)*: Δράμα – 2004
- *The Departed (DEP)*: Δράμα, Περιπέτεια, Θρίλερ – 2006
- *Finding Nemo (FNE)*: Περιπέτεια, Κινούμενα Σχέδια – 2003
- *Gladiator (GLA)*: Δράμα – 2000
- *Lord of the Rings (LOR)*: Φαντασίας, Περιπέτεια – 2003

Ο ρυθμός δειγματοληψίας των ταινιών είναι 25 καρέ ανά δευτερόλεπτο. Όπως είναι αντιληπτό, το ενδιαφέρον της παρούσας Διπλωματικής Εργασίας επικεντρώνεται στα οπτικά γεγονότα (visual events) που διαδραματίζονται από ανθρώπους. Επομένως, δεν λαμβάνεται υπόψιν η ταινία κινουμένων σχεδίων *Finding Nemo*. Αντί αυτής, για λόγους ύπαρξης επαρκούς όγκου δεδομένων και διαφορετικής ποικιλομορφίας, συμπεριλαμβάνεται η ταινία *Gone with the Wind (GWW)*. Πρόκειται για ένα αισθηματικό δράμα του 1939, το οποίο λόγω της παλαιότητας του διαθέτει αρκετά χαμηλότερη ανάλυση βίντεο σε σχέση με τις προαναφερθείσες ταινίες, γεγονός που επιβάλλει μία ιδιαίτερη δυσκολία στην επεξεργασία της. Επίσης, ο ρυθμός δειγματοληψίας της είναι διαφορετικός και αντιστοιχεί σε 23,976 καρέ το δευτερόλεπτο, ενώ το απόσπασμα της ταινίας που χρησιμοποιείται είναι περίπου 1 ώρα και 40 λεπτά. Στον Πίνακα 1.1 αναγράφεται η διάρκεια σε λεπτά και καρέ των επτά ταινιών που συμπεριλαμβάνονται στη βάση.

Για την επισημείωση των οπτικών γεγονότων των ταινιών (visual event annotation) χρησιμοποιήθηκαν κλάσεις ανθρώπινων δράσεων παρόμοιες με αυτές άλλων αντίστοιχων βάσεων δεδομένων, όπως η HMDB51 και η Hollywood2. Οι κλάσεις αυτές είναι 63 και διαχωρίζονται σε 6 ευρύτερες κατηγορίες:

1. Γενικές Κινήσεις Προσώπου: *chew, cry, laugh, smile, talk*
2. Κινήσεις Προσώπου Αλληλεπιδρούμενες με Αντικείμενα: *eat, drink, smoke*

	Duration (min)	**Total Frames**
BMI	31:17	46937
CHI	30:08	45202
CRA	26:37	39926
DEP	30:28	45707
GLA	30:02	45062
LOR	37:33	56339
GWW	104:10	449568

Πίνακας 1.1: Διάρκεια των 7 ταινιών της Cognimuse σε λεπτά και καρέ

3. Γενικές Κινήσεις Σώματος: *cartwheel, clap hands, climb, climb stairs, dance, dive, fall on the floor, backhand flip, handstand, jump, pull up, push up, running, sitting down, sitting up, somersault, standing up, turn, walk*

4. Χειρονομίες: *pantomime, point at something, wave hands*

5. Κινήσεις Σώματος Αλληλεπιδρούμενες με Αντικείμενα: *answering phone, brush hair, catch, draw sword, dribble, driving car, getting out of the car, golf, hit something, kick ball, open car door, open door, pick, pour, push something, ride bike, ride horse, shoot ball, shoot bow, shoot gun, swing baseball bat, sword exercise, throw*

6. Κινήσεις Σώματος Αλληλεπιδρούμενες με Ανθρώπους: *fencing, fighting, grab hand, hugging, kick someone, kissing, punch, shake hands, sword fight, threaten person*

Προφανώς, σε μία ταινία δεν εμφανίζονται όλες οι παραπάνω κατηγορίες δράσεων, παρά μόνο μερικές από αυτές, ενώ κάποιες δράσεις εμφανίζονται ελάχιστα ή καθόλου στις επτά ταινίες της Cognimuse. Αντίστοιχα, σε κάθε ταινία εμφανίζονται ανθρώπινες δράσεις οι οποίες δεν ανήκουν σε καμία από τις 63 κλάσεις που λαμβάνονται υπόψη, οι οποίες πρέπει να εντοπιστούν και να απορριφθούν. Γι΄ αυτό το λόγο, στη διαδικασία της επισημείωσης των επτά ταινιών της Cognimuse, παράλληλα με τις 63 κλάσεις δράσεων, χρησιμοποιήθηκε και μία κλάση *other* για κάθε μία από τις 6 ευρύτερες κατηγορίες, προκειμένου να επισημειωθούν όλα τα οπτικά γεγονότα που δεν ανήκουν σε καμία από τις ενδιαφερόμενες κλάσεις. Στην πράξη, για τη διεξαγωγή πειραμάτων (βλ. Κεφάλαιο 4) υπολογίζονται μόνο οι κατηγορίες δράσεων που εμφανίζονται αρκετές φορές (π.χ. ≥ 30), αθροιστικά και στις επτά ταινίες, και είναι πολύ λιγότερες από 63.

Για τις ανάγκες της παρούσας Διπλωματικής Εργασίας, προκειμένου να κατασκευάσουμε μία βάση ανθρώπινων δράσεων χρησιμοποιώντας τη Cognimuse, για κάθε μία

από τις επτά ταινίες έγινε χρονική κατάτμηση σε κλιπς με βάση τις πληροφορίες από την επισημείωση των οπτικών γεγονότων. Στη συνέχεια, απορρίψαμε τα βίντεο που ανήκουν στις κλάσεις *other*, συλλέγοντας έτσι συνολικά 4323 βίντεο από όλες τις ταινίες, ενώ κάθε βίντεο υπέστη μία χωρική επεξεργασία για να διαμορφωθεί η ανάλυση του στις διαστάσεις 256×340. Κατανέμοντας τα βίντεο αυτά σε κάθε κλάση ξεχωριστά, παρατηρήθηκε μία σημαντική ιδιαιτερότητα: Τα κλιπς που ανήκουν στην κλάση *talk* είναι πολύ περισσότερα από κάθε άλλη κλάση και καλύπτουν σχεδόν το 1/2 των διαθέσιμων δεδομένων όπως φαίνεται στον Πίνακα 1.2.

	Total Clips	***talk* Clips**
BMI	491	236
CHI	531	258
CRA	382	147
DEP	526	267
GLA	367	109
LOR	319	115
GWW	1707	806
Total	4323	1931

Πίνακας 1.2: Το πλήθος των βίντεο για κάθε ταινία: Για όλες τις κλάσεις (δεύτερη στήλη) και για την κλάση *talk* ξεχωριστά (τρίτη στήλη)

Η ιδιαιτερότητα αυτή δημιουργεί πρόβλημα στην επεξεργασία των δεδομένων καθώς είναι φανερό ότι υπάρχει υπερ-πληροφορία για μία κατηγορία δράσεων η οποία εκμηδενίζει όλες τις υπόλοιπες. Μία λύση θα ήταν, ενδεχομένως, η διαλογή ενός μικρού αριθμού κλιπς από την κατηγορία *talk*, ανάλογου με το πλήθος των βίντεο των υπόλοιπων κλάσεων. Παρ΄ όλα αυτά μία τέτοια διαδικασία θα ήταν ιδιαιτέρως χρονοβόρα, ενώ θα υπήρχε ο κίνδυνος να χαθεί σημαντική πληροφορία από τη διαφορετικότητα των συνθηκών που πραγματοποιείται η δράση μέσα σε κάθε ταινία. Επίσης, καθότι η κατηγορία *talk* αναφέρεται στην κίνηση του προσώπου ενός ανθρώπου που μιλάει, πολλές σύγχρονες μέθοδοι ανίχνευσης και αναγνώρισης προσώπου είναι σε θέση να αναγνωρίζουν τη δράση αυτή με πολύ υψηλά ποσοστά επιτυχίας [13]. Επομένως, ερευνητικά, προκύπτει μεγαλύτερη ανάγκη για αναγνώριση άλλων μορφών ανθρώπινων δράσεων, πιο πολύπλοκων και πιο ολοκληρωμένων σημασιολογικά. Συνεπώς, για τους σκοπούς αυτής της Διπλωματικής Εργασίας, η κατηγορία ανθρώπινων δράσεων *talk* δεν συμπεριλαμβάνεται στον χώρο κλάσεων.

Τέλος, όπως προαναφέραμε κάποιες κατηγορίες ανθρώπινων δράσεων εμφανίζονται ελάχιστες φορές ή καθόλου στις 7 ταινίες της παρούσας βάσης, άρα για ευνόητους λόγους δεν λαμβάνονται υπόψη. Θεσπίζουμε ένα κατώφλι $N_{thres} = 30$ για το πλήθος

των βίντεο κάθε κλάσης και απορρίπτουμε όσες κατηγορίες δράσεων διαθέτουν αριθμό κλιπς $N < N_{thres}$, κρατώντας έτσι μόνο 20. Αυτές είναι οι εξής: *climb stairs, cry, dance, fall on the floor, grab hand, hugging, laugh, open door, pick, point at something, ride horse, running, sitting down, sitting up, smile, standing up, throw, turn, walk, wave hands*. Οι κλάσεις αυτές περιλαμβάνουν τουλάχιστον 30 βίντεο η καθεμία, τα οποία είναι αρκετά για την κατασκευή συνόλων εκπαίδευσης και αξιολόγησης. Έτσι, το πλήθος των βίντεο της βάσης Cognimuse μειώνεται σε 2238. Διαμορφώνονται, λοιπόν, κλάσεις που είναι ανισοκατανεμημένες. Υπάρχουν, δηλαδή, κάποιες που έχουν «λίγα» δεδομένα ($\sim 30-40$) και κάποιες που έχουν «πολλά» δεδομένα ($\sim 150-200$). Αυτό είναι ένα φαινόμενο που είναι λογικό να συμβεί όταν γίνεται μία επισημείωση όπως αυτή της Cognimuse, μιας και η συχνότητα εμφάνισης μίας ανθρώπινης δράσης εξαρτάται αποκλειστικά από το περιεχόμενο της κάθε ταινίας. Σε κάθε περίπτωση, όμως, αποφύγαμε μία ακραία μορφή ανισοκατανομής όπως αυτή που δημιουργούσε η κλάση *talk*.

Στο εξής, με τον όρο Cognimuse θα αναφερόμαστε στη βάση ανθρώπινων δράσεων που περιλαμβάνει 2238 βίντεο με ανάλυση 256×340, καθένα από τα οποία ανήκει σε μία από τις 20 κλάσεις που προαναφέρθηκαν. Η Cognimuse παρουσιάζει ιδιαίτερες προκλήσεις καθώς στα βίντεο της εμφανίζονται συχνά επικαλύψεις, απότομες κινήσεις της κάμερας, αλλαγές στο φωτισμό και πολύ «ακατάστατο» φόντο. Ακόμη, η μη περαιτέρω επεξεργασία των δεδομένων της και η άνιση κατανομή τους μέσα στις κλάσεις προσθέτει επιπλέον δυσκολία και την καθιστά πολύ απαιτητική.

1.4 Διάρθρωση της Διπλωματικής Εργασίας

Στην παρούσα Διπλωματική Εργασία εξετάζουμε το πρόβλημα της αναγνώρισης ανθρώπινων δράσεων και χειρονομιών. Αναλύουμε και συγκρίνουμε τις διάφορες προσεγγίσεις για την εξαγωγή χαρακτηριστικών και την ταξινόμηση, συνδυάζοντας τόσο παραδοσιακές τεχνικές Όρασης Υπολογιστών όσο και σύγχρονες αρχιτεκτονικές Νευρωνικών Δικτύων. Πιο συγκεκριμένα, στο Κεφάλαιο 2 συνοψίζεται η σχετική έρευνα που έχει πραγματοποιηθεί μέχρι σήμερα, περιγράφοντας περιληπτικά τις πιο αντιπροσωπευτικές μεθόδους, ενώ στο Κεφάλαιο 3 περιγράφονται λεπτομερώς οι μέθοδοι που χρησιμοποιήθηκαν στα πλαίσια της παρούσας εργασίας. Δίνουμε έμφαση στις τεχνικές εξαγωγής χαρακτηριστικών τις οποίες διακρίνουμε σε hand-crafted (Ενότητα 3.1) και τεχνικές βαθιάς μάθησης (Ενότητα 3.2). Ειδικότερα, στις Ενότητες 3.1.1 και 3.1.2 αναλύονται διεξοδικά δύο ευρέως χρησιμοποιούμενες hand-crafted τεχνικές εξαγωγής χαρακτηριστικών: τα χωρο-χρονικά σημεία ενδιαφέροντος και οι πυκνές τροχιές, ενώ στην Ενότητα 3.1.4 μελετούνται δύο πολύ δημοφιλείς αναπαραστάσεις βίντεο: το μοντέλο συνόλου οπτικών λέξεων (bag-of-words) και οι χωρο-χρονικές πυραμίδες.

Αντίστοιχα, στην Ενότητα 3.2.1 γίνεται μία εισαγωγή στα νευρωνικά δίκτυα και στην Ενότητα 3.2.2 περιγράφεται ο ρόλος και η λειτουργία κάθε επιπέδου ενός Συνελικτικού Νευρωνικού Δικτύου, η τοπολογία των οποίων αναπτύσσεται στην Ενότητα 3.2.3 μαζί με τη μελέτη ενδεικτικών αρχιτεκτονικών της βιβλιογραφίας. Η Ενότητα 3.3 αναφέρεται στην ταξινόμηση των βίντεο με Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines).

Επίσης, στο Κεφάλαιο 4 παρουσιάζονται πιο λεπτομερώς συγκεκριμένες μεθοδολογίες που χρησιμοποιήθηκαν και καταγράφονται τα πειραματικά αποτελέσματα που αυτές επέφεραν τόσο πάνω στις δημοφιλείς βάσεις ανθρώπινων δράσεων της βιβλιογραφίας όσο και πάνω στην Cognimuse. Στην Ενότητα 4.1 αξιολογούνται διάφορα σχήματα διαμερίσεων σε υποσύνολα εκπαίδευσης και αξιολόγησης της Cognimuse, ενώ στις Ενότητες 4.2 και 4.3 καταγράφονται αποτελέσματα από τη χρήση διάφορων τεχνικών εξαγωγής χαρακτηριστικών σε όλες τις διαθέσιμες βάσεις ανθρώπινων δράσεων. Στην Ενότητα 4.4 περιγράφεται αναλυτικά η προτεινόμενη μέθοδος αναγνώρισης ανθρώπινων δράσεων, η οποία δημιουργήθηκε βήμα προς βήμα στις προηγούμενες ενότητες και επιτυγχάνει υψηλά ποσοστά αναγνώρισης. Τέλος, στο Κεφάλαιο 5 στοιχειοθετείται η συνεισφορά της παρούσας Διπλωματικής Εργασίας καθώς και η προέκταση της που βρίσκεται εν εξελίξει, ενώ προτείνονται κατευθύνσεις για μελλοντική έρευνα.

Κεφάλαιο 2

Σχετική Βιβλιογραφία

Η βιβλιογραφία σχετικά με το πρόβλημα της αναγνώρισης ανθρώπινων δράσεων είναι ιδιαίτερα εκτενής και συνεχίζει να επεκτείνεται με ραγδαίους ρυθμούς. Τα τελευταία χρόνια, η σχετική έρευνα έχει επηρεαστεί σε μεγάλο βαθμό από το πεδίο της ταξινόμησης εικόνων. Πολύ σημαντικό ρόλο διαδραμάτισε η εμφάνιση των τοπικών (local) αναπαραστάσεων [14], οι οποίες περιγράφουν ένα βίντεο μέσω μίας συλλογής «σημαντικών» (salient) σημείων του, ή πιο πρόσφατα μέσω τροχιών [15]. Πρόσφατα, επίσης, ακολουθώντας τις τάσεις της Όρασης Υπολογιστών και της Μηχανικής Μάθησης, υιοθετούνται όλο και πιο συχνά βαθιές αρχιτεκτονικές νευρωνικών δικτύων [16, 17]. Μάλιστα, δεν θα ήταν υπερβολή να ισχυριστούμε ότι η πλειοψηφία των ερευνητών έχουν στρέψει την προσοχή τους αποκλειστικά στην χρήση και την εξέλιξη των βαθιών νευρωνικών δικτύων, για την αναγνώριση ανθρώπινων δράσεων.

Οι δημοσιευμένες εργασίες μπορούν να διακριθούν με διάφορους τρόπους, ανάλογα με τις τεχνικές τις οποίες αναπτύσσουν. Μία αρκετά συνηθισμένη κατηγοριοποίηση [18], την οποία ακολουθούμε και στο κεφάλαιο αυτό, διαχωρίζει τις μεθόδους σε *τοπικές (local)*, *ολικές (global)*, καθώς και παραλλαγές τους, οι οποίες με τη σειρά τους περιλαμβάνουν χωρικές, χρονικές και χωρο-χρονικές αναπαραστάσεις.

2.1 Τοπικές (Local) αναπαραστάσεις

Οι τοπικές αναπαραστάσεις περιγράφουν ένα βίντεο ως μία συλλογή τοπικών περιγραφητών ή περιοχών ενδιαφέροντος και είναι συνήθως ανεπηρέαστες από τη γωνία λήψης, το φόντο ή πιθανές επικαλύψεις. Οι περιοχές ενδιαφέροντος εξάγονται συνήθως γύρω από χωρο-χρονικά σημεία ενδιαφέροντος τα οποία θεωρούνται «σημαντικά» (salient) για μία δράση και μπορούν να την χαρακτηρίσουν. Τα σημεία αυτά αναφέρονται ως STIPs (Spatio-Temporal Interest Points) και η πρώτη απόπειρα εντοπισμού τους έγινε από τους Laptev et al. [14], οι οποίοι επέκτειναν τον ανιχνευτή γωνιών σε 2Δ εικόνες Harris [19] στον 3Δ χώρο (Harris3D), βρίσκοντας, έτσι, σημεία στα οποία

η γειτονιά τους αλλάζει απότομα τόσο στο χωρικό όσο και στο χρονικό πεδίο. Πρότειναν, μάλιστα, την αυτόματη επιλογή κλίμακας της χωρο-χρονικής γειτονιάς, την οποία ωστόσο εγκατέλειψαν αργότερα [6], χρησιμοποιώντας προκαθορισμένες κλίμακες για την ανίχνευση των σημείων ενδιαφέροντος, ενώ στην ίδια εργασία χρησιμοποίησαν την κωδικοποίηση Bag-of-Words και μη-γραμμικά SVM για την ταξινόμηση των βίντεο. Η εργασία αυτή επεκτάθηκε έπειτα [20] προκειμένου να συμπεριληφθεί η αντιστάθμιση της σχετικής κίνησης της κάμερας, ενώ ένα χρόνο αργότερα οι Laptev et al. [10] χρησιμοποίησαν τους περιγραφητές HoG [21] και HoF, τον οποίο ουσιαστικά εμπνεύστηκαν από την προσέγγιση των Laptev & Lindenberg [22], για να περιγράψουν τις γειτονιές γύρω από τα σημεία ενδιαφέροντος. Την επέκταση του HoG στον 3Δ χώρο (HoG3D) ανέπτυξαν λίγο πιο μετά οι Kläser et al. [23].

Παράλληλα, οι Dollár et al. [24] πρότειναν την εφαρμογή Gabor φίλτρων ξεχωριστά στη χωρική και τη χρονική διάσταση για την ανίχνευση σημείων ενδιαφέροντος, το πλήθος των οποίων διαμορφώνεται από το μέγεθος των διαστάσεων της γειτονιάς στην οποία εντοπίζονται τα τοπικά ελάχιστα. Για την περιγραφή των σημείων ενδιαφέροντος έφεραν στο προσκήνιο τα Cuboids, τα οποία αποτελούν ουσιαστικά 3Δ προκαθορισμένες περιοχές γύρω από κάθε σημείο, εντός των οποίων εξάγονται περιγραφητές όμοιοι με τον SIFT [25], δηλαδή ιστογράμματα των τιμών έντασης των pixels, της παραγώγου και της οπτικής ροής. Αργότερα, οι Scovanner et al. [26] υλοποίησαν την επέκταση του SIFT στον 3Δ χώρο (SIFT3D) για την αναγνώριση δράσεων.

Πιο πρόσφατα, οι Willems et al. [27] ταυτίζουν τη σημαντικότητα των σημείων ενδιαφέροντος με την ορίζουσα του 3Δ Hessian πίνακα ενός βίντεο, ενώ χρησιμοποιούν μία επέκταση του περιγραφητή SURF [28] στον χρόνο, τον eSURF, ο οποίος βασίζεται στον υπολογισμό κυματιδίων Haar εντός μιας ορθογώνιας περιοχής γύρω από το σημείο ενδιαφέροντος. Σε μία ακόμη πιο πρόσφατη εργασία, οι Wang et al. [29] συγκρίνουν διάφορους ανιχνευτές σημείων ενδιαφέροντος σε μια σειρά από βάσεις δεδομένων και καταλήγουν στο συμπέρασμα πως μια απλή *πυκνή δειγματοληψία (dense sampling)* σημείων σε ένα σταθερό χωρο-χρονικό πλέγμα μέσα στο βίντεο, υπερτερεί σημαντικά όλων των προηγούμενων. Επίσης, οι Maninis et al. [30], βασιζόμενοι στις ιδέες των Georgakis et al. [31], εφαρμόζουν στο σήμα βίντεο 3Δ Gabor φίλτρα σε πολλαπλές συχνότητες και χρησιμοποιούν τον τελεστή ενέργειας Teager-Kaiser για τη δημιουργία ενός ενεργειακού «χάρτη σημαντικότητας», όπου τα μέγιστα του αντιστοιχούν σε σημεία ενδιαφέροντος, στη γειτονιά των οποίων υπολογίζονται διάφοροι περιγραφητές.

Παρ΄ όλα αυτά, τα χωρο-χρονικά σημεία ενδιαφέροντος παρουσιάζουν ένα βασικό μειόνεκτημα το οποίο έγκειται στην αντιμετώπιση του χώρου και του χρόνου με ενιαίο τρόπο, χωρίς να ενσωματώνουν ουσιαστικά πληροφορία για την κίνηση. Εξελίσσοντας την ιδέα των STIPs, πολλές εργασίες προσπαθούν να ενσωματώσουν πληροφορία για

την κίνηση μέσα στο βίντεο, παρακολουθώντας τα σημεία ενδιαφέροντος στο χρόνο, σχηματίζοντας τις τροχιές τους οι οποίες περιγράφουν μία δράση. Πρώτοι οι Messing et al. [32] εξάγουν τροχιές από το βίντεο, εντοπίζοντας τα σημεία ενδιαφέροντος με τον ανιχνευτή Harris και παρακολουθώντας τα στο χρόνο με τον Kanade-Lucas-Tomasi (KLT) tracker. Ως χαρακτηριστικά του βίντεο εξάγουν τις σχετικές ταχύτητες των σημείων των τροχιών. Επίσης, οι Sun et al. [33] παρακολουθούν σημεία ενδιαφέροντος υπολογίζοντας την ομοιότητα μεταξύ διαδοχικών καρέ μέσω SIFT περιγραφητών (SIFT matching) και χρησιμοποιούν στατιστικά μεγέθη των τροχιών για την αναπαράσταση του βίντεο. Αργότερα, οι ίδοι [34] συνδυάζουν τις δύο παραπάνω μεθόδους και εξάγουν τροχιές μεγάλης διάρκειας, ενώ χρησιμοποιούν το μέτρο της τοπικής κλίσης και τη μεταβλητότητα της φωτεινότητας της εικόνας ως κριτήριο «σημαντικότητας» (saliency criterion), βάσει του οποίου ρυθμίζουν την πυκνότητα των σημείων που παρακολουθούνται. Συνδυάζοντας τις ιδέες της πυκνής δειγματοληψίας και της χρήσης τροχιών για την περιγραφή της κίνησης, οι Wang et al. [35] εξάγουν *πυκνές τροχιές (dense trajectories)*, οι οποίες παρακολουθούν τα πυκνά δειγματοληπτημένα σημεία μέσω ενός πυκνού πεδίου οπτικής ροής. Γύρω από τα σημεία παρακολούθησης εξάγονται διάφοροι περιγραφητές μεταξύ των οποίων και ο MBH [36]. Αργότερα, οι ίδιοι βελτίωσαν τη μέθοδο τους, εξάγοντας, όπως τις ονόμασαν, *βελτιωμένες τροχιές (improved trajectories)* [15], οι οποίες λαμβάνουν υπόψην την αντιστάθμιση της κίνησης της κάμερας για τον υπολογισμό της οπτικής ροής.

2.2 Ολικές (Global) αναπαραστάσεις

Οι ολικές αναπαραστάσεις περιγράφουν μία δράση χρησιμοποιώντας ολόκληρο το βίντεο, ή μία ολόκληρη περιοχή ενδιαφέροντος εντός αυτού. Συνήθως, εντοπίζεται η σιλουέτα του ανθρώπου που κινείται και χρησιμοποιείται η περιοχή που την περικλύει, ενώ επιπλέον πληροφορίες όπως οι ακμές ή η οπτική ροή συμβάλλουν στον υπολογισμό των περιγραφητών. Οι αναπαραστάσεις αυτές είναι συχνά ευαίσθητες στο θόρυβο, στις επικαλύψεις ή στις αλλαγές της γωνίας λήψης, γι΄ αυτό πολλές φορές η περιοχή ενδιαφέροντος διαιρείται σε ένα χωρο-χρονικό πλέγμα, εντός του οποίου υπολογίζονται τοπικά τμήματα τις τελικής αναπαράστασης.

Σε μια από τις πιο παλιές εργασίες, οι Darrell & Pentland [37] υπολογίζουν την ομοιότητα ανάμεσα σε βίντεο χειρονομιών μέσω της συσχέτισης ανάμεσα στα καρέ τους, χωρίς εξαγωγή χαρακτηριστικών. Αντίστοιχα, οι Bobick & Davis [38] απομονώνουν τη σιλουέτα του ανθρώπου σε κάθε καρέ, από μία συγκεκριμένη οπτική γωνία, και συγκεντρώνουν τις διαφορές ανάμεσα σε συνεχόμενα frames σε μία τελική εικόνα, σχηματίζοντας έτσι τις Motion Energy Image (MEI) και Motion History Image (MHI), που υποδεικνύουν τα pixels που εμφανίζεται η κίνηση, ενώ η σύγκριση τους

γίντεται με βάση τις Hu Moments τους [39]. Έπειτα, η ιδέα αυτή γενικεύεται στο 3Δ χώρο από τους Weiland et al. [40], με τον υπολογισμό των Motion Energy Volumes, τους οποίους συγκρίνουν χρησιμοποιώντας περιγραφητές βασισμένους στο Fourier μετασχηματισμό τους σε κυλινδρικές συντεταγμένες. Οι Gorelick et al. [41] και οι Yilmaz & Shah [42] θεωρούν το περίγραμμα του ανθρώπου ως ένα ενιαίο τρισδιάστατο σχήμα, το οποίο παρακολουθούν στο χρόνο και υπολογίζουν διάφορα χαρακτηριστικά. Όπως παρατηρούμε, όλες οι παραπάνω μέθοδοι εξάγουν στατικές αναπαραστάσεις και δεν κωδικοποιούν καθόλου την πληροφορία της κίνησης εντός του βίντεο. Αντίθετα, οι Efros et al. [43] υπολογίζουν την οπτική ροή σε «ανθρωποκεντρικές» εικόνες, χρησιμοποιώντας σκηνές από βίντεο αθλητικών αγώνων.

Επίσης, όσον αφορά την περιγραφή μιας ολικής αναπαράστασης από μικρότερες τοπικές αναπαραστάσεις, οι Raptis et al. [44] εξάγουν πυκνές τροχιές [35] από το βίντεο, τις οποίες ομαδοποιούν βάσει ενός κριτηρίου χωρο-χρονικής εγγύτητας, διαιρώντας, έτσι, τον 3Δ όγκο του βίντεο σε μέρη που ανήκουν στον ίδιο άνθρωπο ή αντικείμενο. Σε κάθε περιοχή υπολογίζουν διάφορους δημοφιλείς περιγραφητές (HoG, HoF κ.α.) και μοντελοποιούν τις δράσεις με ένα Markov Random Field (MRF). Αντίστοιχα, οι Yang et al. [45] εφαρμόζουν μία ιεραρχική ομαδοποίηση για να εντοπίσουν θεμελειώδεις δράσεις (action primitives) τις οποίες ενώνουν για τον σχηματισμό μίας ολοκληρωμένης ανθρώπινης δράσης. Ειδικότερα, αφού απομονώσουν την ανθρώπινη σιλουέτα από το βίντεο, υπολογίζουν την οπτική ροή (u, v) σε κάθε σημείο της (x, y) και ομαδοποιούν τα διανύσματα (x, y, u, v) σε K-means clusters, κάθενα από τα οποία αντιστοιχεί στην εκτέλεση ενός primitive, τα οποία ενώνονται με τη βοήθεια αλγόριθμων ταξινόμησης όπως ο string matching ή τα HMMs.

2.3 Άλλες αναπαραστάσεις

Τα τελευταία χρόνια εμφανίζονται στο προσκήνιο όλο και περισσότερες εργασίες που προσπαθούν να μοντελοποιήσουν ρητά τη χρονική πληροφορία και να οδηγήσουν σε πλήρεις σημασιολογικά χωρο-χρονικές αναπαραστάσεις για τα βίντεο ανθρώπινων δράσεων. Οι Bhattacharya et al. [46], χρησιμοποιώντας ένα χρονικά κυλιόμενο παράθυρο, εξάγουν σκορ ταξινόμησης από μικρούς όγκους μέσα στο βίντεο, οι οποίοι περιέχουν ένα τμήμα της δράσης (υπο-δράση), ως μέρος μιας αλληλουχίας που μοντελοποιείται με ένα γραμμικό δυναμικό μοντέλο (Linear Dynamical System - LDS). Παρόμοια, οι Kuehne et al. [11], δημιουργοί της HMDB51, ταξινομούν μικρές περιοχές υπο-δράσεων (action units) με τη χρήση Bag-of-Features ιστογραμμάτων, εκπαίδευοντας ένα Κρυφό Μαρκοβιανό Μοντέλο (Hidden Markov Model - HMM). Πιο πρόσφατα, οι Fernando et al. [47] χρησιμοποιούν γραμμικές μηχανές κατάταξης (ranking machines) οι οποίες μοντελοποιούν τη χρονική εξέλιξη μιας δράσης και εκπαιδεύονται σε διαδοχικά frames

έτσι ώστε να σταθμίζουν με μικρότερο σκορ αυτό που προηγείται χρονικά.

Ακόμα, οι υψηλές επιδόσεις των νευρωνικών δικτύων σε συνδυασμό με τη μείωση των χρόνων εκπαίδευσης και αξιολόγησης μέσω των σύγχρονων υπολογιστικών πόρων, έχουν δημιουργήσει ένα νέο ερευνητικό πεδίο για επίλυση προβλημάτων Όρασης Υπολογιστών με τη χρήση βαθιών αρχιτεκτονικών Συνελικτικών Νευρωνικών Δικτύων (ΣΝΔ) (deep Convolutional Neural Networks), το οποίο εμπλουτίζεται συνεχώς, με ταχύτατους ρυθμούς μέχρι και σήμερα. Η πρώτη απόπειρα χρήσης ΣΝΔ για την αναγνώριση ανθρώπινων δράσεων έγινε από τους Ji et al. [16], οι οποίοι πρότειναν την επέκταση των υπάρχοντων 2Δ ΣΝΔ στον 3Δ χώρο, με την εφαρμογή 3Δ συνελίξεων σε μία στοίβα από συνεχόμενα καρέ ενός βίντεο, ώστε να επιτευχθεί η εξαγωγή χωρο-χρονικών χαρακτηριστικών. Λίγο αργότερα, οι Tran et al. [48] εκπαίδευσαν 3Δ ΣΝΔ σε δημοφιλείς βάσεις ανθρώπινων δράσεων και εν συνεχεία χρησιμοποιούν τα εκπαιδευμένα μοντέλα για την εξαγωγή χωρο-χρονικών χαρακτηριστικών από βίντεο, τα οποία ονόμασαν C3D features. Συνδυάζοντας τα C3D features με τις improved trajectories επιτυγχάνουν υψηλά ποσοστά αναγνώρισης με έναν γραμμικό ταξινομητή SVM.

Επιπλέον, οι Simonyan & Zisserman [17] εμπνευσμένοι από τη φυσική διάσπαση του βίντεο στο χωρικό και το χρονικό του μέρος, έφεραν στο προσκήνιο τα ΣΝΔ Διπλής Ροής (Two-Stream ConvNets), τα οποία αποτελούνται από ένα χωρικό δίκτυο που μοντελοποιεί την πληροφορία για την εμφάνιση των ανθρώπων ή αντικειμένων και εφαρμόζεται σε στατικά frames, και από ένα χρονικό δίκτυο που περιγράφει την κίνηση μέσα στο βίντεο και εφαρμόζεται πάνω σε 3Δ όγκους οπτικής ροής γύρω από το καρέ ενδιαφέροντος. Τα δύο δίκτυα συνενώνονται με μία σύμμειξη τελικού σταδίου (late fusion) στα softmax scores τους επιτυγχάνοντας πολύ υψηλά ποσοστά ταξινόμησης. Η ιδέα των Simonyan & Zisserman, αποτέλεσε οδηγό για πολλές εργασίες μετέπειτα, όπως αυτή των Wang et al. [49] οι οποίοι δημιούργησαν έναν νέο περιγραφητή βίντεο συνδυάζοντας τις πυκνές τροχιές με τα two-stream δίκτυα, τον οποίο ονόμασαν Trajectory-pooled Deep-convoutional Descriptor (TDD), και τον εφάρμοσαν στην ταξινόμηση δράσεων με ένα γραμμικό SVM με μεγάλη επιτυχία. Ακόμη πιο πρόσφατα, οι Feichtenhofer, Pinz & Zisserman [50] επέκτειναν την two-stream αρχιτεκτονική τους, χρησιμοποιώντας μία πρώιμη σύμμειξη των δύο δικτύων μέσα στο χωρικό δίκτυο, μετατρέποντας το σε ένα χωρο-χρονικό δίκτυο, κρατώντας παράλληλα ενεργό το αμιγώς χρονικό δίκτυο για πιο πλήρεις αναπαραστάσεις. Παράλληλα, οι Wang et al. [51] δημιούργησαν τα Δίκτυα Χρονικών Τμημάτων (Temporal Segment Networks), τα οποία εφαρμόζονται σε μία ακολουθία 3 αποσπασμάτων, αραιά δειγματοληπτημένων από όλο το βίντεο, καθένα από τα οποία παράγει τη δική του πρόβλεψη για τις κλάσεις δράσεων, ενώ ο μέσος όρος τους (average consensus) δίνει την τελική πρόβλεψη για όλο το βίντεο. Μία πρόσφατη παραλλαγή αυτής της αρχιτεκτονικής πραγματοποιήθηκε

από τους Chen & Zhang [52], οι οποίοι προσπαθώντας να διατηρήσουν τη χρονική πληροφορία μέχρι τέλους εφαρμόζουν μία απλή συνένωση (concatenation consensus) των προβλέψεων των επιμέρους αποσπασμάτων. Τέλος, οι Feichtenhofer et al. [53] συνδύαζουν τα two-stream ΣΝΔ με τα residual δίκτυα, εντάσσοντας τις χαρακτηριστικές residual συνδέσεις ανάμεσα στο χωρικό και το χρονικό δίκτυο, ενώ οι Diba et al. [54] προτείνουν μία νέα αναπαράσταση βίντεο, ονόματι Temporal Linear Encoding (TLE), η οποία είναι ενσωματωμένη στο ΣΝΔ ως ένα επιπλέον επίπεδο και κωδικοποιεί τους χάρτες χαρακτηριστικών μέσω bilinear models επιτυγχάνοντας ιδιαιτέρως υψηλά ποσοστά αναγνώρισης ανθρώπινων δράσεων.

Κεφάλαιο 3

Θεωρητικό υπόβαθρο

Στο κεφάλαιο αυτό αναλύονται οι θεωρητικές πτυχές όλων των επιμέρους σταδίων που είναι απαραίτητα για την αναγνώριση ανθρώπινων δράσεων. Ειδικότερα, εστιάζουμε σε κάποιες βασικές μεθόδους και τεχνικές που χρησιμοποιήθηκαν τόσο για την εξαγωγή και κωδικοποίηση χαρακτηριστικών όσο και για την ταξινόμηση. Πιο συγκεκριμένα, ακολουθώντας τη νόρμα της βιβλιογραφίας, θεωρούμε ότι υπάρχουν δύο τύποι χαρακτηριστικών, οι οποίοι συνοδεύονται από τις αντίστοιχες τεχνικές εξαγωγής: α) τα hand-crafted χαρακτηριστικά (Ενότητα 3.1), τα οποία εξάγονται με τεχνικές όρασης υπολογιστών και β) τα χαρακτηριστικά βαθιάς μάθησης (deep learned features) (Ενότητα 3.2), τα οποία εξάγονται μέσω σύγχρονων αρχιτεκτονικών βαθιών Συνελικτικών Νευρωνικών Δικτύων.

Όσον αφορά τις hand-crafted τεχνικές, στις Ενότητες 3.1.1 και 3.1.2 περιγράφονται δύο πολύ δημοφιλείς κατηγορίες χαρακτηριστικών: τα χωρο-χρονικά σημεία ενδιαφέροντος (STIPs) και οι πυκνές τροχιές (Dense Trajectories). Τα χαρακτηριστικά αυτά αναπαριστώνται από κατάλληλους «περιγραφητές» (Ενότητα 3.1.3), οι οποίοι «κβαντίζουν» την οπτική πληροφορία εντός μιας γειτονιάς γύρω από τα σημεία ενδιαφέροντος ή τις τροχιές. Οι «περιγραφητές» αποτελούν «χαμηλού επιπέδου» πληροφορία που εξάγεται από το βίντεο και, συνήθως, δεν είναι σε αξιοποιήσιμη μορφή για περαιτέρω επεξεργασία από κάποιον ταξινομητή. Για το σκοπό αυτό έχουν προταθεί διάφορες αναπαραστάσεις βίντεο που βασίζονται στην κωδικοποίηση των χαρακτηριστικών, οι οποίες αναλύονται στην Ενότητα 3.1.4.

Αντίστοιχα, οι τεχνικές βαθιάς μάθησης βασίζονται στη δυναμική και διαχρονικότητα των Νευρωνικών Δικτύων και, πιο συγκεκριμένα, στη ραγδαία ανάπτυξη που παρουσιάζουν τα τελευταία χρόνια τα Συνελικτικά Νευρωνικά Δίκτυα (Convolutional Neural Networks - CNNs). Η ταχεία εξέλιξη των καρτών γραφικών σε πολυπύρηνες μονάδες επεξεργασίας (Multi-core GPUs) - οι οποίες εκτελούν υπολογισμούς πολύ ταχύτερα από τις CPUs - βοήθησε σημαντικά στη χρήση των Συνελικτικών Νευρωνικών Δικτύων, καθώς οι χρόνοι εκπαίδευσης και αξιολόγησης τους έγιναν πολύ μικρότεροι

και διαχειρίσιμοι. Στην Ενότητα 3.2.1, λοιπόν, γίνεται μία περιγραφή των νευρωνικών δικτύων, από την αρχική μοντελοποίηση ενός νευρώνα και την αρχιτεκτονική ενός νευρωνικού δικτύου, μέχρι την επεξεργασία των δεδομένων για την αρχικοποίηση του και, σε τελικό στάδιο, την εκπαίδευση και αξιολόγηση του. Έπειτα, στην Ενότητα 3.2.2 γίνεται μια ειδικότερη περιγραφή των Συνελικτικών Νευρωνικών Δικτύων (ΣΝΔ), καταγράφοντας ένα προς ένα τα επίπεδα κατασκευής τους, ενώ στην Ενότητα 3.2.3 αναλύονται εκτενώς οι κατηγορίες ΣΝΔ που χρησιμοποιούνται συνηθέστερα. Τέλος, στην Ενότητα 3.3 περιγράφεται η ταξινόμηση των βίντεο, με βάση τα εξαγόμενα χαρακτηριστικά τους (είτε hand-crafted είτε deep-learned), χρησιμοποιώντας Μηχανές Διανυσμάτων Υποστήριξης.

3.1 Hand-crafted τεχνικές

Πρόκειται για μαθηματικές τεχνικές εξαγωγής χαρακτηριστικών, οι οποίες απαιτούν πολύ καλή γνώση των εργαλείων της Όρασης Υπολογιστών και εφαρμόζονται συνήθως σε μία μικρή περιοχή της εικόνας ή του βίντεο. Για πολλά χρόνια ήταν οι κυριότεροι τρόποι εξαγωγής χαρακτηριστικών, ενώ χρησιμοποιούνται μέχρι και σήμερα με μεγάλη αποτελεσματικότητα. Χαρακτηριστικές περιπτώσεις είναι τα χωρο-χρονικά σημεία ενδιαφέροντος και οι πυκνές τροχιές, τα οποία αποτελούν βασικά σημεία αναφοράς για τη σύγκριση και αξιολόγηση νέων μεθόδων.

3.1.1 Χωρο-χρονικά σημεία ενδιαφέροντος (STIPs)

Τα χωρο-χρονικά σημεία ενδιαφέροντος (Spatio-Temporal Interest Points - STIPs) αποτελούν μια από τις πρώτες κατηγορίες χαρακτηριστικών που αναπτύχθηκαν για το σκοπό της αυτόματης αναγνώρισης ανθρώπινων δράσεων. Βασίζονται στην ιδέα της αναπαράστασης ενός βίντεο από ένα σύνολο «σημαντικών» ή «εξέχοντων» (salient) σημείων του.

Ο ανιχνευτής Harris3D

Ο ανιχνευτής Harris3D είναι ίσως ο δημοφιλέστερος ανιχνευτής χωρο-χρονικών σημείων ενδιαφέροντος και χρησιμοποιείται κατά κόρον στη βιβλιογραφία. Αναπτύχθηκε από τους Laptev και Lindberg [14], με αφορμή τη επιτυχία του ανιχνευτή γωνιών Harris [19], του οποίου αποτελεί ουσιαστικά επέκταση στον τρισδιάστατο χώρο. Η βασική ιδέα του ανιχνευτή γωνιών Harris σε εικόνες είναι η ανίχνευση των σημείων εκείνων στα οποία οι τιμές των pixel μεταβάλλονται πάνω από ένα κατώφλι και στις δύο κατευθύνσεις. Αντίστοιχα, ο Harris3D ανιχνεύει τα σημεία εκείνα των οποίων οι τιμές μεταβάλλονται επαρκώς και στις τρεις διαστάσεις. Τα σημεία αυτά θεωρούνται χωρικά

σημεία ενδιαφέροντος με διακριτή θέση στο χρόνο, τα οποία παρουσιάζουν μη σταθερή κίνηση εντός μία χωρο-χρονικής γειτονιάς. Βασική διαφορά του τρισδιάστατου ανιχνευτή είναι ο τρόπος που αντιμετωπίζει τη διάσταση του χρόνου. Συγκεκριμένα, έστω μία ακολουθία εικόνων (βίντεο) $f : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$, για την οποία κατασκευάζεται η αναπαράσταση σε πολλαπλές κλίμακες (scale-space representation) $L : \mathbb{R}^2 \times \mathbb{R} \times \mathbb{R}^2_+ \mapsto \mathbb{R}$ συνελίσσοντας το σήμα βίντεο με έναν ανισοτροπικό Γκαουσιανό πυρήνα με διαφορετική χωρική (σ_l^2) και χρονική (τ_l^2) μεταβλητότητα (variance):

$$L(\cdot; \sigma_l^2, \tau_l^2) = g(\cdot; \sigma_l^2, \tau_l^2) * f(\cdot), \tag{3.1}$$

όπου $g(\cdot)$ ο τρισδιάστατος διαχωρίσιμος Γκαουσιανός πυρήνας:

$$g(x, y, t; \sigma_l^2, \tau_l^2) = \frac{exp(-(x^2 + y^2)/2\sigma_l^2 - t^2/2\tau_l^2)}{\sqrt{(2\pi)^3 \sigma_l^4 \tau_l^2}} \tag{3.2}$$

Όπως και στην περίπτωση του ανιχνευτή δύο διαστάσεων, κατασκευάζεται ο 3×3 πίνακας δεύτερων «στιγμών» (second-moment matrix), αποτελούμενος από τις πρώτες παραγώγους της L σταθμιζόμενες από έναν Γκαουσιανό πυρήνα $g(\cdot; \sigma_l^2, \tau_l^2)$:

$$M = g(\cdot; \sigma_i^2, \tau_i^2) * \begin{pmatrix} L_x^2 & L_x L_y & L_x L_t \\ L_x L_y & L_y^2 & L_y L_t \\ L_x L_t & L_y L_t & L_t^2 \end{pmatrix}, \tag{3.3}$$

όπου $\sigma_i^2 = s\sigma_l^2$ και $\tau_i^2 = s\tau_l^2$ η χωρική και χρονική κλίμακα ολοκλήρωσης αντίστοιχα. Οι πρώτες παράγωγοι ορίζονται ως $L_u(\cdot; \sigma_l^2, \tau_l^2) = \theta_u(g * f)$. Οι ιδιοτιμές $\lambda_1, \lambda_2, \lambda_3$ του πίνακα M περιέχουν την πληροφορία για τη μεταβολή του βίντεο στις τρεις διαστάσεις. Ως σημεία ενδιαφέροντος θεωρούνται τα σημεία στα οποία μεγιστοποιούνται και οι τρεις ιδιοτιμές, όπου υπάρχει δηλαδή έντονη μεταβολή στο χώρο και στο χρόνο.

Παρόλα αυτά, λόγω της μεγάλης διάστασης του πίνακα M ο υπολογισμός των ιδιοτιμών του είναι μία διαδικασία υπολογιστικά πολύπλοκη. Για την αποφυγή της, ομοίως με τον δισδιάστατο ανιχνευτή Harris, οι συγγραφείς προτείνουν ένα κριτήριο «γωνιότητας» για την εύρεση των σημείων ενδιαφέροντος. Σύμφωνα με αυτό, ως σημεία ενδιαφέροντος ορίζονται τα σημεία τοπικού μεγίστου της ποσότητας:

$$H = det(M) - ktrace^3(M) = \lambda_1 \lambda_2 \lambda_3 - k(\lambda_1 + \lambda_2 + \lambda_3)^3 \tag{3.4}$$

Τα τοπικά μέγιστα της H αποδεικνύεται ότι μεγιστοποιούν και τις τρεις ιδιοτιμές $\lambda_1, \lambda_2, \lambda_3$ και άρα αποτελούν σημεία ενδιαφέροντος του βίντεο.

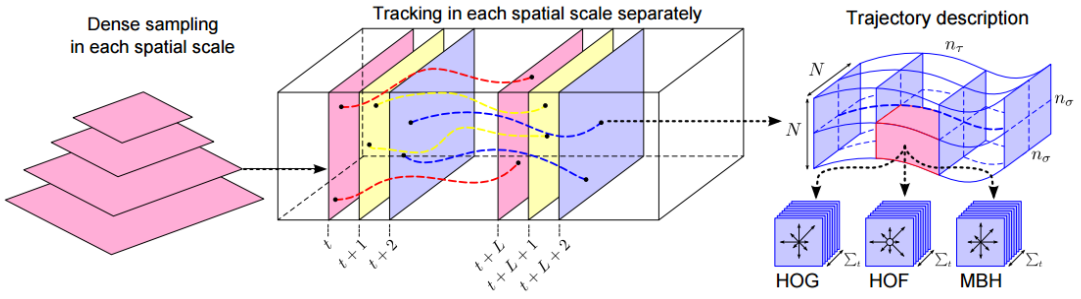
Πυκνή Δειγματοληψία

Σε μία πρόσφατη έρευνα των Wang et al. το 2009 [29], αξιολογήθηκαν οι δημοφιλέστερες τεχνικές εξαγωγής χωρο-χρονικών σημείων ενδιαφέροντος και παρατηρήθηκε

ότι, υπό τις ίδιες συνθήκες, όλες οι τεχνικές παρουσιάζουν παρεμφερή αποτελέσματα, με τον Cubois ανιχνευτή να εμφανίζει ελαφρώς καλύτερες επιδόσεις στις πιο απαιτητικές βάσεις δεδομένων. Επίσης, στην ίδια εργασία, διεξήχθησαν πειράματα δειγματοληπτώντας σημεία ενδιαφέροντος από το βίντεο σε ένα σταθερό χωρο-χρονικό πλέγμα, εξάγοντας περιγραφητές σε προκαθορισμένες χωρο-χρονικές γειτονιές γύρω από αυτά. Αποδείχθηκε, πως η απλή αυτή μέθοδος οδηγεί σε καλύτερα αποτελέσματα σε σχέση με κάθε προηγούμενη υλοποίηση.

3.1.2 Πυκνές Τροχιές (Dense Trajectories)

Η τεχνική των Πυκνών Τροχιών (Dense Trajectories) [15] αποτελεί τη δημοφιλέστερη hand-crafted μέθοδο εξαγωγής χαρακτηριστικών σε βίντεο, με εξαιρετικά αποτελέσματα ακόμη και σε πολύ απαιτητικές βάσεις δεδομένων. Επηρεασμένη από τον εντοπισμό χωρο-χρονικών σημείων ενδιαφέροντος, συνδυάζει την πυκνή δειγματοληψία με μεθόδους περιγραφής της κίνησης μέσω των τροχιών που ακολουθούν τα σημεία ενδιαφέροντος εντός του βίντεο (π.χ. [34]). Πιο συγκεκριμένα, η βασική μεθοδολογία συνίσταται, αρχικά, στην πυκνή δειγματοληψία σημείων ενδιαφέροντος μέσα σε ένα ορθογώνιο πλέγμα στο χώρο της εικόνας και, έπειτα, στην παρακολούθηση τους (tracking) κατά τη διάρκεια του βίντεο μέσω της οπτικής ροής. Η διαδικασία αυτή πραγματοποιείται ξεχωριστά για κάθε μία από 8 χωρικές κλίμακες και για κάθε τροχιά που σχηματίζεται υπολογίζονται 5 διαφορετικοί περιγραφητές. Στη συνέχεια περιγράφουμε αναλυτικά τα επιμέρους βήματα που περιλαμβάνει η διαδικασία εξαγωγής των Πυκνών Τροχιών, όπως φαίνονται στο Σχήμα 3.1.



Σχήμα 3.1: Η διαδικασία εξαγωγής των Πυκνών Τροχιών. *Αριστερά:* Πυκνή δειγματοληψία σε πολλές χωρικές κλίμακες για τον εντοπισμό σημείων ενδιαφέροντος. *Μέση:* Παρακολούθηση των σημείων για L καρέ σε κάθε κλίμακα. *Δεξιά:* Αναπαράσταση της τροχιάς μέσω περιγραφητών που υπολογίζονται σε έναν χωρο-χρονικό όγκο $N \times N \times L$ γύρω από την τροχιά. Για πιο αναλυτική περιγραφή των δομών της, η γειτονιά γύρω από την τροχιά διαιρείται σε ένα χωρο-χρονικό πλέγμα μεγέθους $n_\sigma \times n_\sigma \times n_\tau$. Σχήμα από [35]

Εντοπισμός σημείων ενδιαφέροντος

Το πρώτο βήμα για την εξαγωγή των Πυκνών Τροχιών είναι η εύρεση σημείων ενδιαφέροντος της εικόνας μέσω δειγματοληψίας σε ένα ορθογώνιο πλέγμα με βήμα W pixels. Πειραματικά παρατηρήθηκε ότι ένα βήμα δειγματοληψίας $W = 5$ είναι αρκετά «πυκνό» και οδηγεί σε καλά αποτελέσματα. Η δειγματοληψία γίνεται ανεξάρτητα σε 8 κλίμακες οι οποίες διαφέρουν μεταξύ τους κατά έναν παράγοντα $1/\sqrt{2}$. Έπειτα, πραγματοποιείται ένας έλεγχος των δειγματοληπτημένων σημείων ώστε να απορριφθούν εκείνα που ανήκουν σε ομοιογενείς περιοχές της εικόνας, καθώς είναι αδύνατον να ακολουθηθεί η τροχιά τους μέσω της οπτικής ροής. Η επιλογή τους γίνεται με το κριτήριο των Shi & Tomasi [55], το οποίο βασίζεται στον ανιχνευτή γωνιών Harris [19]. Ο τελευταίος υπολογίζει έναν πίνακα αυτοσυσχέτισης $M(x, y)$ όμοιο με τον (3.3) - χωρίς την ύπαρξη της διάστασης του χρόνου - και εν συνεχεία οι Shi & Tomasi προτείνουν το εξής κριτήριο «γωνιότητας»: *Αν η μικρότερη ιδιοτιμή του πίνακα αυτοσυσχέτισης του σημείου είναι μικρότερη από ένα κατώφλι τότε το σημείο αυτό απορρίπτεται.*

Σε μαθηματικούς όρους, το κριτήριο «γωνιότητας» ορίζεται ως εξής: Ένα σημείο (x, y) του πλέγματος θεωρείται σημείο ενδιαφέροντος αν ικανοποιείται η ακόλουθη συνθήκη:

$$H(x, y) > \kappa \cdot \underset{x,y}{max} H(x, y) \tag{3.5}$$

όπου $H(x, y) = min(\lambda_1, \lambda_2)$. Οι λ_1, λ_2 είναι οι ιδιοτιμές του πίνακα $M(x, y)$, ενώ για την τιμή του κατωφλίου, πειραματικά επιλέχθηκε $\kappa = 0.001$.

Υπολογισμός οπτικής ροής

Έπειτα από τον εντοπισμό των σημείων ενδιαφέροντος ακολουθεί η παρακολούθησή τους με τη χρήση της οπτικής ροής. Η οπτική ροή είναι το μέγεθος εκείνο που περιγράφει τη σχετική κίνηση μεταξύ της κάμερας και της σκηνής που απαθανατίζεται σε μία ακολουθία εικόνων. Ποσοτικοποιεί, ουσιαστικά, τη «στιγμιαία» ταχύτητα κάθε pixel ή αλλιώς τη σχετική μετατόπιση που υφίσταται μεταξύ δύο διαδοχικών καρέ I_t και I_{t+1}. Για την εκτίμηση της οπτικής ροής έχουν αναπτυχθεί πολλές μέθοδοι, με πιο διαδεδομένη αυτή των Lucas & Kanade [56]. Για τους σκοπούς των Πυκνών Τροχιών, ωστόσο, στην παρακολούθηση των σημείων ενδιαφέροντος χρησιμοποιείται η μέθοδος του Farnebäck [57], η οποία υπολογίζει ένα πυκνό πεδίο οπτικής ροής $\mathbf{d} = (u_t, v_t)$, όπου u_t, v_t η οριζόντια και κατακόρυφη διάσταση της αντίστοιχα. Η εκτίμηση της μετατόπισης κάθε pixel γίνεται με τη χρήση πολυωνυμικών αναπτυγμάτων.

Σχηματισμός και αναπαράσταση τροχιών

Έπειτα από τον εντοπισμό των σημείων ενδιαφέροντος και τον υπολογισμό της οπτικής ροής, ακολουθεί η περιγραφή της τροχιάς τους μέσα στο βίντεο. Ειδικότερα,

κάθε δειγματοληπτημένο σημείο ενδιαφέροντος $P_t = (x_t, y_t)$ στο καρέ I_t ακολουθείται στο καρέ I_{t+1}, μέσω ενός φιλτραρίσματος διάμεσης τιμής (median filtering) σε ένα πυκνό πεδίο οπτικής ροής $\mathbf{d}_t = (u_t, v_t)$, ως εξής:

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * d_t)|_{(x_t, y_t)} \tag{3.6}$$

όπου M είναι ο 3×3 πυρήνας του φίλτρου ενδιάμεσης τιμής. Οι θέσεις, λοιπόν, από τις οποίες περνάει ένα σημείο κατά μήκος διαδοχικών καρέ διαμορφώνουν μία τροχιά (trajectory): $(P_t, P_{t+1}, P_{t+2}, ...)$.

Ένα κοινό πρόβλημα στην παρακολούθηση των τροχιών είναι η ολίσθηση. Οι τροχιές τείνουν να ολισθαίνουν από την αρχική τους θέση κατά τη διάρκεια του βίντεο. Για την αποφυγή αυτού του προβλήματος, οι συγγραφείς περιορίζουν το μήκος τους σε $L = 15$ καρέ, έτσι ώστε όταν μία τροχιά ξεπεράσει το όριο αυτό, το σημείο παύει να παρακολουθείται. Ακόμη, για να διασφαλιστεί η πυκνή κάλυψη του βίντεο, σε κάθε frame, αν δεν παρακολουθείται κανένα σημείο εντός μίας γειτονιάς $W \times W$ τότε ένα νέο σημείο δειγματοληπτείται και προστίθεται στη διαδικασία παρακολούθησης. Τέλος, τροχιές που είναι στατικές ή, αντίθετα, τροχιές που μεταβάλλονται σε διαδοχικά καρέ περισσότερο από το 70% της ολικής μετατόπισης της τροχιάς, απορρίπτονται.

Για την αναπαράσταση των Πυκνών Τροχιών χρησιμοποιούνται οι ακόλουθοι 5 περιγραφητές: Περιγραφητής Τροχιάς (Trajectory Descriptor - TD), HoG, HoF, MBHx, MBHy καθώς και ο MBH, η ένωση, δηλαδή, των δύο τελευταίων (βλ. Ενότητα 3.1.3). Υπολογίζονται εντός ενός χωρο-χρονικού όγκου γύρω από την τροχιά, μεγέθους $N \times N \times L$. Για να ενσωματωθεί πληροφορία σχετικά με τις δομές που υπάρχουν μέσα στην τροχιά, η γειτονιά γύρω από αυτήν διαιρείται σε ένα χωρο-χρονικό πλέγμα μεγέθους $n_\sigma \times n_\sigma \times n_\tau$. Κάθε περιγραφητής υπολογίζεται στους επιμέρους όγκους που ορίζει το χωρο-χρονικό πλέγμα και ο τελικός περιγραφητής της τροχιάς προκύπτει από την ένωση (concatenation) των περιγραφητών των επιμέρους όγκων. Πειραματικά, για τις παραμέτρους της γειτονιάς γύρω από την τροχιά και του πλέγματος επιλέχθηκαν οι τιμές: $N = 32, n_\sigma = 2, n_\tau = 3$.

Όσον αφορά τις παραμέτρους των περιγραφητών, για τους HoG, MBHx, MBHy χρησιμοποιούνται 8 bins (θέσεις στο ιστόγραμμα) ενώ για τον HoF χρησιμοποιείται ένα επιπλέον bin (σύνολο 9 bins). Το μέγεθος, επομένως, των τελικών περιγραφητών για κάθε τροχιά είναι $8 \times 2 \times 2 \times 3 = 96$ και $9 \times 2 \times 2 \times 3 = 108$ αντίστοιχα. Προφανώς, το μέγεθος του MBH προκύπτει ως $2 \times 96 = 192$. Ο TD είναι ένας απλός περιγραφητής που κωδικοποιεί το σχήμα της τροχιάς. Απαρτίζεται από τις σχετικές μετατοπίσεις των σημείων της τροχιάς, κανονικοποιημένες με το άθροισμα τους. Συγκεκριμένα, ως σχήμα μίας τροχιάς θεωρείται η ακολουθία $S = (\Delta P_t, ..., \Delta P_{t+L-1})$ των σχετικών μετατοπίσεων της $\Delta P_t = (P_{t+1} - P_t) = (x_{t+1} - x_t, y_{t+1} - y_t)$. Έτσι, ο Trajectory

Descriptor ορίζεται ως:

$$TD = \frac{(\Delta P_t, ..., \Delta P_{t+L-1})}{\sum_{j=t}^{t+L-1} \|\Delta P_j\|} \tag{3.7}$$

Όπως είναι αντιληπτό, το μέγεθος του TD εξαρτάται άμεσα από το μήκος L της τροχιάς και εφόσον υπολογίζει τις δισδιάστατες χωρικές μετατοπίσεις κάθε σημείου, το μέγεθος του είναι ίσο με $2 \times L = 30$.

Ο HoG περιγράφει το σχήμα και την εμφάνιση γύρω από κάθε τροχιά. Οι HoF και MBH περιγράφουν την κίνηση, με τον δεύτερο να λαμβάνει υπόψη του και την κίνηση της κάμερας για μεγαλύτερη ευρωστία. Οι περιγραφητές αυτοί, λοιπόν, κωδικοποιούν διαφορετική πληροφορία για κάθε τροχιά. Λόγω αυτής της συμπληρωματικότητας τους, ενώνονται συνήθως για την δημιουργία ενός «Συνδυαστικού» περιγραφητή (Combined Descriptor) ο οποίος χρησιμοποιείται κατά κόρον στην αναγνώριση ανθρώπινων δράσεων.

3.1.3 Περιγραφητές (Descriptors)

Ιστογράμματα κατευθυνόμενων παραγώγων

Τα ιστογράμματα κατευθυνόμενων παραγώγων (Histograms of Oriented Gradients - HoG) ήρθαν στο προσκήνιο από τους Dalal & Triggs [21] στην προσπάθεια τους για εντοπισμό ανθρώπων (human detection) σε στατικές εικόνες. Στο πρόβλημα της αναγνώρισης ανθρώπινων δράσεων εφαρμόστηκαν πρώτη φορά από τους Laptev et al. [10]. Η βασική ιδέα πίσω από τον περιγραφητή HoG είναι ότι το σχήμα και η εμφάνιση των αντικειμένων σε μία μικρή περιοχή της εικόνας μπορεί να περιγραφεί επαρκώς από την κατανομή των τοπικών κλίσεων (gradients), δηλαδή των κατευθύνσεων των ακμών μέσα στην περιοχή αυτή. Για τον υπολογισμό του HoG ακολουθούνται τα παρακάτω βήματα:

1. Εφαρμογή διόρθωσης του εκθετικού κανόνα (gamma correction) και κανονικοποίησης χρώματος στην εικόνα.

2. Υπολογισμός της κλίσης της εικόνας (gradient) μέσω συνέλιξης με τον πυρήνα $\begin{bmatrix} -1 & 0 & 1 \end{bmatrix}$ για τον οριζόντιο άξονα και με τον ανάστροφο του για τον κατακόρυφο άξονα.

3. Υπολογισμός για κάθε pixel του μέτρου $m(x, y)$ και της διεύθυνσης $\theta(x, y)$ της κλίσης:

$$m(x, y) = \sqrt{I_x^2 + I_y^2} \tag{3.8}$$

$$\theta(x, y) = \arctan\left(\frac{I_y}{I_x}\right) \tag{3.9}$$

4. Διαίρεση της εικόνας σε κελιά (cells), συνήθως μεγέθους 8×8 pixels, και κατασκευή, για το σύνολό των pixels εντός του κελιού, του ιστογράμματος των διευθύνσεων του gradient, σταθμισμένο με το αντίστοιχο πλάτος του. Συγκεκριμένα, η διεύθυνση $\theta(x, y)$ της κλίσης κβαντίζεται σε ισοκατανεμημένες στάθμες (bins) στο διάστημα $0^o - 180^o$ (χωρίς πρόσημο) ή $0^o - 360^o$ (με πρόσημο). Έπειτα, κάθε pixel (x, y) του κελιού συνεισφέρει μία «ψήφο», πλάτους $m(x, y)$, στο διάστημα (bin) του ιστογράμματος που ανήκει η τιμή του $\theta(x, y)$. Ο αριθμός των bins είναι συνήθως 8 ή 9.

5. Ομαδοποίηση των κελιών σε μπλοκ (συνήθως 2×2 κελιών) και ενοποίηση (concatenation) των επιμέρους ιστογραμμάτων των κελιών σε έναν ενιαίο περιγραφητή για κάθε μπλοκ, κανονικοποιημένο, έπειτα, με την $L2$ νόρμα.

6. Ένωση (concatenation) των επιμέρους περιγραφητών κάθε μπλοκ της εικόνας για την δημιουργία του τελικού περιγραφητή HoG. Τα μπλοκ της εικόνας μπορεί να είναι επικαλυπτόμενα και τελικά κάθε pixel να συνεισφέρει παραπάνω από μία φορά στο τελικό διάνυσμα του περιγραφητή.

Στην περίπτωση μίας ακολουθίας εικόνων (βίντεο), η μόνη διαφορά στην εξαγωγή του περιγραφητή HoG είναι ότι τα κελιά (cells) είναι τρισδιάστατες χωρο-χρονικές γειτονιές και ομαδοποιούνται έπειτα σε ένα χωρο-χρονικό μπλοκ $n_x \times n_y \times n_t$ κελιών για τη δημιουργία του περιγραφητή.

Ιστογράμματα οπτικής ροής

Τα ιστογράμματα οπτικής ροής (Histograms of Optical Flow - HoF) αναπτύχθηκαν από τους Laptev et al. [10] και αποτελούν ουσιαστικά συνέχιση μιας παλαιότερης εργασίας των Laptev και Lindeberg [22], όπου πειραματίζονταν με διάφορες παραλλαγές ιστογραμμάτων οπτικής ροής. Ο HoF περιγράφει την κίνηση μέσα σε ένα μικρό χωροχρονικό όγκο του βίντεο ακολουθώντας το ίδιο μοτίβο με τον HoG, με τη διαφορά ότι αντί για το gradient χρησιμοποιείται η οπτική ροή. Ειδικότερα, υπολογίζεται, αρχικά, η οπτική ροή ως προς τους δύο άξονες και κατόπιν το μέτρο και η διεύθυνση της. Για την κατασκευή των ιστογραμμάτων, η διεύθυνση κβαντίζεται συνήθως σε 8+1 στάθμες, με την τελευταία να προορίζεται για τα pixels εκείνα στα οποία το μέτρο της οπτικής ροής είναι μικρότερο από κάποιο κατώφλι. Έπειτα, το βίντεο υποδιαιρείται σε ένα χωροχρονικό πλέγμα $n_x \times n_y \times n_t$ κελιών, για κάθε ένα από τα οποία κατασκευάζεται ένα ιστόγραμμα των διευθύνσεων της οπτικής ροής, ακριβώς όπως στον HoG. Τέλος, τα επιμέρους ιστογράμματα ενώνονται και κανονικοποιούνται σχηματίζοντας τον τελικό περιγραφητή του βίντεο.

Ιστογράμματα περιγράμματος κίνησης

Τα ιστογράμματα περιγράμματος κίνησης (Motion Boundary Histograms - MBH) προτάθηκαν από τους Dalal et al. [36] στην προσπάθεια τους να λάβουν υπόψη την πληροφορία για την κίνηση της κάμερας, προκειμένου ένας περιγραφητής να διαχωρίζει την κίνηση του παρασκηνίου (background) από την πραγματική κίνηση την οποία πρέπει να περιγράφει. Η ιδέα, λοιπόν, που επινοήθηκε για τον MBH αφορά την περιγραφή της κίνησης μέσω της παραγώγου της οπτικής ροής. Συγκεκριμένα, έπειτα από την εξαγωγή της οπτικής ροής, υπολογίζεται η παράγωγος της τόσο στο οριζόντιο όσο και στο κατακόρυφο μέρος της. Στη συνέχεια ακολουθείται η ίδια διαδικασία με αυτή του HoG και προκύπτουν δύο περιγραφητές: ο MBHx για την οριζόντια κίνηση και ο MBHy για την κατακόρυφη, οι οποίοι κανονικοποιούνται με την $L2$ νόρμα.

Ο MBH είναι αρκετά εύρωστος στην κίνηση της κάμερας, αφού όταν η κίνηση είναι ομαλή, η οπτική ροή περιέχει μια σταθερή συνιστώσα, η οποία αφαιρείται κατά τον υπολογισμό της παραγώγου. Έτσι, κωδικοποιούνται μόνο οι μεταβολές στο πεδίο της οπτικής ροής.

3.1.4 Κωδικοποίηση Χαρακτηριστικών

Η εξαγωγή χαρακτηριστικών οδηγεί συνήθως σε αναπαραστάσεις που είναι ανόμοιες μεταξύ τους, τόσο ως προς το πλήθος των χαρακτηριστικών όσο και ως προς το εύρος τιμών των περιγραφητών τους. Οι διαφορές αυτές παρατηρούνται όχι μόνο μεταξύ διαφορετικών δράσεων αλλά, σε πολλές περιπτώσεις, και μεταξύ διαφορετικών εκτελέσεων της ίδιας δράσης. Ένας ταξινομητής, για να μπορέσει να διακρίνει τις διαφορές μεταξύ των ανθρώπινων δράσεων, πρέπει να είναι σε θέση να επεξεργαστεί όμοιες αναπαραστάσεις χαρακτηριστικών (π.χ. ίδιες διαστάσεις, κοινό εύρος τιμών κλπ). Προκύπτει, λοιπόν, η ανάγκη καθορισμού μιας ενιαίας αναπαράστασης που να κωδικοποιεί το σύνολο των χαρακτηριστικών που εντοπίζονται σε ένα βίντεο. Προς αυτή την κατεύθυνση έχουν κινηθεί μέθοδοι που καταγράφουν την κατανομή των χαρακτηριστικών ενός βίντεο με βάση ένα «λεξικό» (codebook), όπως η πολύ δημοφιλής μέθοδος Bag-of-Words.

Δημιουργία του οπτικού λεξικού

Το πρώτο βήμα για την κατασκευή μιας ενιαίας αναπαράστασης χαρακτηριστικών για ένα σύνολο βίντεο ανθρώπινων δράσεων είναι η δημιουργία του «οπτικού λεξικού» (visual codebook). Το οπτικό λεξικό περιλαμβάνει τις «οπτικές λέξεις» (visual codewords) κάθε μία από τις οποίες εκπροσωπεί ένα σύνολο παρόμοιων χαρακτηριστικών. Συγκεκριμένα, έστω $\mathbf{X} = \{x_1, x_2, ..., x_N\}$ ένα σύνολο N χαρακτηριστικών εκπαίδευσης, όπου $x_i \in \mathbb{R}^D$ και D η διάσταση κάθε χαρακτηριστικού. Αναζητούμε

ένα αντιπροσωπευτικό υποσύνολο $\mathbf{C} = \{c_1, c_2, ..., c_K\}$ του $\mathbf{X}$, το οποίο θα αποτελεί τη βάση για την κωδικοποιήση των χαρακτηριστικών που εξάγονται από κάθε βίντεο. Το υποσύνολο αυτό αποτελεί το οπτικό λεξικό και συνήθως υπολογίζεται μέσω ενός αλγόριθμου ομαδοποίησης (clustering), όπως τον δημοφιλή K-means [58] ή την μέθοδο ομαδοποίησης με Μείγμα Γκαουσιανών Κατανομών (Gaussian Mixture Model - GMM). Παρακάτω κάνουμε μία σύντομη περιγραφή των δύο αυτών αλγορίθμων.

K-means clustering: Ο K-means είναι ο δημοφιλέστερος αλγόριθμος διανυσματικής κβάντισης (vector quantization). Χρησιμοποιείται για να διαιρέσει το χώρο $\mathbf{X}$ των χαρακτηριστικών σε K περιοχές, κάθε μία από τις οποίες αντιπροσωπεύεται από ένα διάνυσμα $\mathbf{c}_k$, το οποίο ονομάζεται «κεντροειδές». Το σύνολο των «κεντροειδών» $C = \{\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_K\}$ αποτελεί το οπτικό λεξικό. Για τον υπολογισμό τους, ο K-means χρησιμοποιεί ένα σύνολο βοηθητικών boolean μεταβλητών $r_{i,k} \in \{0, 1\}, i = 1, ..., N, k = 1, ..., K$, οι οποίες υποδεικνύουν αν το χαρακτηριστικό $\mathbf{x}_i$ ανήκει ή όχι ($r_{i,k} = 1$ ή 0 αντίστοιχα) στην περιοχή που αντιπροσωπεύεται από το κέντρο $\mathbf{c}_k$. Κάθε στοιχείο $\mathbf{x}_i$ αντιστοιχεί σε ένα μόνο κέντρο $\mathbf{c}_k$, έτσι ο υπολογισμός των «κεντροειδών» γίνεται ελαχιστοποιώντας το συναρτησιακό:

$$\sum_{i=1}^{N} \sum_{k=1}^{K} r_{i,k} \|\mathbf{x}_i - \mathbf{c}_k\|^2 \tag{3.10}$$

Για την ελαχιστοποίηση ακολουθείται μία επαναληπτική διαδικασία αποτελούμενη από δύο βήματα που εναλλάσσονται μέχρι τη σύγκλιση:

- Ελαχιστοποίηση του συναρτησιακού ως προς $r_{i,k}$ με σταθερά τα $\mathbf{c}_k$
- Υπολογισμός των νέων κέντρων ως η μέση τιμή των διανυσμάτων που ανήκουν στην ίδια περιοχή

Περισσότερες πληροφορίες και πιο διεξοδική ανάλυση για τον αλγόριθμο K-means περιέχονται στο βιβλίο [59].

GMM clustering: Τα Μείγματα Γκαουσιανών Κατανομών (GMMs) αποτελούν αναγεννητικά μοντέλα (generative models) που περιγράφουν μία κατανομή σε ένα χώρο χαρακτηριστικών και χρησιμοποιούνται για την εύρεση ομάδων εντός αυτού. Ένα GMM με K γκαουσιανές κατανομές ορίζεται ως εξής:

$$p(x, \theta) = \sum_{k=1}^{K} \pi_k N\left(x; \mu_k, \Sigma_k\right) \tag{3.11}$$

όπου $\theta = \{\pi_k, \mu_k, \Sigma_k\}$ είναι οι παράμετροι των γκαουσιανών $N\left(x; \mu_k, \Sigma_k\right), k = 1, ..., K$. Αρχικά εφαρμόζεται μία κατανομή GMM πάνω στα γνωστά χαρακτηριστικά εκπαίδευσης $\mathbf{X} = \{x_1, x_2, ..., x_N\}$, μέσω των οποίων υπολογίζονται οι παράμετροι θ. Εν συνεχεία, η ομαδοποίηση (clustering) πραγματοποιείται ανιστοιχίζοντας σε κάθε διάνυσμα

$\mathbf{x}$ του χώρου μία posterior πιθανότητα $p(k|\mathbf{x})$, η οποία ποσοτικοποιεί το κατά πόσο «ανήκει» ή περιγράφεται από τη γκαουσιανή k. Κάθε γκαουσιανή του GMM, λοιπόν, αντιστοιχεί σε ένα «κέντρο» και ορίζει μία περιοχή του χώρου χαρακτηριστικών. Η ανάθεση ενός διανύσματος $\mathbf{x}$ σε μία περιοχή μπορεί να γίνει είτε με «αυστηρό» τρόπο, αναθέτοντας το στο cluster της γκαουσιανής που μεγιστοποιεί την πιθανότητα $p(k|\mathbf{x})$, είτε με «χαλαρό» τρόπο, αναθέτοντας το σε όλα τα clusters, με βάρος $p(k|\mathbf{x})$. Η πρώτη περίπτωση αντιστοιχεί στον K-means, με τη διαφορά ότι για τα GMMs το σχήμα κάθε cluster είναι ευμετάβλητο και καθορίζεται από τον πίνακα συμμεταβλητότητας Σ_k της αντίστοιχης γκαουσιανής. Περισσότερες λεπτομέρειες σχετικά με την ομαδοποίηση με GMMs περιλαμβάνονται στο βιβλίο [59].

Τεχνικές Κωδικοποίησης

Όπως αναφέρθηκε παραπάνω, για να είναι σε θέση ένας ταξινομητής να κατηγοριοποιήσει ένα βίντεο, θα πρέπει αυτό να περιγράφεται με μία αναπαράσταση που να το καθιστά διαχωρίσιμο από τα υπόλοιπα. Αυτό γίνεται μέσω στατιστικών υπολογισμών πάνω στα χαρακτηριστικά του, χρησιμοποιώντας το οπτικό λεξικό. Η διαδικασία αυτή αναφέρεται ως κωδικοποίηση χαρακτηριστικών (feature encoding) και περιλαμβάνει συνήθως δύο στάδια:

(α′) την εξαγωγή του *κώδικα* (code) $\mathbf{s}_i$ κάθε χαρακτηριστικού $\mathbf{x}_i \in \mathbf{X}$ του βίντεο, όπου εκτιμάται η κατανομή του χαρακτηριστικού σχετικά με τις «οπτικές λέξεις»

(β′) τη συσσώρευση (pooling) των επιμέρους κωδικών $\mathbf{s}_i, i = 1, ..., N$ σε ένα ενιαίο διάνυσμα $\mathbf{S}$, που αποτελεί την αναπαράσταση του βίντεο.

Δύο πολύ δημοφιλείς τεχνικές κωδικοποίησης χαρακτηριστικών είναι το μοντέλο συνόλου οπτικών λέξεων (bag-of-words) και οι χωρο-χρονικές πυραμίδες (spatio-temporal pyramids), οι οποίες περιγράφονται παρακάτω. Υπάρχουν και άλλες εξίσου σημαντικές τεχνικές κωδικοποίησης όπως το διάνυσμα τοπικών συσσωρευμένων περιγραφητών (VLAD) [60] και το διάνυσμα Fisher [61], οι οποίες όμως, λόγω της μεγάλης υπολογιστικής πολυπλοκότητας που επιφέρουν, δεν χρησιμοποιούνται στα πλαίσια της παρούσας Διπλωματικής.

Μοντέλο συνόλου οπτικών λέξεων: Το μοντέλο συνόλου οπτικών λέξεων (Bag-of-Words - BoW) [62] είναι η πιο απλή μέθοδος κωδικοποίησης χαρακτηριστικών και είναι εμπνευσμένη από τα πεδία της επεξεργασίας φυσικής γλώσσας και ανάκτησης πληροφορίας. Σύμφωνα με αυτή, κάθε χαρακτηριστικό αντικαθίσταται από την κοντινότερη του οπτική λέξη και το βίντεο αναπαριστάται από τη συχνότητα εμφάνισης κάθε οπτικής λέξης. Ο κώδικας $\mathbf{s}_i$ για κάθε χαρακτηριστικό $\mathbf{x}_i \in \mathbf{X}$ υπολογίζεται από τη

σχέση:

$$\mathbf{s}_i = \begin{cases} 1, & \alpha\nu\ k = \text{argmin}_k \|\mathbf{x}_i - \mathbf{c}_k\| \\ 0, & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases} \tag{3.12}$$

όπου $k = 1, ..., K$. Όπως φαίνεται από την παραπάνω σχέση, ο κώδικας κάθε χαρακτηριστικού έχει διάσταση K και περιέχει μονάδα μόνο στη θέση που αντιστοιχεί στην κοντινότερη οπτική του λέξη και μηδέν στις υπόλοιπες, πραγματοποιώντας ουσιαστικά μία κβάντιση. Για το μέγεθος του οπτικού λεξικού επιλέγονται συνήθως τιμές στο διάστημα $K \in [500, 4000]$. Έπειτα, οι κώδικες $\mathbf{s}_i$ όλων των χαρακτηριστικών αθροίζονται και το διάνυσμα που προκύπτει κανονικοποιείται με το συνολικό τους πλήθος. Δηλαδή:

$$\mathbf{p} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{s}_i \tag{3.13}$$

Το διάνυσμα $\mathbf{p}$, λοιπόν, περιλαμβάνει τη σχετική συχνότητα εμφάνισης κάθε οπτικής λέξης μέσα στο βίντεο. Στο τελικό στάδιο, πραγματοποιείται μία κανονικοποίηση του $\mathbf{p}$ χρησιμοποιώντας την $L1$ ή πιο συχνά την $L2$ νόρμα και το προκύπτον διάνυσμα αποτελεί την αναπαράσταση του βίντεο. Συνεπώς, το μοντέλο κωδικοποίησης Bag-of-Words έγκειται στην κατασκευή ενός ιστογράμματος εμφάνισης των οπτικών λέξεων μέσα στο βίντεο, κατάλληλα κανονικοποιημένο ώστε να εξαλειφθεί η εξάρτηση από το πλήθος των χαρακτηριστικών που ανιχνεύθηκαν.

Χωρο-χρονικές πυραμίδες: Η ιδέα των χωρο-χρονικών πυραμίδων (spatio-temporal pyramids) εμπνεύστηκε από τις χωρικές πυραμίδες (spatial pyramids) των Lazebnik et al. [63], τις οποίες πρότειναν στην προσπάθεια τους να ενσωματώσουν στο μοντέλο Bag-of-Words πληροφορία για τη χωρική διάταξη των χαρακτηριστικών, βελτιώνοντας τη διακριτική του ικανότητα για την ταξινόμηση εικόνων (image classification). Γρήγορα, οι Schuldt et al. [6] προσάρμοσαν τη συγκεκριμένο μέθοδο στο πρόβλημα της αναγνώρισης ανθρώπινων δράσεων, επεκτείνοντας τις πυραμίδες στη διάσταση του χρόνου.

Το μοντέλο Bag-of-Words, όπως αναλύθηκε παραπάνω, κωδικοποιεί τα χαρακτηριστικά ενός βίντεο χωρίς να εξετάζει τη διάταξη τους στο χώρο ή στο χρόνο. Έτσι, για παράδειγμα, δύο χαρακτηριστικά μπορεί να αντιστοιχούν στην ίδια οπτική λέξη αλλά να εμφανίζονται σε εντελώς διαφορετικές θέσεις ή με μεγάλη χρονική απόσταση. Για να εξομαλύνουν τέτοια φαινόμενα, οι χωρο-χρονικές πυραμίδες διαιρούν τον τρισδιάστατο όγκο του βίντεο σε ένα χωρο-χρονικό πλέγμα $n_x \times n_y \times n_t$ κελιών και υπολογίζουν μία ξεχωριστή Bag-of-Words αναπαράσταση για κάθε κελί. Έπειτα, τα επιμέρους διανύσματα ενώνονται (concatenation) σε ένα ενιαίο διάνυσμα, το οποίο κανονικοποιείται αρχικά με το πλήθος των χαρακτηριστικών και εν συνεχεία με την $L2$ νόρμα, για να

σχηματιστεί η τελική αναπαράσταση του βίντεο. Σημειώνεται ότι τα ιστογράμματα των επιμέρους κελιών υπολογίζονται με βάση *το ίδιο* οπτικό λεξικό.

3.2 Τεχνικές βαθιάς μάθησης (deep learning)

Η έννοια της «βαθιάς μάθησης» είναι αλληλένδετη με ένα πολύ δημοφιλές πεδίο της Τεχνητής Νοημοσύνης, τα νευρωνικά δίκτυα. Η επεκτάσιμη αρχιτεκτονική τους και τα διαφορετικά επίπεδα μάθησης που τα χαρακτηρίζουν, καθιστούν τα νευρωνικά δίκτυα πολύ σημαντικά εργαλεία για την επίλυση προβλημάτων Όρασης Υπολογιστών. Πιο συγκεκριμένα, οι τεχνικές βαθιάς μάθησης αφορούν τα νευρωνικά δίκτυα τα οποία περιέχουν μία αλληλουχία πολλών μονάδων-κόμβων μη γραμμικής επεξεργασίας δεδομένων για εξαγωγή χαρακτηριστικών, όπου κάθε κόμβος δέχεται σαν είσοδο την έξοδο του αμέσως προηγούμενου. Με αυτόν τον τρόπο, το δίκτυο εκπαιδεύεται από διαφορετικά επίπεδα χαρακτηριστικών, δημιουργώντας μία ιεραρχία, στην οποία χαρακτηριστικά «υψηλότερου» επιπέδου παράγονται από χαρακτηριστικά «χαμηλότερου» επιπέδου. Μία κατηγορία τέτοιων δικτύων είναι τα Βαθιά Συνελικτικά Νευρωνικά Δίκτυα (Deep Convolutional Neural Networks - Deep CNNs) τα οποία αναλύονται εκτενώς παρακάτω, αφού πρώτα όμως γίνει μία μικρή περιγραφή της δομής ενός νευρωνικού δικτύου και των μεθόδων εκπαίδευσης και αξιολόγησης του.

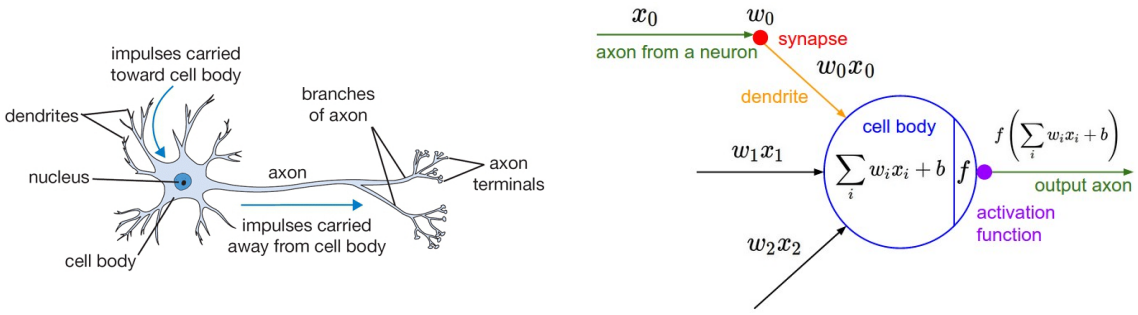
3.2.1 Νευρωνικά Δίκτυα

Μοντελοποίηση ενός νευρώνα

Το ερευνητικό πεδίο των νευρωνικών δικτύων έχει τα θεμέλια του στη βιολογία και εμπνεύστηκε αρχικά από την απόπειρα μοντελοποίησης του ανθρώπινου νευρικού συστήματος. Ειδικότερα, το ανθρώπινο νευρικό σύστημα αποτελείται από περίπου 86 δισεκατομμύρια *νευρώνες (neurons)*, οι οποίοι αποτελούν τις βασικές υπολογιστικές μονάδες του εγκεφάλου και είναι συνδεδεμένοι με περίπου $10^{14} - 10^{15}$ *συνάψεις (synapses)*. Όπως φαίνεται στο Σχήμα 3.2 (αριστερά), κάθε νευρώνας λαμβάνει ως είσοδο σήματα από τους *δενδρίτες (dendrites)* του και παράγει σήματα εξόδου, τα οποία μεταφέρει μέσω του *άξονα (axon)* του, ο οποίος με τη σειρά του διακλαδίζεται και ενώνεται μέσω συνάψεων με τους δενδρίτες άλλων νευρώνων.

Με βάση τα παραπάνω, μπορεί να ορισθεί ένα μαθηματικό μοντέλο για έναν νευρώνα, σύμφωνα με το οποίο τα σήματα x_0 που ταξιδεύουν μέσω των αξόνων αλληλεπιδρούν πολλαπλασιαστικά με τους δενδρίτες ενός άλλου νευρώνα ανάλογα με το «βάρος» w_0 της αντίστοιχης σύναψης. Η ιδέα είναι ότι τα βάρη w είναι εκπαιδεύσιμα και απεικονίζουν τη δύναμη της επιρροής ενός νευρώνα πάνω σε έναν άλλο. Στο θεμελιώδες μοντέλο ενός νευρώνα, όπως φαίνεται στο Σχήμα 3.2 (δεξιά), οι δενδρίτες μεταφέρουν

τα σήματα στον πυρήνα όπου όλα αθροίζονται μεταξύ τους. Αν η τιμή του αθροίσματος είναι πάνω από ένα κατώφλι τότε ο νευρώνας «ενεργοποιείται», στέλνοντας το αντίστοιχο σήμα στον άξονα του. Αυτή η διαδικασία μοντελοποιείται μαθηματικά με μία *συνάρτηση ενεργοποίησης* f η οποία αναπαριστά ουσιαστικά τη συχνότητα με την οποία μεταφέρονται σήματα μέσω του άξονα του νευρώνα. Χαρακτηριστικά παραδείγματα συναρτήσεων ενεργοποίησης, με ευρεία χρήση στη βιβλιογραφία, είναι η *σιγμοειδής συνάρτηση σ (sigmoid function)*, το *τόξο εφαπτομένης tanh* και η *ReLU*. Παρακάτω γίνεται μία σύντομη περιγραφή αυτών των συναρτήσεων.



Σχήμα 3.2: Η απεικόνιση ενός βιολογικού νευρώνα και το μαθηματικό του μοντέλο. Σχήμα από [2]

Sigmoid: Η σιγμοειδής συνάρτηση έχει τη μορφή $\sigma(x) = 1/\left(1 + e^{-x}\right)$. Λαμβάνει ως είσοδο έναν οποιονδήποτε πραγματικό αριθμό και τον αντιστοιχεί στην έξοδο της στο διάστημα $[0, 1]$. Πρακτικά, μεγάλοι αρνητικοί αριθμοί αντιστοιχούν στο 0 και μεγάλοι θετικοί αριθμοί αντιστοιχούν στο 1.

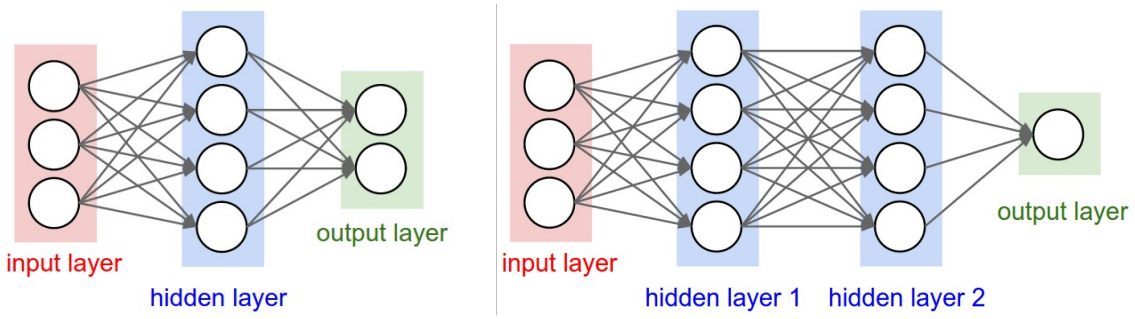
Tanh: Το τόξο εφαπτομένης, όπως είναι γνωστό από την τριγωνομετρία, μετατρέπει έναν οποιονδήποτε πραγματικό αριθμό σε μία τιμή στο διάστημα $[-1, 1]$. Στην πράξη προτιμάται από την σιγμοειδή συνάρτηση γιατί το εύρος τιμών της είναι κεντραρισμένο γύρω από το 0 (zero-centered) με αρνητικό και θετικό κατώφλι τη μονάδα. Αποτελεί ουσιαστικά μία επέκταση της σιγμοειδούς συνάρτησης στον αρνητικό άξονα με μεγαλύτερη κλίση, δηλαδή $\tanh(x) = 2\sigma(2x) - 1$.

ReLU: Η Ανορθωμένη Γραμμική Μονάδα (Rectified Linear Unit - ReLU) είναι ίσως η πιο δημοφιλής συνάρτηση ενεργοποίησης. Υπολογίζεται από τον τύπο $f(x) = max(0, x)$. Η έξοδος της, δηλαδή, είναι κατωφλιοποιημένη στο 0 ώστε να μην λαμβάνει αρνητικές τιμές για $x < 0$, ενώ είναι γραμμική για $x > 0$.

Αρχιτεκτονική ενός νευρωνικού δικτύου

Ένα νευρωνικό δίκτυο αποτελείται από συλλογές νευρώνων οι οποίες συνδέονται μεταξύ τους σε έναν (μη-κυκλικό) γράφο, δηλαδή, οι έξοδοι κάποιων νευρώνων είναι

οι είσοδοι κάποιων άλλων. Οι νευρώνες είναι οργανωμένοι σε διαφορετικά επίπεδα (layers) καθένα από τα οποία εκτελεί μία συγκεκριμένη διεργασία. Ο πιο κοινός τύπος επιπέδου είναι το *πλήρως συνδεδεμένο επίπεδο (fully-connected layer)*, στο οποίο όλοι νευρώνες του συνδέονται με κάθε νευρώνα του γειτονικού του layer ανά ζεύγη. Στο Σχήμα 3.3 απεικονίζονται δύο τοπολογίες νευρωνικών δικτύων, οι οποίες χρησιμοποιούν αποκλειστικά πλήρως συνδεδεμένα επίπεδα.



Σχήμα 3.3: Δύο τοπολογίες ενός νευρωνικού δικτύου που χρησιμοποιούν μόνο fully-connected layers. *Αριστερά:* Ένα νευρωνικό δίκτυο 3 εισόδων και 2 επιπέδων (1 «κρυμμένα» (hidden) layer με 4 νευρώνες & 1 layer εξόδου (output) με 2 νευρώνες). *Δεξιά:* Ένα νευρωνικό δίκτυο 3 εισόδων και 3 επιπέδων (2 «κρυμμένα» layers με 4 νευρώνες το καθένα & 1 νευρώνας εξόδου). Οι συνδέσεις γίνονται μεταξύ νευρώνων διαφορετικών επιπέδων και όχι ανάμεσα σε νευρώνες του ίδιου επιπέδου. Σχήμα από [2]

Ένα απλό νευρωνικό δίκτυο, χωρίς κρυμμένα επίπεδα (οι είσοδοι συνδέονται αμέσως στην έξοδο), αναφέρεται στη βιβλιογραφία με τον όρο *perceptron* και είναι ουσιαστικά ένας δυαδικός ταξινομητής, καθώς απεικονίζει την είσοδο $\mathbf{x} \in \mathbb{R}^n$ σε μία τιμή εξόδου $f(x) \in \{0, 1\}$. Δηλαδή:

$$f(x) = \begin{cases} 1, & \alpha\nu\ \mathbf{w} \cdot \mathbf{x} + \mathbf{b} > 0 \\ 0, & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases} \tag{3.14}$$

όπου $\mathbf{w} \in \mathbb{R}^n$ είναι τα βάρη των συνάψεων του νευρώνα και $\mathbf{b} \in \mathbb{R}^n$ μία σταθερά (bias) που καθορίζει το κατώφλι απόφασης. Τοπολογίες όπως αυτές του Σχήματος 3.3 μπορούν να θεωρηθούν σαν perceptrons με επιπλέον κρυμμένα επίπεδα, έτσι ένα τέτοιου είδους νευρωνικό δίκτυο εμφανίζεται συχνά στη βιβλιογραφία με τον όρο *perceptron πολλαπλών επιπέδων (multilayer perceptron - MLP)*.

Ο χρόνος εκπαίδευσης ενός νευρωνικού δικτύου εξαρτάται άμεσα από τον αριθμό των νευρώνων του. Η «γνώση» κάθε νευρώνα αντιπροσωπεύεται από τα βάρη w των συνάψεων του και τη σταθερά μετατόπισης b. Αυτές είναι οι παράμετροι που κάθε νευρώνας καλείται να «μάθει» κατά τη διαδικασία εκπαίδευσης του. Για παράδειγμα, το

πρώτο νευρωνικό δίκτυο (αριστερά) του Σχήματος 3.3 καλείται να «μάθει» $[3\times4]+[4\times2]=20$ βάρη και $4+2=6$ σταθερές μετατόπισης, δηλαδή συνολικά 26 παραμέτρους, ενώ το δεύτερο (δεξιά) $[3\times4]+[4\times4]+[4\times1]=32$ βάρη και $4+4+1=9$ σταθερές μετατόπισης, σύνολο 41 παραμέτρους.

Προ-επεξεργασία των δεδομένων

Τα δεδομένα, πριν δοθούν σαν είσοδοι σε ένα νευρωνικό δίκτυο, επεξεργάζονται κατάλληλα προκειμένου να κανονικοποιηθούν και να είναι αξιοποιήσιμα από τους νευρώνες. Έστω $X=[x_1,x_2,...,x_N]^T$, ένας $N\times D$ πίνακας δεδομένων, όπου N είναι το πλήθος των δεδομένων και D η διάσταση τους. Συνηθισμένες μέθοδοι επεξεργασίας είναι οι εξής:

Αφαίρεση μέσης τιμής (mean subtraction): Χρησιμοποιείται κατά κόρον στο στάδιο της προ-επεξεργασίας των δεδομένων και αφορά την αφαίρεση από κάθε δεδομένο $x_i, i=1,...,N$ της μέσης τιμής του. Γεωμετρικά, αυτή η αφαίρεση έχει την ερμηνεία της συλλογής των δεδομένων κοντά και γύρω από την κεντρική τους τιμή. Σε περίπτωση που τα δεδομένα είναι γκρίζες εικόνες, η μέση τιμή τους αφαιρείται από όλα τα pixels, ενώ για τις έγχρωμες η αφαίρεση αυτή πραγματοποιείται σε κάθε κανάλι χρώματος.

Κανονικοποίηση (normalization): Χρησιμοποιούνται δύο τρόποι για την κανονικοποίηση των δεδομένων: α) η διαίρεση κάθε διάστασης με την τυπική της απόκλιση και β) η κανονικοποίηση σε κάθε διάσταση ώστε η μικρότερη και η μεγαλύτερη τιμή της να είναι -1 και 1 αντίστοιχα. Αν τα δεδομένα είναι εικόνες, μία τέτοιου είδους κανονικοποίηση δεν είναι απαραίτητη.

Σε κάποιες περιπτώσεις, πραγματοποιείται μία μείωση της διάστασης των δεδομένων μέσω Ανάλυσης σε Πρωτεύουσες Συνιστώσες (PCA), έτσι ώστε να αφαιρεθεί τυχόν πλεονάζουσα πληροφορία και να είναι πιο διαχωρίσιμα.

Αρχικοποίηση του μοντέλου

Έπειτα, λοιπόν, από την προ-επεξεργασία των δεδομένων, πριν αρχίσει η εκπαίδευση του νευρωνικού δικτύου πρέπει να γίνει μια αρχικοποίηση στις παραμέτρους του. Απο τη στιγμή που τα δεδομένα είναι κανονικοποιημένα, θα μπορούσε κάποιος να κάνει μία «χονδροειδή» εκτίμηση ότι, έπειτα από την εκπαίδευση του δικτύου, περίπου τα μισά βάρη θα έχουν θετική τιμή και τα άλλα μισά αρνητική τιμή. Μία ιδέα, επομένως, θα ήταν να αρχικοποιηθούν όλα τα βάρη με μηδενική τιμή. Κάτι τέτοιο, ωστόσο, αποδεικνύεται λάθος, καθώς κάθε νευρώνας υπολογίζει αρχικά την ίδια έξοδο και κατ΄ επέκταση την ίδια παράγωγο κατά την εκτέλεση του αλγόριθμου backpropagation (βλ.

παρακάτω), οδηγώντας σε ίδιες τιμές παραμέτρων για όλους τους νευρώνες, το οποίο δεν είναι επιθυμητό. Αντίθετα, για τις σταθερές μετατόπισης, η αρχικοποίηση τους με μηδενικές τιμές δεν επιφέρει κάποιο πρόβλημα και είναι η πλέον συνηθισμένη.

Μία λύση στο παραπάνω πρόβλημα θα μπορούσε να είναι η αρχικοποίηση των βαρών με πολύ μικρές (κοντά στο 0) τυχαίες τιμές. Έτσι, κάθε νευρώνας θα έχει διαφορετικά βάρη και θα παράγει διαφορετική έξοδο. Η επιλογή των τιμών των βαρών μπορεί να γίνεται από μία γκαουσιανή κατανομή μηδενικής μέσης τιμής και μοναδιαίας τυπικής απόκλισης πολλαπλασιασμένη με έναν παράγοντα κοντά στο 0 (π.χ. $0.01 \times N(0,1)$). Το πρόβλημα, ωστόσο, με αυτή τη λύση είναι ότι οι τιμές των εξόδων του δικτύου αποκλίνουν όλο και περισσότερο μεταξύ τους όσο αυξάνεται το πλήθος των εισόδων.

Για να αντιμετωπιστεί αυτό, όπως πρότειναν οι Glorot & Bengio [64], μπορεί να εφαρμοστεί μια κανονικοποίηση στους νευρώνες του τελευταίου επιπέδου, έτσι ώστε το διάνυσμα των βαρών τους να διαιρείται με την τετραγωνική ρίζα του πλήθους των εισόδων του. Αυτό οδηγεί σε μία περίπου ίδια κατανομή των τιμών των εξόδων του δικτύου και βελτιώνει το ποσοστό σύγκλισης. Στο ίδιο μήκος κύματος, μία πιο πρόσφατη έρευνα των He et al. [65], στην οποία καταπιάστηκαν με τους νευρώνες που έχουν συνάρτηση ενεργοποίησης την ReLU, έδειξε ότι η διακύμανση των τιμών των νευρώνων πρέπει να είναι $2/n$, όπου n το πλήθος των εισόδων. Έτσι, έχουμε την αρχικοποίηση $w = N(0,1) \cdot \sqrt{2/n}$, η οποία συνοδευόμενη με ReLU νευρώνες χρησιμοποιείται κατά κόρον στις σύγχρονες αρχιτεκτονικές νευρωνικών δικτύων.

Μία τεχνική που αναπτύχθηκε πρόσφατα από τους Ioffe & Szegedy [66] και ονομάζεται *κανονικοποίηση παρτίδας*[1] *(batch normalization)*, λύνει πολλά προβλήματα που αφορούν την αρχικοποίηση των νευρωνικών δικτύων, υποχρεώνοντας, ουσιαστικά, τις ενεργοποιήσεις σε όλο το δίκτυο να υιοθετούν μία μοναδιαία γκαουσιανή κατανομή στην αρχή της εκπαίδευσης. Όσον αφορά την τοπολογία του δικτύου, η τεχνική αυτή υλοποιείται με την τοποθέτηση ενός BatchNorm επιπέδου αμέσως μετά τα fully-connected (ή τα συνελικτικά, όπως θα δούμε παρακάτω) επίπεδα και ακριβώς πριν τις συναρτήσεις ενεργοποίησης. Η τεχνική αυτή είναι πολύ δημοφιλής καθώς τα νευρωνικά δίκτυα που χρησιμοποιούν batch normalization είναι πολύ πιο εύρωστα σε λανθασμένες αρχικοποιήσεις.

Εκπαίδευση και αξιολόγηση του μοντέλου

Η εκπαίδευση ενός νευρωνικού δικτύου εξαρτάται τόσο από τα δεδομένα εισόδου όσο και από τις επιθυμητές τιμές εξόδου. Σε προβλήματα ταξινόμησης, όπως αυτό που εξετάζει η παρούσα Διπλωματική Εργασία, οι έξοδοι του δικτύου αντιπροσωπεύουν το χώρο κλάσεων και καθεμία περιέχει τη βεβαιότητα με την οποία το δεδομένο εισόδου

[1] Η ονομασία προκύπτει από την θεώρηση ενός layer ως μία μονάδα παραγωγής μίας παρτίδας εξόδων

να ανήκει στην αντίστοιχη κατηγορία (confidence score). Έτσι, τα επίπεδα του νευρωνικού δικτύου μπορούν να θεωρηθούν ως μέρη μίας συνάρτησης (score function), με παραμέτρους τα βάρη $\mathbf{w}$ και τις σταθερές μετατόπισης $\mathbf{b}$, η οποία αντιστοιχεί τις εισόδους σε βεβαιότητες κλάσεων (class scores). Το δίκτυο καλείται να μάθει αυτές τις παραμέτρους στο στάδιο της εκπαίδευσης. Η «ποιότητα» του συνόλου των παραμέτρων εξαρτάται από το κατά πόσον τα παραγόμενα confidence scores συμφωνούν με τις επισημειωμένες ετικέτες (ground-truth labels) των δεδομένων εκπαίδευσης και ποσοτικοποιείται από μία συνάρτηση κόστους (cost ή loss function). Σκοπός της εκπαίδευσης του δικτύου είναι η εύρεση, μέσω ενός αλγόριθμου βελτιστοποίησης, των παραμέτρων εκείνων που ελαχιστοποιούν τη συνάρτηση κόστους.

Για προβλήματα ταξινόμησης, δύο πολύ δημοφιλείς επιλογές για τη συνάρτηση κόστους είναι η *squared hinge loss* και ο *ταξινομητής Softmax*. Συγκεκριμένα, έστω $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ το σύνολο των δεδομένων εκπαίδευσης και $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N\}$ οι κλάσεις στις οποίες ανήκουν. Το συνολικό κόστος μπορεί να υπολογιστεί ως ο μέσος όρος των κόστεων όλων των δεδομένων, δηλαδή $L = \frac{1}{N}\sum_i L_i, i = 1, ..., N$. Αν θεωρήσουμε ότι $f = f_j(\mathbf{x}_i; \mathbf{W}), j = 1, ..., K$ (όπου K είναι το πλήθος των κλάσεων) είναι οι ενεργοποιήσεις του επιπέδου εξόδου του δικτύου, τότε οι δύο προαναφερθείσες συναρτήσεις κόστους υπολογίζονται από τους τύπους:

- **Squared hinge loss:** $L_i = \sum_{j \neq y_i} max(0, f_j - f_{y_i} + 1)^2$
- **Softmax classifier:** $L_i = -\log\left(\frac{e^{f_{y_i}}}{\sum_j e^{f_j}}\right) \iff L_i = -f_{y_i} + \log\sum_j e^{f_j}$

Για την ελαχιστοποίηση της συνάρτησης κόστους και κατ΄ επέκταση την εκπαίδευση ενός νευρωνικού δικτύου χρησιμοποιείται ο επαναληπτικός αλγόριθμος *απότομης καθόδου (gradient descent)*. Σε κάθε επανάληψη υπολογίζεται η παράγωγος της συνάρτησης κόστους και γίνεται μία ανανέωση των παραμέτρων. Σε πολλές απαιτητικές εφαρμογές, ωστόσο, το πλήθος των δεδομένων εκπαίδευσης είναι πάρα πολύ μεγάλο ($\sim 1M$) με συνέπεια ο υπολογισμός της συνάρτησης κόστους για όλα τα δεδομένα, σε κάθε επανάληψη, να είναι πολύ χρονοβόρος. Σε τέτοιες περιπτώσεις, προτιμάται ο υπολογισμός της παραγώγου σε μικρές ομάδες δεδομένων (mini batches) (συχνά ένα batch περιέχει 256 δείγματα από το σύνολο εκπαίδευσης), το οποίο οδηγεί σε μία καλή εκτίμηση της παραγώγου της συνολικής συνάρτησης κόστους με πολύ ταχύτερη σύγκλιση του αλγόριθμου απότομης καθόδου. Αυτή η διαδικασία ονομάζεται *Mini-Batch Gradient Descent (MBGD)*, ενώ στην ακραία περίπτωση που το mini-batch περιέχει μόνο ένα δείγμα εκπαίδευσης ονομάζεται *Stochastic Gradient Descent (SGD)*. Παρ'ολα αυτά, συχνά στη βιβλιογραφία οι δύο όροι συγχέονται, καθώς καθώς πολλοί συγγραφείς χρησιμοποιούν τον όρο SGD όταν αναφέρονται στην MBGD.

Ο υπολογισμός της παραγώγου της συνάρτησης κόστους γίνεται με τη μέθοδο *backpropagation* [67], η οποία αποτελεί μία αναδρομική εφαρμογή του *κανόνα της αλυ-*

σίδας. Σε συνδυασμό με τον αλγόριθμο SGD πραγματοποιεί τη διαδικασία εκπαίδευσης ενός νευρωνικού δικτύου, η οποία αποτελείται από δύο φάσεις:

- **Πρώτη φάση: Διάδοση (Propagation)**

  Το στάδιο της διάδοσης περιλαμβάνει τα ακόλουθα βήματα:

  1. *Εμπρόσθια διάδοση (forward propagation)* μίας εισόδου εκπαίδευσης μέσα στο νευρωνικό δίκτυο για την παραγωγή των εξόδων.
  2. *Οπίσθια διάδοση (backward propagation)* των ενεργοποιήσεων εξόδου διαμέσου του δικτύου προκειμένου να υπολογιστούν οι αποκλίσεις μεταξύ των επιθυμητών και των παραγόμενων εξόδων σε όλους τους νευρώνες (Δw_i).

- **Δεύτερη φάση: Ανανέωση των βαρών (Weight update)**

  Για κάθε βάρος w_i, ακολουθούνται τα ακόλουθα βήματα:

  1. Η απόκλιση του Δw_i πολλαπλασιάζεται με την ενεργοποίηση εισόδου του νευρώνα για να βρεθεί η παράγωγος του.
  2. Ένα ποσοστό από την παράγωγο του βάρους αφαιρείται από την τιμή του.

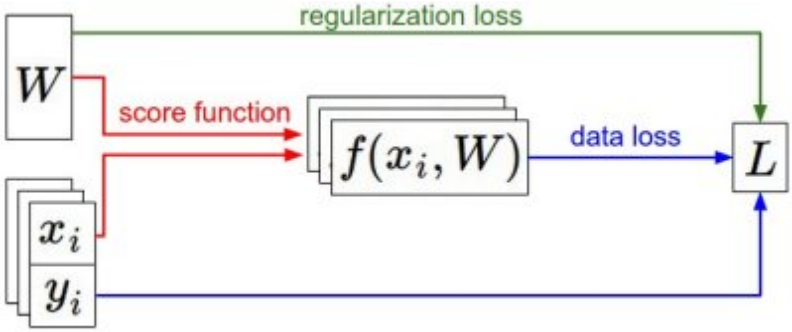
  Το ποσοστό αυτό επηρεάζει την ταχύτητα και την «ποιότητα» της εκπαίδευσης και ονομάζεται *ρυθμός μάθησης (learning rate)*. Όσο μεγαλύτερος είναι ο ρυθμός μάθησης τόσο ταχύτερη είναι η εκπαίδευση αλλά όσο μικρότερος είναι τόσο πιο ακριβής είναι η εκπαίδευση.

Οι δύο φάσεις επαναλαμβάνονται συνεχώς μέχρι όλα τα δείγματα του συνόλου εκπαίδευσης να ταξινομηθούν σωστά ή να ικανοποιηθεί κάποιο άλλο κριτήριο.

Έπειτα από την εκπαίδευση του δικτύου, έχει υπολογιστεί ένα σύνολο βαρών $\mathbf{W}$ το οποίο ιδανικά ταξινομεί σωστά κάθε δείγμα του συνόλου εκπαίδευσης (δηλαδή $L_i = 0$ για κάθε i). Το σύνολο αυτό $\mathbf{W}$ όμως δεν είναι απαραίτητα μοναδικό. Για παράδειγμα, αν πολλαπλασιαστεί με οποιονδήποτε παράγοντα $\lambda > 1$ η συνάρτηση κόστους παραμένει μηδενική οπότε και το διάνυσμα $\lambda\mathbf{W}$ ταξινομεί σωστά όλο το σύνολο εκπαίδευσης. Οι μεγάλες τιμές βαρών ωστόσο δεν είναι επιθυμητές καθώς οι αντίστοιχες είσοδοι επηρεάζουν πολύ περισσότερο τη διαμόρφωση των εξόδων και οδηγούν σε *υπερπροσαρμογή (overfitting)* του δικτύου στα δεδομένα, δηλαδή σε μείωση της *γενίκευσης* του δικτύου, που είναι μία ιδιότητα πολύ σημαντική. Έτσι, προκειμένου να προτιμηθούν μικρές τιμές βαρών, συνήθως λαμβάνεται υπόψη στη συνάρτηση κόστους και μία «ποινή» κανονικοποίησης (regularization penalty) $R(\mathbf{W})$ ως εξής:

$$L = \underbrace{\frac{1}{N}\sum_i L_i}_{data\ loss} + \underbrace{kR(\mathbf{W})}_{regularization\ loss} \tag{3.15}$$

όπου k μία παράμετρος η οποία συνήθως καθορίζεται πειραματικά. Στο Σχήμα 3.4 φαίνεται το διάγραμμα ροής της πληροφορίας μέσα στο δίκτυο έπειτα και από την προσθήκη του παράγοντα κανονικοποίησης στη συνάρτηση κόστους. Οι πιο δημοφιλείς



Σχήμα 3.4: Διάγραμμα ροής της πληροφορίας μέσα σε ένα νευρωνικό δίκτυο. Τα ζεύγη δεδομένων $(\mathbf{x}, \mathbf{y})$ είναι γνωστά. Τα βάρη $\mathbf{W}$ αρχικοποιούνται με τυχαίες τιμές και μπορούν να αλλάξουν. Κατά τη διάρκεια της εμπρόσθιας διάδοσης οι νευρώνες υπολογίζουν τα class scores στο διάνυσμα $\mathbf{f}$. Η συνάρτηση κόστους περιλαμβάνει δύο όρους: τις αποκλίσεις των σκορ του $\mathbf{f}$ από τα δεδομένα $\mathbf{y}$ (data loss) και την «ποινή» κανονικοποίησης (regularization loss) η οποία εξαρτάται μόνο από τα βάρη. Σχήμα από [2]

τεχνικές κανονικοποίησης είναι οι ακόλουθες:

L2 regularization: Είναι ίσως η πιο κοινή μορφή κανονικοποίησης και πραγματοποιείται με την πρόσθεση της «ποινής» $R(\mathbf{W}) = \frac{1}{2}\sum_k \sum_l W_{k,l}^2$ στη συνάρτηση κόστους. Δηλαδή, για κάθε βάρος $\mathbf{w}$ προστίθεται ο όρος $\frac{1}{2}k\mathbf{w}$ στη συνάρτηση κόστους. Με αυτό τον τρόπο, κατά τη διάρκεια της ανανέωσης των παραμέτρων στον αλγόριθμο εκπαίδευσης, η τιμή κάθε βάρους w φθίνει γραμμικά προς το μηδέν: $w = w - kw$.

L1 regularization: Είναι μία λιγότερο δημοφιλής μορφή κανονικοποίησης σύμφωνα με την οποία, για κάθε βάρος $\mathbf{w}$ προστίθεται ο όρος $k\,|\mathbf{w}|$ στη συνάρτηση κόστους. Μπορεί να συνδυαστεί με την $L2$ κανονικοποίηση ως $k_1\,|\mathbf{w}| + \frac{1}{2}k_2\mathbf{w}$, η οποία ονομάζεται elastic net regularization [68]. Η $L1$ κανονικοποίηση οδηγεί σε διανύσματα $\mathbf{w}$ τα οποία γίνονται αραιά (sparse) κατά τη διάρκεια της βελτιστοποίησης. Δηλαδή, τελικά οι νευρώνες καταλήγουν να χρησιμοποιούν μόνο ένα αραιό υποσύνολο με τις πιο σημαντικές εισόδους τους και να είναι σχεδόν αδιάφοροι για τις υπόλοιπες. Παρόλα αυτά, αν δεν προκύπτει ενδιαφέρον για τη διάκριση συγκεκριμένων χαρακτηριστικών, προτιμάται η $L2$ σαν μέθοδος κανονικοποίησης λόγω πολύ μεγαλύτερης αποδοτικότητας.

Max norm constraints: Είναι μια μορφή κανονικοποίησης που έχει επιδείξει βελτιώσεις με τη χρήση της και αφορά την επιβολή ενός ανώτατου ορίου στο μέτρο του

διανύσματος βαρών κάθε νευρώνα. Έτσι, έπειτα από την ανανέωση των παραμέτρων κατά την εκπαίδευση, κάθε διάνυσμα βαρών $\vec{\mathbf{w}}$ τροποποιείται έτσι ώστε να ικανοποιείται $\|\vec{\mathbf{w}}\|_2 < c$, όπου η τιμή του c είναι συνήθως παράγωγο του 3 ή του 4. Με αυτόν τον τρόπο, το δίκτυο δεν μπορεί να «εκραγεί» ακόμα και σε πολύ υψηλούς ρυθμούς μάθησης, γιατί οι παράμετροι είναι συνεχώς οριοθετημένες.

Dropout: Είναι μία απλή και πάρα πολύ αποτελεσματική τεχνική κανονικοποίησης η οποία προτάθηκε πρόσφατα από τους Srivastana et al. [69] και συμπληρώνει τις τρεις παραπάνω μεθόδους. Σύμφωνα με αυτή, κατά τη διάρκεια της εκπαίδευσης, ένας νευρώνας διατηρείται ενεργός με μία πιθανότητα $\mathbf{p}$, αλλιώς τίθεται στο μηδέν. Η έξοδος $\mathbf{y}$ ενός ενεργού νευρώνα, μετά το dropout διαμορφώνεται ως $\mathbf{py} + (1 - \mathbf{p})0$. Έτσι κατά τη διαδικασία αξιολόγησης του δικτύου, όπου όλοι οι νευρώνες είναι ενεργοί, πρέπει να γίνει η αντιστοίχηση $\mathbf{y} \rightarrow \mathbf{py}$, δηλαδή οι ενεργοποιήσεις εξόδου να κλιμακωθούν με την πιθανότητα $\mathbf{p}$. Αντί αυτού, για λόγους αποδοτικότητας, προτιμάται η τεχνική του *ανεστραμμένου dropout*, η οποία πραγματοποιεί την αντίστροφη κλιμάκωση κατά τη διάρκεια της εκπαίδευσης διατηρώντας την εμπρόσθια διάδοση κατά την αξιολόγηση αμετάβλητη.

3.2.2 Συνελικτικά Νευρωνικά Δίκτυα

Οι τεχνικές βαθιάς μάθησης χρησιμοποιούνται κατά κόρον τα τελευταία χρόνια για προβλήματα Όρασης Υπολογιστών λόγω της ραγδαίας ανάπτυξης των Συνελικτικών Νευρωνικών Δικτύων (ΣΝΔ). Οι μεγάλες απαιτήσεις τους σε υπολογιστικό χρόνο και κόστος αποτελούσαν για πολλά χρόνια εμπόδιο στη χρήση και την εξέλιξη τους. Οι σύγχρονες κάρτες γραφικών, όμως, κατάφεραν να υπερκεράσουν αυτό το εμπόδιο με την πολυπύρηνη αρχιτεκτονική τους η οποία οδηγεί σε πολύ ταχύτερη εκτέλεση υπολογισμών, μειώνοντας δραστικά τους χρόνους εκπαίδευσης και αξιολόγησης ενός ΣΝΔ.

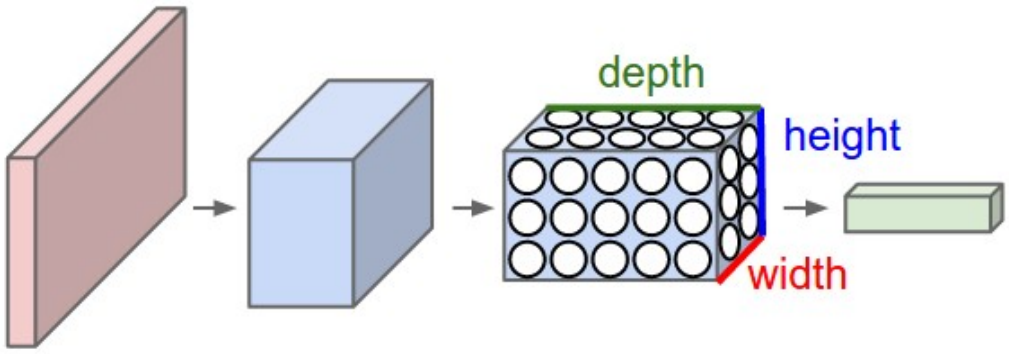
Τα συνελικτικά νευρωνικά δίκτυα είναι όμοια με τα κοινά νευρωνικά δίκτυα που περιγράψαμε παραπάνω: είναι φτιαγμένα από νευρώνες που διαθέτουν εκπαιδεύσιμα βάρη και σταθερές μετατόπισης. Κάθε νευρώνας λαμβάνει εισόδους και υπολογίζει ένα εσωτερικό γινόμενο, στο οποίο προαιρετικά εφαρμόζεται μία μη-γραμμική ενεργοποίηση. Το σύνολο του δικτύου εκφράζει μία παραγωγίσιμη συνάρτηση (score function), η οποία αντιστοιχεί τα pixels μίας εικόνας σε βεβαιότητες κλάσεων (class scores), ενώ μία συνάρτηση κόστους (loss function) στο τελευταίο (fully-connected) επίπεδο ποσοτικοποιεί την «ποιότητα» του συνόλου των παραμέτρων. Η διαφορά σε σχέση με τα κοινά νευρωνικά δίκτυα έγκειται στη σαφή παραδοχή ότι οι είσοδοι των ΣΝΔ είναι εικόνες, το οποίο επιτρέπει την κωδικοποίηση ορισμένων ιδιοτήτων μέσα στην αρχιτεκτονική του δικτύου, οι οποίες καθιστούν την εμπρόσθια διάδοση πολύ πιο αποδοτική

και μειώνουν σημαντικά το πλήθος των παραμέτρων του δικτύου.

Αρχιτεκτονική ενός Συνελικτικού Νευρωνικού Δικτύου

Όπως περιγράψαμε παραπάνω, στα κοινά νευρωνικά δίκτυα κάθε νευρώνας ενός κρυμμένου επιπέδου (hidden layer) είναι πλήρως συνδεδεμένος (fully connected) με όλους τους νευρώνες του προηγούμενου επιπέδου και λειτουργεί εντελώς ανεξάρτητα από τους υπόλοιπους νευρώνες του ίδιου επιπέδου χωρίς να συνδέεται μαζί τους. Μία τέτοια αρχιτεκτονική, ωστόσο, όταν καλείται να επεξεργαστεί στην είσοδο της εικόνες, ο αριθμός των παραμέτρων που πρέπει να εκπαιδεύσει είναι συνήθως πάρα πολύ μεγάλος. Για παράδειγμα, για μία έγχρωμη εικόνα $200 \times 200 \times 3$ ένας νευρώνας του πρώτου κρυμμένου επιπέδου θα έχει $200 \cdot 200 \cdot 3 = 120000$ βάρη, ενώ η παρουσία αρκετών τέτοιων νευρώνων σε επόμενα επίπεδα θα είναι απαραίτητη για την γρήγορη ανανέωση των παραμέτρων. Είναι ξεκάθαρο, λοιπόν, ότι αυτή η πλήρης συνδεσιμότητα των νευρώνων είναι μη αποδοτική ενώ ο τεράστιος αριθμός παραμέτρων οδηγεί πολύ γρήγορα σε υπερπροσαρμογή (overfitting).

Τα συνελικτικά νευρωνικά δίκτυα, αντίθετα, διαθέτουν επίπεδα στα οποία οι νευρώνες είναι οργανωμένοι σε 3 διαστάσεις: *πλάτος, ύψος, βάθος*, κάτι το οποίο έρχεται σε συνάρτηση με την τρισδιάστατη φύση των εικόνων. Όπως θα αναλύσουμε και παρακάτω, οι νευρώνες ενός επιπέδου, αντίθετα με τη λογική της πλήρους συνδεσιμότητας, είναι συνδεδεμένοι μόνο με μία μικρή περιοχή του προηγούμενου επιπέδου, ενώ το τελικό layer εξόδου έχει διαστάσεις $1 \times 1 \times K$, όπου K το πλήθος των κλάσεων, προκειμένου η εικόνα εισόδου να αντιστοιχηθεί σε ένα διάνυσμα με class scores, διατεταγμένα στη διάσταση του βάθους. Στο Σχήμα 3.5 απεικονίζεται η τοπολογία ενός ΣΝΔ.



Σχήμα 3.5: Η αρχιτεκτονική ενός Συνελικτικού Νευρωνικού Δικτύου. Το μοβ επίπεδο εισόδου περιέχει την εικόνα, άρα το πλάτος και το ύψος του είναι οι διαστάσεις της, ενώ το βάθος του είναι 3, όσα και τα κανάλια χρώματος. Σχήμα από [2]

Επίπεδα κατασκευής ενός Συνελικτικού Νευρωνικού Δικτύου

Τα επίπεδα ενός Συνελικτικού Νευρωνικού Δικτύου μετατρέπουν έναν 3Δ όγκο εισόδου σε έναν 3Δ όγκο εξόδου μέσω μίας διαφορίσιμης συνάρτησης. Για την κατασκευή ενός ΣΝΔ χρησιμοποιούνται κυρίως τρεις τύποι επιπέδων: *Συνελικτικό Επίπεδο (Convolutional Layer)*, *Συγκεντρωτικό Επίπεδο (Pooling Layer)* και *Πλήρως-Συνδεδεμένο Επίπεδο (Fully-Connected Layer)*. Τα επίπεδα αυτά «στοιβάζονται» με συγκεκριμένο τρόπο για να συνθέσουν ένα ΣΝΔ. Ένα παράδειγμα μίας τυπικής αρχιτεκτονικής ενός ΣΝΔ είναι η εξής: [INPUT - CONV - RELU - POOL - FC], όπου:

- INPUT: περιλαμβάνει τις τιμές των pixels της εικόνας εισόδου. Ας θεωρήσουμε π.χ. μία είσοδο με πλάτος 32, ύψος 32 και 3 κανάλια χρώματος R,G,B ($32 \times 32 \times 3$).

- CONV: είναι το συνελικτικό επίπεδο που παράγει τις εξόδους των νευρώνων οι οποίοι είναι συνδεδεμένοι σε μικρές περιοχές της εισόδου. Κάθε νευρώνας υπολογίζει το εσωτερικό γινόμενο μεταξύ των βαρών του και της περιοχής της εικόνας με την οποία συνδέεται. Οι διαστάσεις του όγκου εξόδου αυτού του επιπέδου εξαρτώνται από το πλήθος D των φίλτρων που επιλέγονται και στο συγκεκριμένο παράδειγμα είναι $32 \times 32 \times D$.

- RELU: είναι το επίπεδο που εφαρμόζεται η συνάρτηση ενεργοποίησης, π.χ. $max(0, x)$. Το μέγεθος του 3Δ όγκου παραμένει αναλλοίωτο.

- POOL: είναι το συγκεντρωτικό επίπεδο, το οποίο πραγματοποιεί μία υποδειγματοληψία στις χωρικές διαστάσεις (πλάτος, ύψος), εξάγωντας έναν 3Δ όγκο διαστάσεων $16 \times 16 \times D$.

- FC: είναι το πλήρως-συνδεδεμένο επίπεδο το οποίο υπολογίζει τις βεβαιότητες κλάσεων, δίνοντας μία έξοδο διαστάσεων $1 \times 1 \times K$, όπου K το πλήθος των κλάσεων.

Τα CONV/FC επίπεδα περιέχουν παραμέτρους (τα βάρη και τις σταθερές μετατόπισης των νευρώνων τους), οι οποίες εκπαιδεύονται με τη μέθοδο της απότομης καθόδου (gradient descent) έτσι ώστε τα class scores που υπολογίζονται από το δίκτυο να είναι συνεπή με τις επισημειωμένες ετικέτες των εικόνων του συνόλου εκπαίδευσης. Τα RELU/POOL επίπεδα εφαρμόζουν μία προκαθορισμένη, μη παραμετρική συνάρτηση. Παρακάτω γίνεται μία αναλυτική περιγραφή της δομής και της λειτουργίας κάθε επιπέδου.

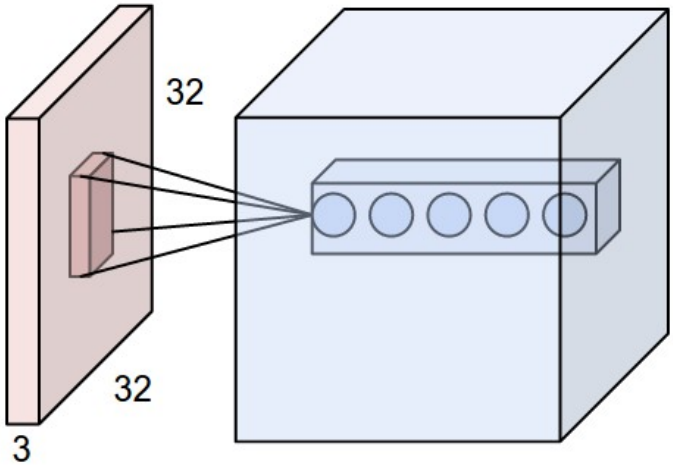
Συνελικτικό Επίπεδο (Convolutional Layer)

Το συνελικτικό επίπεδο είναι ο πυρήνας του Συνελικτικού Νευρωνικού Δικτύου και αναλαμβάνει το μεγαλύτερο υπολογιστικό βάρος. Οι παράμετροι του αφορούν ένα σύνολο εκπαιδεύσιμων φίλτρων, τα οποία είναι μικρά χωρικά αλλά εκτείνονται σε όλο το βάθος του όγκου εισόδου (π.χ. ένα φίλτρο $5 \times 5 \times 3$). Κατά την εμπρόσθια διάδοση (forward pass) κάθε φίλτρο «σαρώνει» την εικόνα και υπολογίζονται τα γινόμενα ανάμεσα στις τιμές του φίλτρου και των αντίστοιχων θέσεων της εικόνας (εκτελείται δηλαδή μία πράξη *συνέλιξης* ανάμεσα στο φίλτρο και την εικόνα). Το αποτέλεσμα αυτής της διαδικασίας είναι ένας δισδιάστατος χάρτης ενεργοποίησης ο οποίος περιέχει τις αποκρίσεις του φιλτραρίσματος σε κάθε θέση. Έπειτα από την εξαγωγή όλων των 2Δ χαρτών ενεργοποίησης, αυτοί «στοιβάζονται» κατά μήκος της διάστασης του βάθους και παράγουν τον 3Δ όγκο εξόδου. Κάθε εγγραφή του 3Δ αυτού όγκου μπορεί να ερμηνευτεί ως η έξοδος ενός νευρώνα ο οποίος «βλέπει» μόνο σε μία μικρή περιοχή της εισόδου και μοιράζεται τις ίδιες παραμέτρους με τους *χωρικά* γειτονικούς του νευρώνες (αφού κάθε τιμή προκύπτει εφαρμόζοντας το ίδιο φίλτρο). Στη συνέχεια αναλύονται λεπτομερώς η συνδεσιμότητα των νευρώνων, η διάταξη τους στο χώρο και ο τρόπος διαμοιρασμού των παραμέτρων.

Τοπική συνδεσιμότητα: Όπως αναφέραμε παραπάνω, σε περιπτώσεις εισόδων μεγάλων διαστάσεων όπως οι εικόνες, είναι ανώφελο να χρησιμοποιούνται πλήρως συνδεδεμένοι νευρώνες. Έτσι, ένα συνελικτικό επίπεδο περιέχει νευρώνες που συνδέονται με μία μικρή περιοχή του όγκου εισόδου. Η χωρική έκταση αυτή της σύνδεσης είναι μία υπερ-παράμετρος που ονομάζεται *ευαίσθητο πεδίο (receptive field)* του νευρώνα (αντίστοιχα το μέγεθος του φίλτρου). Η έκταση της ως προς το βάθος είναι ίση με το βάθος του όγκου εισόδου. Για παράδειγμα, αν η εικόνα εισόδου είναι διαστάσεων $32 \times 32 \times 3$ και το ευαίσθητο πεδίο είναι 5×5, τότε κάθε νευρώνας στο συνελικτικό επίπεδο θα περιέχει βάρη σε μία περιοχή $5 \times 5 \times 3$, δηλαδή 75 βάρη ($+1$ σταθερά μετατόπισης), ενώ αν η είσοδος έχει διαστάσεις $16 \times 16 \times 20$, τότε με το ίδιο ευαίσθητο πεδίο κάθε νευρώνας θα έχει $5 \times 5 \times 20 = 500$ βάρη. Στο Σχήμα 3.6 απεικονίζεται σχηματικά η έννοια του ευαίσθητου πεδίου.

Χωρική διάταξη: Η διάταξη των νευρώνων του συνελικτικού επιπέδου εξαρτάται από το μέγεθος του όγκου εξόδου. Τρεις υπερ-παράμετροι καθορίζουν το πλήθος των νευρώνων της εξόδου: το *βάθος (depth)*, το *βήμα (stride)* και το *μηδενικό περιθώριο (zero-padding)*, όπως αναλύονται παρακάτω:

- Το *βάθος* του όγκου εξόδου αντιστοιχεί στο πλήθος των φίλτρων που χρησιμοποιείται, καθένα από τα οποία εκπαιδεύεται για να εντοπίζει κάτι διαφορετικό στην εικόνα. Για παράδειγμα, αν το πρώτο συνελικτικό επίπεδο λαμβάνει ως



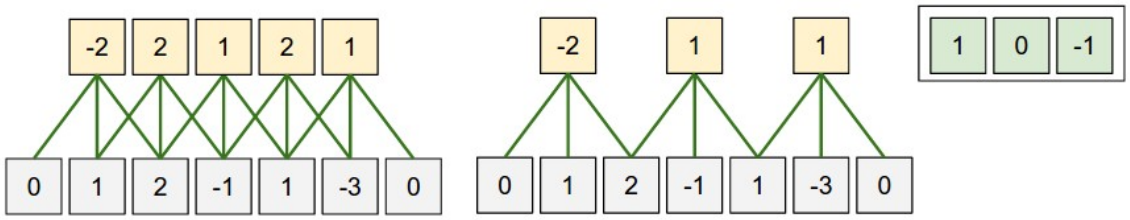
Σχήμα 3.6: Ένα παράδειγμα σύνδεσης των νευρώνων με το ευαίσθητο πεδίο τους. Κάθε νευρώνας στο συνελικτικό επίπεδο συνδέεται χωρικά μόνο με μία μικρή περιοχή του 3Δ όγκου εισόδου αλλά σε όλη την έκταση του βάθους του. Στην ίδια περιοχή «βλέπουν» όλοι οι νευρώνες της στοίβας (εδώ 5) που εφαρμόζουν διαφορετικά φιλτραρίσματα. Σχήμα από [2]

είσοδο μία εικόνα, τότε διαφορετικοί νευρώνες μπορεί να ενεργοποιηθούν στη διάσταση του βάθους, κατά την παρουσία διαφορετικών χαρακτηριστικών. Το σύνολο των νευρώνων που «βλέπουν» στην ίδια περιοχή της εισόδου ονομάζεται *στήλη ή στοίβα βάθους*.

- Το *βήμα* αφορά το ρυθμό «σάρωσης» της εικόνας από το φίλτρο. Αν η τιμή του είναι 1, τότε τα φίλτρα μετακινούνται κατά ένα pixel κάθε φορά, ενώ αν είναι 2 τότε τα φίλτρα προσπερνάνε 2 pixels σε κάθε «σάρωση», οδηγώντας σε μικρότερους, χωρικά, όγκους στην έξοδο. Τιμές μεγαλύτερες του 2 χρησιμοποιούνται σπάνια στην πράξη.

- Αρκετές φορές είναι χρήσιμο να επεκτείνονται τα όρια της εισόδου με μηδενικά. Το μέγεθος αυτής της επέκτασης είναι το *μηδενικό περιθώριο* και επιτρέπει στο χρήστη να καθορίσει τις χωρικές διαστάσεις της εξόδου (συνήθως χρησιμοποιείται για να διατηρηθεί το μέγεθος της εισόδου ατόφιο και στην έξοδο).

Οι χωρικές διαστάσεις της εξόδου μπορούν να υπολογιστούν ως συνάρτηση του μεγέθους της εισόδου $\mathbf{M}$, του ευαίσθητου πεδίου $\mathbf{F}$ των νευρώνων του συνελικτικού επιπέδου, του βήματος $\mathbf{S}$ με το οποίο εφαρμόζονται και της ποσότητας του μηδενικού περιθωρίου $\mathbf{P}$ ως $(\mathbf{M} - \mathbf{F} + \mathbf{2P})/\mathbf{S} + \mathbf{1}$. Για παράδειγμα, για μία 7×7 είσοδο και ένα 3×3 φίλτρο με βήμα 1 και «γέμισμα» 0 θα πάρουμε μία 5×5 έξοδο. Αν το βήμα είναι 2 τότε θα πάρουμε μία 3×3 έξοδο. Μία γραφική αναπαράσταση για το πως επηρεάζει το *βήμα* την έξοδο φαίνεται στο Σχήμα 3.7.



Σχήμα 3.7: Απεικόνιση της χωρικής διάταξης της εξόδου σε μία διάσταση. Στο συγκεκριμένο παράδειγμα, υπάρχει ένας νευρώνας με ευαίσθητο πεδίο $F = 3$, ενώ η είσοδος έχει μέγεθος $M = 5$ και zero padding $P = 1$. Τα βάρη του νευρώνα είναι $[1, 0, -1]$, όπως φαίνονται στην άκρη δεξιά. *Αριστερά:* Για βήμα $S = 1$ παίρνουμε μία έξοδο μεγέθους $(5-3+2)/1+1 = 5$. *Δεξιά:* Για βήμα $S = 2$ παίρνουμε μία έξοδο μεγέθους $(5 - 3 + 2)/2 + 1 = 3$. Σχήμα από [2]

Στο παράδειγμα του Σχήματος 3.7 στα αριστερά, η διάσταση της εξόδου είναι ίδια με αυτή της εισόδου, κάτι που είναι επιθυμητό στην πλειονότητα των εφαρμογών. Αυτό συνέβη λόγω της προσθήκης μηδενικού περιθωρίου $P = 1$. Γενικά, όταν το βήμα είναι $S = 1$, για να εξασφαλίσουμε ότι οι χωρικές διαστάσεις της εισόδου και της εξόδου θα είναι ίδιες, προσθέτουμε στην είσοδο μηδενικό περιθώριο μεγέθους $P = (F - 1)/2$.

Ένα πιο ρεαλιστικό παράδειγμα είναι η αρχιτεκτονική των Krizhevsky et al.[2] [70], η οποία στην είσοδο της δέχεται εικόνες μεγέθους $227 \times 227 \times 3$. Το πρώτο συνελικτικό επίπεδο χρησιμοποιεί νευρώνες με ευαίσθητο πεδίο $F = 11$, βήμα $S = 4$ και χωρίς zero padding ($P = 0$), ενώ το βάθος του είναι ίσο με $D = 96$. Αφού $(227-11)/4+1 = 55$, ο όγκος εξόδου έχει μέγεθος $55 \times 55 \times 96$, δηλαδή 290400 νευρώνες. Κάθε ένας από αυτούς τους 290400 νευρώνες είναι συνδεδεμένος με μία μικρή περιοχή της εικόνας μεγέθους $11 \times 11 \times 3$, ενώ όλοι οι 96 νευρώνες κάθε στοίβας βάθους είναι συνδεδεμένοι στην ίδια περιοχή, με διαφορετικά όμως βάρη.

Διαμοιρασμός παραμέτρων: Στα Συνελικτικά Νευρωνικά Δίκτυα χρησιμοποιείται μία μέθοδος διαμοιρασμού των παραμέτρων μεταξύ των νευρώνων προκειμένου να μειωθεί σημαντικά ο αριθμός τους. Αναλύοντας το παραπάνω ρεαλιστικό παράδειγμα παρατηρούμε ότι υπάρχουν 290400 νευρώνες στο πρώτο συνελικτικό επίπεδο και καθένας περιέχει $11 \cdot 11 \cdot 3 = 363$ βάρη και 1 σταθερά μετατόπισης. Συνολικά, δηλαδή, το πρώτο συνελικτικό επίπεδο περιλαμβάνει $290400 \cdot 363 = 105.705.600$ παραμέτρους. Αυτός ο πολύ μεγάλος αριθμός παραμέτρων, ωστόσο, μπορεί να μειωθεί δραματικά κάνοντας μία εύλογη παραδοχή: Αν ένα χαρακτηριστικό είναι χρήσιμο να υπολογιστεί σε μία θέση (x, y), τότε θα είναι εξίσου χρήσιμο να υπολογιστεί και σε μία διαφορετική θέση (x_2, y_2). Με άλλα λόγια, θεωρώντας ότι το βάθος ενός συνελικτικού δικτύου αποτελείται από στοιβαγμένες δισδιάστατες επιφάνειες (*depth slices*) (ένας όγκος δια-

[2]Νικητές του διαγωνισμού Imagenet το 2012

στάσεων $55 \times 55 \times 96$ περιλαμβάνει 96 επιφάνειες βάθους μεγέθους 55×55 η καθεμία), οι νευρώνες κάθε επιφάνειας βάθους περιορίζονται έτσι ώστε να περιέχουν ίδια βάρη και σταθερές μετατόπισης. Με αυτόν τον τρόπο, το πρώτο συνελικτικό επίπεδο του προηγούμενου παραδείγματος θα έχει μόνο 96 διαφορετικά σύνολα παραμέτρων (1 για κάθε επιφάνεια βάθους), δηλαδή $96 \cdot 11 \cdot 11 \cdot 3 = 34848$ βάρη και 96 σταθερές μετατόπισης, συνολικά 34944 παραμέτρους. Στην πράξη, κατά την εκτέλεση του αλγόριθμου backpropagation, κάθε νευρώνας του συνελικτικού επιπέδου υπολογίζει την παράγωγο των βαρών του η οποία όμως προστίθεται κατά μήκος όλης της επιφάνειας βάθους που ανήκει και πραγματοποιείται ανανέωση των τιμών ενός μόνο συνόλου βαρών ανά επιφάνεια. Παρατηρούμε, λοιπόν, ότι η εμπρόσθια διάδοση ενός συνελικτικού επιπέδου σε κάθε επιφάνεια βάθους μπορεί να υπολογιστεί ως η *συνέλιξη* των βαρών ενός νευρώνα με την είσοδο. Γι΄ αυτό, πολλές φορές στη βιβλιογραφία το σύνολο των βαρών ενός νευρώνα αναφέρεται ως *φίλτρο*, το οποίο συνελίσσεται με την είσοδο.

Ανακεφαλαιώνοντας, το συνελικτικό επίπεδο:

- Δέχεται ως είσοδο έναν 3Δ όγκο $\mathbf{W_1} \times \mathbf{H_1} \times \mathbf{D_1}$
- Χρειάζεται τέσσερις υπερ-παραμέτρους:
  - το πλήθος των φίλτρων $\mathbf{K}$
  - τη χωρική τους έκταση $\mathbf{F}$
  - το βήμα $\mathbf{S}$
  - την ποσότητα του μηδενικού περιθώριου $\mathbf{P}$
- Παράγει έναν 3Δ όγκο $\mathbf{W_2} \times \mathbf{H_2} \times \mathbf{D_2}$ όπου:
  - $\mathbf{W_2} = (\mathbf{W_1} - \mathbf{F} + \mathbf{2P})/\mathbf{S} + \mathbf{1}$
  - $\mathbf{H_2} = (\mathbf{H_1} - \mathbf{F} + \mathbf{2P})/\mathbf{S} + \mathbf{1}$
  - $\mathbf{D_2} = \mathbf{K}$
- Λόγω του διαμοιρασμού των παραμέτρων, εισάγει στο δίκτυο $\mathbf{F} \cdot \mathbf{F} \cdot \mathbf{D_1}$ βάρη ανά φίλτρο, δηλαδή συνολικά $(\mathbf{F} \cdot \mathbf{F} \cdot \mathbf{D_1}) \cdot \mathbf{K}$ βάρη και $\mathbf{K}$ σταθερές μετατόπισης.
- Στον όγκο εξόδου, η $\mathbf{d}$-οστή επιφάνεια βάθους (μεγέθους $W_2 \times H_2$) είναι το αποτέλεσμα της συνέλιξης του $\mathbf{d}$-οστού φίλτρου με τον όγκο εισόδου με βήμα $\mathbf{S}$, σταθμισμένο με την $\mathbf{d}$-οστή σταθερά μετατόπισης.

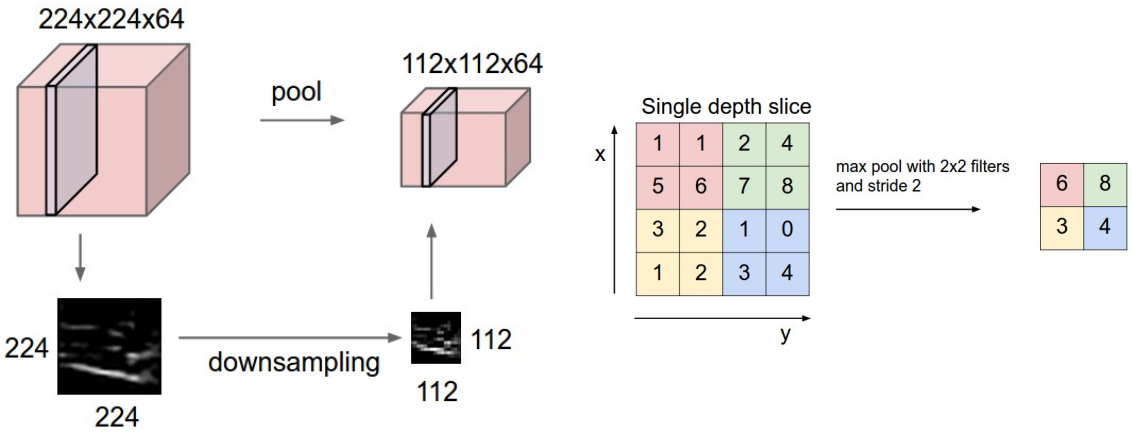
Συγκεντρωτικό Επίπεδο (Pooling Layer)

Μία κοινή τεχνική κατασκευής ενός Συνελικτικού Νευρωνικού Δικτύου είναι η περιοδική τοποθέτηση ενός συγκεντρωτικού επιπέδου ανάμεσα σε διαδοχικά συνελικτικά επίπεδα. Ο ρόλος του είναι να μειώνει προοδευτικά τις χωρικές διαστάσεις της αναπαράστασης και, συνεπώς, να ελαττώνει το πλήθος των παραμέτρων του δικτύου, υποβαθμίζοντας έτσι την πιθανότητα της υπερπροσαρμογής (overfitting). Το συγκεντρωτικό επίπεδο εφαρμόζεται ξεχωριστά σε κάθε επιφάνεια βάθους της εισόδου και τροποποιεί το χωρικό της μέγεθος, χρησιμοποιώντας συνήθως την εύρεση της μέγιστης τιμής μιας γειτονιάς. Η πιο δημοφιλής μορφή είναι ένα συγκεντρωτικό επίπεδο με φίλτρα μέγιστης τιμής, μεγέθους 2×2 με βήμα 2, τα οποία υποδειγματοληπτούν κάθε επιφάνεια βάθους της εισόδου, απορρίπτοντας το 75% των ενεργοποιήσεων της, όπως φαίνεται στο Σχήμα 3.8. Η διάσταση του βάθους παραμένει αναλλοίωτη. Δηλαδή, το συγκεντρωτικό επίπεδο:

- Δέχεται ως είσοδο έναν 3Δ όγκο $\mathbf{W_1} \times \mathbf{H_1} \times \mathbf{D_1}$
- Χρειάζεται δύο υπερ-παραμέτρους:
  - τη χωρική έκταση των φίλτρων του $\mathbf{F}$
  - το βήμα $\mathbf{S}$
- Παράγει έναν 3Δ όγκο $\mathbf{W_2} \times \mathbf{H_2} \times \mathbf{D_2}$ όπου:
  - $\mathbf{W_2} = (\mathbf{W_1} - \mathbf{F})/\mathbf{S} + \mathbf{1}$
  - $\mathbf{H_2} = (\mathbf{H_1} - \mathbf{F})/\mathbf{S} + \mathbf{1}$
  - $\mathbf{D_2} = \mathbf{D_1}$
- Δεν επιβαρύνει με επιπλέον παραμέτρους το δίκτυο αφού εκτελεί μία συγκεκριμένη πράξη στην είσοδο

Πλήρως-Συνδεδεμένο Επίπεδο (Fully-connected Layer)

Οι νευρώνες ενός πλήρους συνδεδεμένου επιπέδου συνδέονται με όλες τις ενεργοποιήσεις του προηγούμενου επιπέδου, όπως στα κοινά νευρωνικά δίκτυα (βλ. παραπάνω). Οι έξοδοι τους υπολογίζονται ως το γινόμενο των βαρών με τις εισόδους τους, μετατοπισμένο κατά μία σταθερά. Τα πλήρως συνδεδεμένα επίπεδα χρησιμοποιούνται συνήθως στο τελικό στάδιο εξόδου ενός Συνελικτικού Νευρωνικού Δικτύου, είτε σαν διανύσματα χαρακτηριστικών είτε σαν ενεργοποιήσεις του ταξινομητή Softmax για τον υπολογισμό των class scores.



Σχήμα 3.8: Παράδειγμα εφαρμογής του συγκεντρωτικού επιπέδου, το οποίο υποδειγματοληπτεί κάθε επιφάνεια βάθους του όγκου εισόδου. *Αριστερά:* Ο όγκος εισόδου $224 \times 224 \times 64$ συγκεντρώνεται με μέγεθος φίλτρου 2 και βήμα 2 στον όγκο εξόδου $112 \times 112 \times 64$. *Δεξιά:* Η εφαρμογή του max pooling με βήμα 2. Εξάγεται το μέγιστο από κάθε γειτονιά 2×2. Σχήμα από [2]

3.2.3 Δημοφιλείς Αρχιτεκτονικές Συνελικτικών Νευρωνικών Δικτύων

Όπως περιγράψαμε παραπάνω, τα Συνελικτικά Νευρωνικά Δίκτυα κατασκευάζονται από μόνο τρεις τύπους επιπέδων. Στο εξής θα χρησιμοποιούμε το συμβολισμό *CONV* για το συνελικτικό επίπεδο, *POOL* για το συγκεντρωτικό επίπεδο και *FC* για το πλήρως συνδεδεμένο επίπεδο. Επίσης, θα χειριζόμαστε τη συνάρτηση ενεργοποίησης ReLU ως ένα επίπεδο με συμβολισμό *RELU*.

Μοτίβα τοποθέτησης επιπέδων (Layer Patterns)

Μία πολύ κλασική μορφή της αρχιτεκτονικής ενός Συνελικτικού Νευρωνικού Δικτύου στοιβάζει αρκετά CONV-RELU επίπεδα, ακολουθούμενα από POOL layers και επαναλαμβάνει αυτή την τριάδα μέχρι η εικόνα να έχει συγχωνευτεί χωρικά σε μία αναπαράσταση μικρού μεγέθους. Έπειτα, συνηθίζεται η εφαρμογή ενός FC επιπέδου το οποίο περιέχει την έξοδο του δικτύου, π.χ. τα class scores. Δηλαδή, το μοτίβο αρχιτεκτονικής διαμορφώνεται ως:

$$INPUT \rightarrow [[CONV \rightarrow RELU] * N \rightarrow POOL] * M \rightarrow [FC \rightarrow RELU] * K \rightarrow FC$$

όπου ο παράγοντας $*$ υποδηλώνει επανάληψη και $N \geq 0$, $M \geq 0$, $K \geq 0$. Όσο αυξάνεται η τιμή του N και του M οδηγούμαστε σε μεγαλύτερα και πιο βαθιά δίκτυα, τα οποία μπορούν να εξάγουν πιο πολύπλοκα χαρακτηριστικά από την εικόνα εισόδου.

Ειδικότερα, σε βαθιές αρχιτεκτονικές, πριν το pooling, είναι προτιμότερο να υπάρχει μία ακολουθία από CONV επίπεδα μικρών φίλτρων παρά ένα μοναδικό CONV επίπεδο

μεγάλου ευαίσθητου πεδίου, δηλαδή σχετικά μεγάλα N. Για παράδειγμα, έστω ότι έχουμε 3 συνεχόμενα CONV layers με φίλτρα 3×3 (με τις αντίστοιχες ReLU μη γραμμικότητες ενδιάμεσα). Με αυτή την τοπολογία, κάθε νευρώνας του πρώτου CONV επιπέδου «βλέπει» μία περιοχή 3×3 της εισόδου, κάθε νευρώνας του δεύτερου CONV επιπέδου «βλέπει» μία περιοχή 3×3 του πρώτου CONV επιπέδου, δηλαδή μία 5×5 περιοχή της εισόδου, ενώ ένας νευρώνας του τρίτου CONV επιπέδου «βλέπει» μία περιοχή 3×3 του δεύτερου CONV επιπέδου, δηλαδή μία 7×7 περιοχή της εισόδου. Αυτή η ακολουθία συνελικτικών επιπέδων, οι έξοδοι των οποίων περνάνε από μία μη γραμμική συνάρτηση ενεργοποίησης, παράγει χαρακτηριστικά πολύ πιο περιγραφικά σε σχέση με την ύπαρξη ενός μόνο CONV επιπέδου με ευαίσθητο πεδίο νευρώνων 7×7, το οποίο υπολογίζει μία απλή γραμμική συνάρτηση της εισόδου. Επίσης, αν θεωρήσουμε ότι, σε κάθε περίπτωση, τόσο η είσοδος όσο και όλα τα συνελικτικά επίπεδα έχουν C κανάλια βάθους, τότε ένα 7×7 CONV layer περιέχει $C \cdot (7 \cdot 7 \cdot C) = 49C^2$ παραμέτρους ενώ η ακολουθία τριών 3×3 CONV layers περιλαμβάνει $3 \cdot (C \cdot (3 \cdot 3 \cdot C)) = 27C^2$ παραμέτρους.

Μοτίβα μεγέθους επιπέδων (Layer Sizing Patterns)

Όσον αφορά τις διαστάσεις κάθε επιπέδου υπάρχει ένας οδηγός κανόνων ο οποίος ακολουθείται στις περισσότερες περιπτώσεις. Συγκεκριμένα:

- Το *επίπεδο εισόδου (INPUT)* που περιέχει την εικόνα πρέπει να διαιρείται με το 2 πολλές φορές. Δημοφιλείς επιλογές διαστάσεων είναι τα 32, 64, 96, 224, 384, 512 pixels.

- Τα *συνελικτικά επίπεδα (CONV)* πρέπει να χρησιμοποιούν μικρά φίλτρα (π.χ. 3×3 ή το πολύ 5×5), με βήμα $\mathbf{S} = \mathbf{1}$, συμπληρώνοντας τον όγκο εισόδου με μηδενικά όποτε χρειάζεται, με τέτοιο τρόπο που να μην αλλάζει τις χωρικές διαστάσεις του. Δηλαδή, π.χ. αν $\mathbf{F} = \mathbf{3}$, τότε χρησιμοποιώντας $\mathbf{P} = \mathbf{1}$ το μέγεθος της εισόδου παραμένει αναλλοίωτο. Αντίστοιχα, για $\mathbf{F} = \mathbf{5}, \mathbf{P} = \mathbf{2}$. Γενικότερα, για οποιοδήποτε $\mathbf{F}$, το μηδενικό περιθώριο $\mathbf{P} = (\mathbf{F} - \mathbf{1})/\mathbf{2}$ διατηρεί το μέγεθος της εισόδου. Μεγαλύτερα μεγέθη φίλτρων όπως 7×7 συνίσταται να χρησιμοποιούνται μόνο στο πρώτο συνελικτικό επίπεδο που «βλέπει» την εικόνα και μόνο αν είναι απαραίτητο.

- Τα *συγκεντρωτικά επίπεδα (POOL)* είναι υπεύθυνα για υποδειγματοληψία των χωρικών διαστάσεων της εισόδου. Η πιο συνήθης επιλογή συγκέντρωσης είναι η max-pooling με ευαίσθητο πεδίο 2×2 ($\mathbf{F} = \mathbf{2}$) και βήμα $\mathbf{S} = \mathbf{2}$. Έτσι απορρίπτεται ακριβώς το 75% των ενεργοποιήσεων του όγκου εισόδου. Μία λιγότερο δημοφιλής επιλογή για το ευαίσθητο πεδίο είναι η $\mathbf{F} = \mathbf{3}$, ενώ μεγαλύτερες τιμές οδηγούν σε μεγάλη απώλεια πληροφορίας και δεν χρησιμοποιούνται.

Μελέτη συγκεκριμένων περιπτώσεων (Case studies)

Με την πάροδο των ετών, και ειδικότερα τα τελευταία χρόνια, δημιουργήθηκαν αρκετές χαρακτηριστικές αρχιτεκτονικές Συνελικτικών Νευρωνικών Δικτύων οι οποίες πήραν συγκεκριμένο όνομα και αποτελούν σημεία αναφοράς σε πολλές σύγχρονες εργασίες. Αυτές είναι:

- *LeNet*: Αποτελεί μία από τις πρώτες αρχιτεκτονικές ΣΝΔ, η οποία δημιουργήθηκε το 1998 από τους LeCun et al. [71] για την αναγνώριση ταχυδρομικών κωδικών, ψηφίων κλπ.

- *AlexNet*: Είναι η αρχιτεκτονική που εισήγαγε τα ΣΝΔ στην Όραση Υπολογιστών και υλοποιήθηκε από τους Krizhevsky et al. [70] στο πλαίσιο του διαγωνισμού Imagenet ILSVRC[3] για την ταξινόμηση εικόνων το 2012, στον οποίο υπερκέρασε σημαντικά τις υπόλοιπες εργασίες. Το δίκτυο είχε όμοια αρχιτεκτονική με το LeNet, με τη διαφορά ότι ήταν πιο βαθύ, πιο μεγάλο και εισήγαγε τη χρήση ακολουθιών συνελικτικών επιπέδων αντί για ένα μόνο συνελικτικό επίπεδο πριν το pooling σε κάθε στάδιο (κάτι που όπως αναλύσαμε παραπάνω βελτιώνει σημαντικά την επίδοση του δικτύου).

- *ZF Net*: Αρχιτεκτονική που προτάθηκε από τους Zeiler & Fergus [72] και κέρδισε τον διαγωνισμό ILSVRC το 2013. Ήταν ουσιαστικά μία βελτίωση του AlexNet της προηγούμενης χρονιάς, μέσω της επέκτασης του μεγέθους των μεσαίων συνελικτικών επιπέδων και της μείωσης του βήματος και του μεγέθους του φίλτρου του πρώτου συνελικτικού επιπέδου.

- *GoogleNet*: Αποτελεί τον επόμενο νικητή του διαγωνισμού ILSVRC το 2014 και δημιουργήθηκε από τους Szegedy et al. [73] της Google. Η κύρια συνεισφορά αυτού του δικτύου είναι κατασκευή μιας πολυκλαδικής αρχιτεκτονικής (*Inception Module*) η οποία μείωσε δραματικά το πλήθος των παραμέτρων ($4M \rightarrow 15$ φορές λιγότερες από τις $60M$ του AlexNet). Επίσης, στο τελικό στάδιο χρησιμοποιήθηκε pooling μέσου όρου αντί για πλήρως συνδεδεμένα επίπεδα απορρίπτοντας έναν μεγάλο αριθμό παραμέτρων που δεν φαίνεται να είχαν σημαντική επίδραση στο δίκτυο. Βασιζόμενες πάνω στην πολυκλαδική λογική του GoogleNet ακολούθησαν κι άλλες αρχιτεκτονικές όπως το πρόσφατο δίκτυο Inception-v4 [74].

- *VGGNet*: Ήταν ο δεύτερος νικητής του διαγωνισμού ILSVRC το 2014 και αποτελεί δημιουργία των Simonyan & Zisserman [75]. Με την αρχιτεκτονική αυτή έδειξαν ότι το βάθος ενός ΣΝΔ παίζει πολύ σημαντικό ρόλο στην επίδοση του.

[3] http://image-net.org/challenges/LSVRC/2012

Το τελικό τους δίκτυο περιλαμβάνει 16 ομοιογενή CONV/FC layers, τα οποία εκτελούν μόνο 3×3 συνελίξεις και 2×2 pooling από την αρχή μέχρι το τέλος. Ένα μειονέκτημα του VGG δικτύου είναι ο μεγάλος αριθμός των παραμέτρων του ($\sim 140M$) που το καθιστά πολύ απαιτητικό υπολογιστικά. Οι περισσότερες από αυτές τις παραμέτρους ανήκουν στο πρώτο FC επίπεδο. Πρόσφατα αποδείχτηκε ότι τα FC layers του VGG μπορούν να αφαιρεθούν χωρίς να επηρεαστεί η επίδοση του δικτύου, μειώνοντας σημαντικά το πλήθος των παραμέτρων του.

- *ResNet*: Το Residual δίκτυο αναπτύχθηκε από τους He et al. [76] και κέρδισε τον διαγωνισμό ILSVRC το 2015. Χαρακτηρίζεται από *παρακαμπτόμενες συνδέσεις (skip connections)*, μεγάλη χρήση της *κανονικοποίησης παρτίδας (batch normalization)* και απουσία FC επιπέδων. Πολλές αρχιτεκτονικές residual δικτύων αναπτύσσονται μέχρι και σήμερα παρουσιάζοντας εξαιρετικές επιδόσεις σε πολλά προβλήματα. Χαρακτηριστικό παράδειγμα αποτελεί η μετέπειτα παραλλαγή του πρωτότυπου ResNet από τους ίδιους τους He et al. [77].

Μοτίβα εκπαίδευσης (Learning patterns)

Στην πράξη, λίγοι είναι αυτοί που εκπαιδεύουν ένα ΣΝΔ εξ΄ αρχής (from scratch), καθώς είναι σχετικά σπάνιο να έχουν στη διάθεση τους ένα επαρκές σύνολο δεδομένων. Αντίθετα, συνήθως ένα ΣΝΔ προ-εκπαιδεύεται σε ένα πολύ μεγάλο σύνολο δεδομένων (π.χ. στο ImageNet[4], το οποίο περιλαμβάνει 1,2 εκ. εικόνες κατανεμημένες σε 1000 κατηγορίες) και έπειτα χρησιμοποιείται είτε ως αρχικοποίηση για μία νέα εκπαίδευση είτε ως μέσο εξαγωγής χαρακτηριστικών για ένα πρόβλημα. Με αυτό τον τρόπο αποθηκεύουν τη «γνώση» τους όσο επιλύουν ένα πρόβλημα και την εφαρμόζουν σε ένα διαφορετικό αλλά σχετικό πρόβλημα, μεταβιβάζοντας έτσι τη μάθηση τους (transfer learning). Τα δύο κύρια σενάρια μεταβίβασης της μάθησης αναλύονται παρακάτω:

- *Το ΣΝΔ ως μέσο εξαγωγής χαρακτηριστικών:* Αν από ένα προ-εκπαιδευμένο ΣΝΔ αφαιρεθεί το τελευταίο πλήρως συνδεδεμένο επίπεδο (στην περίπτωση εκπαίδευσης στο ImageNet, οι έξοδοι αυτού του επιπέδου είναι τα 1000 class scores) τότε το υπόλοιπο δίκτυο μπορεί να χρησιμοποιηθεί για την εξαγωγή χαρακτηριστικών σε ένα νέο σύνολο δεδομένων. Για παράδειγμα, στο AlexNet [70], το υπόλοιπο δίκτυο θα υπολογίζει ένα διάνυσμα χαρακτηριστικών 4096 θέσεων για κάθε εικόνα εισόδου, το οποίο θα περιέχει τις ενεργοποιήσεις του κρυφού επιπέδου ακριβώς πριν τον ταξινομητή. Έπειτα από την εξαγωγή των χαρακτηριστικών συνηθίζεται η εκπαίδευση ενός γραμμικού ταξινομητή για το νέο σύνολο δεδομένων.

[4] http://www.image-net.org/

- *Το ΣΝΔ ως αρχικοποίηση για εκ΄ νέου εκπαίδευση:* Η δεύτερη στρατηγική αφορά την προσαρμογή των βαρών του προ-εκπαιδευμένου δικτύου σε ένα νέο σύνολο δεδομένων, συνεχίζοντας τον αλγόριθμο εκπαίδευσης backpropagation. Αυτή η διαδικασία ονομάζεται *fine-tuning* και είναι δυνατόν να εφαρμοστεί σε όλα τα layers του δικτύου ή μόνο σε κάποια «υψηλότερα» layers (προς το τέλος του δικτύου) κρατώντας μερικά από τα πρώτα επίπεδα αναλλοίωτα, μεριμνώντας για το ενδεχόμενο της υπερπροσαρμογής. Το δεύτερο είναι και το πλέον σύνηθες, καθώς έχει παρατηρηθεί ότι τα αρχικά επίπεδα παράγουν πιο γενικά χαρακτηριστικά (π.χ. ανίχνευση ακμών) τα οποία είναι χρήσιμα σε πολλές περιπτώσεις, ενώ τα επόμενα επίπεδα γίνονται προοδευτικά πιο συγκεκριμένα σε ειδικότερες λεπτομέρειες κάθε κλάσης.

3Δ Συνελικτικά Νευρωνικά Δίκτυα (3D ConvNets)

Τα παραπάνω μοτίβα αρχιτεκτονικών αλλά και οι case studies αναφέρονται σε δισδιάστατα Συνελικτικά Νευρωνικά Δίκτυα, τα οποία λαμβάνουν ως είσοδο μία εικόνα, εξάγουν πολύπλοκα χαρακτηριστικά και την ταξινομούν με βάση αυτά[5]. Ωστόσο, για την επίλυση προβλημάτων που επεξεργάζονται πιο σύνθετες δομές δεδομένων, όπως αυτό της αναγνώρισης ανθρώπινων δράσεων, προκύπτει η ανάγκη εξαγωγής χαρακτηριστικών πάνω σε ακολουθίες εικόνων, οι οποίες εκτός από τις 2 διαστάσεις του πλάτους και του ύψους, χαρακτηρίζονται κι από τη διάσταση του χρόνου. Οι πρώτες προσεγγίσεις για την ταξινόμηση βίντεο ανθρώπινων δράσεων με τη χρήση ΣΝΔ αφοσιώθηκαν στην προσαρμογή των 2Δ ΣΝΔ πάνω στις 3Δ εισόδους ακολουθιών εικόνων, μέσω της επέκτασης των επιπέδων τους στον τρισδιάστατο χώρο.

Η πρώτη απόπειρα χρήσης τρισδιάστατων Συνελικτικών Νευρωνικών Δικτύων για αναγνώριση ανθρώπινων δράσεων έγινε το 2013 από τους Ji et al. [16], οι οποίοι πρότειναν την εφαρμογή 3Δ συνελίξεων στα συνελικτικά επίπεδα ενός δικτύου, προκειμένου να επιτευχθεί η εξαγωγή χαρακτηριστικών τόσο από τη χωρική όσο και από τη χρονική διάσταση της εισόδου. Δηλαδή, επέκτειναν τα φίλτρα των συνελικτικών επιπέδων στην 3η διάσταση δημιουργώντας έναν 3Δ πυρήνα ο οποίος συνελίσσεται με τον όγκο που σχηματίζεται στοιβάζοντας συνεχόμενα καρέ ενός βίντεο στην είσοδο. Με αυτόν τον τρόπο, οι έξοδοι των συνελικτικών επιπέδων, δηλαδή οι *χάρτες χαρακτηριστικών (feature maps)*, είναι συνδεδεμένοι με πολλά γειτονικά frames, λαμβάνοντας, έτσι, υπόψη και την πληροφορία της κίνησης. Το μοτίβο διαμοιρασμού των παραμέτρων σε κάθε συνελικτικό επίπεδο παραμένει αναλλοίωτο, έτσι ώστε κάθε 3Δ πυρήνας να εφαρμόζεται σε όλο το μήκος του όγκου εισόδου. Βασιζόμενοι, λοιπόν,

[5]Θεωρούμε ότι μία εικόνα ότι έχει 2 διαστάσεις, μη λαμβάνοντας υπόψη τα κανάλια χρώματος σαν επιπλέον διάσταση, αφού κάθε φίλτρο ενός συνελικτικού δικτύου εκτελεί διαφορετικές 2Δ συνελίξεις για κάθε κανάλι

στην πράξη της τρισδιάστατης συνέλιξης και στη γενική αρχή των ΣΝΔ ότι οι χάρτες χαρακτηριστικών πρέπει να αυξάνονται πηγαίνοντας προς τα τελευταία layers, παράγοντας διαφορετικούς τύπους χαρακτηριστικών από ένα σύνολο αντίστοιχων χαρτών χαμηλότερου επιπέδου, οι Ji et al. δημιούργησαν μία αρχιτεκτονική ενός τρισδιάστατου ΣΝΔ για την αναγνώριση ανθρώπινων δράσεων στη βάση δεδομένων TRECVID[6], η οποία φαίνεται στο Σχήμα 3.9, και έχει τα εξής χαρακτηριστικά:

- Η είσοδος στο δίκτυο τη χρονική στιγμή t αποτελείται από 7 καρέ διαστάσεων 60×40 κεντραρισμένα στο τρέχων frame f_t.

- Στο πρώτο επίπεδο, πριν την εφαρμογή συνελίξεων, εξάγονται 33 χάρτες χαρακτηριστικών για 5 κανάλια πληροφορίας: *gray, gradient-x, gradient-y, optflow-x, optflow-y* για τις γκρίζες τιμές των pixels (7 κανάλια), το gradient για την οριζόντια και κάθετη διεύθυνση (7+7 κανάλια) και την οπτική ροή για τις δύο διευθύνσεις (6+6 κανάλια) αντίστοιχα (H1 layer).

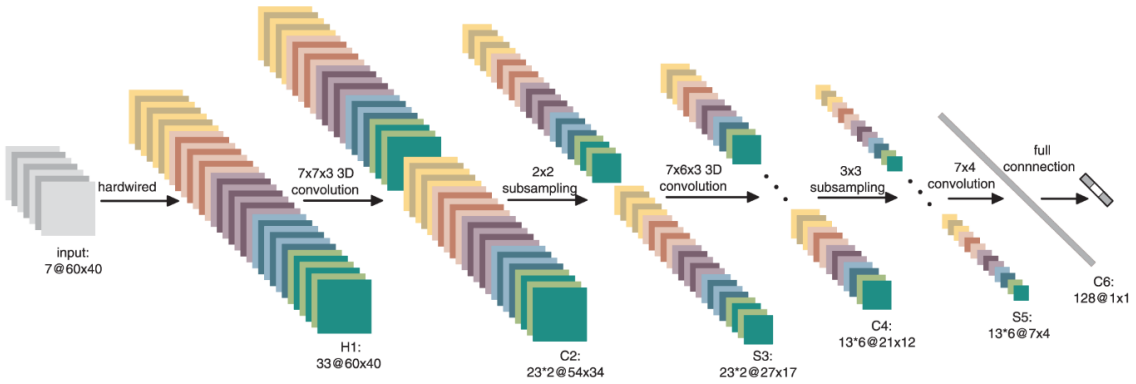
- Έπειτα εφαρμόζεται το πρώτο συνελικτικό επίπεδο με δύο διαφορετικούς πυρήνες διαστάσεων $7 \times 7 \times 3$ οι οποίοι εφαρμόζονται και στα 5 κανάλια πληροφορίας. Παράγονται έτσι 2 σύνολα από 23 χάρτες χαρακτηριστικών διαστάσεων 54×34 (C2 layer).

- Στο επόμενο επίπεδο πραγματοποιείται μία 2×2 υποδειγματοληψία μειώνοντας τις διαστάσεις των αναπαραστάσεων σε 27×17 (S3 layer).

- Στη συνέχεια ένα ακόμη συνελικτικό επίπεδο εφαρμόζει 3 διαφορετικούς πυρήνες διαστάσεων $7 \times 6 \times 3$ σε όλα τα κανάλια πληροφορίας οδηγώντας έτσι σε 6 διαφορετικά σύνολα χαρτών χαρακτηριστικών διαστάσεων 21×12, με 13 αναπαραστάσεις το καθένα (C4 layer).

- Ένα pooling επίπεδο εφαρμόζει 3×3 υποδειγματοληψία σε κάθε αναπαράσταση παράγοντας μικρότερους χάρτες διαστάσεων 7×4 (S5 layer). Σε αυτό το στάδιο η χρονική διάσταση για κάθε κανάλι πληροφορίας είναι μικρή (3 για gray, gradient-x, gradient-y και 2 για optflow-x, optflow-y).

- Άρα, στο τελευταίο συνελικτικό επίπεδο χρησιμοποιούνται 2Δ χωρικές συνελίξεις με έναν πυρήνα 7×4 έτσι ώστε το μέγεθος των χαρτών χαρακτηριστικών που θα παραχθούν στην έξοδο να είναι 1×1 (C6 layer). Επειδή, οι συγγραφείς θέλουν η έξοδος του δικτύου να είναι ένα διάνυσμα χαρακτηριστικών 128 θέσεων, ακολουθεί ένα πλήρως συνδεδεμένο επίπεδο 128 νευρώνων καθένας από τους οποίους συνδέεται με όλους τους 78 χάρτες χαρακτηριστικών του προηγούμενου επιπέδου.



Σχήμα 3.9: Η αρχιτεκτονική ενός 3Δ ΣΝΔ για αναγνώριση ανθρώπινων δράσεων. Αποτελείται από 1 hardwired, 3 συνελικτικά, 2 συγκεντρωτικά και 1 πλήρως συνδεδεμένο επίπεδο. Σχήμα από [16]

Οι συγγραφείς πειραματίστηκαν και στη βάση δεδομένων KTH εκμεταλλευόμενοι την ίδια αρχιτεκτονική με κάποιες διαφοροποιήσεις. Χρησιμοποιήσαν έναν όγκο από 9 καρέ στην είσοδο, τα οποία περιείχαν μία περιοχή 80×60 από το προσκήνιο (foreground) κάθε εικόνας, ενώ τα 3 συνελικτικά επίπεδα διέθεταν πυρήνες $9 \times 7 \times 3$, $7 \times 7 \times 3$ και 6×4 αντίστοιχα, και τα δύο συγκεντρωτικά επίπεδα εφάρμοζαν υποδειγματοληψία με μία περιοχή 3×3.

Στηριζόμενοι στην ιδέα των Ji et al., λίγο αργότερα οι Tran et al. [48] πρότειναν μία τεχνική εκπαίδευσης χωρο-χρονικών χαρακτηριστικών χρησιμοποιώντας τρισδιάστατα βαθιά Συνελικτικά Νευρωνικά Δίκτυα. Αρχικά, πειραματίστηκαν πάνω στη βάση δεδομένων UCF101, για διάφορες τιμές των ευαίσθητων πεδίων των νευρώνων των συνελικτικών επιπέδων, προκειμένου να βρουν τους 3Δ πυρήνες των φίλτρων που αποδίδουν καλύτερα (Πείραμα 1). Έπειτα, χρησιμοποιώντας την καλύτερη επιλογή για το μέγεθος των πυρήνων, κατασκεύασαν ένα δίκτυο για την εξαγωγή χωρο-χρονικών χαρακτηριστικών σε αναπαράστασεις βίντεο, το οποίο εκπαίδευσαν με διαφορετικές τεχνικές και το αξιολόγησαν πάνω στην UCF101 (Πείραμα 2). Πιο συγκεκριμένα:

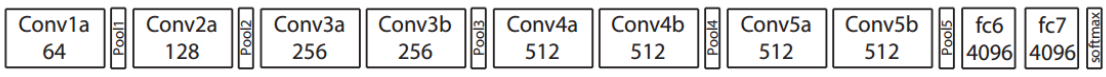
Πείραμα 1: Η αρχιτεκτονική τους γι΄ αυτό το πείραμα αποτελούνταν από 5 συνελικτικά επίπεδα και 5 συγκεντρωτικά επίπεδα (κάθε συνελικτικό επίπεδο ακολουθούνταν από ένα συγκεντρωτικό), 2 πλήρως συνδεδεμένα επίπεδα και ένα επίπεδο ταξινόμησης βασισμένο στη συνάρτηση κόστους Softmax (softmax loss layer). Ο αριθμός των φίλτρων, δηλαδή το βάθος κάθε συνελικτικού επιπέδου, από το πρώτο μέχρι το τελευταίο, ήταν 64, 128, 256, 256, 256 αντίστοιχα. Σύμφωνα με τα ευρήματα του [75] για τα δισδιάστατα ΣΝΔ, κράτησαν σταθερό, και ίσο με 3×3, το χωρικό ευαίσθητο πεδίο όλων των νευρώνων των συνελικτικών επιπέδων και άλλαζαν μόνο τη χρονική του διάσταση **d**. Όλα τα frames των βίντεο της UCF101 τροποποιήθηκαν έτσι ώστε να έχουν

[6] http://www-nlpir.nist.gov/projects/trecvid/

διαστάσεις 128×171, περίπου δηλαδή τη μισή ανάλυση, ενώ κάθε βίντεο χωρίστηκε σε μικρά μη-επικαλυπτόμενα κλιπ των 16 καρέ, τα οποία δίνονταν σαν είσοδο στο δίκτυο. Κατά τη διάρκεια της εκπαίδευσης χρησιμοποιήθηκε η τεχνική του *jittering*, περικόπτοντας μία τυχαία περιοχή 112×112 του κλιπ εισόδου. Όλα τα συνελικτικά επίπεδα εφαρμόστηκαν με το κατάλληλο padding και με βήμα 1, ώστε η διάσταση των χαρτών χαρακτηριστικών να μην αλλάζει στην έξοδο. Τα συγκεντρωτικά επίπεδα ήταν max pooling σε έναν πυρήνα $2 \times 2 \times 2$ (εκτός του πρώτου $2 \times 2 \times 1$) με βήμα 1, μειώνοντας τη διάσταση της αναπαράστασης στο 1/8, ενώ τα δύο πλήρως συνδεδεμένα επίπεδα είχαν 2048 ενεργοποιήσεις εξόδου. Η εκπαίδευση του δικτύου έγινε εξ΄ αρχής (from scratch), για 16 εποχές, χρησιμοποιώντας mini-batches των 30 κλιπς, με αρχικό ρυθμό μάθησης 0.003, ο οποίος διαιρούνταν με το 10 κάθε 4 εποχές.

Χρησιμοποιώντας αυτή την αρχιτεκτονική, δοκίμασαν διάφορες κατανομές τιμών για το βάθος $\mathbf{d}$ των νευρώνων και κατέληξαν στο συμπέρασμα ότι ο πυρήνας $3 \times 3 \times 3$ είναι η καλύτερη επιλογή για όλα τα συνελικτικά επίπεδα, οδηγώντας έτσι σε ένα ομοιογενές δίκτυο. Με βάση αυτό το σημαντικό εύρημα, σχεδίασαν ένα νέο ΣΝΔ το οποίο αποτελείται από 8 συνελικτικά επίπεδα, 5 pooling επίπεδα, ακολουθούμενα από 2 πλήρως συνδεδεμένα επίπεδα και ένα softmax loss επίπεδο, όπως φαίνεται στο Σχήμα 3.10. Όλα τα 3Δ φίλτρα των συνελικτικών επιπέδων είναι $3 \times 3 \times 3$ με βήμα $1 \times 1 \times 1$, ενώ όπως και στην παραπάνω αρχιτεκτονική, τα συγκεντρωτικά επίπεδα εφαρμόζουν max pooling σε μία περιοχή $2 \times 2 \times 2$ με βήμα $2 \times 2 \times 2$, εκτός από το πρώτο (pool1) το οποίο έχει ευαίσθητο πεδίο $2 \times 2 \times 1$ και βήμα $2 \times 2 \times 1$, προκειμένου να διατηρηθεί αναλλοίωτη η πληροφορία για τη χρονική διάσταση στο πρώτο στάδιο επεξεργασίας. Κάθε πλήρως συνδεδεμένο επίπεδο διαθέτει 4096 μονάδες εξόδου. Το δίκτυο αυτό ονομάστηκε από τους συγγραφείς *C3D*.

Έπειτα από την εκπαίδευση του, το C3D μπορεί να χρησιμοποιηθεί σαν περιγραφητής βίντεο. Συγκεκριμένα, για την εξαγωγή C3D χαρακτηριστικών, ένα βίντεο χωρίζεται σε κλιπς των 16 καρέ (συνήθως με μία επικάλυψη 8 καρέ μεταξύ δύο συνεχόμενων κλιπς), τα οποία «περνούν» από το C3D δίκτυο και εξάγονται οι ενεργοποιήσεις του πρώτου πλήρους συνδεδεμένου επιπέδου *fc6*. Έπειτα, υπολογίζεται ο μέσος όρος των fc6 ενεργοποιήσεων όλων των κλιπς του βίντεο, ακολουθούμενος από L2 κανονικοποίηση για να σχηματιστεί το τελικό διάνυσμα C3D χαρακτηριστικών 4096 θέσεων.



Σχήμα 3.10: Η αρχιτεκτονική του δικτύου C3D. Σε κάθε κουτάκι αναγράφεται ο αριθμός των φίλτρων κάθε συνελικτικού επιπέδου. Σχήμα από [48]

Πείραμα 2: Στο πείραμα αυτό, χρησιμοποιήσαν το δίκτυο C3D, το οποίο εκπαίδευσαν σε δύο βάσεις δεδομένων: α) εκπαίδευση στη δική τους βάση I380K, β) εκπαίδευση

στη βάση δεδομένων Sports-1M[7] [78] και γ) εκπαίδευση στην I380K και fine-tuning στην Sports-1M. Όσον αφορά την προετοιμασία των εισόδων, από κάθε βίντεο εξάγονται με τυχαίο τρόπο 5 κλιπς διάρκειας 2 δευτερολέπτων, τα frames των οποίων, όπως και στο 1^o πείραμα, υπόκεινται σε μετατροπή του μεγέθους τους σε 128×171. Έπειτα, χρησιμοποιώντας χωρικό και χρονικό jittering, κάθε κλιπ χωρίζεται τυχαία σε επιμέρους όγκους $16 \times 112 \times 112$, οι οποίοι αναστρέφονται ως προς τον οριζόντιο άξονα με πιθανότητα 50%. Η εκπαίδευση του δικτύου γίνεται για 13 εποχές (1.9M επαναλήψεις), χρησιμοποιώντας mini-batches των 30 κλιπς, με αρχικό ρυθμό μάθησης 0.003, ο οποίος διαιρείται με το 2 κάθε 150K επαναλήψεις.

Για την αξιολόγηση των τριών εκπαιδευμένων μοντέλων χρησιμοποιήθηκε η βάση ανθρώπινων δράσεων UCF101, εξάγοντας C3D χαρακτηριστικά για τα βίντεο του συνόλου εκπαίδευσης και αξιολόγησης και των 3 splits. Χρησιμοποιώντας έναν γραμμικό ταξινομητή SVM (βλ. Ενότητα 3.3), έδειξαν ότι τα C3D χαρακτηριστικά από το fine-tuned δίκτυο (γ) είναι αρκετά αποδοτικά με $82,3\%$ ακρίβεια αναγνώρισης, ενώ ενώνοντας τις fc6 ενεργοποιήσεις και των τριών δικτύων (α), (β), (γ) σε ένα ενιαίο διάνυσμα χαρακτηριστικών 12288 θέσεων, η ακρίβεια ενισχύεται σημαντικά ($85,2\%$). Τέλος, συνδυάζοντας τα C3D χαρακτηριστικά με τις πυκνές τροχιές η ακρίβεια αναγνώρισης εκτοξεύεται στο $90,4\%$. Φαίνεται λοιπόν ότι το C3D δίκτυο μπορεί να περιγράψει αρκετά καλά τόσο τη χωρική πληροφορία όσο και την πληροφορία της κίνησης μέσα σε ένα βίντεο, ενώ παράγει υψηλού επιπέδου, σημασιολογικά, αναπαραστάσεις που είναι συμπληρωματικές με τα ιστογράμματα της οπτικής ροής και τα χαμηλού επιπέδου gradients που εξάγονται από τις πυκνές τροχιές. Αυτά τα πλεονεκτήματα καθιστούν τα C3D χαρακτηριστικά ιδιαιτέρως σημαντικά και πολύ χρήσιμα. Στο εξής της παρούσας Διπλωματικής, ως C3D χαρακτηριστικά θα θεωρούμε αυτά τα οποία εξάγονται από το εκπαιδευμένο δίκτυο (γ).

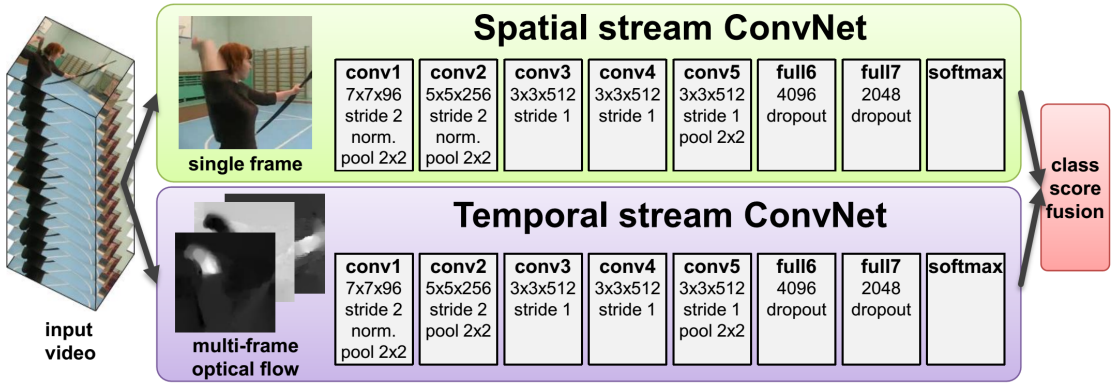
Συνελικτικά Νευρωνικά Δίκτυα Διπλής Ροής (Two-stream ConvNets)

Μία άλλη προσέγγιση για την ταξινόμηση βίντεο ανθρώπινων δράσεων, η οποία εξελίσσεται ραγδαία τα τελευταία χρόνια, είναι η χρήση διαφορετικών 2Δ Συνελικτικών Νευρωνικών Δικτύων για την επεξεργασία της χωρικής και της χρονικής πληροφορίας ενός βίντεο και η ταξινόμηση του με βάση τον συνδυασμό των εξόδων κάθε δικτύου. Η ιδέα ήρθε στο προσκήνιο το 2014 από τους Simonyan & Zisserman [17] οι οποίοι χρησιμοποίησαν ως εισόδους στατικά 2Δ καρέ για το χωρικό δίκτυο (spatial stream) και 3Δ όγκους οπτικής ροής για το χρονικό δίκτυο (temporal stream), τα οποία αποτελούν μέρη ενός ενιαίου δικτύου το οποίο ονομάζεται *Συνελικτικό Νευρωνικό Δίκτυο Διπλής Ροής (Two-Stream CNN).*

Συγκεκριμένα, οι συγγραφείς εμπνεύστηκαν από τη φυσική διάσπαση του βίντεο

[7] 1,1 εκ. αθλητικά βίντεο, 487 αθλητικές κλάσεις: http://cs.stanford.edu/people/karpathy/deepvideo/

στο χωρικό και χρονικό του μέρος, όπου το χωρικό τμήμα αποτελείται από κάθε καρέ ξεχωριστά και περιέχει πληροφορία για το περιβάλλον και τα αντικείμενα που απεικονίζονται στο βίντεο, ενώ το χρονικό κομμάτι περιγράφει την κίνηση ανάμεσα στα frames, ποσοτικοποιώντας τις μετακινήσεις της κάμερας και των αντικειμένων. Με αυτή τη συλλογιστική, σχεδίασαν μία αρχιτεκτονική ταξινόμησης βίντεο, διαιρώντας την, αντίστοιχα, σε δύο τμήματα όπως φαίνεται στο Σχήμα 3.11. Οι δύο πυλώνες ταξινόμησης υλοποιούνται χρησιμοποιώντας βαθιά ΣΝΔ, τα softmax scores των οποίων συνδυάζονται στο τέλος με μία μέθοδο σύμμειξης (late fusion): είτε α) με απλό μέσο όρο είτε β) με την εκπαίδευση ενός γραμμικού SVM ταξινομητή πάνω στα L2-κανονικοποιημένα softmax scores.



Σχήμα 3.11: Η αρχιτεκτονική ενός two-stream δικτύου για ταξινόμηση βίντεο. Σχήμα από [17]

Το χωρικό δίκτυο εφαρμόζεται πάνω σε ξεχωριστά frames από το βίντεο, πραγματοποιώντας αναγνώριση δράσεων από στατικές εικόνες. Αποτελεί ουσιαστικά μία αρχιτεκτονική ενός ΣΝΔ για ταξινόμηση εικόνων, όπως τα case studies που αναφέραμε παραπάνω, και μπορεί να προ-εκπαιδευτεί σε ένα μεγάλο σύνολο εικόνων όπως το Imagenet. Όσον αφορά το χρονικό δίκτυο, η είσοδος του αποτελείται από τα πεδία μετατόπισης της οπτικής ροής $\mathbf{d}$ ανάμεσα σε διαδοχικά καρέ, ομαδοποιημένα γύρω από το frame ενδιαφέροντος. Ειδικότερα, έστω w, h το πλάτος και το ύψος του βίντεο και L το μήκος των διαδοχικών καρέ που επιλέγεται για την αναπαράσταση της κίνησης. Τότε ένας όγκος εισόδου $I_\tau \in \mathbb{R}^{w \times h \times 2L}$ για ένα frame τ σχηματίζεται ως εξής:

$$\begin{aligned} I_\tau(u, v, 2k-1) &= d^x_{\tau+k-1}(u, v) \\ I_\tau(u, v, 2k) &= d^y_{\tau+k-1}(u, v), \quad u = [1; w], v = [1; h], k = [1; L] \end{aligned} \tag{3.16}$$

όπου d^x_τ, d^y_τ είναι το οριζόντιο και κάθετο πεδίο μετατόπισης για το frame τ.

Εκτός από την κλασική χρήση όλου του πεδίου μετατόπισης της οπτικής ροής στο σχηματισμό του όγκου εισόδου (optical flow stacking), οι συγγραφείς δοκίμασαν και μία εναλλακτική αναπαράσταση της κίνησης χρησιμοποιώντας τις πυκνές τροχιές [15]

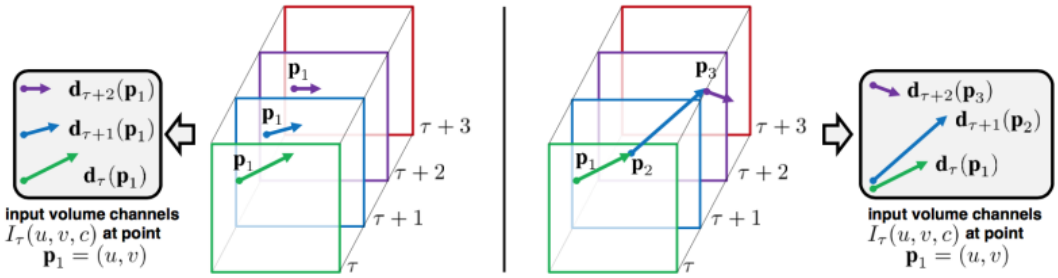
για να δειγματοληπτήσουν την οπτική ροή μόνο κατά μήκος των σημείων της τροχιάς (trajectory stacking), δηλαδή:

$$\begin{aligned} I_\tau(u,v,2k-1) &= d^x_{\tau+k-1}(\mathbf{p}_k) \\ I_\tau(u,v,2k) &= d^y_{\tau+k-1}(\mathbf{p}_k), \quad u=[1;w], v=[1;h], k=[1;L] \end{aligned} \tag{3.17}$$

όπου $\mathbf{p}_k$ είναι το k-οστό σημείο της τροχιάς, η οποία ξεκινάει από το σημείο (u,v) στο frame τ και ορίζεται από την ακόλουθη αναδρομική σχέση:

$$\mathbf{p}_1 = (u,v), \quad \mathbf{p}_k = \mathbf{p}_{k-1} + \mathbf{d}_{\tau+k-2}(\mathbf{p}_{k-1}), k > 1 \tag{3.18}$$

Η διαφορά των δύο προσεγγίσεων φαίνεται στο Σχήμα 3.12. Σε κάθε περίπτωση, η οπτική ροή ανάμεσα σε δύο διαδοχικά καρέ μπορεί να περιλαμβάνει μετατοπίσεις που δεν είναι πραγματικές και οφείλονται στην κίνηση της κάμερας. Για να αντισταθμιστεί αυτή η κίνηση και να εξαλειφθούν λανθασμένα πεδία μετατοπίσεων, οι συγγραφείς προτείνουν μία απλή τεχνική: από κάθε πεδίο μετατόπισης $\mathbf{d}$ αφαιρείται το διάνυσμα μέσης τιμής του (mean flow subtraction).



Σχήμα 3.12: Η κατασκευή του όγκου εισόδου για το χρονικό δίκτυο. *Αριστερά:* Το optical flow stacking δειγματοληπτεί τα διανύσματα μετατόπισης $\mathbf{d}$ στην ίδια θέση σε όλα τα frames της στοίβας. *Δεξιά:* Το trajectory stacking δειγματοληπτεί τα διανύσματα μετατόπισης κατά μήκος της τροχιάς. Σχήμα από [17]

Σε αντίθεση με τον χωρικό πυλώνα επεξεργασίας, το χρονικό δίκτυο εκπαιδεύεται σε δεδομένα βίντεο. Δυστυχώς όμως, οι διαθέσιμες βάσεις δεδομένων για ταξινόμηση ανθρώπινων δράσεων δεν είναι αρκετά μεγάλες για αυτό το σκοπό. Χρειάζεται επομένως μία ειδική μεταχείριση μέσω μίας μορφής συνδυασμού δεδομένων από διαφορετικές βάσεις προκειμένου να αποφευχθεί η υπερπροσαρμογή (overfitting) του δικτύου κατά την εκπαίδευση του. Για το σκοπό αυτό, οι Simonyan & Zisserman χρησιμοποίησαν τις δύο βάσεις ανθρώπινων δράσεων HMDB51 και UCF101 και τροποποίησαν ελαφρώς την αρχιτεκτονική του δικτύου προσθέτοντας ένα ακόμη softmax επίπεδο ταξινόμησης έτσι ώστε: το ένα softmax επίπεδο να υπολογίζει τα class scores της HMDB51 και το άλλο τα class scores της UCF101. Κάθε layer είναι εξοπλισμένο με τη δική του συνάρτηση κόστους, η οποία εφαρμόζεται μόνο για τα βίντεο που ανήκουν στο αντίστοιχο

σύνολο δεδομένων. Το συνολικό κόστος εκπαίδευσης υπολογίζεται ως το άθροισμα των δύο επιμέρους κόστεων, ενώ οι παράγωγοι των βαρών του δικτύου βρίσκονται με τη μέθοδο backpropagation (multi-task learning). Λεπτομέρειες για τη διαδικασία της εκπαίδευσης αλλά και της αξιολόγησης του two-stream ΣΝΔ αναφέρονται παρακάτω:

Εκπαίδευση: Η εκπαίδευση του δικτύου γίνεται με τη μέθοδο Mini-Batch Gradient Descent. Σε κάθε επανάληψη, διαλέγονται 256 βίντεο εκπαίδευσης (ομοιόμορφα κατανεμημένα στις κλάσεις), σε καθένα από τα οποία απομονώνεται τυχαία ένα frame, κατασκευάζοντας έτσι ένα mini-batch 256 δειγμάτων εισόδου. Για το χωρικό δίκτυο, περικόπτεται μία περιοχή 224×224 από το επιλεγμένο frame, η οποία υποβάλλεται σε οριζόντια αναστροφή και σε RGB jittering, ενώ για το χρονικό δίκτυο εξάγεται ο όγκος της οπτικής ροής I για το επιλεγμένο καρέ, όπως περιγράφηκε παραπάνω, και στη συνέχεια απομονώνεται τυχαία ένας όγκος εισόδου $224 \times 224 \times 2L$, ο οποίος επίσης αναστρέφεται. Ο ρυθμός μάθησης αρχικοποιείται στην τιμή 10^{-2} και εν συνεχεία μειώνεται σύμφωνα με ένα σταθερό χρονοδιάγραμμα, ίδιο για όλα τα σύνολα εκπαίδευσης. Συγκεκριμένα, όταν ένα ΣΝΔ εκπαιδεύεται εξ΄ αρχής, ο ρυθμός μάθησης μειώνεται στο 10^{-3} μετά από 50K επαναλήψεις, έπειτα στο 10^{-4} έπειτα από 70K επαναλήψεις και σταματάει μετά από 80K επαναλήψεις. Στην περίπτωση του fine-tuning, ο ρυθμός μάθησης αλλάζει σε 10^{-3} έπειτα από 14K επαναλήψεις και έπειτα σταματάει στις 20K επαναλήψεις.

Αξιολόγηση: Για την αξιολόγηση του δικτύου, δοθέντος ενός βίντεο, επιλέγονται τυχαία 25 frames με ίση χρονική απόσταση μεταξύ τους, σε καθένα από τα οποία εξάγονται 10 διαφορετικές είσοδοι στο δίκτυο, περικόπτοντας και αναστρέφοντας τις 4 γωνίες και το κέντρο του. Τα class scores για όλο το βίντεο υπολογίζονται ως ο μέσος όρος των επιμέρους scores όλων των περικοπτόμενων περιοχών των επιλεγμένων καρέ.

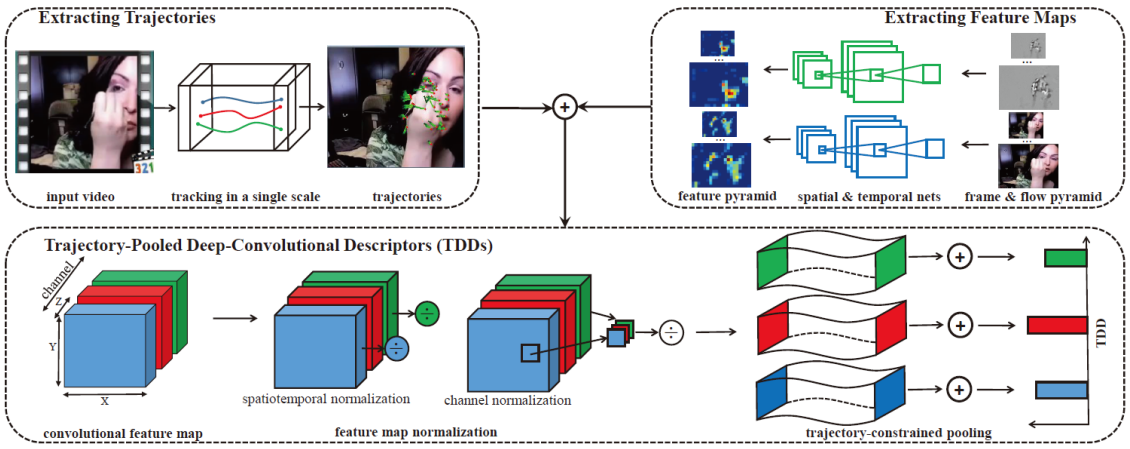
Έπειτα από πολλά πειράματα, με διαφορετικούς τρόπους εκπαίδευσης, χρησιμοποιώντας τις δύο βάσεις ανθρώπινων δράσεων HMDB51 και UCF101 αλλά και τη βάση εικόνων ILSVRC-2012 του Imagenet οι συγγραφείς έφτασαν σε πολύ υψηλά ποσοστά αναγνώρισης - τα υψηλότερα μέχρι τότε (state-of-the-art) - ενώ κατέληξαν στο συμπέρασμα ότι η αρχιτεκτονική με την καλύτερη επίδοση έχει τα εξής χαρακτηριστικά:

- Χωρικό δίκτυο προ-εκπαιδευμένο στην ILSVRC-2012, με το τελευταίο επίπεδο ταξινόμησης εκπαιδευμένο στην UCF101 ή στην HMDB51.
- Χρονικό δίκτυο εκπαιδευμένο με multi-task learning στις δύο βάσεις UCF101 και HMDB51, χρησιμοποιώντας στην είσοδο optical flow stacking με αφαίρεση της μέσης τιμής της (mean flow subtraction).
- Συνδυασμό των softmax scores των δύο δικτύων με έναν SVM ταξινομητή.

Η ιδέα του ΣΝΔ διπλής ροής των Simonyan & Zisserman αποτέλεσε έμπνευση για πολλούς ερευνητές, οι οποίοι στρέφουν την προσοχή τους στη χρήση παρόμοιων αρχιτεκτονικών για την επίλυση του προβλήματος της αναγνώρισης ανθρώπινων δράσεων. Η εξέλιξη του πεδίου τα τελευταία χρόνια είναι ραγδαία με πολλές εργασίες να έχουν συνδράμει σημαντικά σε αυτή. Ωστόσο, είναι πρακτικά αδύνατον να κάνουμε αναφορά σε κάθε εργασία της σχετικής βιβλιογραφίας, στα πλαίσια της παρούσας Διπλωματικής, για αυτό το λόγο, επιλέγουμε να αναφέρουμε, παρακάτω, τέσσερις από τις πιο σημαντικές αρχιτεκτονικές two-stream ΣΝΔ που ακολούθησαν της αρχικής.

Trajectory-pooled Deep-convolutional Descriptor (TDD)

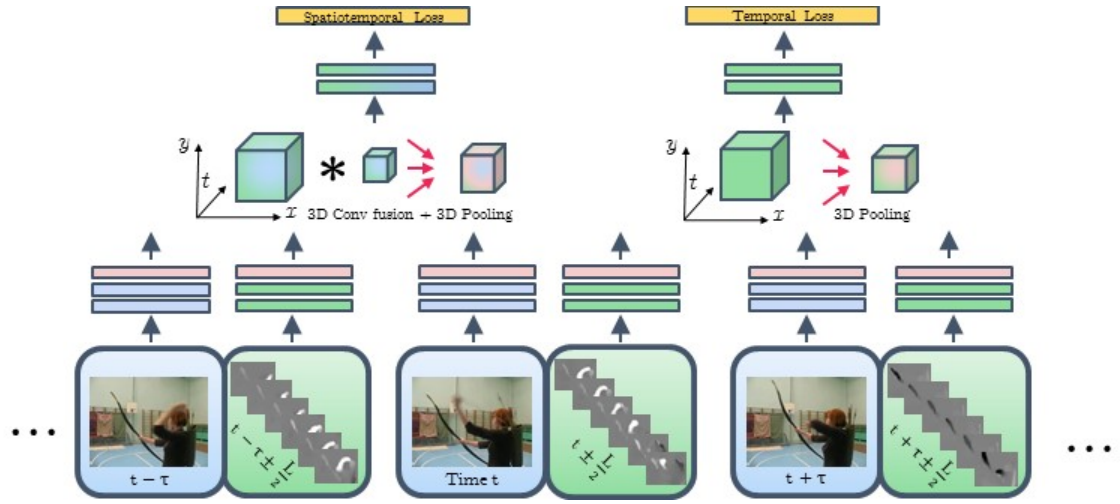
Το 2015 οι Wang et al. [49] πρότειναν έναν νέο περιγραφητή βίντεο, συνδυάζοντας τις πυκνές τροχιές με τα two-stream ΣΝΔ, τον οποίο ονόμασαν *Trajectory-pooled Deep-convolutional Descriptor (TDD)*. Χρησιμοποιώντας την καλύτερη αρχιτεκτονική των Simonyan & Zisserman, έπειτα από την εκπαίδευση της, τα δύο επιμέρους δίκτυα χρησιμοποιούνται για την εξαγωγή πολυκλιμακωτών χαρτών χαρακτηριστικών για κάθε βίντεο. Επίσης, χρησιμοποιώντας την τεχνική των πυκνών τροχιών, ανιχνεύονται σημεία ενδιαφέροντος για την κίνηση μέσα στο βίντεο, γύρω από τα οποία συγκεντρώνονται οι τοπικές αποκρίσεις των δύο ΣΝΔ. Πραγματοποιείται, ουσιαστικά, ένα pooling των χαρακτηριστικών μέσα στον χωρο-χρονικό όγκο της τροχιάς, σχηματίζοντας έτσι τον TDD περιγραφητή, όπως φαίνεται στο Σχήμα 3.13. Τέλος, εφαρμόζεται η Fisher Vector αναπαράσταση για τη συσσώρευση όλων των τοπικών TDDs του βίντεο σε ένα συνολικό διάνυσμα-περιγραφητή, το οποίο χρησιμοποιείται για την αναγνώριση δράσεων μέσω ενός γραμμικού SVM ταξινομητή.



Σχήμα 3.13: Η διαδικασία εξαγωγής του TDD, η οποία χωρίζεται σε τρία βήματα: (i) εξαγωγή των τροχιών, (ii) εξαγωγή πολυκλιμακωτών χαρτών χαρακτηριστικών από τα δύο ΣΝΔ και (iii) υπολογισμός του TDD. Σχήμα από [49]

Two-stream Δίκτυα με πρώιμη χωρο-χρονική σύμμειξη

Το 2016, οι Feichtenhofer, Pinz & Zisserman [50] επέκτειναν την αρχική τους two-stream αρχιτεκτονική [17] εφαρμόζοντας διαφορετικές τεχνικές συνδυασμού του χωρικού και χρονικού δικτύου, χρησιμοποιώντας εναλλακτικές μορφές pooling. Συγκεκριμένα, πρότειναν την συνένωση των δύο δικτύων, στο τελευταίο συνελικτικό τους επίπεδο (έπειτα από την εφαρμογή της ReLU) μέσα στο χωρικό δίκτυο μετατρέποντας το σε ένα χωρο-χρονικό δίκτυο, χρησιμοποιώντας 3Δ συνελικτική σύμμειξη ακολουθούμενη από 3Δ pooling, όπως φαίνεται στο Σχήμα 3.14 (αριστερά). Ταυτόχρονα, εκμεταλλεύονται περαιτέρω τη χρονική πληροφορία εφαρμόζωντας επίσης 3Δ pooling στο χρονικό δίκτυο, όπως φαίνεται στο Σχήμα 3.14 (δεξιά). Τα κόστη και των δύο δικτύων χρησιμοποιούνται στην εκπαίδευση ενώ κατά την αξιολόγηση υπολογίζεται ο μέσος όρος των προβλέψεων των δύο δικτύων.



Σχήμα 3.14: Η σύμμειξη των two-stream ΣΝΔ για την εξαγωγή χωρο-χρονικών και αμιγώς χρονικών χαρακτηριστικών. Τα δύο δίκτυα περιγράφουν βραχυπρόθεσμη πληροφορία σε μία μικρή χρονική κλίμακα ($t \pm L/2$), ανάμεσα σε γειτονικές εισόδους μεγαλύτερης χρονικής κλίμακας ($t + T\tau$). Τα δύο δίκτυα συνενώνονται μέσω ενός 3Δ φίλτρου το οποίο μαθαίνει τις αντιστοιχίες ανάμεσα στα χαρακτηριστικά του χωρικού (μπλε) και του χρονικού δικτύου (πράσινο), ως τοπικούς σταθμισμένους συνδυασμούς των x, y, t. Στα παραγόμενα χαρακτηριστικά από το χωρο-χρονικό και το αμιγώς χρονικό δίκτυο εφαρμόζεται ένα 3Δ pooling, προκειμένου να εξαχθούν χωρο-χρονικά (πάνω αριστερά) και αμιγώς χρονικά (πάνω δεξιά) χαρακτηριστικά για το βίντεο εισόδου. Σχήμα από [50]

Δίκτυα Χρονικών Τμημάτων (Temporal Segment Networks)

Αργότερα, την ίδια χρονιά, οι Wang et al. [51], βασισμένοι στην two-stream αρχιτεκτονική, σχεδίασαν πολύ αποτελεσματικά Συνελικτικά Νευρωνικά Δίκτυα για την αναγνώριση ανθρώπινων δράσεων και πέτυχαν την εκπαίδευση τους με σχετικά μικρό σύνολο δεδομένων. Χρησιμοποιώντας την ιδέα της μακροπρόθεσμης μοντελοποίησης της χρονικής πληροφορίας μέσω σχετικά αραιής δειγματοληψίας, δημιούργησαν τα Δίκτυα Χρονικών Τμημάτων (Temporal Segment Networks - TSN), με τα οποία πέτυχαν τα υψηλότερα μέχρι τότε (state-of-the-art) ποσοστά αναγνώρισης στις δύο πολύ δημοφιλείς βάσεις ανθρώπινων δράσεων HMDB51 και UCF101. Συγκεκριμένα, όσον αφορά τη μοντελοποίηση της κίνησης, παρατήρησαν ότι διαδοχικά frames, όπως αυτά που εξάγονται από μία πυκνή δειγματοληψία, περιέχουν πλεονάζουσα πληροφορία, έτσι μία στρατηγική πιο αραιής δειγματοληψίας είναι προτιμότερη. Έτσι, τα TSN εφαρμόζονται σε μία ακολουθία μικρών αποσπασμάτων, αραιά δειγματοληπτημένων από *όλο το βίντεο*, καθένα από τα οποία παράγει τη δική του πρόβλεψη για τις κλάσεις δράσεων, ενώ ένας συνδυασμός τους (consensus) δίνει την πρόβλεψη για όλο το βίντεο. Στη διαδικασία της εκπαίδευσης, ελαχιστοποιείται το κόστος από τα class scores της πρόβλεψης για *όλο το βίντεο*, για την επαναληπτική ανανέωση των παραμέτρων του μοντέλου.

Ειδικότερα, δοθέντος ενός βίντεο V, διαιρείται σε K τμήματα $\{S_1, S_2, ..., S_K\}$ ίσης διάρκειας και έπειτα το TSN εφαρμόζεται σε μία ακολουθία αποσπασμάτων ως εξής:

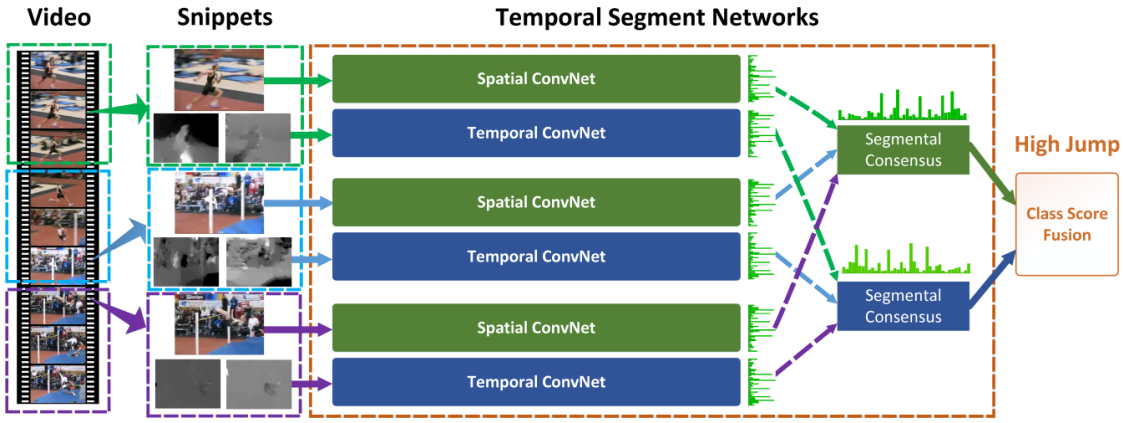
$$TSN(T_1, T_2, ..., T_K) = H(G(F(T_1; \mathbf{W}), F(T_2; \mathbf{W}), ..., F(T_K; \mathbf{W}))) \tag{3.19}$$

όπου $(T_1, T_2, ..., T_K)$ είναι μία ακολουθία αποσπασμάτων, καθένα (T_k) από τα οποία έχει δειγματοληπτηθεί τυχαία από το αντίστοιχο τμήμα S_k του βίντεο. $F(T_1; \mathbf{W})$ είναι η συνάρτηση που αντιπροσωπεύει ένα ΣΝΔ με παραμέτρους $\mathbf{W}$ το οποίο εφαρμόζεται στο μικρό απόσπασμα T_k και παράγει scores για όλες τις κλάσεις. Η συνάρτηση συνένωσης των τμημάτων (segmental consensus) G συνδυάζει τις εξόδους από τα διάφορα μικρά αποσπάσματα και παράγει τα τελικά class scores από κάθε δίκτυο, ενώ η συνάρτηση πρόβλεψης H υπολογίζει την τελική πιθανότητα κάθε κλάσης δράσεων για όλο το βίντεο, όπως φαίνεται στο Σχήμα 3.15. Για τη μορφή της H επιλέγεται η πολύ δημοφιλής συνάρτηση Softmax, έτσι η συνολική συνάρτηση κόστους, λαμβάνοντας υπόψη τον τμηματικό συνδυασμό $\mathbf{G} = G(F(T_1; \mathbf{W}), F(T_2; \mathbf{W}), ..., F(T_K; \mathbf{W}))$, έχει την ακόλουθη μορφή:

$$L(y, \mathbf{G}) = -\sum_{i=1}^{C} y_i \left(G_i - \log \sum_{j=1}^{C} \exp G_j \right) \tag{3.20}$$

όπου C είναι το πλήθος των κλάσεων και y_i είναι η επισημειωμένη ετικέτα (ground-truth label) που αφορά την κλάση i. Για τον αριθμό των αποσπασμάτων, στα οποία

χωρίζεται το βίντεο εισόδου, οι συγγραφείς επέλεξαν $K = 3$, βασισμένοι σε προηγούμενες εργασίες για τη μοντελοποίηση της χρονικής πληροφορίας. Για τον συνδυασμό G χρησιμοποιούν μία απλή συνάρτηση, που δεν εξαρτάται από τις παραμέτρους του μοντέλου, δηλαδή $G_i = g(F_i(T_1), ..., F_i(T_K))$, όπου G_i το class score που προκύπτει από τα επιμέρους scores της ίδιας κλάσης για όλα τα αποσπάσματα βίντεο, υπολογισμένο από τη συνάρτηση συνένωσης g. Για την επιλογή της συνάρτησης g πειραματίστηκαν με διάφορες μορφές και κατέληξαν στη χρήση του απλού μέσου όρου.



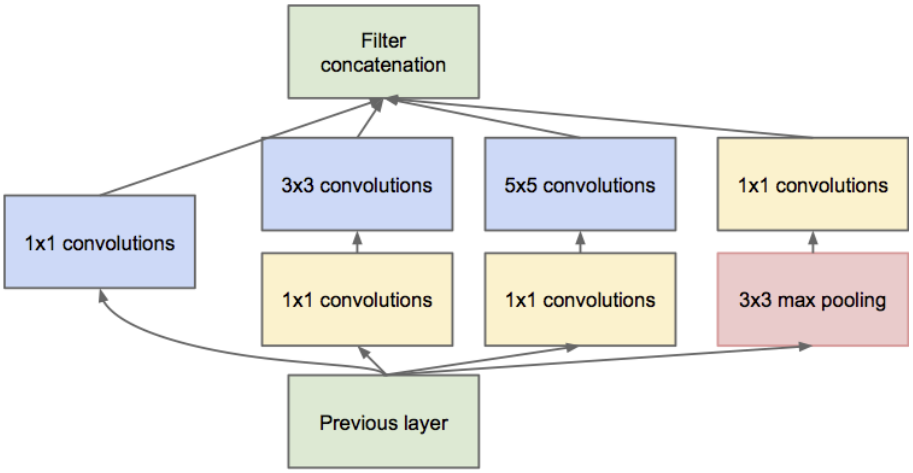
Σχήμα 3.15: Δίκτυο Χρονικών Τμημάτων (Temporal Segment Network). Ένα βίντεο εισόδου χωρίζεται σε K τμήματα και ένα μικρό απόσπασμα επιλέγεται τυχαία από κάθε τμήμα. Τα class scores από διαφορετικά αποσπάσματα συνδυάζονται από μία συνάρτηση συνένωσης η οποία παράγει την πρόβλεψη για όλο το βίντεο. Οι προβλέψεις κάθε δικτύου συνδυάζονται σε τελικό στάδιο για την εξαγωγή της τελικής πρόβλεψης. Τα ΣΝΔ για όλα τα αποσπάσματα έχουν κοινές παραμέτρους. Σχήμα από [51]

Η αρχιτεκτονική ενός TSN επιτρέπει τη χρήση των διαφόρων αποσπασμάτων για την βελτιστοποίηση των παραμέτρων $\mathbf{W}$ του μοντέλου μέσω του κλασικού αλγόριθμου backpropagation. Συνυπολογίζοντας και τον τμηματικό συνδυασμό $\mathbf{G}$, οι παράγωγοι των παραμέτρων $\mathbf{W}$ ως προς τη συνάρτηση κόστους L υπολογίζονται από τη σχέση:

$$\frac{\partial L(y, \mathbf{G})}{\partial \mathbf{W}} = \frac{\partial L}{\partial \mathbf{G}} \sum_{k=1}^{K} \frac{\partial G}{\partial F(T_k)} \frac{\partial F(T_k)}{\partial \mathbf{W}} \tag{3.21}$$

Χρησιμοποιώντας μία βελτιστοποίηση μέσω απότομης καθόδου, όπως την Stochastic Gradient Descent, η Σχέση 3.21 εξασφαλίζει ότι οι ανανεώσεις των παραμέτρων λαμβάνουν υπόψη την συνάθροιση $\mathbf{G}$ των προβλέψεων όλων των τμημάτων βίντεο. Με αυτόν τον τρόπο, ένα TSN μπορεί να εκπαιδεύει τις παραμέτρους του με πληροφορία από όλο το βίντεο, αντί μόνο από ένα μικρό απόσπασμα του, κάτι ιδιαιτέρως σημαντικό. Στη συνέχεια, περιγράφονται αναλυτικά η αρχιτεκτονική ενός TSN, καθώς και η διαδικασία εκπαίδευσης και αξιολόγησης του.

Αρχιτεκτονική: Για την κατασκευή ενός Δικτύου Χρονικών Τμημάτων οι Wang et al. χρησιμοποίησαν την αρχιτεκτονική του GoogleNet των Szegedy et al. [73] με το πολυκλαδικό μπλοκ κατασκευής (Inception Module), το οποίο φαίνεται στο Σχήμα 3.16, ακολουθούμενο από κανονικοποίηση παρτίδας (Batch Normalization), όπως το υλοποίησαν οι Ioffe & Szegedy [66] και το ονόμασαν *BN-Inception* δίκτυο. Όπως και στην αρχική two-stream αρχιτεκτονική [17], το χωρικό δίκτυο εφαρμόζεται σε στατικές RGB εικόνες ενώ το χρονικό δίκτυο δέχεται στην είσοδο του μία στοίβα από συνεχόμενα πεδία οπτικής ροής.



Σχήμα 3.16: Σχηματική απεικόνιση του Inception module το οποίο αντικαθιστά τα παραδοσιακά συνελικτικά επίπεδα. Σχήμα από [73]

Εκπαίδευση: Όπως έχουμε προαναφέρει, οι βάσεις δεδομένων για την αναγνώριση ανθρώπινων δράσεων είναι σχετικά μικρές, επομένως η εκπαίδευση βαθιών ΣΝΔ με αυτές αντιμετωπίζει τον κίνδυνο της υπερπροσαρμογής (overfitting). Για να περιορίσουν αυτό το πρόβλημα, οι συγγραφείς σχεδίασαν διάφορες τεχνικές για την εκπαίδευση των δικτύων ενός TSN όπως περιγράφονται παρακάτω:

- *Cross Modality Pre-training*: Η προ-εκπαίδευση ενός ΣΝΔ είναι ένας αποτελεσματικός τρόπος για την αρχικοποίηση των παραμέτρων του, ειδικά όταν το σύνολο δεδομένων δεν αποτελείται από αρκετά δείγματα εκπαίδευσης. Η φύση των εισόδων των χωρικών δικτύων, τους επιτρέπουν να προ-εκπαιδεύονται στο πολύ μεγάλο σύνολο στατικών εικόνων του Imagenet, ενώ αντίθετα η αρχικοποίηση των χρονικών δικτύων αποτελεί μια πιο πολύπλοκη διαδικασία. Οι Wang et al. εφάρμοσαν μία τεχνική διασταύρωσης των δύο πυλώνων (cross modality pre-training) χρησιμοποιώντας το χωρικό δίκτυο για την αρχικοποίηση του χρονικού δικτύου. Ειδικότερα, αρχικά μετέτρεψαν τα πεδία οπτικής ροής σε ένα 2Δ διακριτό σήμα με τιμές από 0 έως 255, μέσω ενός γραμμικού μετασχηματισμού, καθιστώντας το εύρος τιμών τους ίδιο με αυτό των RGB εικόνων. Έπειτα, τρο-

ποποίησαν τα βάρη του πρώτου συνελικτικού επιπέδου του χωρικού δικτύου, έτσι ώστε να μπορεί να εφαρμοστεί πάνω στα πεδία οπτικής ροής, υπολογίζοντας τον μέσο όρο τους κατά μήκος των RGB καναλιών και αναπαράγοντας τις τιμές του σε όλα τα κανάλια της εισόδου του χρονικού δικτύου.

- *Regularization*: Το πρόβλημα της μεταβλητής μετατόπισης των παραμέτρων αντιμετωπίζεται όπως προαναφέραμε με την τεχνική του batch normalization. Κατά τη διάρκεια της εκπαίδευσης, η κανονικοποίηση παρτίδας υπολογίζει τη μέση τιμή και τη διακύμανση των ενεργοποιήσεων κάθε batch και τις τροποποιεί έτσι ώστε να ακολουθούν μία μοναδιαία γκαουσιανή κατανομή. Αυτή η στρατηγική επιταχύνει σημαντικά την σύγκλιση του αλγόριθμου εκπαίδευσης αλλά ταυτόχρονα οδηγεί σε υπερπροσαρμογή λόγω της ντετερμινιστικής εκτίμησης της κατανομής των ενεργοποιήσεων από ένα μικρό σύνολο δεδομένων εκπαίδευσης. Γι΄ αυτό το λόγο, έπειτα από την αρχικοποίηση των δικτύων, οι συγγραφείς επέλεξαν να κρατούν αναλλοίωτες τις παραμέτρους της μέσης τιμής και της διακύμανσης όλων των BatchNorm επιπέδων πλην του πρώτου. Αφού η κατανομή των τιμών της οπτικής ροής είναι διαφορετική από αυτή των RGB εικόνων, γίνεται διαφορετική εκτίμηση της μέσης τιμής και της διακύμανσης των ενεργοποιήσεων του πρώτου συνελικτικού επιπέδου για τα δύο δίκτυα. Αυτή η τεχνική ονομάζεται *μερική κανονικοποίηση παρτίδας (partial batch normalization)*. Επίσης, οι Wang et al. πρόσθεσαν ένα επιπλέον dropout επίπεδο έπειτα από το ολικό επίπεδο συγκέντρωσης (global pooling layer) στην αρχιτεκτονική του BN-Inception δικτύου, προκειμένου να ελαττώσουν περαιτέρω τον κίνδυνο του overfitting. Για την πιθανότητα απόσυρσης (dropout ratio) των νευρώνων επιλέχθηκαν αρκετά μεγάλες τιμές, συγκεκριμένα 0.8 για τα χωρικά και 0.7 για τα χρονικά δίκτυα.

- *Data Augmentation*: Για την επαύξηση των δεδομένων, οι Wang et al. πρότειναν δύο νέες μεθόδους: την περικοπή γωνιών (corner cropping) και το κλιμακωτό jittering. Για την περικοπή γωνιών, οι εξαγόμενες περιοχές επιλέγονται μόνο από τις γωνίες ή από το κέντρο της εικόνας. Για το πολυκλιμακωτό jittering, τροποποιείται το μέγεθος της εικόνας εισόδου ή των πεδίων οπτικής ροής στις διαστάσεις 256×340 και έπειτα το πλάτος και το ύψος της περικοπτόμενης περιοχής επιλέγεται τυχαία από το $\{256, 224, 192, 168\}$. Τέλος, το μέγεθος όλων των περικοπτόμενων περιοχών επανατροποποιείται στις διαστάσεις 224×224.

Αξιολόγηση: Εφόσον όλα τα ΣΝΔ ενός TSN έχουν κοινές παραμέτρους, τα εκπαιδευμένα μοντέλα μπορούν να αξιολογηθούν σε επίπεδο frames όπως τα κοινά ΣΝΔ. Συνεπώς, για την αξιολόγηση των TSN οι συγγραφείς ακολούθησαν την ίδια διαδικασία με την αρχική two-stream αρχιτεκτονική [17], όπου δοθέντος ενός βίντεο, επιλέγονται

τυχαία 25 frames με ίση χρονική απόσταση μεταξύ τους, σε καθένα από τα οποία εξάγονται 10 διαφορετικές είσοδοι στο δίκτυο, περικόπτοντας και αναστρέφοντας τις 4 γωνίες και το κέντρο του. Τα class scores για όλο το βίντεο υπολογίζονται ως ο μέσος όρος των επιμέρους scores όλων των περικοπτόμενων περιοχών των επιλεγμένων καρέ. Τέλος, για τη σύμμειξη του χωρικού και του χρονικού δικτύου, υπολογίζεται ο σταθμισμένος μέσος όρος των τελικών class scores κάθε δικτύου, με βάρη 1 για το χωρικό και 1.5 για το χρονικό δίκτυο. Η σύμμειξη των class scores τόσο για τα 25 καρέ, όσο και μετέπειτα για τα δύο δίκτυα γίνεται πριν την εφαρμογή του κανονικοποιητή Softmax, όπως και στην εκπαίδευση.

3.3 Μηχανές Διανυσμάτων Υποστήριξης (SVMs)

Οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines - SVMs) αποτελούν τους δημοφιλέστερους ταξινομητές σε προβλήματα επιβλεπόμενης μάθησης (supervised learning), με πολύ μεγάλη δυνατότητα γενίκευσης. Η πιο απλή τους εφαρμογή είναι η δυαδική ταξινόμηση. Συγκεκριμένα, έστω ένα σύνολο D-διάστατων γραμμικά διαχωρίσιμων δεδομένων εκπαίδευσης $X = \{\mathbf{x}_i, y_i\}, \mathbf{x}_i \in \mathbb{R}^D$, όπου $y_i \in \{-1, 1\}$ η επισημειωμένη ετικέτα (label) του $\mathbf{x}_i$. Η διαδικασία εκπαίδευσης του SVM αφορά την εύρεση του βέλτιστου υπερεπιπέδου στο χώρο $\mathbb{R}^D$ που διαχωρίζει τα δεδομένα, δηλαδή για το $\mathbf{w}\mathbf{x} + b$ υπερεπίπεδο, θα πρέπει να ισχύει:

$$\begin{aligned} \mathbf{w}\mathbf{x}_i + b < 0 \quad \alpha\nu \quad y_i < 0 \\ \mathbf{w}\mathbf{x}_i + b > 0 \quad \alpha\nu \quad y_i > 0 \end{aligned} \tag{3.22}$$

Μετά την εκπαίδευση του SVM προκύπτουν τα «διανύσματα υποστήριξης» (support vectors), τα οποία είναι ένα σημαντικό σημασιολογικά υποσύνολο των δεδομένων εκπαίδευσης, καθώς βρίσκονται πιο κοντά στο βέλτιστο υπερεπίπεδο. Η απόσταση ενός οποιουδήποτε διανύσματος από το υπερεπίπεδο, το πρόσημο της οποίας μαρτυράει την κατηγορία στην οποία ανήκει, μπορεί να εκφραστεί ως γραμμικός συνδυασμός εσωτερικών γινομένων μεταξύ των διανυσμάτων υποστήριξης. Ωστόσο, στις περισσότερες περιπτώσεις τα δεδομένα δεν είναι γραμμικά διαχωρίσιμα. Έτσι, προκειμένου να εκπαιδευτεί ο ταξινομητής, εισάγονται στο δεξί μέλος των ανισοτήτων 3.22 κάποιες «μεταβλητές χαλάρωσης» (slack variables), επιτρέποντας ορισμένα λάθη στον ταξινομητή. Το βέλτιστο υπερεπίπεδο που προκύπτει είναι αυτό που ελαχιστοποιεί αυτά τα λάθη.

Επίσης, ένα επιπλέον πολύ σημαντικό χαρακτηριστικό των SVM είναι η δυνατότητα εύρεσης *μη-γραμμικών* διαχωριστικών επιφανειών, μέσω μιας απεικόνισης $\Phi(\cdot)$ των δεδομένων $\mathbf{x}_i$ από τον ευκλείδειο χώρο σε έναν άλλο χώρο εσωτερικού γινομένου. Η $\Phi(\mathbf{x}_i)$ δεν είναι αναγκαίο να είναι προκαθορισμένη, αρκεί να είναι γνωστός ο ορισμός του εσωτερικού γινομένου στο νέο χώρο. Έτσι, για τον προσδιορισμό του βέλτιστου

υπερεπιπέδου, τα εσωτερικά γινόμενα των διανυσμάτων δίνονται από τον πυρήνα $\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j) = K(\mathbf{x}_i, \mathbf{x}_j)$. Για παράδειγμα, στην ταξινόμηση δεδομένων με αναπαραστάσεις ιστογραμμάτων, όπως στην περίπτωση του μοντέλου Bag-of-Words, χρησιμοποιείται συχνά ο πυρήνας x^2 (chi-squared kernel), ο οποίος ποσοτικοποιεί τη διαφοροποίηση μεταξύ δύο ιστογραμμάτων $\mathbf{h}_i, \mathbf{h}_j$ υπολογίζοντας την εξής «απόσταση»:

$$D(\mathbf{h}_i, \mathbf{h}_j) = \sum_{k=1}^{K} \frac{(\mathbf{h}_i^k - \mathbf{h}_j^k)^2}{\mathbf{h}_i^k + \mathbf{h}_j^k} \tag{3.23}$$

όπου $\mathbf{h}_i^k$ το k-οστό στοιχείο του ιστογράμματος $\mathbf{h}_i$.

Επεκτείνοντας την απλή δυαδική ταξινόμηση, το SVM μπορεί να εφαρμόσει και ταξινόμηση σε πολλαπλές κλάσεις, για την οποία έχουν προταθεί διάφορες τεχνικές όπως η «ένας εναντίον ενός» (one-versus-one) ή η «ένας εναντίον όλων (one-versus-all) [59]. Στα πλαίσια της παρούσας Διπλωματικής χρησιμοποιείται η τελευταία, η οποία πραγματοποιείται με την εκπαίδευση ενός ταξινομητή ξεχωριστά για κάθε κλάση. Κάθε ταξινομητής προβλέπει κατά πόσο ένα δείγμα αξιολόγησης ανήκει στην αντίστοιχη κατηγορία, υπολογίζοντας μία πιθανότητα (confidence score), και εν συνεχεία το δείγμα ταξινομείται στην κλάση με την μεγαλύτερη πιθανότητα.

Σύμμειξη πολλαπλών καναλιών πληροφορίας

Όπως έχει γίνει σαφές από τα παραπάνω, ένα βίντεο μπορεί να αναπαρασταθεί με πολλούς τρόπους (π.χ. διαφορετικά χαρακτηριστικά, διαφορετική κωδικοποίηση κ.α.) περιγράφοντας πολλές φορές διαφορετική, συμπληρωματική πληροφορία. Για παράδειγμα, όπως αναλύσαμε παραπάνω, ένα δίκτυο C3D παράγει υψηλού επιπέδου αναπαραστάσεις οι οποίες περιγράφουν διαφορετική πληροφορία και συμπληρώνουν τα χαμηλού επιπέδου ιστογράμματα που εξάγονται από τις πυκνές τροχιές. Τα διαφορετικά αυτά κανάλια πληροφορίας μπορούν να συμβάλλουν από κοινού στη δημιουργία μίας ενιαίας πληρέστερης αναπαράστασης και κατ΄ επέκταση στην καλύτερη ταξινόμηση του βίντεο. Για τον συνδυασμό των διαφορετικών αναπαραστάσεων, έχουν προταθεί στη βιβλιογραφία διάφορες μέθοδοι, όπως οι ακόλουθες:

1. *Πρώιμη σύμμειξη (early fusion)*: Γίνεται συνδυασμός στο επίπεδο των αναπαραστάσεων, με απλή συνένωση τους (concatenation) πριν την τροφοδότηση του SVM. Στο πεδίο των hand-crafted χαρακτηριστικών, η συνένωση μπορεί να αφορά είτε σε επίπεδο περιγραφητών (π.χ. MBHx και MBHy) πριν την κωδικοποίηση, είτε σε επίπεδο BoW ιστογραμμάτων των επιμέρους κελιών στην κωδικοποίηση με χωρο-χρονικές πυραμίδες.

2. *Σύμμειξη τελικού σταδίου (late fusion)*: Οι πιθανότητες (confidence scores) που προκύπτουν από την ταξινόμηση ενός βίντεο, με διαφορετικό μοντέλο SVM για

κάθε κανάλι πληροφορίας, συνδυάζονται σε επόμενο στάδιο, συνήθως με κάποιον γραμμικό συνδυασμό, για τη διαμόρφωση της τελικής απόφασης.

3. *Σύμμειξη στο επίπεδο του μοντέλου*: Το κάθε κανάλι πληροφορίας χρησιμοποιείται για τον υπολογισμό των αποστάσεων μεταξύ των βίντεο (στη γραμμική περίπτωση) ή του πυρήνα του SVM (στη μη-γραμμική περίπτωση) και οι επιμέρους αποστάσεις ή πυρήνες συνδυάζονται σχηματίζοντας τον τελικό πυρήνα που χρησιμοποιείται. Πιο συγκεκριμένα, αν $\mathbf{x}_i^c, \mathbf{x}_j^c$ οι αναπαραστάσεις δύο διαφορετικών βίντεο i και j, υπολογισμένες για το c-οστό κανάλι πληροφορίας (π.χ. c-οστό περιγραφητή), τότε ο συνδυαστικός πυρήνας προκύπτει από τη σχέση:

$$K(\mathbf{x}_i^c, \mathbf{x}_j^c) = \exp\left(-\sum_{c=1}^{N_c} \frac{1}{A^c} D(\mathbf{x}_i^c, \mathbf{x}_j^c)\right) \tag{3.24}$$

όπου N_c ο αριθμός των διαφορετικών καναλιών, $D(\cdot;\cdot)$ η μετρική της ομοιότητας μεταξύ των αναπαραστάσεων των δύο βίντεο (π.χ. x^2) και A^c ένας παράγοντας κανονικοποίησης. Στην παραπάνω σχέση, αρχικά αθροίζονται οι αποστάσεις μεταξύ $\mathbf{x}_i^c$ και $\mathbf{x}_j^c$ για τα διάφορα κανάλια και εν συνεχεία υπολογίζεται ο πυρήνας του SVM. Εναλλακτικά, η άθροιση μπορεί να γίνει μετά τον υπολογισμό του πυρήνα, δηλαδή:

$$K(\mathbf{x}_i^c, \mathbf{x}_j^c) = \sum_{c=1}^{N_c} \exp\left(-\frac{1}{A^c} D(\mathbf{x}_i^c, \mathbf{x}_j^c)\right) \tag{3.25}$$

Στην παρούσα Διπλωματική χρησιμοποιείται η Σχέση 3.24 για τον συνδυασμό των διαφόρων περιγραφητών, ενώ για τον βέλτιστο συνδυασμό των επιμέρους πυρήνων του SVM χρησιμοποιείται η *Μάθηση Πολλαπλών Πυρήνων (Multiple Kernel Learning - MKL)* η οποία βασίζεται στη Σχέση 3.25. Για την MKL έχουν προταθεί διάφορες τεχνικές με πιο δημοφιλείς την Simple MKL των Rakotomamonjy et al. [79] και την Generalized MKL των Varma & Babu [80], βελτιωμένη έκδοση της οποίας παρουσιάσαν έπειτα οι Jain et al. [81] την οποία ονόμασαν SPG-GMKL.

Κεφάλαιο 4

Πειραματικά Αποτελέσματα

Στο κεφάλαιο αυτό παρουσιάζουμε τις μεθοδολογίες που ακολουθήσαμε και τα πειραματικά αποτελέσματα που αυτές επέφεραν στην ταξινόμηση ανθρώπινων δράσεων και χειρονομιών. Στην πρώτη ενότητα εφαρμόζουμε διάφορες διαμερίσεις των δεδομένων της Cognimuse τις οποίες αξιολογούμε μέσω υπαρχουσών μεθόδων, ενώ στη δεύτερη ενότητα αξιολογούνται διάφορες τεχνικές ταξινόμησης με τη χρήση ή την επανεκπαίδευση Συνελικτικών Νευρωνικών Δικτύων. Στην τρίτη ενότητα περιγράφουμε τη διαδικασία εξαγωγής deep learned χαρακτηριστικών από αρχιτεκτονικές two-stream ΣΝΔ, τα οποία αξιολογούνται, τόσο ανεξάρτητα όσο και σε συνδυασμό με hand-crafted ή άλλα deep learned χαρακτηριστικά, σε όλες τις διαθέσιμες βάσεις ανθρώπινων δράσεων.

4.1 Αξιολόγηση διαφορετικών διαμερίσεων της Cognimuse

Οι διαμερίσεις της Cognimuse γίνονται επιλέγοντας αρχικά τις κλάσεις που επιθυμούμε να πειραματιστούμε (όλες ή κάποιες από αυτές) και έπειτα δημιουργώντας τα υποσύνολα εκπαίδευσης και αξιολόγησης με δύο τρόπους: α) επιλέγοντας ένα ποσοστό από όλα τα δεδομένα για εκπαίδευση και το υπόλοιπο για αξιολόγηση (partition-based splitting) ή β) επιλέγοντας τις δράσεις μίας ταινίας για αξιολόγηση και αυτές των υπόλοιπων ταινιών για εκπαίδευση (movie-based splitting). Επίσης, σε κάθε περίπτωση εξετάζουμε τις κλάσεις ανισοκατανεμημένες, με το πλήθος των δεδομένων τους όπως έχει διαμορφωθεί από την υπάρχουσα προ-επεξεργασία που περιγράψαμε στην Ενότητα 1.3.5 (raw distribution).

4.1.1 Πειραματικό Πλαίσιο

Χρησιμοποιώντας υποσύνολα των 20 κλάσεων της Cognimuse χωρίζουμε τα δεδομένα σε διαφορετικά splits ακολουθώντας το μοτίβο των βάσεων UCF101 και HMDB51. Σε κάθε περίπτωση, για την εξαγωγή χαρακτηριστικών χρησιμοποιούμε τις improved trajectories (iDT) και τα C3D features. Για τις iDT εφαρμόζουμε Bag-of-Words κωδικοποίηση με ένα οπτικό λεξικό K=4000 κέντρων για κάθε περιγραφητή: TD, HoG, HoF, MBHx, MBHy, MBH. Ο υπολογισμός των κέντρων γίνεται με τον αλγόριθμο K-means ενώ στα ιστογράμματα κάθε περιγραφητή εφαρμόζεται L1 κανονικοποίηση. Η εξαγωγή των πυκνών τροχιών γίνεται μέσω της υλοποίησης των συγγραφέων[1] [15] χρησιμοποιώντας τις ίδιες παραμέτρους, δηλαδή $L = 15$ για το μήκος της τροχιάς, $N = 32$ για τον χωρο-χρονικό όγκο $N \times N \times L$ γύρω από την τροχιά και $n_\sigma = 2, n_\tau = 3$ για το χωρο-χρονικό πλέγμα $n_\sigma \times n_\sigma \times n_\tau$ διαίρεσης της γειτονιάς. Για τα C3D features χρησιμοποιείται το fine-tuned δίκτυο[2], το οποίο εφαρμόζεται πάνω σε *μη-επικαλυπτόμενα* κλιπς των 16 καρέ για κάθε βίντεο, μέσω του framework που παρέχουν οι συγγραφείς[3]. Έπειτα, υπολογίζεται ο μέσος όρος των διανυσμάτων εξόδου των επιμέρους κλιπς ακολουθούμενος από L2 κανονικοποίηση, σχηματίζοντας έτσι το διάνυσμα χαρακτηριστικών του βίντεο. Για την ταξινόμηση χρησιμοποιείται μία μη γραμμική SVM πολλαπλών κλάσεων με πυρήνα x^2 (chi-squared kernel) και στις δύο περιπτώσεις. Όσον αφορά τις iDT, πραγματοποιείται μία σύμμειξη στο επίπεδο του μοντέλου για τους περιγραφητές TD, HoG, HoF, MBHx, MBHy, σχηματίζοντας έναν ενιαίο πυρήνα x^2 που αναπαριστά τον Combined περιγραφητή. Επίσης, εφαρμόζεται σύμμειξη στο επίπεδο του μοντέλου των C3D features με τον περιγραφητή Combined των iDT, συνδυάζοντας τους x^2 πυρήνες των δύο ροών πληροφορίας, με μία απλή άθροιση, σε έναν ενιαίο πυρήνα, ο οποίος χρησιμοποιείται για την ταξινόμηση από το SVM[4].

4.1.2 Αποτελέσματα

Πλήρες πείραμα

Για αυτό το πείραμα χρησιμοποιούμε όλες τις κλάσεις της Cognimuse, ανισοκατανεμημένες - όπως έχουν προκύψει από την προ-επεξεργασία των δεδομένων - και εφαρμόζουμε και τα δύο είδη διαμερίσεων. Για την movie-based διαμέριση, κρατάμε τις δράσεις μίας ταινίας κάθε φορά για αξιολόγηση και αυτές των υπόλοιπων ταινιών για εκπαίδευση. Δημιουργούμε έτσι ένα training set με τις δράσεις από 6 διαφορετικές ταινίες,

[1] `http://lear.inrialpes.fr/~wang/improved_trajectories`

[2] `https://drive.google.com/open?id=0Bx-2rTokRAt1Vm9nLWItQTgtSVE`

[3] `http://vlg.cs.dartmouth.edu/c3d/`

[4] Ουσιαστικά πρόκειται για μία πρώιμη σύμμειξη των δύο αναπαραστάσεων (early fusion) μέσω μίας μη-γραμμικής συνένωσης

τις οποίες επιθυμούμε να εντοπίσουμε σε μία νέα, άγνωστη στον ταξινομητή, ταινία. Συμβολίζουμε κάθε split με το όνομα της ταινίας που χρησιμοποιείται για testing, δηλαδή *movie_out* όπου $movie = \{BMI, CHI, CRA, DEP, GLA, LOR, GWW\}$. Τα αποτελέσματα που πήραμε φαίνονται στον Πίνακα 4.1.

Split	iDT	C3D	C3D+iDT
BMI_out	47.7	41.7	**48.2**
CHI_out	21.5	15.9	**23.9**
CRA_out	43.4	34.8	**45.9**
DEP_out	43.1	33.1	**44.5**
GLA_out	29.4	27.8	**31.2**
LOR_out	32.3	30.4	**35.6**
GWW_out	**32.9**	26.1	31.7
Average	35.8	30.0	**37.3**

Πίνακας 4.1: Αποτελέσματα ταξινόμησης σε 20 κλάσεις για κάθε movie-based split της Cognimuse. Κάθε μέτρηση αναφέρεται στο ποσοστό σωστών ταξινομήσεων των δράσεων της ταινίας. Για τις iDT χρησιμοποιείται ο Combined περιγραφητής, που αποδίδει καλύτερα.

Με μια πρώτη ανάγνωση των αποτελεσμάτων παρατηρούμε ότι η επιτυχής αναγνώριση των ανθρώπινων δράσεων μέσα σε μία ταινία, χρησιμοποιώντας τη «γνώση» από μόνο 6 ταινίες είναι ένα δύσκολο και φιλόδοξο task. Όπως παρατηρούμε από τα ποσοστά αναγνώρισης του Πίνακα 4.1, ταινίες που ανήκουν σε συγγενικά είδη (π.χ. περιπέτεια, δράμα) εμφανίζουν κοντινή ακρίβεια ταξινόμησης. Συγκεκριμένα, οι ταινίες εποχής *Gladiator (GLA)* και *Gone with the wind (GWW)*, καθώς και η περιπέτεια φαντασίας *Lord of the rings (LOR)* εμφανίζουν σε πολλές περιπτώσεις ομοιότητες στο background τους ενώ περιέχουν σκηνές ανθρώπινων δράσεων που δεν ανήκουν σε άλλες ταινίες όπως π.χ. σκηνές ιππασίας (κλάση *ride horse*). Επίσης, έχουν σημαντικές διαφορές με τις υπόλοιπες ταινίες της βάσης οι οποίες αναφέρονται σε μία πιο σύγχρονη εποχή με την ακρίβεια ταξινόμησης τους να κυμαίνεται από $\sim 31\%$ έως $\sim 36\%$. Αντίστοιχα, οι ταινίες *Beatiful Mind (BMI)*, *Crash (CRA)* και *The Departed (DEP)*, αποτελούν περιπέτειες της σύγχρονης εποχής με ανθρώπινες δράσεις που πραγματοποιούνται υπό παρόμοιες συνθήκες και πολύ διαφορετικό φόντο από τις ταινίες εποχής/φαντασίας. Παρουσιάζουν λοιπόν υψηλότερα ποσοστά αναγνώρισης της τάξεως του $\sim 44-48\%$. Μία ειδική περίπτωση αποτελεί το μιούζικαλ *Chicago (CHI)* το οποίο είναι μεν ταινία που εκτυλίσσεται στη σύγχρονη περίοδο αλλά με πολύ διαφορετικές σκηνές από όλες τις υπόλοιπες ταινίες της βάσης. Χαρακτηριστικά αναφέρουμε ότι περιέχει τη μεγάλη πλειοψηφία σκηνών χορού (κλάση *dance*) της βάσης με αποτέλεσμα

όταν χρησιμοποιείται για αξιολόγηση να μην υπάρχουν αρκετά δεδομένα εκπαίδευσης γι΄ αυτή την κλάση. Συγκεκριμένα, η μόνη ταινία από τις υπόλοιπες που περιέχει σκηνές χορού είναι η *GWW* που όμως περιέχει περίπου το 1/5 από τις αντίστοιχες του *CHI* και οι οποίες πραγματοποιούνται υπό πολύ διαφορετικές συνθήκες. Συνεπώς, είναι λογικό κάποιες κατηγορίες δράσεων να έχουν χαμηλά ποσοστά αναγνώρισης, κάτι που οφείλεται στη μη ύπαρξη άλλης ταινίας αντίστοιχου είδους (μιούζικαλ) στη βάση δεδομένων και οδηγεί σε αρκετά χαμηλά ποσοστά ακρίβειας ταξινόμησης.

Βλέπουμε, λοιπόν, πόσο σημαντικό είναι να υπάρχουν στο σύνολο εκπαίδευσης δεδομένα από σχετικά είδη ταινιών, όταν στόχος είναι η αναγνώρισης των ανθρώπινων δράσεων σε μία νέα, άγνωστη ταινία αξιολόγησης. Οι παρόμοιες συνθήκες εκτέλεσης των δράσεων (φωτισμός, φόντο κ.α.), καθώς και η συχνότητα και ο τρόπος εκτέλεσης τους παίζουν σημαντικό ρόλο για τον ταξινομητή. Προκειμένου τα σύνολα εκπαίδευσης και αξιολόγησης να περιέχουν δεδομένα από όλες τις ταινίες εφαρμόζουμε δύο partition-based διαμερίσεις, επιλέγοντας για κάθε κλάση ένα ποσοστό δεδομένων για εκπαίδευση και τα υπόλοιπα δεδομένα για αξιολόγηση. Για την πρώτη διαμέριση χρησιμοποιούμε το 70% των δειγμάτων κάθε κλάσης για εκπαίδευση ενώ για τη δεύτερη διαμέριση αυξάνουμε το ποσοστό αυτό στο 80%. Τα αντίστοιχα ποσοστά των δεδομένων αξιολόγησης είναι προφανώς 30% και 20% αντίστοιχα. Κάθε πείραμα εκτελείται 3 φορές για 3 διαφορετικές τυχαίες επιλογές δεδομένων και η ακρίβεια ταξινόμησης του υπολογίζεται ως ο μέσος όρος των επιμέρους ακριβειών των 3 επαναλήψεων και φαίνεται στον Πίνακα 4.2.

Training ratio	iDT	C3D	C3D+iDT
70%	40.2	37.8	**42.9**
80%	40.4	36.1	**43.6**

Πίνακας 4.2: Αποτελέσματα ταξινόμησης σε 20 κλάσεις για δύο partition-based splits της Cognimuse με συντελεστές 0,7 και 0,8

Παρατηρούμε ότι η αύξηση των δεδομένων του training set δεν έδωσε μεγάλη ώθηση στην αναγνώριση, παρόλο που θεωρητικά ο ταξινομητής αποκτάει περισσότερη «γνώση» και γίνεται πιο εύρωστος σε διαφορετικές συνθήκες πραγματοποίησης των δράσεων. Αυτό συμβαίνει γιατί πρακτικά, κάποιες κλάσεις με λίγα δεδομένα δεν συμβάλλουν σημαντικά στην περαιτέρω αύξηση του συνόλου εκπαίδευσης. Τα ποσοστά ακρίβειας ταξινόμησης είναι υψηλότερα σε σχέση με τον μέσο όρο των movie-based διαμερίσεων του Πίνακα 4.1, κάτι που είναι λογικό καθώς πλέον η κάθε κλάση εκπαιδεύεται με δεδομένα από όλες τις ταινίες. Παρ΄ όλα αυτά, η επίδοση του ταξινομητή παραμένει σχετικά χαμηλή κάτι που οφείλεται τόσο στη φύση των δεδομένων όσο και στο περιορισμένο πλήθος ταινιών από τις οποίες αυτά προέρχονται.

Παρόμοια ζητήματα έχουν να αντιμετωπίσουν οι ερευνητές όταν χρησιμοποιούν και άλλες βάσεις ανθρώπινων δράσεων βασισμένες σε ταινίες. Ενδεικτικά, χρησιμοποιούμε τις iDT και τα C3D features με τον ίδιο τρόπο πάνω στις δημοφιλείς βάσεις δεδομένων Hollywood2 και HMDB51, τα βίντεο των οποίων όμως προέρχονται από πολλές διαφορετικές ταινίες και είναι σχεδόν ισοκατανεμημένα στις κλάσεις τους, προσφέροντας έτσι μεγαλύτερη ευρωστία από τη Cognimuse. Τα αποτελέσματα φαίνονται στον Πίνακα 4.3, όπου για την HMDB51 καταγράφεται ο μέσος όρος των 3 splits.

Database	iDT	C3D	C3D+iDT
Hollywood2	60.3	46.2	**61.4**
HMDB51	51.2	52.8	**59.6**

Πίνακας 4.3: Ποσοστά ακρίβειας ταξινόμησης στις Hollywood2 και HMDB51

Όπως παρατηρούμε, παρόλο που τα δεδομένα τους διαθέτουν αρκετά πλεονεκτήματα, τα ποσοστά αναγνώρισης στις δύο δημοφιλείς βάσεις ανθρώπινων δράσεων HMDB51 και Hollywood2, αν και σημαντικά μεγαλύτερα από τα αντίστοιχα της Cognimuse, δεν υπερβαίνουν το $\sim 50-60\%$. Γίνεται επομένως πιο κατανοητή η δυσκολία αναγνώρισης ανθρώπινων δράσεων στη βάση Cognimuse, καθώς α) διαθέτει σχετικά λίγα βίντεο (2238) αναλογικά με τον αριθμό των κλάσεων της (20), β) τα βίντεο της προέρχονται από πολύ λίγες ταινίες (7), (γ) οι ταινίες είναι πολύ διαφορετικές μεταξύ τους και (δ) το πλήθος των δειγμάτων είναι μικρό για κάποιες κλάσεις ($\sim 30-50$) και μεγάλο για κάποιες άλλες ($\sim 100-200$). Γι΄ αυτό το λόγο, πειραματιζόμαστε με λιγότερες κλάσεις σε μικρότερα πειράματα που είναι ευκολότερο να εξαχθούν συμπεράσματα, καθώς και πιο εφικτό να επιτευχθούν ικανοποιητικά ποσοστά αναγνώρισης, ειδικά όταν στόχος είναι να ταξινομηθούν σωστά οι ενδιαφερόμενες δράσης μίας άγνωστης ταινίας (movie-based splitting).

Πείραμα των «μικρών» κλάσεων

Για αυτό το πείραμα χρησιμοποιούμε τις 10 κλάσεις της Cognimuse με το μικρότερο πλήθος δειγμάτων ($\sim 30-60$), προκειμένου αφ΄ ενός να μειωθούν οι κλάσεις στο μισό και αφ΄ ετέρου να είναι πιο ισοκατανεμημένες. Οι κλάσεις αυτές είναι οι εξής: *climb stairs, dance, fall on the floor, grab hand, hugging, laugh, open door, sitting down, sitting up, throw.* Χρησιμοποιούμε μία partition-based διαμέριση με ποσοστό δεδομένων στο σύνολο εκπαίδευσης 80%. Λόγω της στοχαστικότητας του πειράματος, εφαρμόζουμε την τυχαία επιλογή των δεδομένων 5 φορές, επιλέγοντας σε κάθε επανάληψη έναν αριθμό $0,8 \cdot N_{class}$ τυχαίων δειγμάτων από κάθε κλάση για εκπαίδευση και τα υπόλοιπα για αξιολόγηση. Παίρνουμε έτσι 5 διαφορετικά υποσύνολα εκπαίδευσης

και αξιολόγησης για το ίδιο σύνολο δεδομένων και υπολογίζουμε την ακρίβεια ταξινόμησης του πειράματος ως τον μέσο όρο των ποσοστών των 5 επαναλήψεων. Τα αποτελέσματα, όπως φαίνονται στον Πίνακα 4.4, είναι σαφώς πιο ενθαρρυντικά.

Split	iDT	C3D	C3D+iDT
10_small_classes	55.2	47.3	**58.4**

Πίνακας 4.4: Αποτελέσματα ταξινόμησης στις 10 μικρότερες κλάσεις της Cognimuse με partition-based splitting συντελεστή $0,8$. Τα ποσοστά αναγνώρισης υπολογίζονται ως ο μέσος όρος της ακρίβειας ταξινόμησης 5 διαφορετικών επαναλήψεων του splitting

Πείραμα των «μεγάλων» κλάσεων

Για αυτό το πείραμα χρησιμοποιούμε τις υπόλοιπες 10 κλάσεις της Cognimuse, οι οποίες περιέχουν σαφώς μεγαλύτερο πλήθος δειγμάτων ($\sim 60 - 200$), οι οποίες είναι οι εξής: *cry, pick, point at something, ride horse, running, smile, standing up, turn, walk, wave hands*. Επαναλαμβάνουμε το ίδιο πειραματικό μοτίβο με παραπάνω και παίρνουμε τα αποτελέσματα του Πίνακα 4.5.

Split	iDT	C3D	C3D+iDT
10_big_classes	51.2	44.7	**54.1**

Πίνακας 4.5: Αποτελέσματα ταξινόμησης στις 10 μεγαλύτερες κλάσεις της Cognimuse με partition-based splitting με συντελεστή $0,8$. Τα ποσοστά αναγνώρισης υπολογίζονται ως ο μέσος όρος της ακρίβειας ταξινόμησης 5 διαφορετικών επαναλήψεων του splitting

Από αυτές τις κλάσεις, οι *turn* και *walk* έχουν αρκετά περισσότερα δείγματα από τις υπόλοιπες (~ 200). Προκειμένου να έχουμε ένα πιο ισοκατανεμημένο training set τις αφαιρούμε και κρατάμε τις υπόλοιπες οκτώ. Με την αντίστοιχη πειραματική μέθοδο έχουμε σημαντικά καλύτερα αποτελέσματα ταξινόμησης όπως φαίνονται στον Πίνακα 4.6.

Για τις ίδιες 8 κλάσεις εφαρμόζουμε τις movie-based διαμερίσεις όπως και παραπάνω για να διερευνήσουμε αν αυξάνεται η ακρίβεια ταξινόμησης των δράσεων ανά ταινία. Τα αποτελέσματα του Πίνακα 4.7, μαρτυρούν ότι όσο λιγότερες είναι κλάσεις τόσο πιο πιθανό είναι να επιτευχθούν καλύτερα ποσοστά αναγνώρισης ανθρώπινων δράσεων, ενώ σε συνάρτηση με τα αντίστοιχα αποτελέσματα των Πινάκων 4.4 και 4.6 συμπεραίνουμε ότι όσο πιο ισοκατανεμημένες είναι οι κλάσεις τόσο καλύτερα εκπαιδεύεται ο ταξινομητής.

Split	iDT	C3D	C3D+iDT
8_classes	52.7	51.6	**59.8**

Πίνακας 4.6: Αποτελέσματα ταξινόμησης σε 8 κλάσεις (αφαιρώντας τις δύο μεγαλύτερες κλάσεις) της Cognimuse με partition-based splitting με συντελεστή $0,8$. Τα ποσοστά αναγνώρισης υπολογίζονται ως ο μέσος όρος της ακρίβειας ταξινόμησης 5 διαφορετικών επαναλήψεων του splitting

Split	iDT	C3D	C3D+iDT
BMI_out	38.9	42.8	**48.7**
CHI_out	24.4	23.2	**29.2**
CRA_out	45.9	41.1	**49.8**
DEP_out	**55.4**	34.4	52.6
GLA_out	33.2	23.6	**36.1**
LOR_out	**44.3**	32.5	41.7
GWW_out	35.2	27.5	**38.0**
Average	39.6	32.2	**42.3**

Πίνακας 4.7: Ποσοστά ακρίβειας ταξινόμησης σε 8 κλάσεις για κάθε movie-based split της Cognimuse

4.2 Χρήση 3Δ Συνελικτικών Δικτύων

Στην προηγούμενη ενότητα χρησιμοποιήθηκαν για τη διεξαγωγή πειραμάτων τα C3D features όπως εξήχθησαν από τις L2-κανονικοποιημένες ενεργοποιήσεις του *fc6* επιπέδου του 3Δ ΣΝΔ εκπαιδευμένου στην I380K και fine-tuned στην Sports-1M, εφαρμοσμένο σε *μη-επικαλυπτόμενα* κλιπς των 16 καρέ για κάθε βίντεο εισόδου. Στην ενότητα αυτή διερευνούμε τις δυνατότητες που μας παρέχει η αρχιτεκτονική 3Δ συνελίξεων των Tran et al. [48], χρησιμοποιώντας διαφορετικές τεχνικές συσσώρευσης των features ή ξεχωριστή επανεκπαίδευση του δικτύου πάνω στα δεδομένα κάθε πειράματος (task-specific fine-tuning), προκειμένου να βελτιώσουμε το αποτέλεσμα της ταξινόμησης των δράσεων. Οι διάφορες μέθοδοι αξιολογούνται τόσο στη Cognimuse όσο και στις δημοφιλείς βάσεις ανθρώπινων δράσεων.

4.2.1 Πειραματικό Πλαίσιο

Χρησιμοποιώντας το framework που παρέχουν οι Tran et al. για την 3Δ αρχιτεκτονική τους, επανεκπαιδεύουμε το fine-tuned δίκτυο τους με τη χρήση των διαθέσιμων

training sets κάθε βάσης δεδομένων η οποία μας ενδιαφέρει. Λόγω του σχετικά μικρού όγκου δεδομένων των δημοφιλών βάσεων ανθρώπινων δράσεων, δεν είναι δυνατή η εξ΄ αρχής εκπαίδευση του δικτύου καθώς θα οδηγήσει σε overfitting. Για την αξιολόγηση του επανεκπαιδευμένου δικτύου, χρησιμοποιούμε end-to-end την αρχιτεκτονική του δικτύου πάνω στα αντίστοιχα validation sets. Επίσης, χρησιμοποιώντας το fine-tuned C3D δίκτυο των Tran et al. εξάγουμε τα *fc6* διανύσματα ενεργοποιήσεων από τα 16-frame κλιπς κάθε βίντεο, στα οποία εκτός από average pooling εφαρμόζουμε και max, multiplication pooling, πριν την L2 κανονικοποίηση, για το σχηματισμό του τελικού διανύσματος χαρακτηριστικών του βίντεο. Επίσης, εξετάζουμε το ενδεχόμενο τα κλιπς διάσπασης των βίντεο εισόδου να είναι κατά το ήμισυ (8 καρέ) επικαλυπτόμενα (8-frames overlap).

4.2.2 Αποτελέσματα

Fine-tuning της C3D αρχιτεκτονικής

Για κάθε βάση δεδομένων, χρησιμοποιούμε το παρεχόμενο training set για το fine-tuning του δικτύου, το οποίο εφαρμόζουμε στη συνέχεια πάνω στο αντίστοιχο testing set. Αξιολογούμε έτσι την αποτελεσματικότητα της 3Δ αρχιτεκτονικής του C3D δικτύου, τα ποσοστά αναγνώρισης της οποίας φαίνονται στον Πίνακα 4.8. Για τις UCF101 και HMDB51 καταγράφουμε τον μέσο όρο της ακρίβειας ταξινόμησης και των 3 splits ενώ για τη Cognimuse, λόγω του μικρού όγκου δεδομένων, η εκπαίδευση του δικτύου οδήγησε σε overfitting και κατ΄ επέκταση σε πολύ χαμηλά ποσοστά ταξινόμησης.

Database	C3D fine-tuned	C3D+iDT
UCF101	83.4	**86.7**
HMDB51	53.9	**59.6**
Hollywood2	50.7	**61.4**

Πίνακας 4.8: Ποσοστά ακρίβειας ταξινόμησης για την 3Δ αρχιτεκτονική του C3D δικτύου πάνω στις δημοφιλείς βάσεις δεδομένων χρησιμοποιώντας τον ταξινομητή Softmax. Στη δεξιά στήλη φαίνεται η αντίστοιχη ακρίβεια ταξινόμησης με τη χρήση των C3D features από το αρχικό δίκτυο συνδυασμένα με τις iDT και ταξινομημένα με ένα μη-γραμμικό SVM με πυρήνα x^2.

Όπως βλέπουμε η 3Δ αρχιτεκτονική του C3D δικτύου, αν επανεκπαιδευτεί για συγκεκριμένο σκοπό αποδίδει αρκετά καλά. Παρ΄ όλα αυτά, η εξαγωγή deep-learned χαρακτηριστικών από ένα «γενικού σκοπού» δίκτυο παρέχει τη δυνατότητα σύμμειξης τους με τις iDT, κάτι που βελτιώνει σημαντικά την ταξινόμηση λόγω της συμπληρωματικότητας τους.

Συσσώρευση C3D χαρακτηριστικών

Για την εξαγωγή των C3D χαρακτηριστικών, το βίντεο χωρίζεται σε κλιπς των 16 frames από τα οποία εξάγονται οι *fc6* ενεργοποιήσεις και στη συνέχεια υπολογίζεται ο μέσος όρος τους (average pooling) ακολουθούμενος από L2 κανονικοποίηση για το σχηματισμό του τελικού διανύσματος χαρακτηριστικών για όλο το βίντεο. Παρ΄ όλα αυτά, η πράξη του μέσου όρου αντιμετωπίζει τα επιμέρους διανύσματα των 16-frame κλιπς ως ξεχωριστές οντότητες και όχι ως μέρη μιας χρονικής ακολουθίας. Έτσι, προκειμένου να μην καταρριφθεί η χρονική συνέχεια ανάμεσα στα κλιπς εφαρμόζουμε δύο διαφορετικούς τρόπους συσσώρευσης των επιμέρους χαρακτηριστικών πριν την L2 κανονικοποίηση: α) την εύρεση της μέγιστης τιμής για κάθε κελί (max pooling) και β) τον πολλαπλασιασμό των τιμών των επιμέρους κελιών μεταξύ τους (multiplication pooling). Δοκιμάζουμε τις δύο τεχνικές στη Hollywood2, στα 3 splits της HMDB51 και στο 8-classes split της Cognimuse, χρησιμοποιώντας τον μέσο όρο των 5 επαναλήψεων της partition-based διαμέρισης της προηγούμενης ενότητας. Τα αποτελέσματα φαίνονται στον Πίνακα 4.9.

Database	average	max	multipl.
HMDB51	52.8	**53.4**	47.2
Hollywood2	46.2	**48.1**	29.3
Cognimuse	51.6	**53.8**	51.4

Πίνακας 4.9: Ακρίβεια ταξινόμησης για διαφορετικές μεθόδους συσσώρευσης των C3D χαρακτηριστικών

Όπως παρατηρούμε, η εύρεση της μέγιστης τιμής για κάθε κελί των επιμέρους διανυσμάτων χαρακτηριστικών αποδίδει καλύτερα, όπως αναμέναμε, καθώς διατηρεί σε έναν βαθμό την χρονική ακολουθία ανάμεσα στα κλιπς. Αντίθετα ο πολλαπλασιασμός τους, αποδεικνύεται η χειρότερη επιλογή, κάτι που εν μέρει δικαιολογείται από το γεγονός ότι αν ένα διάνυσμα από τα κλιπς έχει σε μία θέση του την τιμή 0 τότε αυτή μεταφέρεται και στο τελικό διάνυσμα, απορρίπτοντας έτσι τις αντίστοιχες τιμές από τα υπόλοιπα κλιπς, οι οποίες μπορεί να περιέχουν σημαντική πληροφορία.

Επικαλυπτόμενα κλιπς

Χρησιμοποιώντας το εύρημα της παραπάνω ενότητας, εξάγουμε τα C3D features σε κλιπς τα οποία είναι επικαλυπτόμενα κατά το ήμισυ και τα συγκεντρώνουμε σε ένα ενιαίο διάνυσμα για όλο το βίντεο μέσω max pooling. Προσπαθούμε, με αυτόν τον τρόπο, να υπάρχουν κάποια κοινά τμήματα μεταξύ των επιμέρους κλιπς, δίνοντας

στον ταξινομητή περισσότερη πληροφορία για τη χρονική συνοχή τους, κάτι που όπως φαίνεται στον Πίνακα 4.10 δίνει ώθηση στην αναγνώριση.

Database	C3D	C3D_ovrlp	C3D_ovrlp + iDT
HMDB51	53.4	**53.9**	**60.4**
Hollywood2	48.1	**48.6**	**61.9**
Cognimuse	53.8	**54.3**	**60.2**

Πίνακας 4.10: Ακρίβεια ταξινόμησης για εξαγωγή C3D χαρακτηριστικών σε επικαλυπτόμενα κλιπς με max συσσώρευση σε συνδυασμό με τις iDT

Παρατηρούμε, λοιπόν, ότι τα C3D χαρακτηριστικά, εξαγμένα από επικαλυπτόμενα κλιπς και συσσωρευμένα με max pooling σε συνδυασμό με τις iDT αποδίδουν βέλτιστα. Έτσι, εφαρμόζουμε τη συγκεκριμένη σύμμειξη χαρακτηριστικών στις movie-based διαμερίσεις της Cognimuse, η οποία επιβεβαιώνει το συμπέρασμα μας, όπως φαίνεται στον Πίνακα 4.11.

Split	C3D + average	C3D_ovrlp + max pool	C3D + iDT	C3D_ovrlp + iDT
BMI_out	42.8	44.7	48.7	48.9
CHI_out	23.2	23.2	29.2	29.2
CRA_out	41.1	46.8	49.8	51.0
DEP_out	34.4	35.9	52.6	52.6
GLA_out	23.6	26.5	36.1	36.8
LOR_out	32.5	34.0	41.7	46.3
GWW_out	27.5	28.9	38.0	38.4
Average	32.2	**34.3**	42.3	**43.4**

Πίνακας 4.11: Ποσοστά ακρίβειας ταξινόμησης σε 8 κλάσεις για κάθε movie-based split της Cognimuse σε σύγκριση με την αρχική μέθοδο

4.3 Χρήση Συνελικτικών Δικτύων Δύο Ροών

Αφού έχουμε αναλύσει και χρησιμοποιήσει εκτενώς τεχνικές που αφορούν τα 3Δ Συνελικτικά Νευρωνικά Δίκτυα, στην παρούσα ενότητα εξετάζουμε τις δυνατότητες που μας προσφέρουν τα ΣΝΔ Διπλής Ροής (επαν-)εκπαιδευμένα σε μεγάλες βάσεις ανθρώπινων δράσεων όπως οι UCF101 και HMDB51. Χαρακτηριστικό παράδειγμα

τέτοιων δικτύων είναι η αρχιτεκτονική των Temporal Segment Networks [51] η οποία επιτυγχάνει από μόνη της πολύ υψηλά ποσοστά αναγνώρισης (end-to-end evalutation). Εκπαιδεύοντας την στις δύο προαναφερθείσες δημοφιλείς βάσεις δεδομένων, εξάγουμε μοντέλα spatial και temporal streams τα οποία έπειτα χρησιμοποιούμε ως δίκτυα εξαγωγής χαρακτηριστικών εμφάνισης και κίνησης αντίστοιχα. Η τεχνική αυτή αποδεικνύεται καλύτερη από την end-to-end χρησιμοποίηση των δύο δικτύων καθώς τα παραγόμενα χαρακτηριστικά μπορούν να συνδυαστούν με άλλα συμπληρωματικά και να δώσουν πολύ μεγάλη ώθηση στην ακρίβεια ταξινόμησης.

4.3.1 Πειραματικό Πλαίσιο

Χρησιμοποιώντας το framework που παρέχουν οι Wang et al.[5] για τα Temporal Segment Networks (TSN) εκπαιδεύουμε το χωρικό και το χρονικό δίκτυο σε καθένα από τα 3 σπλιτς της UCF101 και της HMDB51. Κάθε εκπαιδευμένο μοντέλο αξιολογείται end-to-end στο αντίστοιχο validation set κάθε split, χρησιμοποιώντας την μέθοδο που ακολουθούν και οι συγγραφείς. Σύμφωνα με αυτή, από κάθε βίντεο δειγματοληπτούνται 25 καρέ με ίση χρονική απόσταση μεταξύ τους, καθένα από τα οποία παρέχει 10 εισόδους στο δίκτυο μέσω περικοπής και αναστροφής των 4 γωνιών και του κέντρου του. Για το χωρικό δίκτυο οι 10 αυτές είσοδοι είναι στατικές εικόνες ενώ για το χρονικό δίκτυο είναι στοίβες από 10 καρέ οπτικής ροής (optical flow stacks) (5 για την οριζόντια μετατόπιση και 5 για την κάθετη). Έπειτα υπολογίζεται ο μέσος όρος των επιμέρους class scores των 10 εισόδων όλων των καρέ, ενώ για τα τελικά class scores εφαρμόζεται μία αργή σύμμειξη μέσω ενός σταθμισμένου μέσου όρου, με βάρη 1 για το χωρικό και 1.5 για το χρονικό δίκτυο, πριν την εφαρμογή της κανονικοποίησης Softmax.

Στη συνέχεια, χρησιμοποιώντας δύο εκπαιδευμένα χωρικά και χρονικά δίκτυα, εξάγουμε από καθένα τις ενεργοποιήσεις του τελευταίου global pooling επιπέδου πριν από το fully-connected layer των class scores. Πρόκειται για ένα layer το οποίο εφαρμόζει average pooling με πυρήνα 7×7 και βήμα 1 στην έξοδο του ακριβώς προηγούμενου Inception Module. Η έξοδος του είναι ένα διάνυσμα 1024 θέσεων στο οποίο έπειτα εφαρμόζεται το dropout. Ο λόγος που επιλέγουμε να κρατήσουμε τις ενεργοποιήσεις ακριβώς πριν το dropout είναι γιατί η απόσυρση των νευρώνων θα οδηγήσει σε σημαντική απώλεια πληροφορίας. Για την εξαγωγή χαρακτηριστικών από ένα βίντεο ακολουθούμε την ίδια διαδικασία με το μοτίβο αξιολόγησης των συγγραφέων με τη διαφορά ότι ο μέσος όρος των 10 εισόδων και των 25 καρέ εφαρμόζεται για τις 1024 ενεργοποιήσεις των global pooling εξόδων, ενώ για τον σχηματισμό του τελικού διανύσματος χαρακτηριστικών εφαρμόζεται L2 κανονικοποίηση.

[5] `https://github.com/yjxiong/temporal-segment-networks`

4.3.2 Αποτελέσματα

End-to-end χρήση των TSN

Για κάθε split των βάσεων δεδομένων UCF101 και HMDB51, χρησιμοποιούμε το training set για την εκπαίδευση του χωρικού και του χρονικού δικτύου των TSN, τα οποία εν συνεχεία αξιολογούνται στα αντίστοιχα validation sets. Η ακρίβεια ταξινόμησης που επιτυγχάνουν τα δίκτυα για κάθε βάση φαίνεται στον Πίνακα 4.12 όπου καταγράφονται τα ποσοστά για κάθε split ξεχωριστά αλλά και ο μέσος όρος τους.

Database	CNNs	Split 1	Split 2	Split 3	Average
UCF101	Spatial Stream	85.2	84.8	85.0	85.0
	Temporal Stream	87.5	90.2	90.3	89.3
	Two-Stream	93.2	94.4	93.8	**93.8**
HMDB51	Spatial Stream	53.8	49.9	48.7	50.8
	Temporal Stream	62.2	63.0	63.6	62.9
	Two-Stream	69.2	67.1	68.0	**68.1**

Πίνακας 4.12: Ακρίβεια ταξινόμησης για κάθε split των UCF101 και HMDB51 για το χωρικό, το χρονικό και το δίκτυο διπλής ροής των TSN

Όπως βλέπουμε, τα Δίκτυα Χρονικών Τμημάτων επιτυγχάνουν πολύ υψηλά ποσοστά αναγνώρισης και αποτελούν μία πολύ αξιόπιστη επιλογή για την εξαγωγή χωροχρονικών χαρακτηριστικών.

Εξαγωγή χωρο-χρονικών χαρακτηριστικών από τα TSN

Δοθέντος ενός βίντεο, δειγματοληπτούμε 25 καρέ με ίση χρονική απόσταση μεταξύ τους από τα οποία παράγουμε 10 εισόδους περικόπτοντας και αναστρέφοντας τις 4 γωνίες και το κέντρο τους και εφαρμόζουμε πάνω τους ένα εκπαιδευμένο χωρικό TSN δίκτυο από το οποίο εξάγουμε τις global pool ενεργοποιήσεις, για τις οποίες υπολογίζουμε τον μέσο όρο αρχικά για τις 10 διαφορετικές εισόδους και έπειτα για τα 25 καρέ, ενώ σε τελικό στάδιο εφαρμόζουμε L2 κανονικοποίηση και λαμβάνουμε ένα διάνυσμα χαρακτηριστικών εμφάνισης για το βίντεο. Αντίστοιχα, με την ίδια μέθοδο παίρνουμε στοίβες οπτικής ροής για κάθε μία από τις 10 εισόδους των 25 καρέ και εξάγουμε τις global pool ενεργοποιήσεις του αντίστοιχου χρονικού δικτύου, οι οποίες με παρόμοιο τρόπο σχηματίζουν ένα διάνυσμα χαρακτηριστικών κίνησης για το βίντεο.

Τα χαρακτηριστικά από κάθε δίκτυο χρησιμοποιούνται τόσο ξεχωριστά όσο και συνδυαστικά για την αναγνώριση ανθρώπινων δράσεων μέσω ενός μη γραμμικού x^2 SVM ταξινομητή, όπως και παραπάνω. Αξιολογούμε την επίδοση των χαρακτηριστικών

αυτών τόσο στις δημοφιλείς βάσεις δεδομένων HMDB51, Hollywood2 όσο και στη Cognimuse. Για την HMDB51 η αξιολόγηση γίνεται στο split 1, ενώ για τη Cognimuse στο 8-classes split. Η εξαγωγή χαρακτηριστικών γίνεται με τη χρήση του ίδιου two-stream μοντέλου, εκπαιδευμένου πάνω στο training set του split 1 της HMDB51, και για τις 3 βάσεις δεδομένων, καθώς οι δράσεις και τα δεδομένα τους εμφανίζουν αρκετές ομοιότητες. Τα αποτελέσματα φαίνονται στον Πίνακα 4.13.

Database	TSN rgb	TSN flow	TSN rgb+flow
HMDB51(split1)	55.6	63.1	**70.2**
Hollywood2	51.8	63.7	**67.4**
Cognimuse	50.8	52.2	**60.5**

Πίνακας 4.13: Ποσοστά αναγνώρισης των χαρακτηριστικών εμφάνισης και κίνησης από τα TSN μέσω μη γραμμικής x^2 SVM ταξινόμησης. Ο συνδυασμός τους γίνεται με σύμμειξη των πυρήνων τους στο επίπεδο του μοντέλου ταξινόμησης.

Όπως παρατηρούμε, τα χωρο-χρονικά TSN χαρακτηριστικά, στην περίπτωση της HMDB51 υπερβαίνουν το αντίστοιχο ποσοστό της end-to-end αξιολόγησης ενώ στις Hollywood2 και Cognimuse επιτυγχάνουν πολύ υψηλότερα ποσοστά από αντίστοιχες προηγούμενες μεθόδους. Έχοντας πλέον στη διάθεση μας διανύσματα deep-learned χαρακτηριστικών από μοντέλα Two-Stream CNNs μπορούμε να τα συνδυάσουμε με άλλα παρόμοια χαρακτηριστικά όπως τα C3D ή με τις Bag-of-Words κωδικοποιημένες βελτιωμένες τροχιές οι οποίες περιγράφουν «χαμηλού επιπέδου» συμπληρωματική πληροφορία και μπορούν να αυξήσουν περαιτέρω την ακρίβεια ταξινόμησης, όπως φαίνεται στον Πίνακα 4.14.

Database	TSN rgb+flow	TSN + iDT	TSN + C3D + iDT
HMDB51(split1)	70.2	72.5	**73.8**
Hollywood2	67.4	71.7	**72.4**
Cognimuse	60.5	61.9	**62.6**

Πίνακας 4.14: Ποσοστά ακρίβειας ταξινόμησης για τα TSN features σε συνδυασμό με τις iDT με BoW κωδικοποίηση και τα C3D features εξαγμένα με τη βέλτιστη μέθοδο της προηγούμενης ενότητας. Ο συνδυασμός τους γίνεται με σύμμειξη των πυρήνων τους στο επίπεδο του μοντέλου ταξινόμησης.

Βλέπουμε, λοιπόν, ότι ο συνδυασμός των TSN rgb+flow features με τα C3D features και τις iDT δίνει μεγάλη ώθηση στην ακρίβεια ταξινόμησης, τα ποσοστά της οποίας είναι αρκετά υψηλά. Όπως αναφέραμε και παραπάνω οι βελτιωμένες τροχιές

περιγράφουν «χαμηλού επιπέδου» συμπληρωματική πληροφορία ενώ τα C3D χαρακτηριστικά, με τις 3Δ συνελίξεις του δικτύου τους, περιγράφουν τη σχέση κάθε frame του βίντεο με τα γειτονικά του, εμπλουτίζοντας με επιπλέον, διαφορετικού τύπου, χωροχρονική πληροφορία τα στατικά frames από τα οποία εξάγονται τα TSN rgb χαρακτηριστικά.

Ως ένα επιπλέον βήμα για την περαιτέρω βελτίωση της ακρίβειας ταξινόμησης, εξάγουμε τα TSN features και από τα μοντέλα που είναι εκπαιδευμένα στα άλλα δύο splits της HMDB51 και για το τελικό διάνυσμα χαρακτηριστικών ενώνουμε (concatenation) τις ενεργοποιήσεις και από τα 3 μοντέλα σχηματίζοντας ένα διάνυσμα $3 \cdot 1024 = 3072$ θέσεων, στο οποίο εφαρμόζουμε έπειτα L2 κανονικοποίηση. Επίσης, για τη συσσώρευση των επιμέρους χαρακτηριστικών των 25 καρέ χρησιμοποιούμε max pooling αντί για τον μέσο όρο, τεχνική η οποία μας δίνει τα ποσοστά αναγνώρισης του Πίνακα 4.15.

Method	HMDB51 (split1)	Hollywood2	Cognimuse
TSN(3 nets) + C3D + iDT	**74.4**	**73.1**	**63.7**

Πίνακας 4.15: Ποσοστά ακρίβειας ταξινόμησης με τα TSN features συσσωρευμένα με max pooling και εξαγμένα και από τα 3 εκπαιδευμένα μοντέλα των splits της HMDB51

Εφαρμόζουμε την τελική μας μέθοδο στα movie-based splits της Cognimuse για την περίπτωση των 8 κλάσεων και καταγράφουμε τα αποτελέσματα στον Πίνακα 4.16, στον οποίο φαίνεται η μεγάλη συμβολή των TSN features στην βελτίωση της αναγνώρισης ανθρώπινων δράσεων.

Split	C3D + iDT	TSN + C3D + iDT
BMI_out	48.9	62.3
CHI_out	29.2	32.9
CRA_out	51.0	56.3
DEP_out	52.6	60.4
GLA_out	36.8	46.7
LOR_out	46.3	50.2
GWW_out	38.4	46.1
Average	43.4	**50.7**

Πίνακας 4.16: Ποσοστά ακρίβειας ταξινόμησης σε 8 κλάσεις για κάθε movie-based split της Cognimuse σε σύγκριση με την μέθοδο της προηγούμενης ενότητας

4.4 Προτεινόμενη Μέθοδος

Έπειτα από μία σειρά πειραμάτων με τη χρήση διαφορετικών μεθόδων εξαγωγής χαρακτηριστικών, αξιοποιώντας τις πολύ αποδοτικές σύγχρονες αρχιτεκτονικές Συνελικτικών Νευρωνικών Δικτύων και συνδυάζοντας τις με τις παραδοσιακές hand-crafted τεχνικές της Όρασης Υπολογιστών καταλήξαμε, λοιπόν, σε μία μέθοδο που πετυχαίνει υψηλή ακρίβεια αναγνώρισης ανθρώπινων δράσεων, πλησιάζοντας τις πιο αποδοτικές σύγχρονες μεθόδους της βιλβιογραφίας. Συγκεκριμένα, λοιπόν, η προτεινόμενη μέθοδος χρησιμοποιεί έναν μη γραμμικό x^2 SVM ταξινομητή ο οποίος συνδυάζει τους πυρήνες (kernels) των ακόλουθων χαρακτηριστικών:

- *TSN rgb features*: Είναι τα χαρακτηριστικά που εξάγονται από το global pooling layer του χωρικού δικτύου της αρχιτεκτονικής των Temporal Segment Networks, ακολουθούμενα από L2 κανονικοποίηση. Από το βίντεο εισόδου δειγματοληπτούνται 25 καρέ με ίση χρονική απόσταση μεταξύ τους, από τα οποία παράγονται 10 είσοδοι περικόπτοντας και αναστρέφοντας τις 4 γωνίες και το κέντρο τους. Οι global pooling ενεργοποιήσεις των 10 εισόδων συγκεντρώνονται μέσω του μέσου όρους τους σε ένα διάνυσμα για κάθε καρέ και έπειτα τα διανύσματα αυτά συσσωρεύονται σε ένα ενιαίο διάνυσμα 1024 θέσεων για όλο το βίντεο μέσω max pooling. Το δίκτυο που χρησιμοποιήθηκε ήταν εκπαιδευμένο στο split 1 της βάσης δεδομένων HMDB51, ενώ στη συνέχεια χρησιμοποιήθηκαν και τα εκπαιδευμένα μοντέλα στα άλλα δύο splits της HMDB51 σχηματίζοντας ένα ενιαίο διάνυσμα χαρακτηριστικών 3072 θέσεων.

- *TSN flow features*: Αντίστοιχα με τα παραπάνω, με τη διαφορά ότι εξάγονται από το χρονικό δίκτυο των TSN πάνω σε στοίβες οπτικής ροής μεγέθους 5 frames για κάθε διάσταση.

- *C3D features*: Πρόκειται για τα χαρακτηριστικά που εξάγονται από τις fc6 ενεργοποιήσεις του fine-tuned C3D δικτύου πάνω στην Sports-1M, ακολουθούμενα από L2 κανονικοποίηση. Το βίντεο εισόδου χωρίζεται σε κλιπς των 16 καρέ, *επικαλυπτόμενα μεταξύ τους κατά το ήμισυ (8-frames overlap)*, σε καθένα από τα οποία εφαρμόζεται το C3D δίκτυο και παράγει τις fc6 ενεργοποιήσεις του. Στη συνέχεια, τα επιμέρους διανύσματα ενεργοποιήσεων των κλιπς συσσωρεύονται με max pooling σε ένα ενιαίο διάνυσμα χαρακτηριστικών 4096 θέσεων για όλο το βίντεο.

- *improved trajectories (iDT)*: Συμβολίζουν ουσιαστικά τον Combined περιγραφητή των iDT ο οποίος περιέχει τον συνδυασμό των περιγραφητών TD, HoG, HoF, MBHx, MBHy. Οι πυκνές τροχιές εξάγονται για ένα μήκος $L = 15$ σε

έναν χωρο-χρονικό όγκο $N \times N \times L$, όπου $N = 32$, ο οποίος διαιρείται σε ένα χωρο-χρονικό πλέγμα $n_\sigma \times n_\sigma \times n_\tau$, όπου $n_\sigma = 2, n_\tau = 3$. Για κάθε περιγραφητή πραγματοποιείται Bag-of-Words κωδικοποίηση με ένα οπτικό λεξικό $K = 4000$ κέντρων, υπολογισμένα με τον αλγόριθμο K-means, ενώ στα ιστογράμματα των περιγραφητών εφαρμόζεται L1 κανονικοποίηση. Η σύμμειξη των περιγραφητών γίνεται στο επίπεδο του ταξινομητή αθροίζοντας του πηρήνες τους x^2 ενώ η ίδια διαδικασία ακολουθείται και για τον συνδυασμό τους με τα παραπάνω deep learned features. Ουσιαστικά, δηλαδή, ο τελικός περιγραφητής της μεθόδου μας προκύπτει από τη σύμμειξη των περιγραφητών TSN rgb, TSN flow, C3D, TD, HoG, HoF, MBHx, MBHy.

Στον Πίνακα 4.17 καταγράφουμε τα ποσοστά ακρίβειας ταξινόμησης της μεθόδου μας πάνω στις δύο δημοφιλείς βάσεις ανθρώπινων δράσεων HMDB51 (μέσος όρος και των 3 splits) και Hollywood2, σε σύγκριση με άλλες μεθόδους της βιβλιογραφίας. Όπως φαίνεται, η μέθοδος που αναπτύχθηκε στην παρούσα Διπλωματική Εργασία ξεπερνάει τις επιδόσεις των περισσότερων υπαρχόντων μεθόδων στη Hollywood2, ενώ επιτυγχάνει state-of-the-art ακρίβεια ταξινόμησης στην HMDB51.

Method	HMDB51	Hollywood2
iDT+BoW [15]	52.1	62.2
iDT+FV [15]	57.2	64.3
VideoDarwin [47]	63.7	73.7
Two-stream [17]	59.4	-
TDDs [49]	65.9	-
VGG+iDT [50]	69.2	-
HRP+iDT [82]	69.4	76.7
TSN [51]	68.5	-
TLEs [54]	71.1	-
EPT+iDT [83]	-	**78.6**
SSN [52]	73.8	-
Ours	**74.0**	73.1

Πίνακας 4.17: Σύγκριση της μεθόδου μας με άλλες δημοφιλείς μεθόδους της βιβλιογραφίας

Κεφάλαιο 5

Επίλογος - Συμπεράσματα

5.1 Συμβολή της Διπλωματικής Εργασίας

Στην παρούσα Διπλωματική Εργασία εξετάστηκε το πρόβλημα της αναγνώρισης ανθρώπινων δράσεων χρησιμοποιώντας τεχνικές τόσο της Όρασης Υπολογιστών όσο και των Νευρωνικών Δικτύων. Αντιμετωπίσαμε την αναγνώριση ως ένα πρόβλημα ταξινόμησης, όπου εφαρμόσαμε διάφορες μεθόδους εξαγωγής και κωδικοποίησης χαρακτηριστικών τις οποίες αξιολογήσαμε τόσο στις δημοφιλείς διαθέσιμες βάσεις δεδομένων όσο και σε μία βάση ανθρώπινων δράσεων που δημιουργήθηκε από την επισημείωση των οπτικών γεγονότων πάνω σε αποσπάσματα ταινιών και την κατάτμηση των σχετικών βίντεο.

Αρχικά, περιγράψαμε τις κύριες προσεγγίσεις της διεθνούς βιβλιογραφίας, αναλύοντας τη θετική συνεισφορά τους καθώς και τις αδυναμίες τους, ενώ διακρίναμε δύο κύριες τάσεις μεθόδων, τις «χαμηλού επιπέδου» hand-crafted τεχνικές και τις «υψηλού επιπέδου» τεχνικές βαθιάς μάθησης. Στις πρώτες ανήκουν μαθηματικές μέθοδοι εξαγωγής χαρακτηριστικών που εφαρμόζονται σε μία μικρή περιοχή της εικόνας ή του βίντεο ενώ οι δεύτερες αναπτύχθηκαν εκτενώς πρόσφατα με τη χρήση των Συνελικτικών Νευρωνικών Δικτύων τα οποία αποτελούν πολυεπίπεδες μονάδες μη γραμμικής επεξεργασίας των δεδομένων για την εξαγωγή χαρακτηριστικών. Συγκρίναμε τις πιο χαρακτηριστικές μεθόδους των δύο αυτών τάσεων, ενώ εφαρμόσαμε κάποιες νέες τεχνικές, συνδυάζοντας τα hand-crafted με τα deep-learned χαρακτηριστικά, βελτιώνοντας έτσι το αποτέλεσμα της ταξινόμησης. Οι βελτιωμένες τροχιές (improved trajectories) αποδεικνύονται μία αξιόπιστη επιλογή για την ταξινόμηση δράσεων ακόμη και σε πιο σύνθετα περιβάλλοντα ενώ ο συνδυασμός τους με τα προερχόμενα από βαθιά μάθηση C3D χαρακτηριστικά οδηγεί σε πιο αποδοτικές αναπαραστάσεις. Επίσης, τα Συνελικτικά Νευρωνικά Δίκτυα Διπλής Ροής (Two-Stream ConvNets) συμβάλλουν σημαντικά στην εξέλιξη του πεδίου πετυχαίνοντας τα υψηλότερα ποσοστά αναγνώρισης ανθρώπινων δράσεων.

Τα πρώτα μας πειράματα εστίασαν στην αξιολόγηση της νέας βάσης ανθρώπινων δράσεων Cognimuse η οποία δημιουργήθηκε στο πλαίσιο αυτής της Διπλωματικής Εργασίας και παρουσιάζει ιδιαίτερες προκλήσεις, καθώς περιέχει βίντεο προερχόμενα από 7 ταινίες, που χαρακτηρίζονται από συχνές επικαλύψεις, απότομες κινήσεις της κάμερας, αλλαγές στο φωτισμό και πολύ «ακατάστατο» φόντο, ενώ τα δεδομένα είναι ανισοκατανεμημένα στις 20 κλάσεις της. Για την αξιολόγηση της χρησιμοποιούμε τις iDT με Bag-of-Words κωδικοποίηση και τα C3D features τόσο ξεχωριστά, όσο και τον συνδυασμό τους, εκπαιδεύοντας έναν μη γραμμικό x^2 SVM ταξινομητή. Λαμβάνοντας υπόψιν όλες τις κλάσεις της Cognimuse, τα ποσοστά αναγνώρισης είναι σχετικά χαμηλά ενώ χωρίζοντας τις κλάσεις σε «μικρές» και «μεγάλες» και εφαρμόζοντας την ίδια μέθοδο, η ακρίβεια ταξινόμησης βελτιώνεται σημαντικά. Για πιο ισοκατανεμήμενο σύνολο δεδομένων απορρίπτουμε τις δύο μεγαλύτερες κλάσεις, κρατώντας 8 «μεγάλες», επιτυγχάνοντας ποσοστά αναγνώρισης συγκρίσιμα με αυτά των αντίστοιχων βάσεων δεδομένων της βιβλιογραφίας.

Στη συνέχεια, επανεκπαιδεύσαμε την αρχιτεκτονική του C3D δικτύου πάνω στα training sets των δημοφιλών βάσεων δεδομένων, το οποίο αξιολογήσαμε στα αντίστοιχα testing sets με τη χρήση του Softmax layer. Συγκρίναμε τα αποτελέσματα με τα αντίστοιχα που πήραμε από τη σύμμειξη των C3D χαρακτηριστικών εξαγμένων από το αρχικό, fine-tuned στην Sports-1M, δίκτυο με τις improved trajectories και την ταξινόμηση τους με τη χρήση ενός x^2 SVM, μέθοδος η οποία αποδείχτηκε αποτελεσματικότερη. Το συγκεκριμένο εύρημα είναι ιδιαιτέρως σημαντικό καθώς αποδεικνύει τη συμπληρωματικότητα των deep-learned C3D features με τις improved trajectories και αναδεικνύει τον σημαντικό ρόλο των hand-crafed τεχνικών στη αναγνώριση ανθρώπινων δράσεων, οι οποίες αν συνδυαστούν με κατάλληλο τρόπο με τις τεχνικές βαθιάς μάθησης μπορούν να οδηγήσουν σε μεθόδους πολύ αποτελεσματικές. Βασισμένοι σε αυτό το συμπέρασμα, επιχειρήσαμε έπειτα να βελτιώσουμε την αποδοτικότητα των C3D χαρακτηριστικών προσπαθώντας να εντάξουμε περισσότερο τη χρονική πληροφορία στη μέθοδο εξαγωγής τους και καταλήξαμε στο συμπέρασμα ότι για βέλτιστη ακρίβεια ταξινόμησης τα C3D features πρέπει να εξάγονται από επικαλυπτόμενα κατά το ήμισυ κλιπς (8 καρέ επικάλυψη) και στη συνέχεια να συγκεντρώνονται μέσω max pooling σε ενιαίο διάνυσμα χαρακτηριστικών για το βίντεο.

Έπειτα, δώσαμε έμφαση στα Συνελικτικά Δίκτυα Διπλής Ροής (Two-Stream CNNs), τα οποία αποτελούν τα πιο αποδοτικά σύγχρονα εργαλεία αναγνώρισης ανθρώπινων δράσεων, χρησιμοποιώντας την αρχιτεκτονική των Temporal Segment Networks. Αφού αξιολογήσαμε την επίδοση τους πάνω σε δύο πολύ δημοφιλείς και αρκετά μεγάλες βάσεις δεδομένων, τις UCF101 και HMDB51, χρησιμοποιήσαμε τα εκπαιδευμένα μοντέλα τους για την εξαγωγή χαρακτηριστικών εμφάνισης και κίνησης από το global pooling επίπεδο τους, ακριβώς πριν το dropout και την εφαρμογή του Softmax classifier. Έ-

πειτα εφαρμόζοντας max pooling μαζί με L2 κανονικοποίηση και με τη βοήθεια ενός μη γραμμικού SVM με πυρήνα x^2 αξιολογήσαμε τα TSN χαρακτηριστικά τόσο στις HMDB51, Hollywood2 όσο και στη Cognimuse. Τα χαρακτηριστικά εμφάνισης προέρχονται από ένα χωρικό δίκτυο με εισόδους στατικά frames, ενώ τα χαρακτηριστικά κίνησης παράγονται από ένα χρονικό δίκτυο με εισόδους στοίβες πεδίων οπτικής ροής. Λόγω της συμπληρωματικότητας τους τα χαρακτηριστικά αυτά συνδυάζονται στο επίπεδο του ταξινομητή για το σχηματισμό ενός χωρο-χρονικού πυρήνα x^2. Η ταξινόμηση μέσω της σύμμειξης αυτής σε χωρο-χρονικά features πέτυχε ακρίβεια ταξινόμησης υψηλότερη από την αντίστοιχη end-to-end αξιολόγηση των μοντέλων, ενώ ο συνδυασμός των deep TSN features με τις hand-crafted iDT οδήγησε στην επίτευξη υψηλών ποσοστών αναγνώρισης ανθρώπινων δράσεων, ξεπερνώντας το state-of-the-art για την βάση HMDB51.

Συνοψίζοντας, λοιπόν, η μέθοδος που πέτυχε την υψηλότερη ακρίβεια ταξινόμησης ανθρώπινων δράσεων, χρησιμοποιεί έναν μη γραμμικό SVM ταξινομητή ο οποίος συνδυάζει τους x^2 πυρήνες (kernels) των deep-learned χαρακτηριστικών *TSN rgb, TSN flow, C3D*, εξαγμένα από κατάλληλα εκπαιδευμένα μοντέλα, με τους x^2 πυρήνες των hand-crafted περιγραφητών *TD, HoG, HoF, MBHx, MBHy* των βελτιωμένων τροχιών, κωδικοποιημένους με BoW.

5.2 Μελλοντικές Κατευθύνσεις

Το ερευνητικό πεδίο της αναγνώρισης ανθρώπινων δράσεων εξελίσσεται με πολύ γρήγορους ρυθμούς και είναι πλούσιο σε διαφορετικές μεθοδολογίες και τεχνικές. Τα διαθέσιμα εργαλεία για έναν ερευνητή είναι πλέον αρκετά και μπορεί να τα χρησιμοποιήσει με διαφορετικούς τρόπους για να προτείνει προεκτάσεις, βελτιώσεις ή και καινοτόμες ιδέες. Μερικές από αυτές αναφέρονται παρακάτω:

1. *Εφαρμογή διαφορετικών κωδικοποιήσεων στους περιγραφητές των βελτιωμένων τροχιών.* Όπως αναφέραμε στο σχετικό Κεφάλαιο, εκτός από την Bag-of-Words κωδικοποίηση που χρησιμοποιούμε εκτενώς στα πλαίσια της παρούσας Διπλωματικής, στη βιβλιογραφία αναφέρονται και άλλες μορφές κωδικοποιήσεων όπως οι χωρο-χρονικές πυραμίδες, το διάνυσμα συσσωρευμένων περιγραφητών (VLAD) και το διάνυσμα Fisher, οι οποίες αρκετές φορές επιτυγχάνουν υψηλότερη ακρίβεια ταξινόμησης από την BoW κωδικοποίηση, που αποτελεί μία αποδοτική επιλογή με μικρή υπολογιστική πολυπλοκότητα. Θα είχε ενδιαφέρον στα πειράματα μας να χρησιμοποιηθούν οι improved trajectories με άλλες μορφές κωδικοποίησης οι οποίες ενδεχομένως θα βελτίωναν ακόμη περισσότερο τα ποσοστά αναγνώρισης ανθρώπινων δράσεων.

2. *Χρήση χωρο-χρονικού εντοπισμού της δράσης μέσα στο βίντεο (action localization).* Όπως περιγράψαμε στην Εισαγωγή της εργασίας, το πρόβλημα της αναγνώρισης ανθρώπινων δράσεων, με τον αυστηρό ορισμό του, συνδυάζει αρχικά το χωρο-χρονικό εντοπισμό μίας δράσης σε ένα βίντεο (action localization) και εν συνεχεία την ταξινόμηση της δράσης αυτής (action classification). Δηλαδή, οι τεχνικές ταξινόμησης που εφαρμόσαμε στα πλαίσια της παρούσας έρευνας, θα μπορούσαν να εφαρμοστούν σε μία περιοχή ενδιαφέροντος μέσα στο βίντεο και όχι σε όλο αυτό καθέ αυτό. Ένας action localizer θα εξασφάλιζε την κατάτμηση του βίντεο στη σκηνή πραγματοποίησης της δράσης και θα βοηθούσε σημαντικά στην απόρριψη περιττής πληροφορίας, ενώ σε τελικό στάδιο, έπειτα από την ταξινόμηση, ο αλγόριθμος θα ήταν σε θέση να «αντιληφθεί» εκτός από το τι δράση πραγματοποιείται, το πότε ξεκινάει και πότε τελειώνει αυτή η δράση.

3. *Χρήση εντοπισμού και παρακολούθησης των ανθρώπων που πραγματοποιούν μια δράση (human detection and tracking).* Καθώς μας ενδιαφέρουν οι δράσεις που πραγματοποιούνται από ανθρώπους, χρήσιμη μπορεί να αποδειχθεί η εφαρμογή ενός human detector σε κάθε frame του βίντεο, δηλαδή ενός human tracker ουσιαστικά. Η απομόνωση της μικρής περιοχής γύρω από τον άνθρωπο ή τους ανθρώπους που πραγματοποιούν μία δράση μπορεί να μειώσει σημαντικά την αρνητική επίδραση ενός «ακατάστατου» φόντου (cluttered background) και να οδηγήσει στην εξαγωγή πιο «ποιοτικών» χαρακτηριστικών, που αφορούν αποκλειστικά την κίνηση των ενδιαφερόμενων ατόμων, βελτιώνοντας έτσι την ακρίβεια ταξινόμησης των δράσεων.

Βιβλιογραφία

[1] N. Kardaris, *Feature Extraction and Encoding Methods for Human Action and Gesture Recognition*. NTUA, 2015.

[2] A. Karpathy, *CS231n Convolutional Neural Networks for Visual Recognition*, 2015. Available at `http://cs231n.github.io/convolutional-networks`.

[3] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision–based human motion capture and analysis," *Computer Vision and Image Understanding*, vol. 104, pp. 90–126, Oct 2006.

[4] S. Mitra and T. Acharya, "Gesture recognition: A survey," *IEEE Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews*, vol. 37, pp. 311–324, May 2007.

[5] M. Ramanathan, W. Y. Yau, and E. K. Teoh, "Human action recognition with video data: Research and evaluation challenges," *IEEE Transactions on Human–Machine Systems*, vol. 44/5, pp. 650–663, Oct 2014.

[6] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing Human Actions: A Local SVM Approach," in *Proc. Int'l Conf. on Pattern Recognition (ICPR 2004)*, Aug 2004.

[7] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *CoRR*, vol. abs/1212.0402, Nov 2012.

[8] K. K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," *Machine Vision and Applications*, vol. 24/5, pp. 971–981, Jul 2013.

[9] M. Marszalek, I. Laptev, and C. Schmid, "Actions in Context," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2009)*, Jun 2009.

[10] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2008)*, Jun 2008.

[11] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," in *Proc. IEEE Int'l Conf. on Computer Vision (ICCV 2011)*, Nov 2011.

[12] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, "Multimodal saliency and fusion for movie summarization based on aural, visual and textual attention," *IEEE Transactions on Multimedia*, vol. 15/7, pp. 1553–1568, Nov 2013.

[13] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. British Machine Vision Conf. (BMVC 2015)*, Sep 2015.

[14] I. Laptev and T. Lindeberg, "Space-time Interest Points," in *Proc. IEEE Int'l Conf. on Computer Vision (ICCV 2003)*, Oct 2003.

[15] H. Wang and C. Schmid, "Action Recogition with Improved Trajectories," in *Proc. IEEE Int'l Conf. on Computer Vision (ICCV 2013)*, Dec 2013.

[16] S. Ji, W. Xu, M. Yang, and K. Yu, "3D Convolutional Neural Networks for Human Action Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 221–231, Jan 2013.

[17] K. Simonyan and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos," in *Proc. Advances in Neural Information Processing Systems (NIPS 2014)*, Dec 2014.

[18] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28/6, pp. 976–990, Jun 2010.

[19] C. Harris and M. Stephens, "A Combined Corner and Edge Detector," in *Proc. Fourth Alvey Vision Conf.*, Aug-Sep 1988.

[20] I. Laptev, B. Caputo, C. Schüldt, and T. Lindeberg, "Local velocity-adapted motion events for spatio-temporal recognition," *Computer Vision and Image Understanding*, vol. 108/3, pp. 207–229, Dec 2007.

[21] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2005)*, Jun 2005.

[22] I. Laptev and T. Lindeberg, "Local Descriptors for Spatio-Temporal Recognition," *Springer Lecture Notes in Computer Science*, vol. 3667, pp. 91–103, Jun 2006.

[23] A. Kläser, M. Marszalek, and C. Schmid, "A Spatio-Temporal Descriptor Based on 3D-Gradients," in *Proc. British Machine Vision Conf. (BMVC 2008)*, Sep 2008.

[24] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. Int'l Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS 2005)*, Oct 2005.

[25] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60/2, pp. 91–110, Nov 2004.

[26] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional Sift Descriptor and Its Application to Action Recognition," in *Proc. 15th Int'l Conf. on Multimedia (ICM 2007)*, Nov 2007.

[27] G. Willems, T. Tuytelaars, and V. L. Gool, "An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector," in *Proc. IEEE European Conf. on Computer Vision (ECCV 2008)*, Oct 2008.

[28] H. Bay, T. Tuytelaars, and V. L. Gool, "Surf: Speeded up Robust Features," in *Proc. IEEE European Conf. on Computer Vision (ECCV 2006)*, May 2006.

[29] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid, "Evaluation of Local Spatio-Temporal Features for Action Recognition," in *Proc. British Machine Vision Conf. (BMVC 2009)*, Sep 2009.

[30] K. Maninis, P. Koutras, and P. Maragos, "Advances on Action Recognition in Videos using an Interest Point Detector based on Multiband Spatio-Temporal Energies," in *Proc. IEEE Int'l Conf. on Image Processing (ICIP 2014)*, Oct 2014.

[31] C. Georgakis, P. Maragos, G. Evangelopoulos, and D. Dimitriadis, "Dominant Spatio-Temporal Modulations and Energy Tracking in Videos: Application to Interest Point Detection for Action Recognition," in *Proc. IEEE Int'l Conf. on Image Processing (ICIP 2012)*, Sep 2012.

[32] R. Messing, C. Pal, and H. Kautz, "Activity Recognition using the Velocity Histories of Tracked Keypoints," in *Proc. IEEE Int'l Conf. on Computer Vision (ICCV 2009)*, Sep 2009.

[33] J. Sun, X. Wu, S. Yan, L. F. Cheong, T. S. Chua, and J. Li, "Hierarchical Spatio-Temporal Context Modeling for Action Recognition," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2009)*, Jun 2009.

[34] J. Sun, Y. Mu, S. Yan, and L. Cheong, "Activity Recognition using Dense Long-Duration Trajectories," in *Proc. IEEE Int'l Conf. on Multimedia and Expo (ICME 2010)*, Jul 2010.

[35] H. Wang, A. Kläser, and C. Schmid, "Action Recogition by Dense Trajectories," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2011)*, Jun 2011.

[36] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *Proc. IEEE European Conf. on Computer Vision (ECCV 2006)*, May 2006.

[37] T. Darrell and A. Pentland, "Space-time Gestures," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 1993)*, Jun 1993.

[38] A. Bobick and J. Davis, "The Recognition of Human Movement Using Temporal Templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 257–267, Mar 2001.

[39] M. K. Hu, "Visual Pattern Recognition by Moment Invariants," *IRE Transactions on Information Theory*, vol. 8/2, pp. 179–187, Jan 1962.

[40] D. Weinland, R. Ronfard, and E. Boyer, "Free Viewpoint Action Recognition using Motion History Volumes," *Computer Vision and Image Understanding*, vol. 104, pp. 249–257, Nov 2006.

[41] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as Space-Time Shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29/12, pp. 2247–2253, Dec 2007.

[42] A. Yilmaz and M. Shah, "Actions Sketch: a Novel Action Representation," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2005)*, Jun 2005.

[43] A. Efros, A. Berg, G. Mori, and J. Malik, "Recognizing Action at a Distance," in *Proc. IEEE Int'l Conf. on Computer Vision (ICCV 2003)*, Oct 2003.

[44] M. Raptis, I. Kokkinos, and S. Soatto, "Discovering Discriminative Action Parts from Mid-level Video Representations," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2012)*, Jun 2012.

[45] Y. Yang, I. Saleemi, and M. Shah, "Discovering Motion Primitives for Unsupervised Grouping and One-Shot Learning of Human Actions, Gestures, and Expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35/7, pp. 1635–1648, Jul 2013.

[46] S. Bhattacharya, M. M. Kalayeh, R. Sukthankar, and M. Shah, "Recognition of complex events: Exploiting Temporal Dynamics between Underlying Concepts," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2014)*, Jun 2014.

[47] B. Fernando, E. Gavves, J. M. Oramas, A. Ghodrati, and T. Tuytelaars, "Modeling Video Evolution for Action Recognition," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2015)*, Jun 2015.

[48] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks," in *Proc. IEEE Int'l Conf. on Computer Vision (ICCV 2015)*, Dec 2015.

[49] L. Wang, Y. Qiao, and X. Tang, "Action Recognition with Trajectory-Pooled Deep-Convolutional Descriptors," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2015)*, Jun 2015.

[50] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2016)*, Jun 2016.

[51] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool, "Temporal Segment Networks: Towards Good Practices for Deep Action Recognition," in *Proc. IEEE European Conf. on Computer Vision (ECCV 2016)*, Oct 2016.

[52] Q. Q. Chen and Y. J. Zhang, "Sequential segment networks for action recognition," *IEEE Signal Processing Letters*, vol. 24/5, pp. 712–716, May 2017.

[53] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Spatiotemporal Residual Networks for Video Action Recognition," in *Proc. Advances in Neural Information Processing Systems (NIPS 2016)*, Dec 2016.

[54] A. Diba, V. Sharma, and L. V. Gool, "Deep Temporal Linear Encoding Networks," *CoRR*, vol. abs/1611.06678, Nov 2016.

[55] J. Shi and C. Tomasi, "Good features to track," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 1994)*, Jun 1994.

[56] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. 7th Int'l Joint Conf. on Artificial Intelligence*, Aug 1981.

[57] G. Farnebäck, "Two-Frame Motion Estimation Based on Polynomial Expansion," in *Proc. 13th Scandinavian Conf. on Image Analysis (SCIA 2003)*, Jun-Jul 2003.

[58] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28/2, pp. 129–137, Mar 1982.

[59] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2Nd Edition)*. Wiley-Interscience, 2000.

[60] H. Jegou, M. Douze, C. Schmid, and P. Perez, "Aggregating Local Descriptors into a Compact Image Representation," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2010)*, Jun 2010.

[61] F. Perronnin and C. Dance, "Fisher Kernels on Visual Vocabularies for Image Categorization," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2007)*, Jun 2007.

[62] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2005)*, Jun 2005.

[63] C. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2006)*, Jun 2006.

[64] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int'l Conf. on Artificial Intelligence and Statistics (ICAIS 2010)*, May 2010.

[65] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on Imagenet Classification," in *Proc. IEEE Int'l Conf. on Computer Vision (ICCV 2015)*, Dec 2015.

[66] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," in *Proc. 32nd Int'l Conf. on Machine Learning (ICML 2015)*, Jul 2015.

[67] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, Oct 1986.

[68] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society, Series B*, vol. 67, pp. 301–320, Apr 2005.

[69] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *The Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, Jan 2014.

[70] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Advances in Neural Information Processing Systems (NIPS 2012)*, Dec 2012.

[71] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Learning Applied to Document Recognition," *Proceedings of the IEEE*, vol. 86, pp. 2278–2324, Nov 1998.

[72] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," in *Proc. IEEE European Conf. on Computer Vision (ECCV 2014)*, Sep 2014.

[73] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2015)*, Jun 2015.

[74] C. Szegedy, S. Ioffe, and V. Vanhoucke, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," *CoRR*, vol. abs/1602.07261, Feb 2016.

[75] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *CoRR*, vol. abs/1409.1556, Sep 2014.

[76] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2016)*, Jun 2016.

[77] K. He, X. Zhang, S. Ren, and J. Sun, "Identity Mappings in Deep Residual Networks," in *Proc. IEEE European Conf. on Computer Vision (ECCV 2016)*, Oct 2016.

[78] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc.*

IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2014), Jun 2014.

[79] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," *Journal of Machine Learning Research*, vol. 9, pp. 2491–2521, Nov 2008.

[80] M. Varma and B. R. Babu, "More Generality in Efficient Multiple Kernel Learning," in *Proc. 26th Int'l Conf. on Machine Learning (ICML 2009)*, Jun 2009.

[81] A. Jain, S. V. N. Vishwanathan, and M. Varma, "SPG-GMKL: Generalized Multiple Kernel Learning with a Million Kernels," in *Proc. ACM SIGKDD Conf. on Knowledge Discovery and Data Mining*, Aug 2012.

[82] B. Fernando, P. Anderson, M. Hutter, and S. Gould, "Discriminative Hierarchical Rank Pooling for Activity Recognition," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2016)*, Jun 2016.

[83] Y. Wang, V. Tran, and M. Hoai, "Evolution-Preserving Dense Trajectory Descriptors," *CoRR*, vol. abs/1702.04037, Feb 2017.