Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Τομέας Σημάτων, Ελέγχου και Ρομποτικής

Εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας, Λόγου και Επεξεργασίας Σημάτων

Αρχιτεκτονικές Βαθέων Νευρωνικών Δικτύων για την Αναγνώριση και τον Χρονικό Εντοπισμό Δράσεων σε Βίντεο

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΒΑΣΙΛΙΚΗΣ Ι. ΒΑΣΙΛΕΙΟΥ

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π

Αθήνα, Απρίλιος 2021

Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Τομέας Σημάτων, Ελέγχου και Ρομποτικής

Εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας, Λόγου και Επεξεργασίας Σημάτων

Αρχιτεκτονικές Βαθέων Νευρωνικών Δικτύων για την Αναγνώριση και τον Χρονικό Εντοπισμό Δράσεων σε Βίντεο

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΒΑΣΙΛΙΚΗΣ Ι. ΒΑΣΙΛΕΙΟΥ

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π

Εγκρίθηκε από την τριμελή επιτροπή την 9η Απριλίου 2021:

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

Αθήνα, Απρίλιος 2021

.....

Βασιλική Ι. Βασιλείου
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών, Ε.Μ.Π.

Ευχαριστίες

Αισθάνομαι την ανάγκη να ευχαριστήσω εκ βαθέων τον καθηγητή μου Πέτρο Μαραγκό για την ευκαιρία που μου έδωσε να εκπονήσω τη διπλωματική μου εργασία στο Εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας, Λόγου και Επεξεργασίας Σημάτων, για όσα πλουσιοπάροχα μου πρόσφερε κατά την εκπόνηση της διπλωματικής μου εργασίας και να εκφράσω την ευγνωμοσύνη μου για τις επιστημονικές του συμβουλές, καθ' όλη την διάρκεια της συνεργασίας μας. Επίσης θα ήθελα να ευχαριστήσω τον Νίκο Κάρδαρη για την συνεισφορά σε διάφορα θέματα που προέκυπταν κατά την υλοποίηση της παρούσας διπλωματικής εργασίας.

Δεν θα μπορούσα παρά να εκφράσω άπειρες ευχαριστίες στους λαμπρούς μου φίλους με τους οποίους μοιραστήκαμε εμπειρίες, στιγμές χαράς αλλά και άγχη. Ευχαριστίες για τις συνεργασίες μας, την υπομονή και στήριξή τους. Είναι ο Άγγελος Κ., ο Κλεάνθης Α., η Δωροθέα Κ., ο Νίκος Γ., ο Χρήστος Σ., ο Γιώργος Κ., ο Νικήτας Θ. και ο Δημήτρης Γ. για τις συνεργασίες μας, την υπομονή, την στήριξη τους.

Θέλω να ευχαριστήσω επίσης τον παππού μου, Μ. Αυδίκο (Γυμνασιάρχη - Θεολόγο) για την γλωσσική επιμέλεια της διπλωματικής μου εργασίας παρ' όλες τις δυσκολίες της επιστημονικής ορολογίας. Ευχαριστώ την ξαδέρφη μου Φωτεινή Μ. που, παρά το νεαρό της ηλικίας της, μου συμπαραστάθηκε με απεριόριστη αγάπη. Τέλος να εκφράσω την απέραντη ευγνωμοσύνη μου στους γονείς μου Γιάννη και Άννα για τον τρόπο που με γαλούχησαν, την διάπλαση της προσωπικότητας μου, την απέραντη αγάπη τους και στήριξη τους στις επιλογές μου.

Βασιλική Ι. Βασιλείου,
Απρίλιος 2021

Αν σπρώξεις τον εαυτό σου, μέχρι τα όρια του,
τότε θα ανακαλύψεις ότι δεν υπάρχουν όρια
Paulo Coelho

...στους γονείς μου

Περίληψη

Η αναγνώριση των ανθρώπινων δράσεων είναι ένας από τους τομείς της Όρασης Υπολογιστών, που κεντρίζει όλο και περισσότερο το ενδιαφέρον των ερευνητών, καθώς η ανάλυση των ανθρώπινων δράσεων έχει πολλές εφαρμογές στην καθημερινή μας ζωή. Χαρακτηριστικά παραδείγματα αποτελούν τα assistive robots στον τομέα της ρομποτικής και οι accident predictors στον τομέα της βιομηχανίας. Η έρευνα στο αντικείμενο αυτό ξεκίνησε με την εξαγωγή χαρακτηριστικών από περιγραφητές και συνεχίζει σήμερα με την χρήση βαθέων νευρωνικών δικτύων, τα οποία εξετάζουμε και εμείς στην παρούσα διπλωματική εργασία.

Συγκεκριμένα, η εργασία μας έχει δύο βασικές πτυχές πλήρως αλληλένδετες. Η πρώτη παρουσιάζει τη μελέτη της αποτελεσματικότητας των κυριότερων αρχιτεκτονικών βαθέων νευρωνικών δικτύων για την αναγνώριση ανθρώπινων χειρονομιών, σε τριμαρισμένα βίντεο. Αντίθετα η δεύτερη στοχεύει στη χρονική κατάτμηση των ανθρώπινων δράσεων σε μη τριμαρισμένα βίντεο, δίνοντας τη δυνατότητα αξιοποίησης σε online εφαρμογές.

Παράλληλα με την επίδραση των βαθέων νευρωνικών δικτύων εξετάζεται και η συνεισφορά της δυναμικής της ανθρώπινης πόζας, τόσο στην αναγνώριση των ανθρώπινων δράσεων όσο και στο χρονικό εντοπισμό τους.

Για την αξιολόγηση των συστημάτων που υλοποιήθηκαν χρησιμοποιήθηκαν δύο σύνολα δεδομένων. Συγκεκριμένα για το πρόβλημα της αναγνώρισης ανθρώπινων χειρονομιών χρησιμοποιήθηκε η IsoGD βάση δεδομένων, ενώ για το πρόβλημα του χρονικού εντοπισμού ανθρώπινων δράσεων χρησιμοποιήθηκε η THUMOS'14 βάση δεδομένων.

Από την δουλειά μας πάνω στο αντικείμενο του χρονικού εντοπισμού και της αναγνώρισης ανθρώπινων δράσεων σε πραγματικό χρόνο προέκυψαν state-of-the-art αποτελέσματα τα οποία έχουν υποβληθεί στο 29th European Signal Processing Conference (EUSIPCO 2021) με συγγραφείς τους Βασιλική Ι. Βασιλείου, Νικόλαο Κάρδαρη και Πέτρο Μαραγκό.

Λέξεις Κλειδιά – Αναγνώριση ανθρώπινων χειρονομιών και δράσεων, Χρονικός εντοπισμός ανθρώπινων δράσεων, Βαθιά νευρωνικά δίκτυα, Ανθρώπινη πόζα, OpenPose, IsoGD, THUMOS'14

Abstract

Human Action Recognition (HAR) is one of the most challenging tasks of Computer Vision, that increasingly stimulates the interest of researchers, as the analysis of human action has many real world applications. Assistive robots in the field of robotics and accident predictors in the field of industry are some of HAR applications. Research on this subject began with the extraction of characteristics from descriptors and continues today with the use of deep neural networks, which we are also examining in this diploma thesis.

In particular, our work has two main folds that are fully correlated. The first is the study of the effectiveness of the main deep neural network architectures, for human gesture recognition in trimmed videos. On the contrary, the second aims at the temporal segmantation of human actions in untrimmed videos enabling use in online applications.

In addition to the incorporation of deep neural networks, the contribution of the human movement dynamics to the human action recognition and its temporal segmentation is examined.

Two datasets were used to evaluate the systems we implemented. In particular, in the case of human gesture recognition, the IsoGD database was used while the THUMOS'14 database was used in the case of temporal segmentation of an action.

Our work on the subject of temporal segmentation and recognition of human actions in real time achieved state-of-the-art results which have been submitted to the 29th European Signal Processing Conference (EUSIPCO 2021) with authors Vasiliki I. Vasileiou, Nikolaos Kardaris and Petros Maragos.

Keywords – Human action and gesture recognition, Temporal detection of human action, Deep neural networks, Human Pose, OpenPose, IsoGD, THUMOS'14

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Πίνακας Όρων

Στον πίνακα όρων περιλαμβάνουμε τους όρους που χρησιμοποιούνται ευρέως στο κείμενο και αναφέρονται είτε στα ελληνικά είτε στα αγγλικά:

- **Αναγνώριση Ανθρώπινης Δράσεων** Human Action Recognition (HAR)
- **Αναγνώριση Ανθρώπινων Χειρονομιών** Human Gesture Recognition
- **Χρονικός Εντοπισμός Δράσεων** Temporal Action Detection
- **Δισδιάστατος Σκελετός** 2D Skeleton
- **Αναδρομικά Νευρωνικά Δίκτυα** Recurrent Neural Networks (RNNs)
- **Νευρώνες Μακράς-Βραχείας Μνήμης** Long Short Term Memory (LSTM)
- **Στρώμα Προσοχής** Attention Layer
- **Συνελικτικά Νευρωνικά Δίκτυα** Convolutional Neural Networks (CNNs)
- **Συνελικτικά Δίκτυα Διπλής Ροής** Two-Stream Convolutional Networks
- **Τρισδιάστατα Συνελικτικά Δίκτυα** 3D Convolutional Networks
- **Πρότυπο Χρώματος RGB** Red, Green, Blue
- **Χρονικά Συνελικτικά Δίκτυα** Temporal Convolutional Networks (TCNs)
- **Χρονικά Αναδρομικά Δίκτυα** Temporal Recurrent Networks (TRNs)
- **Κυλιόμενο Παράθυρο** Sliding Window
- **Μηχανές Διανυσμάτων Υποστήριξης** Support Vector Machines (SVMs)
- **Στρώμα Συσσώρευσης** Pooling Layer
- **Πλήρως Συνδεδεμένο Στρώμα** Fully-Connected Layer
- **Καρέ** Frame
- **Συν-εμφάνιση** Co-occurrence
- **Εμφάνιση** Appearance
- **Κίνηση** Motion
- **Διογκώνω** Inflate

- **Σύνολο Δεδομένων** Dataset
- **Υπόβαθρο** Background
- **Σε πραγματικό χρόνο** Online
- **Σε μη πραγματικό χρόνο (βίντεο)** Offline
- **Πρώιμη Ανίχνευση Δράσεων** Early Action Detection
- **Συμφραζόμενα** Context
- **Ενισχυτικό Κόστος** Reinforcement Loss
- **Πρόβλεψη (μελλοντικής πληροφορίας)** Anticipation
- **Κωδικοποιητής** Encoder
- **Αποκωδικοποιητής** Decoder
- **Χρονικό Αναδρομικό Δίκτυο** Temporal Recurrent Network (TRN)
- **Παραγωγικό Αντιπαραθετικό Δίκτυο** Generative Adversarial Network
- **Αναδρομική Μονάδα** Recurrent Unit
- **Σύνολο Εκπαίδευσης** Training Set
- **Σύνολο Εκτίμησης** Validation Set
- **Σύνολο Δοκιμής** Test Set
- **Πεδία Μερικής Συνάφειας** Part Affinity Fields (PAFs)
- **Χάρτης Εμπιστοσύνης** Confidence Map
- **Σημεία-Κλειδιά** Keypoints
- **Οπτική Ροή** Optical Flow
- **Αραιή Οπτική Ροή** Sparse Optical Flow
- **Πυκνή Οπτική Ροή** Dense Optical Flow
- **Ολική Διακύμανση** Total Variation
- **Συνάρτηση Σφάλματος** Loss Function
- **Αλγόριθμοι Βελτιστοποίησης** Optimizers
- **Αλγόριθμος Κατάβασης Κλίσης** Gradient Descent Algorithm
- **Αλγόριθμος Οπισθοδιάδοσης** Back-Propagation Algorithm
- **Υπερπροσαρμογή** Overfitting
- **Κανάλι Βάθους** Depth Channel

- **Εικονοστοιχείο** Pixel
- **Διγραμμική Παρεμβολή** Bilinear Interpollation
- **Προ-εκπαίδευση** Pre-training
- **Μέση Ακρίβεια** Average Precision
- **Ανάκληση** Recall
- **Πύλη Χαρακτηριστικών** Feature Gate
- **Γραμμικός** Linear
- **Επεξεργασία Φυσικής Γλώσσας** Natural Language Processing (NLP)

Κεφάλαιο 1

Εισαγωγή

Ένας από τους πιο αναπτυσσόμενους κλάδους της επιστήμης των υπολογιστών, που κεντρίζει όλο και περισσότερο το ενδιαφέρον των ερευνητών είναι η Τεχνητή Νοημοσύνη. Με τον όρο Τεχνητή Νοημοσύνη αναφερόμαστε στον κλάδο της πληροφορικής που ασχολείται με την σχεδίαση και υλοποίηση "έξυπνων" υπολογιστικών συστημάτων που λαμβάνουν ερεθίσματα, επεξεργάζονται πληροφορίες, παίρνουν αποφάσεις και ενεργούν με τρόπους ανάλογους με τους ανθρώπινους. Η χρήση της έχει γίνει ιδιαιτέρως ευρεία στις μέρες μας, ξεκινώντας από την υγεία και την ρομποτική και καταλήγοντας στη δικαιοσύνη και τη βιομηχανία. Η προσπάθεια δημιουργίας συστημάτων με νοημοσύνη συνδέεται άρρηκτα με την προσπάθεια κωδικοποίησης των ανθρώπινων αισθήσεων (όραση, ακοή, αφή) στις μηχανές αυτές. Προκύπτει λοιπόν ένα ευρύ πεδίο, της Τεχνητής Νοημοσύνης που ονομάζεται Όραση Υπολογιστών. Είναι εκείνη η κατεύθυνση της έρευνας που επιχειρεί να αναπαράγει αλγοριθμικά την αίσθηση της όρασης σε ένα ρομπότ ή γενικότερα σε ένα έξυπνο υπολογιστικό σύστημα. Δεδομένα προερχόμενα από ψηφιακές εικόνες, βίντεο, πολλαπλές κάμερες αλλά και κάμερες βάθους λαμβάνονται και αναλύονται καταλλήλως ώστε να αποτελέσουν οπτικά ερεθίσματα στο εκάστοτε σύστημα. Παράλληλα με την Όραση Υπολογιστών αναπτύσσεται ένας ακόμη κάδος της Τεχνητής Νοημοσύνης που είναι η Μηχανική Μάθηση. Συγκεκριμένα με τον όρο Μηχανική Μάθηση εννοούμε την ανάπτυξη της ικανότητας των συστημάτων να μαθαίνουν χωρίς να έχουν ρητά προγραμματιστεί.

1.1 Περιγραφή Προβλήματος

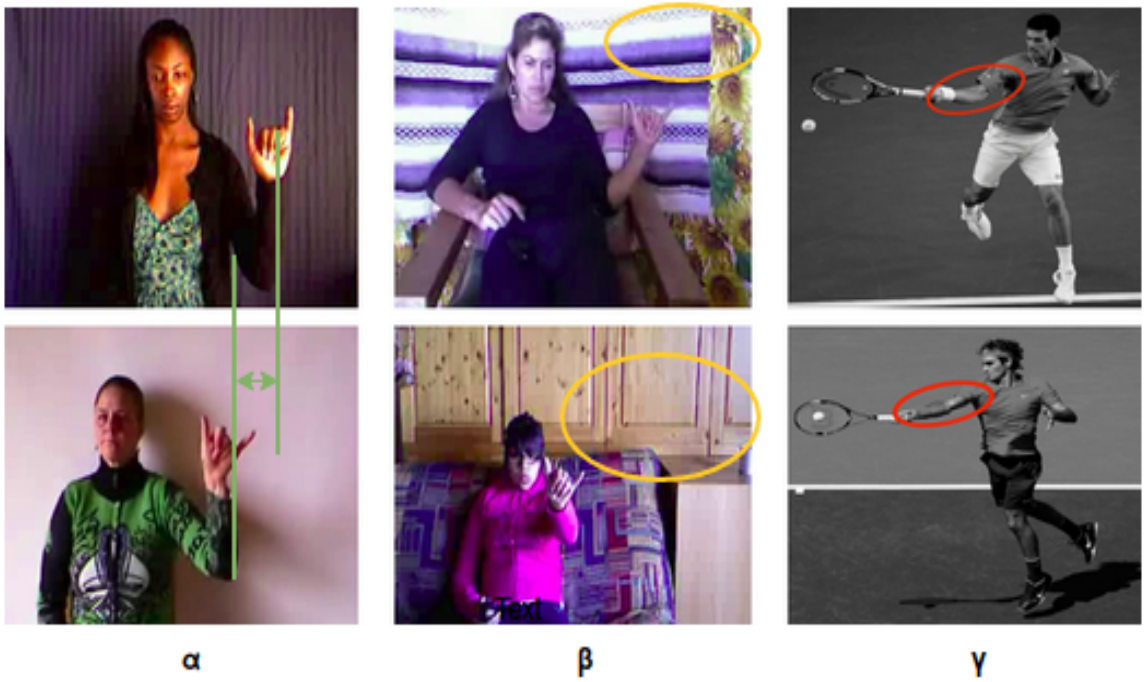
1.1.1 Αναγνώριση Ανθρώπινων Δράσεων και Χειρονομιών

Η αναγνώριση των ανθρώπινων δράσεων είναι πεδίο έρευνας με μεγάλη πρόοδο την τελευταία δεκαετία καθώς οι εφαρμογές της χρησιμοποιούνται όλο και συχνότερα στην καθημερινή ζωή. Πρακτικά περιλαμβάνει την πρόβλεψη μιας ανθρώπινης κίνησης που βασίζεται σε δεδομένα αισθητήρων και παραδοσιακά περιλαμβάνει βαθιά τεχνογνωσία και μεθόδους από τον τομέα της επεξεργασίας σήματος προκειμένου τα ακατέργαστα δεδομένα να προσαρμοστούν κατάλληλα ώστε να ταιριάζουν σε ένα μοντέλο μηχανικής μάθησης. Συχνά συγχέονται οι έννοιες δράση και δραστηριότητα, με την πρώτη να αποτελεί ένα απλό μοτίβο κίνησης π.χ. "Σουτάρισμα στο μπάσκετ" και την δεύτερη να αποτελείται από μια ακολουθία από δράσεις π.χ. "Παίζω μπάσκετ".

Υποσύνολο της αναγνώρισης ανθρώπινων δράσεων αποτελεί η αναγνώριση ανθρώπινων χειρονομιών. Η χειρονομία αποτελεί ένα σύμβολο σωματικής συμπεριφοράς ή συναισθηματικής έκφρασης και διακρίνεται σε χειρονομία χεριού και χειρονομία σώματος. Η πρώτη

περιέχει πληροφορία μόνο από την παλάμη και τα δάχτυλα ενώ η δεύτερη ενσωματώνει πληροφορία και από τον κορμό και το πρόσωπο. Επίσης σημαντικός είναι και ο διαχωρισμός μεταξύ στατικής και δυναμικής χειρονομίας. Στατική είναι η χειρονομία που υποδηλώνει ένα μόνο σημάδι, ενώ δυναμική είναι αυτή που μεταφέρει μερικά μηνύματα.

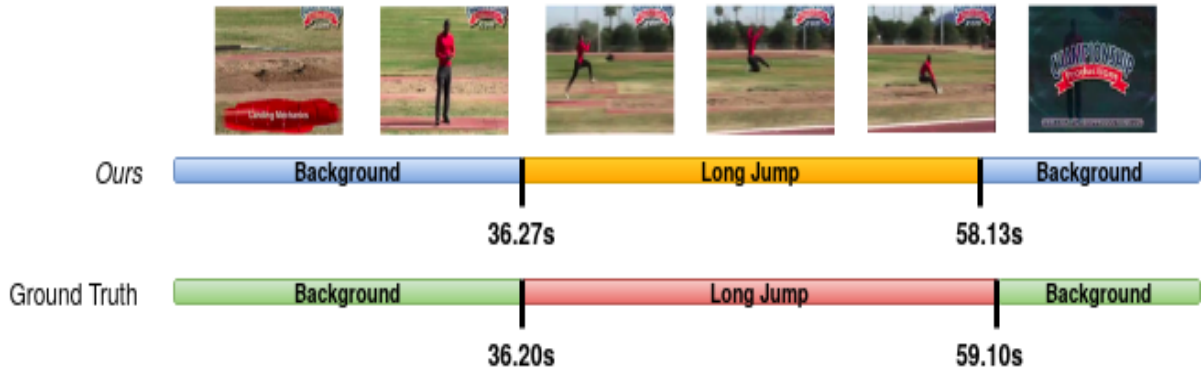
Τόσο στην ευρεία περίπτωση αναγνώρισης δράσεων όσο και στην ειδική αναγνώρισης χειρονομιών, για την εκπαίδευση του εκάστοτε μοντέλου χρησιμοποιούνται τριμαρισμένα βίντεο που περιλαμβάνουν μόνο μία δράση. Μια από τις σημαντικότερες προκλήσεις που αντιμετωπίζουμε σε αυτό το πεδίο έρευνας, είναι ότι χρησιμοποιούνται επισημειωμένα δεδομένα, από δράσεις που έχουν εκτελεστεί από διαφορετικούς ανθρώπους υπό διαφορετικές συνθήκες. Συγκεκριμένα στα διάφορα σύνολα δεδομένων που χρησιμοποιούνται μπορεί να υπάρχουν μεταβολές στην γωνία λήψης και την κλίμακα - π.χ προφίλ - ανφάς, καθιστός-όρθιος, κοντά-μακριά από την κάμερα -, τις συνθήκες του περιβάλλοντος - π.χ. φωτισμός, σκιάσεις, φόντο -, αλλά και στην εκτέλεση μιας δράσης - π.χ. τεχνικές forehand μεταξύ τεννιστών.



Σχήμα 1.1: α) Μεταβολή στη γωνία λήψης, β) Μεταβολή στο φόντο, γ) Μεταβολή στην εκτέλεση

1.1.2 Χρονική Κατάτμηση Ανθρώπινων Δράσεων

Η ευρεία χρήση της αναγνώρισης ανθρώπινης δράσης σε εφαρμογές της καθημερινής ζωής έχει μεταφέρει το ερευνητικό ενδιαφέρον από την μελέτη βίντεο με μεμονωμένες δράσεις - πλήρως χρονικά προσδιορισμένων - στην μελέτη βίντεο με σειριακές δράσεις. Συγκεκριμένα στόχος είναι ο χρονικός εντοπισμός της δράσης - χρονική στιγμή έναρξης και λήξης - και εν συνεχεία η αναγνώριση της. Στην συγκεκριμένη κατεύθυνση υπάρχουν δύο περιπτώσεις μελέτης, η offline μέσω βίντεο και η online σε συνθήκες πραγματικού χρόνου. Στις offline μεθόδους έχουμε στην διάθεση μας ολόκληρο το βίντεο, δηλαδή ολοκληρωμένες τις δράσεις, σε αντίθεση με τις online μεθόδους που κατέχουμε πληροφορία μόνο από το τρέχον και τα παρελθοντικά frames.



Σχήμα 1.2: Παράδειγμα χρονικού εντοπισμού δράσης σε βίντεο. Στο σχήμα αυτό συγκρίνονται και οπτικά τα αποτελέσματα του [65] με τα δεδομένα παρατήρησης. Το εξεταζόμενο βίντεο ξεκινά με τον αθλητή να μιλά (background), συνεχίζει με την δράση (άλμα εις μήκος) και ολοκληρώνεται με μια διαφήμιση (background).

1.2 Εφαρμογές

Το πεδίο της αναγνώρισης των ανθρωπίνων δράσεων και χειρονομιών έχει ποικίλες εφαρμογές που συνδέονται με την αλληλεπίδραση ανθρώπου-μηχανής και την λήψη αποφάσεων χωρίς ανθρώπινη επιτήρηση.

Ξεκινώντας από τη λήψη αποφάσεων, ένα σύστημα που αναγνωρίζει δράσεις μπορεί να χρησιμοποιηθεί για περιλήψεις σε βίντεο (π.χ. αθλητικές ειδήσεις, trailer ταινιών), αλλά και για κατηγοριοποιήσεις οπτικών δεδομένων (εικόνες, βίντεο) με βάση το περιεχόμενο τους. Επιπλέον μπορούν να αποτραπούν παραβατικές συμπεριφορές, καθώς ένα κατάλληλο εκπαιδευμένο σύστημα μπορεί να προειδοποιεί, όταν μια ύποπτη ή μη συνηθισμένη ενέργεια παρατηρηθεί. Παρόμοιας λογικής εφαρμογές χρησιμοποιούνται και για την πρόβλεψη ατυχημάτων στην βιομηχανία, αλλά και στην στην αυτοκινούμενη οδήγηση.

Όσον αφορά την αλληλεπίδραση ανθρώπου-μηχανής το πιο ισχυρό παράδειγμα είναι η χρήση των assistive robots. Οι ρομποτικές αυτές εφαρμογές μπορούν να βοηθήσουν αυτιστικά παιδιά σε θέματα κοινωνικοποίησης αλλά και ηλικιωμένους σε θέματα αυτοεξυπηρέτησης. Παράλληλα ένα σύστημα αυτόματης αναγνώρισης νοηματικής γλώσσας μπορεί να συμβάλει καθοριστικά στην εκπαίδευση, την κοινωνικοποίηση και την ψυχαγωγία κωφάλαλων ανθρώπων. Εξίσου σημαντική είναι και συνεισφορά των συστημάτων αυτόματης αναγνώρισης δράσης στην αποκατάσταση ασθενών με κινητικά προβλήματα ή τραυματισμένων αθλητών.

Από τα προηγούμενα παραδείγματα γίνεται άμεσα αντιληπτό ότι η μοναδική εκ των περιγραφόμενων εφαρμογή που μπορεί να εκτελεστεί offline είναι η δημιουργία περίληψης βίντεο. Επιβεβαιώνεται λοιπόν η ανάγκη για χρονική κατάτμηση και αναγνώριση αναγνώριση δράσεων σε πραγματικό χρόνο, γεγονός που αποτέλεσε κίνητρο για την πρωτότυπη εργασία που αποτελεί το δεύτερο κομμάτι της παρούσας διπλωματικής εργασίας.

1.3 Συνεισφορές

Κινητήριος δύναμη της παρούσας διπλωματικής εργασίας αποτέλεσαν οι προαναφερθείσες προκλήσεις. Συνεπώς στα πλαίσια αυτής της εργασίας προσπαθούμε να τις αντιμετωπίσουμε, στο πλαίσιο βέβαια του εφικτού. Συγκεκριμένα από την έρευνα μας προκύπτουν οι εξής συνεισφορές: [A] Για την αναγνώριση χειρονομιών σε βίντεο - 1ο Μέρος:

- Συνεισφορά του Attention layer στην αναγνώριση χειρονομιών/δράσεων καθώς έως

τώρα χρησιμοποιείται κυρίως σε προβλήματα φυσικής γλώσσας (NLP).

- Επισκόπηση των σημαντικότερων αρχιτεκτονικών νευρωνικών δικτύων για την αναγνώριση χειρονομιών/δράσεων.
- Συνεισφορά της πληροφορίας βάθους στο C3D μοντέλο και συνδυασμός της με την RGB πληροφορία μέσω late fusion.
- Χρήση πληροφορίας βάθους στο I3D μοντέλο.

[B] Για τον χρονικό εντοπισμό και την αναγνώριση δράσεων σε πραγματικό χρόνο - 2ο Μέρος:

- Διερεύνηση της επίδρασης ποικίλων χωροχρονικών χαρακτηριστικών, που έχουν εξαχθεί από τα εξεταζόμενα στο πρώτο μέρος της διπλωματικής μοντέλα, στον εντοπισμό και την αναγνώριση δράσης σε πραγματικό χρόνο.
- Διερεύνηση της επίδρασης της δυναμικής της ανθρώπινης πόζας, στον εντοπισμό και την αναγνώριση δράσεων σε πραγματικό χρόνο.
- Επίτευξη state-of-the-art [1] αποτελεσμάτων τόσο στο task της αναγνώρισης όσο και στο task του anticipation σε πραγματικό χρόνο.

1.4 Δομή Διπλωματικής

Η παρούσα διπλωματική εργασία διαρθρώνεται σε 7 κεφάλαια. Το πρώτο να είναι εισαγωγικό. Τα υπόλοιπα διαμορφώνονται ως εξής:

- **Κεφάλαιο 2:** Αποτελεί το πρώτο κεφάλαιο βιβλιογραφικής ανασκόπησης και αφορά την αναγνώριση ανθρώπινων δράσεων και χειρονομιών. Παρουσιάζονται, με χρονική σειρά εξέλιξης οι κυριότερες μέθοδοι, που έχουν χρησιμοποιηθεί μέχρι σήμερα, ξεκινώντας από τους τοπικούς περιγραφητές και καταλήγοντας στις νεότερες αρχιτεκτονικές βαθιών νευρωνικών δικτύων.
- **Κεφάλαιο 3:** Στο κεφάλαιο αυτό παραθέτεται βιβλιογραφία σχετική με το δεύτερο πεδίο έρευνας που πραγματεύεται αυτή η διπλωματική, το χρονικό εντοπισμό μιας δράσης, τόσο σε offline όσο και σε online συνθήκες.
- **Κεφάλαιο 4:** Στο κεφάλαιο αυτό παρουσιάζονται τα σύνολα δεδομένων που χρησιμοποιήθηκαν για κάθε ένα από τα εξεταζόμενα προβλήματα καθώς και κάποια εργαλεία που χρησιμοποιήθηκαν για την εξαγωγή χαρακτηριστικών.
- **Κεφάλαιο 5:** Το κεφάλαιο περιέχει τα πειράματα που έγιναν σχετικά με το πρόβλημα της αναγνώρισης ανθρώπινων χειρονομιών σε κατακερματισμένα βίντεο μιας μόνο χειρονομίας έκαστο και συνεπώς σε offline συνθήκες. Συγκεκριμένα ξεκινά με τους ορισμούς των παραμέτρων εκπαίδευσης και των μετρικών αξιολόγησης που χρησιμοποιούνται στα πειράματα που ακολουθούν. Συνεχίζει με τις περιγραφές των μοντέλων που υλοποιήθηκαν και ολοκληρώνεται με την σύγκριση των αποτελεσμάτων που προέκυψαν.

[1] [65] V. I. Vasileiou, N. Kardaris, P.Maragos. Exploring Temporal Context and Human Movement Dynamics for Online Action Detection in Videos. Submitted in Proc. EUSIPCO, 2021.

- **Κεφάλαιο 6:** Το κεφάλαιο αφορά το χρονικό εντοπισμό ανθρώπινων δράσεων σε πραγματικές συνθήκες. Η δομή που ακολουθεί είναι ίδια με αυτή του 5ου κεφαλαίου. Συγκεκριμένα ξεκινά με τους ορισμούς των παραμέτρων εκπαίδευσης και των μετρικών αξιολόγησης που χρησιμοποιούνται στα πειράματα που ακολουθούν. Συνεχίζει με τις περιγραφές των μοντέλων που υλοποιήθηκαν και ολοκληρώνεται με την σύγκριση των αποτελεσμάτων που προέκυψαν.
- **Κεφάλαιο 7:** Στο κεφάλαιο αυτό γίνεται μια σύνοψη των αποτελεσμάτων και των συνεισφορών της παρούσας διπλωματικής και προτείνονται πιθανές μελλοντικές κατευθύνσεις έρευνας.

Κεφάλαιο 2

Αναγνώριση Ανθρώπινων Δράσεων/Χειρονομιών στη Βιβλιογραφία

Στο κεφάλαιο αυτό παραθέτουμε τις σημαντικότερες προσεγγίσεις στο αντικείμενο της αναγνώρισης ανθρώπινων δράσεων και χειρονομιών. Το κεφάλαιο έχει διαρθρωθεί με βάση την χρονική εξέλιξη της έρευνας, γεγονός που μας βοηθάει να επεκτείνουμε τις γνώσεις μας στο αντικείμενο και να αποκτήσουμε μια σφαιρική εικόνα για αυτό. Κάποιες από τις νεότερες μεθόδους που περιγράφονται στο κεφάλαιο αυτό θα χρησιμοποιηθούν και στα πειράματα μας, όπως αναλύονται και στο κεφάλαιο 5.

2.1 Τοπικά Χωροχρονικά Χαρακτηριστικά

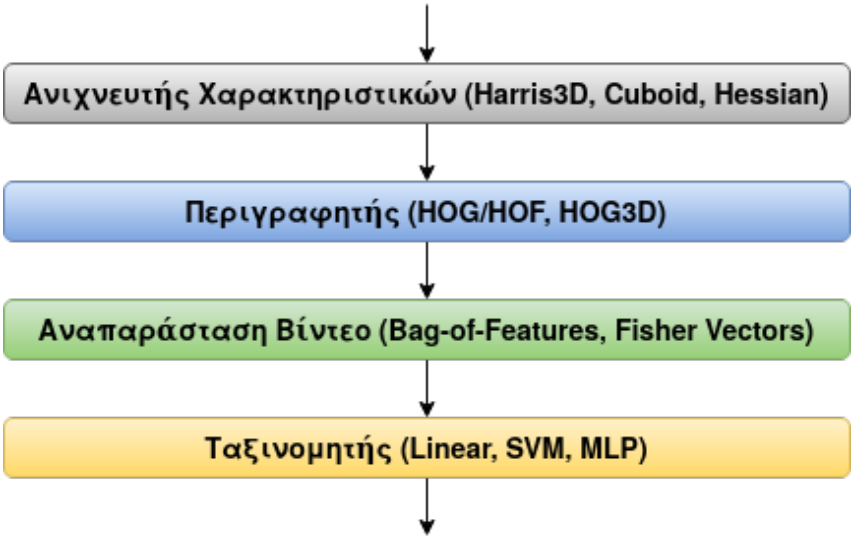
Οι πρώτες προσεγγίσεις της αναγνώρισης ανθρώπινων δράσεων ακολουθούσαν τα εξής βήματα: α) επιλογή χωροχρονικών περιοχών και κλιμάκων μέσω ενός ανιχνευτή χαρακτηριστικών, β) σύλληψη σχήματος και κίνησης στις επιλεχθείσες χωροχρονικές περιοχές μέσω ενός περιγραφητή, γ) κωδικοποίηση των βίντεο μέσω διαφόρων αναπαραστάσεων, και δ) χρήση ταξινομητών για την κατηγοριοποίηση των δράσεων (σχήμα 2.1).

Για καθένα από τα προηγούμενα βήματα παραθέτουμε ενδεικτικά τις πιο δημοφιλείς, στη βιβλιογραφία, επιλογές. Συγκεκριμένα όσον αφορά στους ανιχνευτές χαρακτηριστικών παρουσιάζουμε τους εξής:

- **Harris3D:** Είναι ο δημοφιλέστερος ανιχνευτής στη βιβλιογραφία, προτάθηκε από τους Laptev και Lindeberg [37] ως μια χωροχρονική επέκταση του Harris ανιχνευτή [21]. Δεδομένου ότι ο Harris ανιχνεύει γωνίες σε εικόνες, ο Harris3D ανιχνεύει σημεία που παρουσιάζουν διακριτική εμφάνιση στο χώρο και μη σταθερή κίνηση στον χρόνο.

- **Cuboid:** Εισήχθη από τους Dollar et al. [12] βασίζεται στα φίλτρα Gabor και διαφοροποιείται από τους υπόλοιπους, καθώς αποτελεί δεν αποτελεί επέκταση κάποιου χωρικού δισδιάστατου ανιχνευτή. Τα κριτήρια ανίχνευσης είναι τα διακριτικά χαρακτηριστικά στο χώρο και η σύνθετη κίνηση στο χρόνο.

- **Hessian:** Αντίστοιχα με τον Harris3D ο Hessian αποτελεί την χωροχρονική επέκταση της μετρικής σημαντικότητας Hessian, που χρησιμοποιήθηκε για την ανίχνευση "σταγόνων" σε εικόνες. Αναπτύχθηκε από τους Willems et al. [75] και ανιχνεύει τα πιο αραιά χωροχρονικά τοπικά χαρακτηριστικά.



Σχήμα 2.1: Pipeline αναγνώρισης δράσεων με τη χρήση τοπικών χωροχρονικών χαρακτηριστικών

Από το σύνολο των περιγραφητών αυτοί που προσφέρουν τα πιο ικανοποιητικά αποτελέσματα είναι:

- **HOG/HOF:** Εισήχθη από τους Laptev et al. [38] και αποτελεί επέκταση των περιγραφητών Histogram of Oriented Gradients (HOG) και Histogram of oriented Optical Flow (HOF). Ο περιγραφητής αυτός βασίζεται στον υπολογισμό των ιστογραμμάτων χωρικών gradients και της οπτικής ροής στην χωροχρονική γειτονιά τον επιλεχθέντων σημείων ενδιαφέροντος.
- **HOG3D:** Αποτελεί μια ακόμη επέκταση του HOG προτεινόμενη από τους Klaser et al. [34] και βασίζεται στον υπολογισμό των κατευθύνσεων χωροχρονικών 3D gradients.

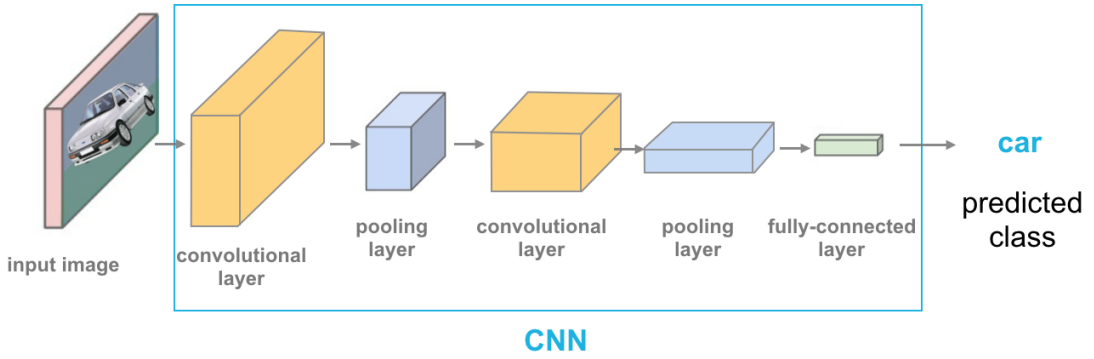
Μια από τις δημοφιλέστερες αναπαραστάσεις ενός βίντεο, είναι η **Bag-of-Features (BoF)**. Αυτή είναι ένα ιστόγραμμα που αποτελεί πρακτικά την ομαδοποίηση των περιγραφητών που έχουν δημιουργηθεί για σημείο ενδιαφέροντος.

Μία από τις κυριότερες τεχνικές ταξινόμησης που συνδυάζεται πολύ αποτελεσματικά με τα ιστογράμματα BoF είναι οι **Μηχανές Διανυσμάτων Υποστήριξης (SVMs)** [9]. Είναι μοντέλα εποπτευόμενης μάθησης που επιτυγχάνουν υψηλά περιθώρια διαχωρισμού.

Αν και οι προηγούμενες μέθοδοι έχουν δείξει ελπιδοφόρα αποτελέσματα η αφθονία των οπτικών δεδομένων και η πρόοδος στο σχεδιασμό υλικού οδήγησε την ερευνητική κοινότητα να αγκαλιάσει τα μοντέλα βαθιών νευρωνικών δικτύων (συνελικτικά και αναδρομικά), τα οποία θα παρουσιάσουμε στις επόμενες δύο ενότητες.

2.2 Συνελικτικά Νευρωνικά Δίκτυα - CNN

Τα Συνελικτικά Νευρωνικά Δίκτυα (Convolutional Neural Networks) [39] είναι κανονικοποιημένες εκδόσεις πολυστρωματικών perceptrons. Με τον όρο πολυστρωματικό perceptron εννοούμε ένα πλήρως συνδεδεμένο δίκτυο, δηλαδή κάθε νευρώνας σε ένα στρώμα συνδέεται με όλους τους νευρώνες στο επόμενο στρώμα. Τα CNN είναι βαθιά νευρωνικά δίκτυα που χρησιμοποιούνται για την επεξεργασία πληροφορίας που μπορεί να αναπαρασταθεί με την μορφή πλέγματος (π.χ. εικόνες και βίντεο).

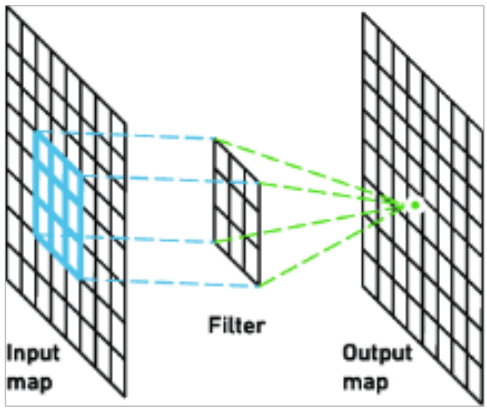


Σχήμα 2.2: Δομή CNN [2]. Απεικονίζεται ένα συνελικτικό νευρωνικό δίκτυο για την αναγνώριση εικόνων, που απαρτίζεται από δύο συνελικτικά στρώματα, δύο στρώματα συσσώρευσης και ένα πλήρως συνδεδεμένο στρώμα.

2.2.1 Δισδιάστατα Συνελικτικά Δίκτυα - 2D-CNN

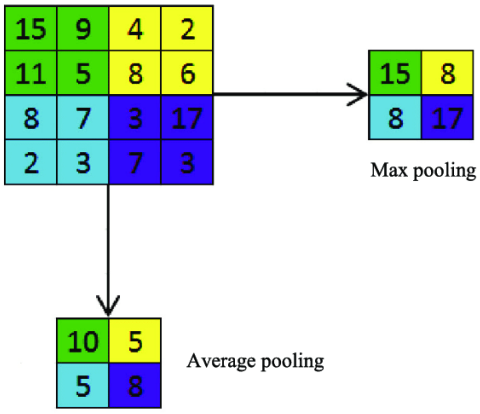
Τα δισδιάστατα συνελικτικά δίκτυα απαρτίζονται από μια ακολουθία στρωμάτων τα οποία μέσω της πράξης της συνέλιξης μειώνουν σταδιακά τις διαστάσεις των δεδομένων εισόδου, διατηρώντας όμως τα χαρακτηριστικά τους. Τα στρώματα αυτά διακρίνονται σε τρεις κατηγορίες Convolutional, Pooling και Fully-Connected:

- **Συνελικτικά - Convolutional:** Είναι τα στρώματα όπου η είσοδος συνελίσσεται με μια σειρά από φίλτρα και παράγει για καθένα από αυτά έναν χάρτη ενεργοποίησης. Τα φίλτρα χρησιμοποιούνται για τον εντοπισμό συγκεκριμένων χαρακτηριστικών.



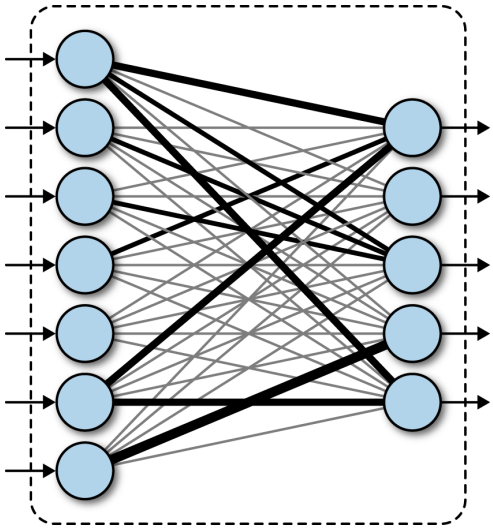
Σχήμα 2.3: Δομή του συνελικτικού στρώματος [79]

- **Συσσώρευσης - Pooling:** Είναι στρώματα που χρησιμοποιούνται μεταξύ των Convolutional στρωμάτων προκειμένου να μειώσουν τις παραμέτρους του δικτύου. Οι βασικές πράξεις που εκτελούνται σε αυτό το επίπεδο είναι: α) *Max-Pooling* - επιλογή μέγιστης τιμής ανά κελί, β) *Average-Pooling* - επιλογή της μέσης τιμής ανά κελί.



Σχήμα 2.4: Δομή του στρώματος συσσώρευσης [50]

- **Πλήρως-Συνδεδεμένα - Fully-Connected:** Είναι τα στρώματα που συνδέουν κάθε νευρώνα του τρέχοντος επιπέδου με το επόμενο επίπεδο. Έχουν παρόμοια λειτουργία με τα Multi-Layer Perceptron. Ως είσοδος δίνεται ένας μονοδιάστατος πίνακας, ο οποίος, αφού πλατυνθνεί, στέλνεται σε ένα πλήρως συνδεδεμένο επίπεδο για να κατηγοριοποιηθούν τα δεδομένα εισόδου.



Σχήμα 2.5: Πλήρως-Συνδεδεμένο Στρώμα [49]

2.2.2 Δισδιάστατα CNN σε Βίντεο

Τα state-of-the-art αποτελέσματα που επιτυγχάνουν τα CNNs στο πεδίο της αναγνώρισης σε εικόνες οδήγησαν στην προσπάθεια γενίκευσης τους στο πεδίο της αναγνώρισης σε βίντεο. Στην περίπτωση της εικόνας, η εισαγόμενη προς επεξεργασία πληροφορία στο δίκτυο είναι ένα δισδιάστατο πλέγμα. Αντίστοιχα γενικεύοντας τα CNNs σε βίντεο προστίθεται μια ακόμα διάσταση, η διάσταση του χρόνου. Για τον λόγο αυτό προτάθηκαν τα τρισδιάστατα συνελικτικά δίκτυα που θα αναλύσουμε σε επόμενη ενότητα. Παράλληλα γίνεται έρευνα και στην χρήση δισδιάστατων συνελικτικών δικτύων για την αναγνώριση δράσεων σε βίντεο.

Προκειμένου να είναι εφικτή η χρήση 2D-CNNs για την αναγνώριση ανθρώπινων δράσεων σε βίντεο πρέπει να θεωρήσουμε το βίντεο ως ένα σύνολο από δισδιάστατες εικόνες/καρέ.

Με την παραδοχή αυτή όμως εγκαταλείπουμε και την χρονική πληροφορία που μας δίνει το βίντεο. Έχουν γίνει αρκετές προσπάθειες για την αξιοποίηση και της χρονικής πληροφορίας, όπως για παράδειγμα από τους Karpathy et al. [30], οι οποίοι βασίστηκαν στο γεγονός ότι κάθε βίντεο περιέχει αρκετά συνεχόμενα καρέ στο χρόνο. Με βάση αυτή την άποψη, η συνδεσιμότητα του δικτύου μπορεί να επεκταθεί στη διάσταση του χρόνου και να αποκτήσουμε τα επιζητούμενα χωροχρονικά χαρακτηριστικά. Στη συνέχεια προτάθηκε η χρήση συνελικτικών δικτύων διπλής ροής [58], όπου η ροή της χωρικής πληροφορίας εκτελεί αναγνώριση με βάση τα RGB frames και η ροή χρονικής πληροφορίας εκτελεί αναγνώριση με βάση την πυκνή οπτική ροή. Το τελικό αποτέλεσμα προκύπτει από την μέση τιμή των αποτελεσμάτων που έχουν εξαχθεί από τα δύο softmax επίπεδα. Επίσης έχει προταθεί η προσθήκη ενός αναδρομικού επιπέδου στο συνελικτικό δίκτυο, όπως στο Long-term Recurrent Convolutional Network [13] μοντέλο που συνδυάζει ένα βαθύ ιεραρχικό οπτικό εξαγωγέα χαρακτηριστικών (π.χ. CNN) με ένα μοντέλο που μπορεί να μάθει, να αναγνωρίσει και να συνθέσει τη χρονική δυναμική για εργασίες που περιλαμβάνουν σειριακά δεδομένα.

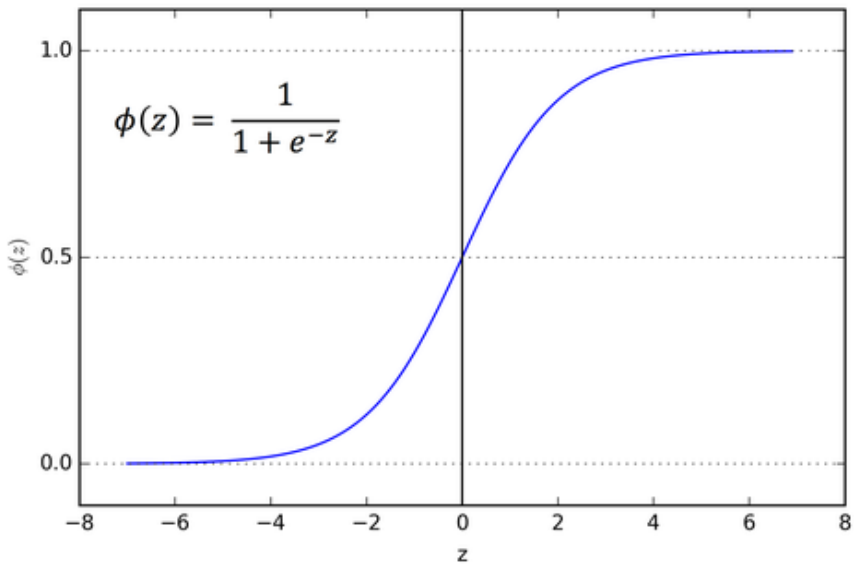
2.3 Αναδρομικά Νευρωνικά Δίκτυα - RNN

Τα Αναδρομικά Νευρωνικά Δίκτυα (Recurrent Neural Networks) χρησιμοποιούνται ευρέως σε εργασίες αναγνώρισης ακολουθιακών δεδομένων, όπως είναι η αναγνώριση δράσεων ή χειρονομιών σε βίντεο. Σε ένα αναδρομικό δίκτυο σε αντίθεση με τα feed-forward δίκτυα κάθε νέα είσοδος εξαρτάται από τις προηγούμενες, γεγονός που γίνεται άμεσα αντιληπτό ακόμη και από το όνομα του. Συνεπώς η τρέχουσα κατάσταση δίνεται από την εξής σχέση:

$$h_t = f(h_{t-1}, x_t).$$

Τα αναδρομικά νευρωνικά δίκτυα χρησιμοποιούν κάποια συνάρτηση ενεργοποίησης, με συνηθέστερες τις:

- **Σιγμοειδής:** Όπως φαίνεται και στην επόμενη εικόνα μοιάζει με ένα τελικό σίγμα.

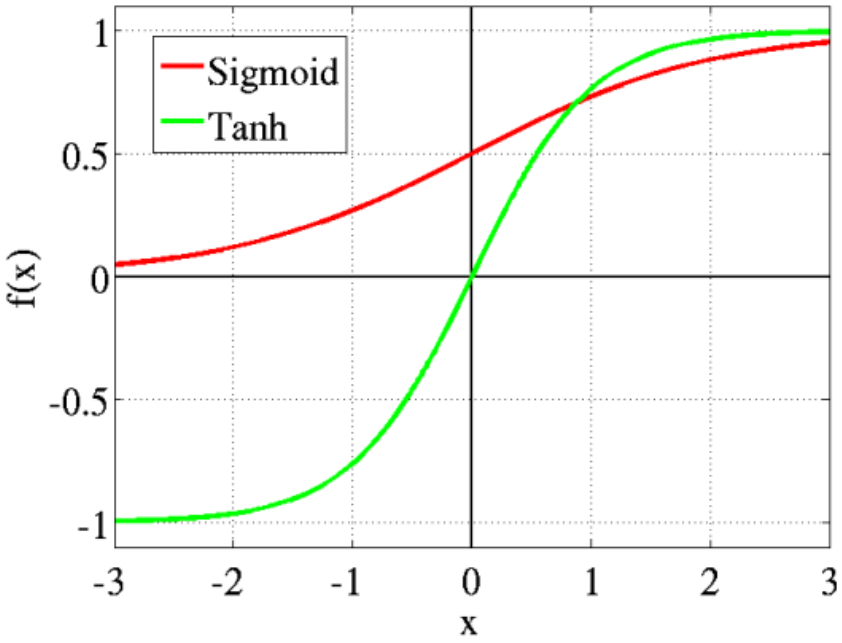


Σχήμα 2.6: Σιγμοειδής συνάρτηση ενεργοποίησης [54]

Δίνεται από την εξίσωση $g(z) = \frac{1}{1+e^{-z}}$ και ο κυριότερος λόγος χρήσης της είναι ότι

κυμαίνεται μεταξύ των τιμών 0 και 1. Συνεπώς, χρησιμοποιείται κυρίως σε μοντέλα τα οποία ως έξοδο προβλέπουν πιθανότητες. Εξ' ορισμού η τιμή μιας πιθανότητας κυμαίνεται στο διάστημα $[0, 1]$, συνεπώς η χρήση αυτής της συνάρτησης ενεργοποίησης είναι ιδανική. Παράλληλα, η σιγμοειδής συνάρτηση είναι διαφοροποιήσιμη και συνεπώς μπορούμε να βρούμε την κλίση της σιγμοειδούς καμπύλης σε οποιοδήποτε σημείο. Τέλος, η συνάρτηση softmax είναι μια πιο γενικευμένη σιγμοειδής συνάρτηση που χρησιμοποιείται για την ταξινόμηση πολλαπλών κλάσεων.

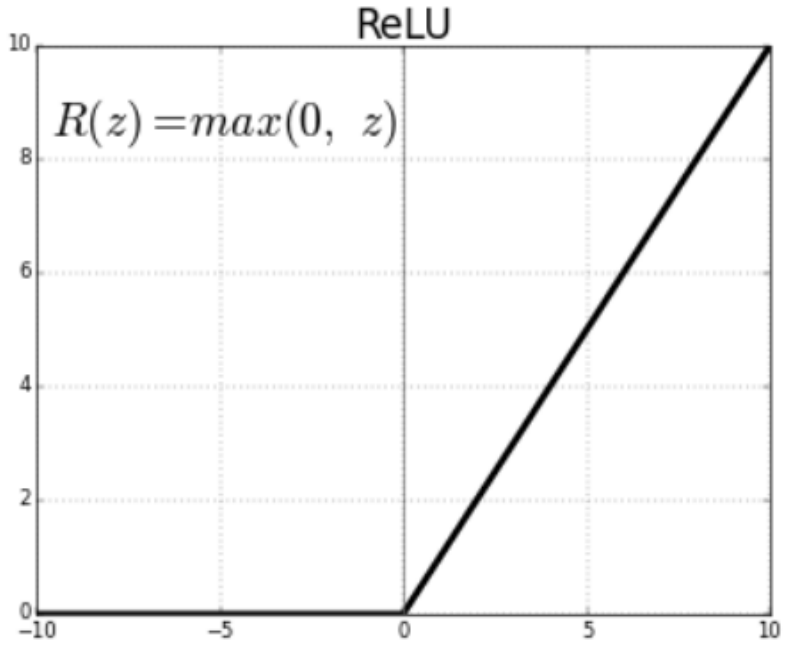
- **Υπερβολική Εφαπτομένη - tanh:** Όπως φαίνεται και στην επόμενη εικόνα έχει και αυτή σιγμοειδή μορφή, αλλά, κυμαίνεται στο διάστημα $[-1, 1]$ και είναι καλύτερη από την σιγμοειδή.



Σχήμα 2.7: Σύγκριση συναρτήσεων ενεργοποίησης σιγμοειδής και υπερβολική εφαπτομένη [54]

Δίνεται από την εξίσωση $g(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$ και είναι διαφοροποιήσιμη όπως και η προηγούμενη. Σε αντίθεση με την απλή σιγμοειδή έχει το πλεονέκτημα ότι οι αρνητικές είσοδοι θα αντιστοιχιστούν (mapping) έντονα αρνητικές ενώ οι μηδενικές είσοδοι θα αντιστοιχιστούν κοντά στο μηδέν. Η tanh συνάρτηση ενεργοποίησης χρησιμοποιείται κυρίως κατηγοριοποίηση μεταξύ δύο κλάσεων.

- **Διορθωμένη Γραμμική Μονάδα - ReLU:** Στην εικόνα 2.8 φαίνεται η πιο συχνά χρησιμοποιούμενη συνάρτηση ενεργοποίησης σήμερα. Δίνεται από την εξίσωση $g(z) = \max(0, z)$ και εκτείνεται στο διάστημα $[0, +\infty)$. Ένα όμως βασικό μειονέκτημα της συνάρτησης αυτής είναι ότι όλες οι αρνητικές τιμές γίνονται αμέσως μηδενικές, γεγονός που μειώνει την ικανότητα του μοντέλου να προσαρμόζεται (fit) ή να εκπαιδεύεται σωστά από τα δεδομένα. Αυτό σημαίνει ότι κάθε αρνητική είσοδος που δίνεται στη συνάρτηση ενεργοποίησης ReLU μετατρέπει την τιμή σε μηδέν αμέσως στο γράφημα, το οποίο με τη σειρά του επηρεάζει το γράφημα που προκύπτει μη αντιστοιχίζοντας κατάλληλα τις αρνητικές τιμές.

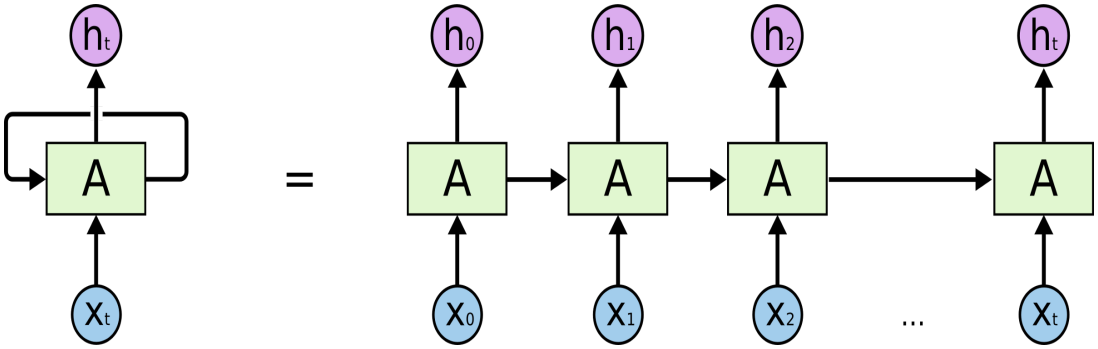


Σχήμα 2.8: ReLU συνάρτηση ενεργοποίησης [54]

Συνεπώς, επιλέγοντας να εφαρμόσουμε την υπερβολική εφαπτομένη, ως συνάρτηση ενεργοποίησης, για την τρέχουσα κατάσταση έχουμε:

$$h_t = \tanh\left(W_{hh}h_{t-1} + W_{xh}x_t\right),$$

όπου h το κρυφό επίπεδο, W_{hh} και W_{xh} είναι τα βάρη της προηγούμενης κρυφής κατάστασης και της τρέχουσας εισόδου αντίστοιχα. Τελικά η έξοδος δίνεται από την σχέση: $y_t = W_{hy}h_t$, όπου W_{hy} τα βάρη της κατάστασης εξόδου.



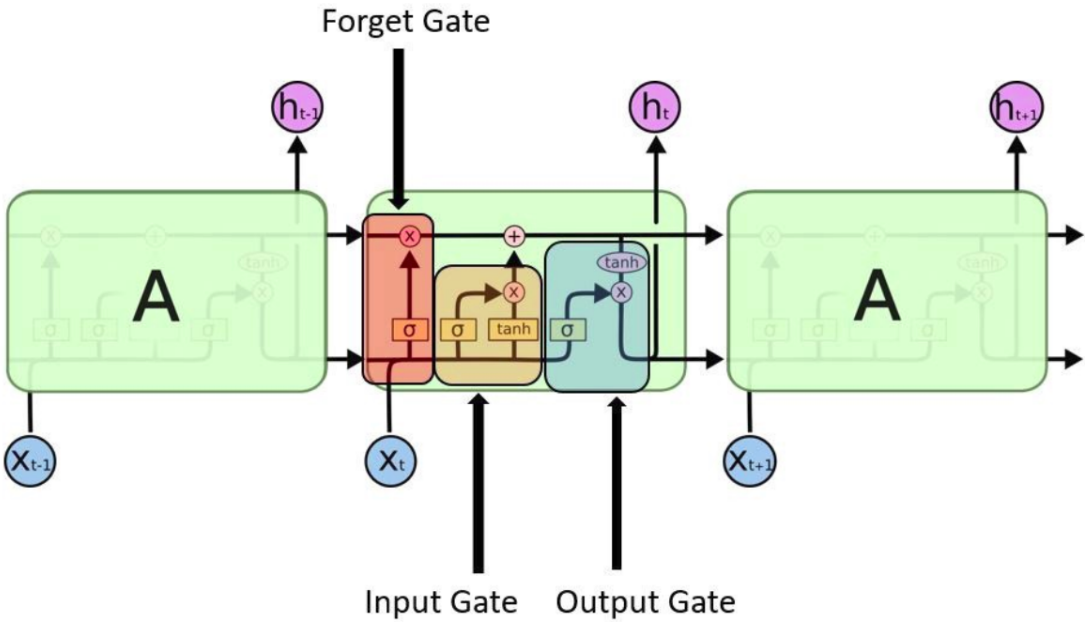
Σχήμα 2.9: Ξεδιπλωμένη δομή ενός RNN [44]

Αν όμως σε κάποιο πρόβλημα προκύψουν υπερβολικά ακραίες τιμές για την κλίση του κόστους, γνωστό ως πρόβλημα vanishing gradient, το αναδρομικό νευρωνικό δίκτυο δεν θα μπορέσει να μοντελοποιήσει επιτυχώς τις εξαρτήσεις μακράς διάρκειας. Το πρόβλημα όμως αυτό αντιμετωπίστηκε όταν εισήχθησαν νευρώνες Μακράς-Βραχείας Μνήμης (Long Short-Term Memory).

2.3.1 Νευρώνες Μακράς-Βραχείας Μνήμης - LSTM

Τα LSTM [24] αποτελούν μια τροποποιημένη εκδοχή των RNNs και έχουν αποδειχθεί ιδιαίτερα αποτελεσματικά στην ταξινόμηση, την επεξεργασία και την πρόβλεψη χρονοσειρών.

Σε αντίθεση με το κλασικό RNN το LSTM δεν εφαρμόζει συνάρτηση ενεργοποίησης στις αναδρομικές συνδέσεις. Συνεπώς όλες οι ενημερώσεις είναι γραμμικές, γεγονός που επιτρέπει να μην εξαφανίζονται οι gradients από την επαναληπτική εφαρμογή των ενημερώσεων, εξασφαλίζοντας έτσι τη ροή της πληροφορίας στο δίκτυο. Χρησιμοποιούνται σιγμοειδείς συναρτήσεις ενεργοποίησης προκειμένου να διευκολύνεται η ανανέωση (τιμή που θυμάται) ή η απόρριψη (τιμή που ξεχνά) της πληροφορίας. Παράλληλα όπως φαίνεται και στην επόμενη εικόνα,



Σχήμα 2.10: Δομή μιας μονάδας LSTM [44]

κάθε μονάδα LSTM διαθέτει έναν μηχανισμό με τρεις θύρες, οι οποίες ρυθμίζουν το πόσο θα ενημερώνεται κάθε διάνυσμα του δικτύου. Επιτυγχάνεται με αυτόν τον τρόπο η καλύτερη αφομοίωση και η διατήρηση των σημαντικότερων πληροφοριών. Η λειτουργία του θα γίνει ακόμα περισσότερο κατανοητή εξετάζοντας κάθε πύλη του ξεχωριστά:

- **Forget Gate:** Η πύλη απόρριψης είναι εκείνη που επιλέγει ποια πληροφορία θα ανανεωθεί και ποια θα απορριφθεί. Συγκεκριμένα όπως φαίνεται και από την επόμενη εξίσωση η σιγμοειδής αποφασίζει ποιες τιμές θα περάσει από 0 (απόρριψη) ή 1 (ανανέωση):

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

- **Input Gate:** Η πύλη εισόδου αποτελείται από δύο τμήματα, το πρώτο διαθέτει μια σιγμοειδή που λειτουργεί όπως η πύλη απόρριψης. Το δεύτερο διαθέτει μια υπερβολική εφαπτομένη η οποία επιλέγει την βαρύτητα που θα δώσει στην μεταβιβαζόμενη πληροφορία με τις τιμές να κυμαίνονται στο διάστημα [-1,1]. Οι αντίστοιχες μαθηματικές εξισώσεις είναι:

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

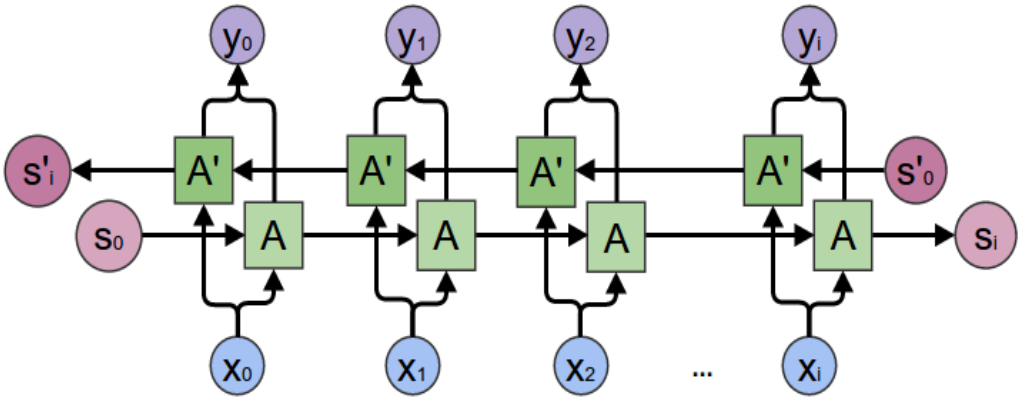
$$\widetilde{C}_i = \tanh(W_C[h_{t-1}, x_t] + b_c)$$

- **Output Gate:** Η πύλη εξόδου χρησιμοποιεί την είσοδο και την μνήμη για να αποφασίσει την έξοδο. Μια σιγμοειδής αποφασίζει εκ νέου ποιες τιμές θα κρατήσει και ποιες θα απορρίψει και η υπερβολική εφαπτομένη επιλέγει την βαρύτητα που θα δώσει στην πληροφορία που θα μεταβιβαστεί με τις τιμές να κυμαίνονται στο διάστημα [-1,1]. Η έξοδος της σιγμοειδούς και της υπερβολικής εφαπτομένης πολλαπλασιάζονται ώστε να προκύψει η κρυφή κατάσταση. Η μαθηματική περιγραφή των προηγούμενων είναι:

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

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

Μια επέκταση των απλών LSTMs είναι τα αμφίδρομα LSTMs (Bi-LSTMs), τα οποία αποτελούνται από δύο ανεξάρτητα LSTMs με αντίθετη ροή πληροφορίας, ώστε το δίκτυο να λαμβάνει τόσο την προς τα εμπρός όσο και την προς τα πίσω πληροφορία, για μια ακολουθία κάθε στιγμή. Με αυτό τον τρόπο τα Bi-LSTMs διαθέτουν πληροφορία τόσο από το παρελθόν



Σχήμα 2.11: Δομή Bi-LSTM [44]

όσο και από το μέλλον γεγονός που τους επιτρέπει να αντιλαμβάνονται καλύτερα το εκάστοτε context.

2.3.2 Αναδρομικά Δίκτυα και Ανθρώπινη Πόζα

Το περιβάλλον στο οποίο εκτελείται μια δράση αλλά και τα αντικείμενα που χρησιμοποιούνται κατά την εκτέλεση της συνήθως λειτουργούν επικουρικά στις παραδοσιακές μεθόδους αναγνώρισης που βασίζονται στην χρήση των RGB frames. Υπάρχουν όμως και περιπτώσεις που αυτοί οι εξωτερικοί παράγοντες μπορεί να έχουν αρνητική επίδραση, όπως συμβαίνει στην αναγνώριση ανθρώπινων χειρονομιών. Αυτή ακριβώς αρνητική επίδραση προκάλεσε προβλήματα, τα οποία αλυσιδωτά αποτέλεσαν κίνητρο για την εισαγωγή νέων μεθόδων όπου το context θα παραμένει ανεπηρέαστο από τους παραπάνω παράγοντες, με μια από τις πιο αξιόλογες προτάσεις να είναι η χρήση της πληροφορίας που προσφέρει ο σκελετός. Ο συνδυασμός των αναδρομικών νευρωνικών δικτύων με την προερχόμενη από τους σκελετούς πληροφορία, έχει χρησιμοποιηθεί σε πολλές δημοσιεύσεις, επιτυγχάνοντας πολλά υποσχόμενα αποτελέσματα στο πεδίο αναγνώρισης ανθρώπινων δράσεων σε βίντεο.

Στο [15] προτάθηκε ένα ιεραρχικό RNN από άκρο σε άκρο, για την αναγνώριση δράσεων, βάσει σκελετού. Στη δουλειά αυτή αντί να χρησιμοποιηθεί ενιαίος ο σκελετός, χωρίστηκε

με βάση την φυσική ανθρώπινη δομή σε πέντε τμήματα, τα οποία τροφοδοτούν πέντε διαφορετικά υποδίκτυα. Καθώς αυξάνεται ο αριθμός των επιπέδων, οι αναπαραστάσεις, που εξάγονται από τα υποδίκτυα, συγχωνεύονται ιεραρχικά, για να αποτελέσουν τις εισόδους υψηλότερων επιπέδων. Η τελική αναπαράσταση των ακολουθιών των σκελετών τροφοδοτεί ένα μονοστρωματικό perceptron.

Οι Zhu et al. [82] πρότειναν ένα βαθύ, πλήρως συνδεδεμένο LSTM δίκτυο για αναγνώριση δράσεων, βάσει σκελετού. Βασική ιδέα είναι η συν-εμφάνιση (co-occurrence) κάποιων αρθρώσεων ανά δράση (π.χ. για την ομιλία σε ένα κινητό τηλέφωνο σχεδόν πάντα εμφανίζονται η παλάμη, ο καρπός και το κεφάλι), ενώ παράλληλα προστέθηκε ένας εσωτερικός dropout [62] μηχανισμός στις πύλες του LSTM.

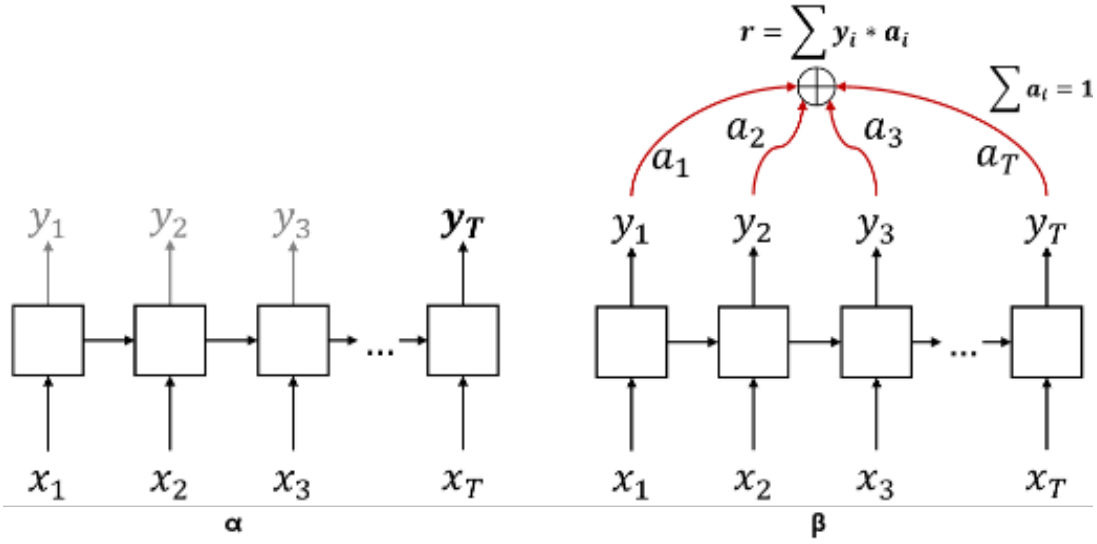
Συγκεκριμένα ο dropout μηχανισμός εφαρμόζει έναν τυχαίο μηδενισμό κάποιων συνδέσεων μεταξύ των νευρώνων του δικτύου. Είναι μια μέθοδος κανονικοποίησης που έχει συμβάλει στη βελτίωση της ικανότητας των μοντέλων να γενικεύονται, καθώς προσεγγίζει την εκπαίδευση ενός μεγάλου αριθμού νευρωνικών δικτύων με διαφορετικές αρχιτεκτονικές παράλληλα. Αναλυτικότερα, κατά τη διάρκεια της εκπαίδευσης, κάποιος αριθμός των στρωμάτων εξόδου αγνοείται τυχαία ή «dropped out». Αυτό έχει ως αποτέλεσμα να κάνει το τρέχον στρώμα να μοιάζει και να αντιμετωπίζεται σαν ένα στρώμα με διαφορετικό αριθμό κόμβων και συνδεσιμότητα με το προηγούμενο στρώμα. Στην πραγματικότητα, κάθε ενημέρωση σε ένα επίπεδο κατά τη διάρκεια της προπόνησης εκτελείται με διαφορετική «εμφάνιση» του διαμορφωμένου επιπέδου. Ο μηχανισμός αυτός μπορεί να εφαρμοστεί σε όλα τα κρυφά στρώματα του δικτύου, όπως επίσης και στα ορατά στρώματα και το στρώμα εισόδου. Δεν μπορεί όμως να εφαρμοστεί στο στρώμα εξόδου. Για την λειτουργία του μηχανισμού αυτού εισάγεται μια νέα υπερπαράμετρος που καθορίζει την πιθανότητα κατά την οποία οι έξοδοι του εκάστοτε στρώματος αποσύρονται, ή αντίστροφα, η πιθανότητα στην οποία διατηρούνται οι έξοδοι του στρώματος. Μια κοινή τιμή είναι η πιθανότητα 0.5 για διατήρηση της εξόδου κάθε κόμβου σε ένα κρυφό στρώμα και μια τιμή κοντά στο 1.0, όπως το 0.8, για τη διατήρηση εισόδων από το ορατό στρώμα.

Μια τρίτη προσπάθεια αποτυπώνεται στο Part-Aware LSTM [53], όπου, αντί να εισαχθεί ολόκληρος ο σκελετός σε ένα ενιαίο κύτταρο LSTM, θεωρήθηκε πιο αποδοτικό να μελετηθεί το context κάθε αυτόνομου ανθρώπινου μέρους ξεχωριστά. Συνεπώς κάθε ανθρώπινο μέρος τροφοδοτεί ένα μικρότερο κύτταρο LSTM, το σύνολο των οποίων συνδυάζεται σε ένα τελικό επίπεδο, ώστε να γίνει η αναγνώριση.

Τέλος ένα συνδυασμό μεταξύ των όσων αναφέραμε στην ενότητα 2.2.2 και στην παρούσα ενότητα, εφάρμοσαν οι Lai et al. στο [36]. Συγκεκριμένα ο συνδυασμός CNN και RNN παρέχει την απαραίτητη χωροχρονική πληροφορία η οποία εμπλουτίζεται από την χρονική πληροφορία που παρέχουν οι κινήσεις του σκελετού όταν τροφοδοτήσουν ένα LSTM δίκτυο.

2.3.3 Μηχανισμός Προσοχής

Τα LSTM όπως και τα RNN χρησιμοποιούν την τελευταία τιμή της εσωτερικής κατάστασης ως διανυσματική αναπαράσταση όλης της ακολουθίας. Αν η ακολουθία που θα δοθεί είναι αρκετά μεγάλη, όπως συμβαίνει και στις περισσότερες περιπτώσεις αναγνώρισης δράσεων σε βίντεο, υπάρχει περίπτωση το δίκτυο να μην μπορεί να συγκρατήσει όλες τις σημαντικές πληροφορίες στην εσωτερική του κατάσταση. Για την αντιμετώπιση αυτής της "μυωπικής" συμπεριφοράς των αναδρομικών νευρωνικών δικτύων μπορούμε να χρησιμοποιήσουμε έναν μηχανισμό προσοχής [66], ο οποίος προσπαθεί να ενισχύσει την συνεισφορά των πιο σημαντικών στοιχείων στην τελική αναπαράσταση. Σε ένα αναδρομικό νευρωνικό δίκτυο, με μηχανισμό προσοχής, η τελική αναπαράσταση της ακολουθίας είναι το σταθμισμένο

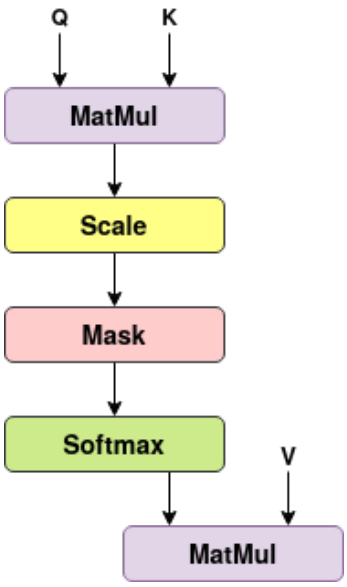


Σχήμα 2.12: α) Απλό RNN, β) RNN με Μηχανισμό Προσοχής

άθροισμα των ενδιάμεσων εξόδων, όπου τα βάρη του αθροίσματος ορίζονται από στρώμα προσοχής. Ο μηχανισμός προσοχής αποτελεί ένα επίπεδο του δικτύου και συνεπώς εκπαιδεύεται όπως και τα υπόλοιπα.

Όσον αφορά την αναγνώριση της δράσεων, βάσει σκελετού, η σημασία των μακροπρόθεσμων πληροφοριών με βάση τα συμφραζόμενα είναι ιδιαίτερα σημαντική. Παρά την επιτυχία που επιδεικνύουν τα Self-Attentive δίκτυα στους τομείς της αναγνώρισης φυσικής γλώσσας, η επέκταση τους στην αναγνώριση δράσεων, βάσει σκελετού, είναι περιορισμένη.

Στο [8] γίνεται μια προσπάθεια να γεφυρωθεί αυτό το χάσμα, όπου χρησιμοποιούνται ένα Self-Attention δίκτυα πάνω στις κωδικοποιημένες αναπαραστάσεις των ακολουθιών θέσης (θέσεις αρθρώσεων) και κίνησης (ταχύτητες αρθρώσεων - διαφορά δύο διαδοχικών θέσεων). Εν συνεχεία χρησιμοποιούνται τρεις διαφορετικές αρχιτεκτονικές δικτύων που βασίζονται στα Self-Attention δίκτυα για την αποτελεσματική καταγραφή των πληροφοριών με βάση τα συμφραζόμενα από τις κωδικοποιημένες δυνατότητες.



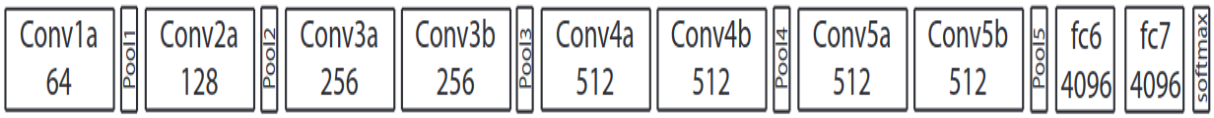
Σχήμα 2.13: Δομή Self-Attention

2.4 Τρισδιάστατα Συνελικτικά Δίκτυα - 3D-CNN

Τα τρισδιάστατα συνελικτικά δίκτυα, όπως έχουμε προαναφέρει, είναι μια γενίκευση των δισδιάστατων συνελικτικών δικτύων και αποτελούνται από τρισδιάστατα convolutional και pooling στρώματα, μέσω των οποίων μοντελοποιείται και η χρονική πληροφορία. Συνεπώς τα τρισδιάστατα συνελικτικά δίκτυα είναι καταλληλότερα από τα δισδιάστατα για την εκμάθηση χωροχρονικών χαρακτηριστικών. Έχουν αναπτυχθεί ποικίλα 3D-CNNs, τα περισσότερα από τα οποία είναι αρκετά ρηχά. Αυτό συνέβη και με το πρώτο 3D-CNN [27] που αποτελούνταν από τρία τρισδιάστατα συνελικτικά επίπεδα, δύο τρισδιάστατα επίπεδα συσσώρευσης και ένα πλήρως συνδεδεμένο επίπεδο και χρησιμοποιήθηκε για αναγνώριση ανθρώπινων δράσεων.

2.4.1 Convolutional 3D - C3D

Το μοντέλο C3D προτάθηκε από τους Tran et al. [63] και είναι ένα βαθύ τρισδιάστατο συνελικτικό δίκτυο. Αποτελείται από οκτώ συνελικτικά επίπεδα, πέντε επίπεδα συσσώρευσης και δύο πλήρως συνδεδεμένα επίπεδα με 4096 νευρώνες έκαστο. Κάθε ένας από τους συνελικτικούς πυρήνες έχει διαστάσεις $3 \times 3 \times 3$ και stride $1 \times 1 \times 1$. Επίσης όλοι οι πυρήνες συσσώρευσης έχουν διαστάσεις $2 \times 2 \times 2$ και stride $2 \times 2 \times 2$ πλην του πρώτου που έχει διαστάσεις $1 \times 2 \times 2$ και stride $1 \times 2 \times 2$. Το C3D λαμβάνει ένα κλιπ μήκους 16 καρέ για την εξαγωγή των χαρακτηριστικών. Αναλυτικά η αρχιτεκτονική του δικτύου φαίνεται στην εικόνα 2.14. Το C3D ξεκινά εστιάζοντας στα πρώτα frames την προσοχή στην εμφάνιση ενώ στα υπό-



Σχήμα 2.14: Η αρχιτεκτονική του C3D δικτύου. Αποτελείται από τρία συνελικτικά επίπεδα, πέντε επίπεδα συσσώρευσης, και δύο πλήρως συνδεδεμένα επίπεδα [63]

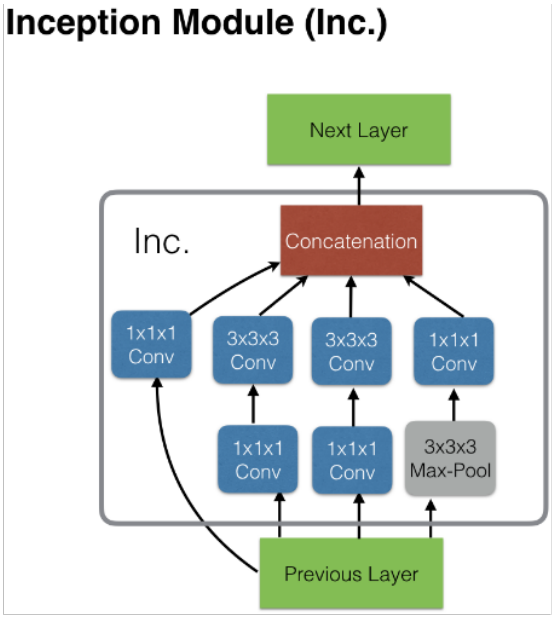
λοιπα παρακολουθεί την κίνηση. Μέρος της παραπάνω εργασίας είναι και η εκπαίδευση του μοντέλου στο Sports-1M σύνολο δεδομένων και η παροχή των βαρών εκπαίδευσης, γεγονός που έχει οδηγήσει πολλούς ερευνητές να χρησιμοποιούν το προ-εκπαιδευμένο C3D μοντέλο.

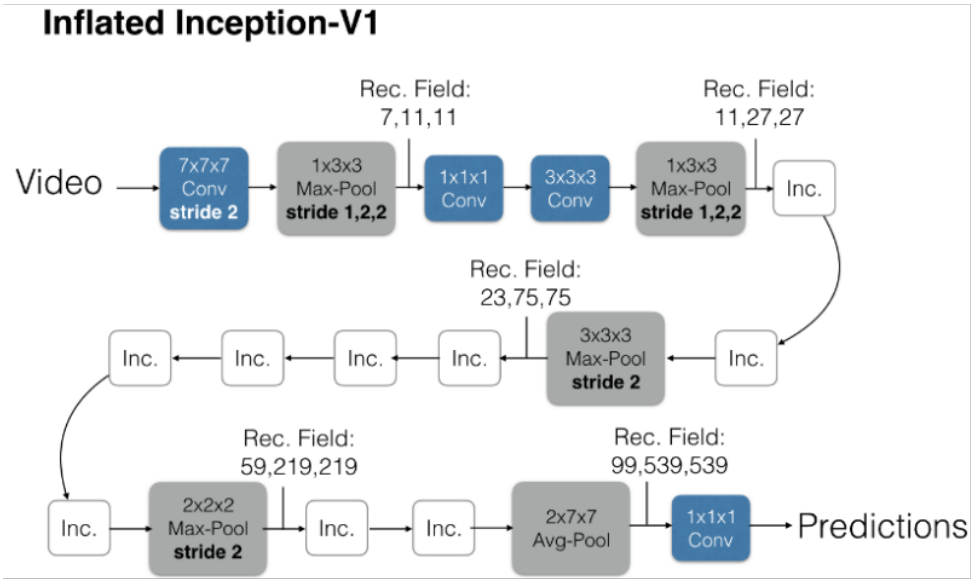
2.4.2 Two-Stream Inflated 3D ConvNets - I3D

Βασιζόμενοι στην έρευνα που έχει γίνει πάνω στην αναγνώριση ανθρώπινων δράσεων σε βίντεο οι Carreira et al. [5] εισήγαγαν ένα νέο μοντέλο που έχει την ικανότητα να επωφεληθεί από την προ-εκπαίδευση στο σύνολο δεδομένων Kinetics και να επιτύχει υψηλή απόδοση. Το μοντέλο αυτό ονομάστηκε Two-Stream Inflated 3D ConvNets (I3D). Χτίστηκε πάνω σε state-of-the-art αρχιτεκτονικές ταξινόμησης εικόνων, με την διαφοροποίηση ότι διογκώνει (inflate) τα φίλτρα και τους πυρήνες συσσώρευσης μετατρέποντας τα σε τρισδιάστατα. Το μοντέλο I3D είναι βασισμένο στο Inception-V1 και επιτυγχάνει απόδοση που ξεπερνά κατά πολύ το state-of-the-art έχοντας εκπαιδευτεί στο Kinetics dataset. Παράλληλα χρησιμοποιείται η τεχνική των δύο ροών καθώς παρότι τα 3D ConvNets μπορούν να μάθουν χρονική πληροφορία μόνο από μια ροή RGB, η απόδοση τους μπορεί να βελτιωθεί σημαντικά προσθέτοντας και μια ροή οπτικής ροής.

Η αρχιτεκτονική του δικτύου προέκυψε μετατρέποντας το επιτυχημένο μοντέλο για ταξινόμηση δισδιάστατης εικόνας σε τρισδιάστατο συνελικτικό δίκτυο. Αυτό επιτεύχθηκε διογκώνοντας όλα τα φίλτρα και τους πυρήνες συσσώρευσης προσθέτοντάς τους μια πρόσθετη

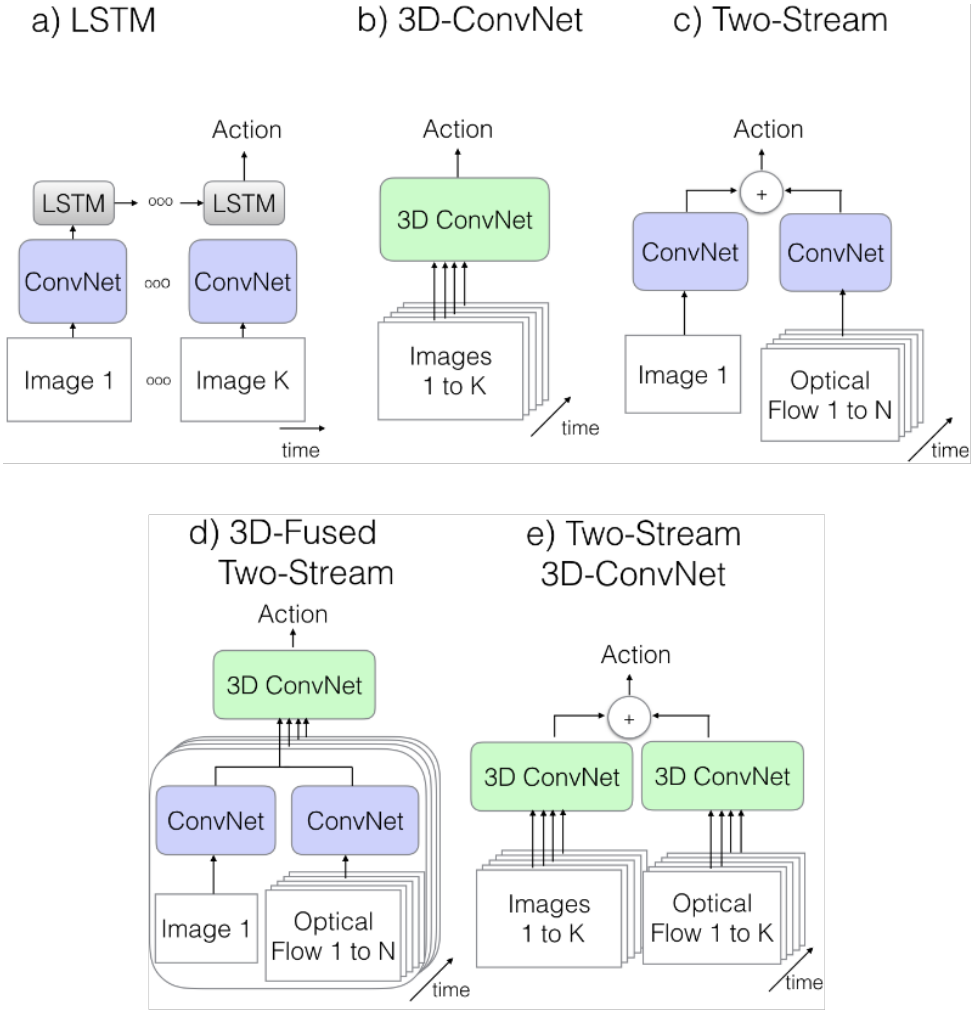
χρονική διάσταση. Τα φίλτρα είναι συνήθως τετράγωνα $N \times N$ και μετατράπηκαν σε κυβικά $N \times N \times N$. Η αρχιτεκτονική του Inflated Inception-V1 φαίνεται στην επόμενη εικόνα:





Σχήμα 2.15: Στην πάνω εικόνα δίνεται η λεπτομερής δομή της κάθε Inception ενότητας. Στην κάτω εικόνα φαίνεται η συνολική αρχιτεκτονική του Inflated Inception-V1 δικτύου. [5]

Ένα δίκτυο I3D εκπαιδεύεται σε εισόδους RGB και ένα άλλο σε εισόδους ροής, που φέρουν βελτιστοποιημένες ομαλές πληροφορίες ροής. Η εκπαίδευση του κάθε δικτύου είναι αυτοτελής και η τελική πρόβλεψη είναι ο μέσος όρος των προβλέψεων των δύο δικτύων. Η δομή του Two-Stream Inflated 3D ConvNet καθώς και πιο σημαντικών μοντέλων που χρησιμοποιούνται στην αναγνώριση ανθρώπινων δράσεων φαίνονται στην επόμενη εικόνα:

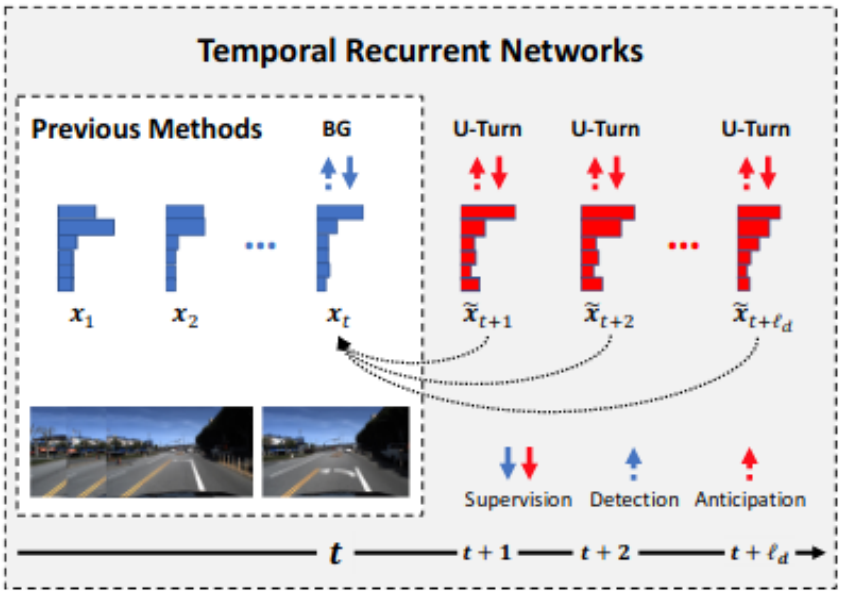


Σχήμα 2.16: Σύνοψη των δικτύων που χρησιμοποιούνται στην αναγνώριση δράσεων και χειρονομιών: α) ConvNet+LSTM [30]: χρήση τόσο συνελικτικών δικτύων όσο και του αναδρομικού δικτύου LSTM προκειμένου να κωδικοποιηθεί η απαραίτητη χωροχρονική πληροφορία, β) 3D-CNN [27]: αποτελείται από τρία τρισδιάστατα συνελικτικά επίπεδα, δύο τρισδιάστατα επίπεδα συσσώρευσης και ένα πλήρως συνδεδεμένο επίπεδο, γ) Two-Stream [58]: χρησιμοποιεί τον μέσο όρο των προβλέψεων ενός RGB καρέ με δέκα καρέ οπτικής ροής, δ) 3D-CNN Διπλής Ροής [33]: αποτελεί επέκταση του β όπου γίνεται fusion στο τελευταίο συνελικτικό επίπεδο ώστε να δημιουργηθεί χωροχρονική ροή, ε) I3D [5]: Το δίκτυο που περιγράφηκε αναλυτικότερα στο σχήμα 2.15.[5]

Κεφάλαιο 3

Χρονικός Εντοπισμός Δράσεων στην Βιβλιογραφία

Στο κεφάλαιο 2 περιγράψαμε τα κυριότερα μοντέλα που χρησιμοποιούνται για την αναγνώριση δράσεων σε βίντεο. Παρόλα αυτά οι περισσότερες από αυτές τις μεθόδους είναι εστι-ασμένες σε τριμαρισμένα βίντεο και δεν μπορούν να εφαρμοστούν σε μεγάλα βίντεο, που περιέχουν ακολουθίες από δράσεις και ποικίλα background. Το πρόβλημα του χρονικού εντοπισμού μιας δράσης ορίζεται ως το πρόβλημα του προσδιορισμού της χρονικής στιγμής έναρξης και της χρονικής στιγμής λήξης και διακρίνεται σε δύο υποπροβλήματα τον offline και τον online εντοπισμό. Η περίπτωση του offline εντοπισμού είναι σαφώς πιο εύκολη καθώς διαθέτουμε εξ' αρχής ολόκληρη την πληροφορία της δράσης σε αντίθεση με την online περίπτωση που διαθέτουμε μόνο την πληροφορία από τα προηγούμενα και το τρέχον καρέ. Επίσης στις offline εφαρμογές ο υπολογιστικός χρόνος δεν μας επηρεάζει, σε αντίθεση με τις online εφαρμογές, όπου πρέπει να αναζητήσουμε μια ισορροπία μεταξύ απόδοσης και υπολογιστικού χρόνου.

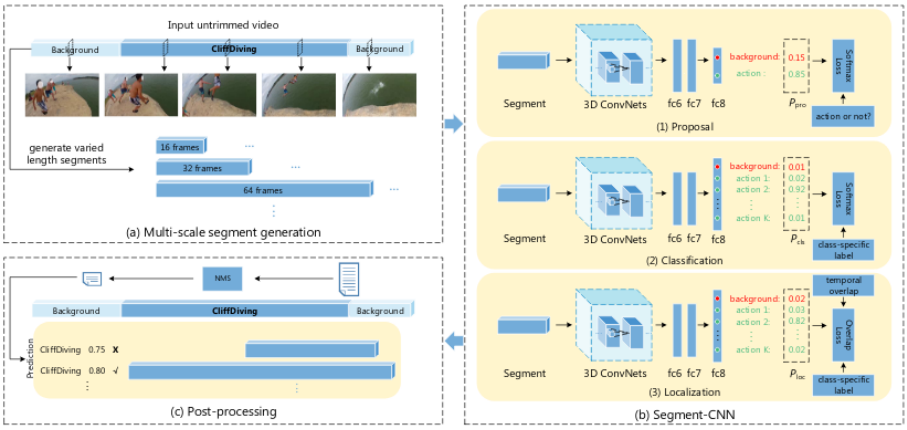


Σχήμα 3.1: Σύγκριση μεθόδων που χρησιμοποιούν μόνο την παρελθοντική και την τρέχουσα πληροφορία με την προτεινόμενη μέθοδο στο [78] που προσβλέπει και στην συνέχεια χρησιμοποιεί την μελλοντική πληροφορία.

3.1 Offline Χρονικός Εντοπισμός Δράσεων

Οι πρώιμες προσεγγίσεις αντιμετωπίζουν το task του χρονικού εντοπισμού δράσεων, εφαρμόζοντας χρονικά κυλιόμενα παράθυρα, που ακολουθούνται από ταξινομητές SVM για την κατηγοριοποίηση μιας δράσης μέσα σε κάθε παράθυρο [46], [29].

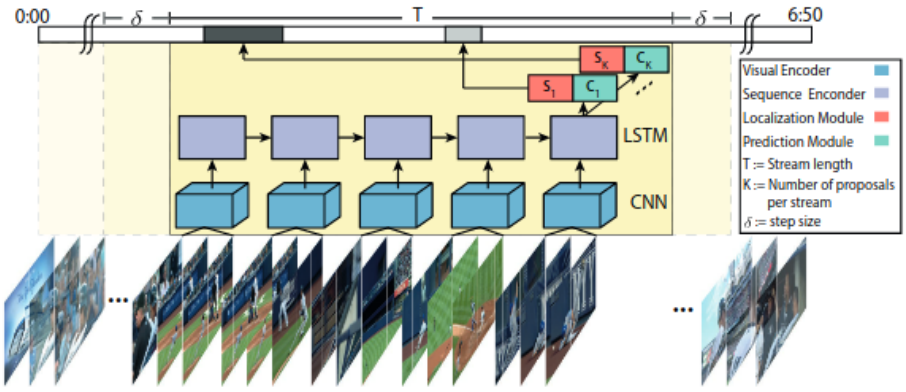
Στη συνέχεια προτάθηκε το S-CNN [57] που είναι ένα μοντέλο εντοπισμού δράσεων δύο σταδίων (σχήμα 3.2). Με το πρώτο να δημιουργεί προτάσεις και το δεύτερο να τις ταξινομεί. Συγκεκριμένα στην πρώτη φάση δημιουργεί αποσπάσματα ποικίλων μεγεθών μέσω κυλιόμενου παραθύρου και στη δεύτερη τροφοδοτεί αυτά τα τμήματα σε C3D δίκτυα, προκειμένου να ταξινομηθούν.



Σχήμα 3.2: Δομή του S-CNN δικτύου. Το δίκτυο αυτό αποτελείται από το στάδιο δημιουργίας προτάσεων (α) και το στάδιο ταξινόμησης των προτάσεων αυτών (β). [57]

Παρόμοιας λογικής είναι το δίκτυο TURN-TAP [19], που αποσυνθέτει το βίντεο σε μικρές ενότητες, προκειμένου να κάνει κάποιες προτάσεις, τις οποίες στη συνέχεια θα βελτιώσει μέσω χρονικής παλινδρόμησης των συντεταγμένων.

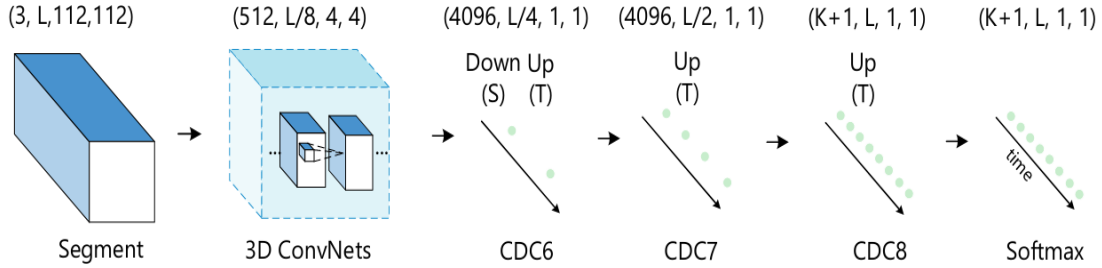
Στόχος είναι όσο δυνατόν να μειωθούν τα παράθυρα που επιλέγονται. Έτσι προτάθηκε το DAP [16] μοντέλο, που δε βασίζεται πλέον στα κυλιόμενα παράθυρα, αλλά χρησιμοποιεί βαθιά νευρωνικά δίκτυα. Συγκεκριμένα αποτελείται από έναν οπτικό κωδικοποιητή, ο οποίος δημιουργεί σημαντικό διάνυσμα μικρής διάστασης, μέσω της χρήσης του ανώτερου επιπέδου του C3D μοντέλου, και έναν κωδικοποιητή ακολουθίας, ο οποίος μοντελοποιεί την χρονική πληροφορία μέσω της χρήσης ενός LSTM δικτύου (σχήμα 3.3).



Σχήμα 3.3: Δομή του DAP δικτύου, που αποτελείται από έναν C3D κωδικοποιητή και έναν LSTM αποκωδικοποιητή. [16]

Ακολούθως, προκειμένου να βελτιωθεί η αποτελεσματικότητά, προτάθηκε το R-C3D [77], στο οποίο τα συνελικτικά χαρακτηριστικά μοιράζονται τόσο στην παραγωγή των προτάσεων όσο και στην ταξινόμηση.

Τέλος οι Shou et al. πρότειναν το CDC φίλτρο [55], - προκειμένου να επιτυγχάνεται χωρική υποδειγματοληψία (για σημασιολογική περίληψη) και χρονική υπερδειγματοληψία (για χρονικό εντοπισμό) - και σχεδίασαν ένα CDC δίκτυο για να κάνει πρόβλεψη δράσης σε επίπεδο καρέ. Η δομή του δικτύου φαίνεται στην επόμενη εικόνα:



Σχήμα 3.4: Δομή CDC. Σε κάθε επίπεδο φαίνονται οι διαστάσεις των δεδομένων με την εξής μορφή: (αριθμός καναλιών, χρονικό μήκος, ύψος, πλάτος) [55]

Ωστόσο, όλες οι προηγούμενες δουλειές εικάζουν ότι όλα τα καρέ μπορούν να παρατηρηθούν εξ' αρχής. Κάτι που δεν ισχύει στην περίπτωση της online αναγνώρισης.

3.2 Πρώιμη Ανίχνευση Δράσεων

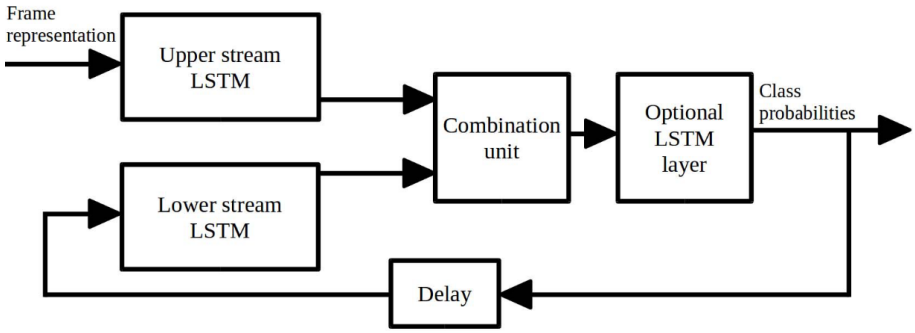
Η πρώιμη ανίχνευση δράσεων είναι η προσπάθεια αναγνώρισης δράσεων, όσο το δυνατόν νωρίτερα, πριν αυτές ολοκληρωθούν. Εμφανίζει πολλές πιθανές εφαρμογές που εκτείνονται από την ασφάλεια (π.χ. πρόβλεψη πανδημικής έκρηξης), τις περιβαλλοντικές επιστήμες (π.χ. προειδοποίηση για τσουνάμι) έως την υγεία (π.χ. ανίχνευση κινδύνου πτώσης) και την ρομποτική (π.χ. ανίχνευση συναισθήματος). Μια από τις μεθόδους που έχουν αναπτυχθεί σε αυτόν τον τομέα έρευνας είναι η χρήση LSTM και τροποποίηση του σφάλματος εκπαίδευσης με βάση την εικασία, ότι η αναλογία μεταξύ σωστών και λανθασμένων κλάσεων δεν πρέπει να είναι φθίνουσα, καθώς ο αριθμός των παρατηρήσεων αυξάνεται [43]. Η πρώιμη ανίχνευση δράσεων και ο online χρονικός εντοπισμός δράσεων μπορούν να θεωρηθούν αρκετά συσχετιζόμενοι τομείς έρευνας.

3.3 Online Χρονικός Εντοπισμός Δράσεων

Δεδομένης μιας ζωντανής ροής βίντεο, online χρονικός εντοπισμός δράσεων ορίζεται ως προσπάθεια εντοπισμού των δράσεων, που εκτελούνται σε κάθε καρέ, μόλις αυτό φτάσει, χωρίς να λαμβάνεται υπόψη το μελλοντικό context. Ένα παράδειγμα εφαρμογής του online χρονικού εντοπισμού δράσεων, θα μπορούσε να είναι ένα assistive ρομπότ, που βοηθά άτομα με κινητικά προβλήματα ή αθλητές υπό θεραπεία, να εκτελέσουν σωστά τις ασκήσεις για την φυσικοθεραπεία τους, αναγνωρίζοντας αρχικά ποια άσκηση κάνουν την κάθε στιγμή και εν συνεχεία δίνοντας τους ένα σκορ απόκλισης. Ιδιαίτερα σημαντική είναι και η συμβολή τους στη χρήση αυτοκινούμενων οχημάτων, με χαρακτηριστικό παράδειγμα την αναγκαστική και άμεση διακοπή πορείας, αν εντοπιστεί ένα παιδί, που κυνηγά μια μπάλα. Η Online αναγνώριση δράσεων αποτελεί ένα πολύ απαιτητικό πρόβλημα για τρεις βασικούς λόγους. Κατ' αρχήν γνωρίζουμε μόνο ένα μέρος της πληροφορίας (η μελλοντική λείπει). Επίσης σημαντικός

παράγοντας είναι και ο υπολογιστικός χρόνος του συστήματος. Τέλος αναφερόμαστε σε εφαρμογές του πραγματικού κόσμου, οπότε, όπως είναι αντιληπτό, οι αποκλίσεις μεταξύ ίδιων δράσεων είναι ποικίλες.

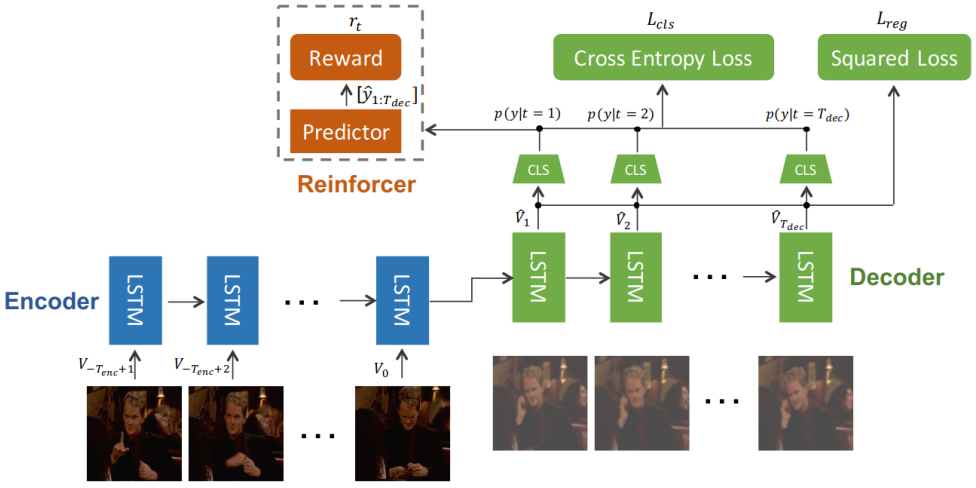
Το εξεταζόμενο πρόβλημα, προσδιορίστηκε για πρώτη φορά από τους De Geest et al. [11], οι οποίοι του έδωσαν ένα συγκεκριμένο ορισμό, ενώ παράλληλα δημιούργησαν και ένα ρεαλιστικό σύνολο δεδομένων (TVSeries) για το πρόβλημα αυτό. Σε επόμενη δουλεία η ίδια ερευνητική ομάδα εισήγαγε ένα διπλής ροής, ανατροφοδοτούμενο δίκτυο (2S-FN) [10]. Η πρώτη ροή εστιάζει ερμηνεία των εισαχθέντων χαρακτηριστικών, ενώ η δεύτερη μοντελοποιεί τις χρονικές εξαρτήσεις μεταξύ των δράσεων. Συγκεκριμένα, η πρώτη ροή λαμβάνει ως



Σχήμα 3.5: Δομή 2S-FN, το οποίο αποτελείται από δύο ροές. Μια για την ερμηνεία των χαρακτηριστικών και μια για την μοντελοποιήση των χρονικών εξαρτήσεων. [10]

είσοδο μια αναπαράσταση του καρέ, που, είτε έχει εξαχθεί από κάποιο συνελικτικό δίκτυο είτε είναι χαρακτηριστικά πόζας και μέσω ενός LSTM, μαθαίνει να ερμηνεύει την είσοδο αυτή. Η επόμενη μονάδα του δικτύου είναι ο συνδετικός κρίκος των δύο ροών που μπορεί είτε να ακολουθείται από ένα ακόμη LSTM είτε να γεννά τις πιθανότητες για κάθε κλάση άμεσα και εν συνεχεία οι πιθανότητες αυτές τροφοδοτούν την δεύτερη ροή. Ως εκ τούτου, αυτή η δεύτερη ροή δεν έχει ποτέ πληροφορίες σχετικά με το πλαίσιο που εξετάζεται επί του παρόντος και πρέπει να λάβει την απόφασή της με βάση ένα καλό χρονικό μοντέλο.

Οι Gao et al. πρότειναν ένα νέο μοντέλο (RED) [18] που χρησιμοποιεί ενισχυτικό κόστος (reinforcement loss), προκειμένου να ενθαρρύνει την, όσο το συντομότερο δυνατό αναγνώριση δράσης. Το μοντέλο αυτό αποτελείται από τρεις ενότητες: α) έναν εξαγωγέα



Σχήμα 3.6: Το δίκτυο RED αποτελείται από έναν κωδικοποιητή και έναν αποκωδικοποιητή και βασίζεται στις βασικές αρχές της ενισχυτικής μάθησης. [18]

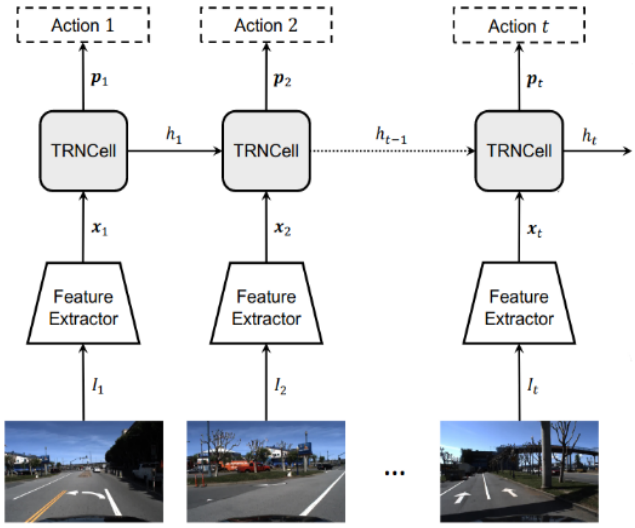
αναπαραστάσεων (είτε two-stream CNN [76], είτε VGG-16 [59]), β) ένα κωδικοποιητή - αποκωδικοποιητή προκειμένου να κωδικοποιεί την παρελθοντική πληροφορία και να προβλέπει τις μελλοντικές αναπαραστάσεις (LSTM) και γ) μια ενισχυτική ενότητα, η οποία θα υπολογίζει τις ανταμοιβές, που ενσωματώνονται στη φάση εκπαίδευσης, χρησιμοποιώντας έναν policy gradient αλγόριθμο. Παρότι το μοντέλο αυτό προτάθηκε για anticipation δράσεων - πρόβλεψη δράσεων κάποια δευτερόλεπτα στο μέλλον - μπορεί να χρησιμοποιηθεί και για online εντοπισμό δράσεων, αν μηδενίσουμε τον χρόνο του anticipation. Η δομή του παραπάνω μοντέλου φαίνεται στην εικόνα 3.6.

Τέλος έχουν ακόμη προταθεί Generative Adversarial Networks - GANs [56] και προσαρμοστική δειγματοληψία για τον προσδιορισμό της χρονικής στιγμής έναρξης μιας δράσης. Με αυτόν τον τρόπο διακρίνονται τα διφορούμενα υπόβαθρα και η σαφής χρονική μοντελοποίηση μεταξύ δράσεων για χρονική συνέπεια.

Όλες οι προηγούμενες μέθοδοι βασίζονται στις τρέχουσες και τις παρελθοντικές παρατηρήσεις. Στην επόμενη ενότητα παρουσιάζουμε μια μέθοδο (TRN) [78], που κάνει online αναγνώριση δράσεων στο άμεσο μέλλον. Συγκεκριμένα δε βασίζεται μόνο στην παρελθοντική και την τρέχουσα πληροφορία, αλλά προβλέπει την μελλοντική πληροφορία και χρησιμοποιεί το σύνολο αυτών, ώστε να βελτιώσει την απόδοση της αναγνώρισης δράσεων στο παρόν. Τα κυριότερα μοντέλα, με τα οποία συγκρίνεται το TRN μοντέλο στο [78], έχουν περιγραφεί αναλυτικά και έχουν αποτυπωθεί στην ενότητα αυτή για λόγους κατανόησης και πληρότητας. Το TRN μοντέλο περιγράφεται αναλυτικά στην επόμενη, αυτόνομη ενότητα καθώς αποτέλεσε και την βάση της υποβληθείσας προς δημοσίευση εργασίας μας.

3.4 Temporal Recurrent Networks for Online Action Detection - TRNs

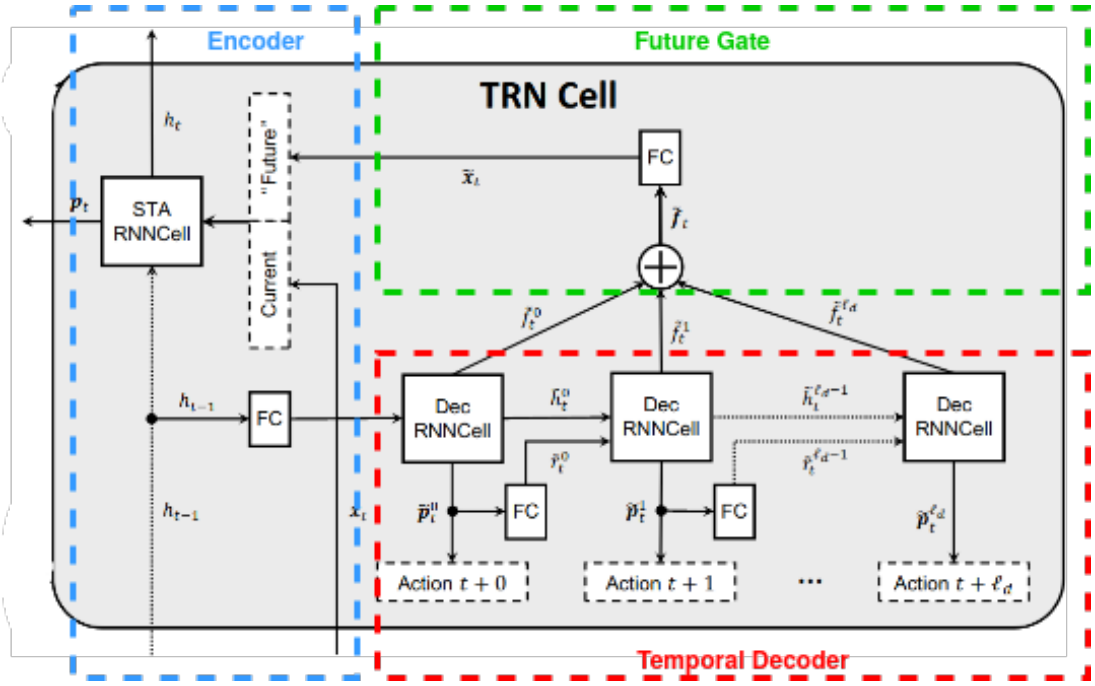
Το χρονικό αναδρομικό νευρωνικό δίκτυο - Temporal Recurrent Network (TRN) είναι ένα καινοτόμο πλαίσιο, που προτάθηκε για την online αναγνώριση δράσεων. Σε αντίθεση με τις προηγούμενες εργασίες, το δίκτυο αυτό προβλέπει δράσεις πολλών μελλοντικών καρέ, και στη συνέχεια χρησιμοποιεί την πληροφορία αυτή για να κατηγοριοποιήσει μια δράση στο παρόν. Ο πυρήνας του δικτύου αυτού, κύτταρο TRN, είναι μια ισχυρή αναδρομική μονάδα που λειτουργεί όπως οποιοδήποτε RNN κύτταρο.



Σχήμα 3.7: Απεικονίζεται η δομή του TRN δικτύου που είναι παρόμοια με ένα RNN. [78]

Συγκεκριμένα κάθε χρονική στιγμή t ένα TRN κύτταρο δέχεται ως είσοδο ένα διάνυσμα χωροχρονικών χαρακτηριστικών x_t και την προηγούμενη κρυφή κατάσταση h_{t-1}, δίνει ως έξοδο μια κατανομή πιθανοτήτων p_t που εκτιμά ποια δράση εκτελείται στο καρέ I_t και τελικά ενημερώνει την τιμή της κρυφής κατάστασης h_t για το επόμενο βήμα. Η καινοτομία του δικτύου αυτού σε σχέση με οποιοδήποτε άλλο RNN είναι ότι χρησιμοποιεί τις χρονικές συσχετίσεις μεταξύ του παρόντος και του μέλλοντος. Συγκεκριμένα προβλέπει (anticipate) τις επερχόμενες δράσεις και χρησιμοποιεί εκτενώς αυτές τις προσεγγίσεις προκειμένου να ενισχύσει την αναγνώριση της τρέχουσας δράσης.

Όσον αφορά την εσωτερική δομή του TRN κυττάρου αποτελείται από έναν χρονικό αποκωδικοποιητή, μια μελλοντική πύλη και έναν χωροχρονικό κωδικοποιητή. Τόσο ο κωδικοποιητής όσο και ο αποκωδικοποιητής απαρτίζονται από LSTMs, ενώ η μελλοντική πύλη αποτελείται από έναν τελεστή μέσης συσσώρευσης, ακολουθούμενο από ένα πλήρως συνδεδεμένο επίπεδο.



Σχήμα 3.8: Εσωτερική Αρχιτεκτονική TRN με επισημειωμένα τα τρία βασικά τμήματα του. Το πρώτο στάδιο του δικτύου, ο χρονικός αποκωδικοποιητής είναι σημειωμένος με κόκκινο χρώμα, ακολουθεί η μελλοντική πύλη σχηματιζόμενη με πράσινο χρώμα. Τέλος ο κωδικοποιητής που εξάγει τις προβλέψεις έχει επισημανθεί με μπλε χρώμα. [78]

Ο χρονικός αποκωδικοποιητής μαθαίνει τις αναπαραστάσεις των χαρακτηριστικών, η μελλοντική πύλη λαμβάνει ένα διάνυσμα από κρυφές καταστάσεις και προωθεί αυτά τα χαρακτηριστικά ως μελλοντική πληροφορία και ο χωροχρονικός κωδικοποιητής συλλαμβάνει τα χωροχρονικά χαρακτηριστικά από την παρελθοντική, την τρέχουσα και την προβλεφθείσα μελλοντική πληροφορία και εκτιμά τη δράση που συμβαίνει στο τρέχον καρέ.

Το μοντέλο εκπαιδεύεται σε offline, συνθήκες αλλά κατά την φάση του testing το μοντέλο χρησιμοποιεί την προβλεφθείσα μελλοντική πληροφορία χωρίς να έχει πρόσβαση στα επόμενα καρέ. Γεγονός που το καθιστά online μοντέλο. Για την εκπαίδευσή του το μοντέλο χρησιμοποιεί χωροχρονικά δεδομένα, τα οποία έχουν εξαχθεί είτε από ένα συνελικτικό δίκτυο διπλής ροής είτε από ένα VGG-16 δίκτυο, όπως συμβαίνει και στο RED δίκτυο, που περιγράψαμε στην προηγούμενη ενότητα. Όσον αφορά το συνελικτικό δίκτυο διπλής ροής

τα appearance χαρακτηριστικά εξάγονται από ένα ResNet-200 [22], ενώ τα motion χαρακτηριστικά εξάγονται από ένα BN-Inception [26] δίκτυο, το οποίο έχει τροφοδοτηθεί από καρέ οπτικής ροής.

Σύμφωνα με τους δημιουργούς αυτής της δουλειάς, το μοντέλο αυτό μπορεί να χρησιμοποιηθεί και σε άλλες εργασίες, όπως ο εντοπισμός αντικειμένου. Στη παρούσα διπλωματική εργασία χρησιμοποιήθηκε το μοντέλο αυτό και συνδυάστηκε με διάφορα βαθιά νευρωνικά δίκτυα για την εξαγωγή των χαρακτηριστικών που το τροφοδοτούν. Κύριος στόχος είναι να εξεταστεί η συνεισφορά του χρονικού context και της δυναμικής των σκελετών στην online αναγνώριση δράσεων αλλά και να βελτιωθεί εν τέλει η απόδοση του μοντέλου. Αναλυτική περιγραφή των πειραμάτων που έγιναν κατά την μελέτη μας πάνω στο μοντέλο αυτό αλλά και σχολιασμός επί των αποτελεσμάτων περιέχονται στο κεφάλαιο 6 της παρούσας διπλωματικής.

Κεφάλαιο 4

Πειράματα, Αποτελέσματα και Συγκρίσεις

Στα πλαίσια της εργασίας αυτής υλοποιήθηκαν κάποια μοντέλα τόσο για την αναγνώριση χειρονομιών όσο και για τον χρονικό προσδιορισμό και την αναγνώριση δράσης σε πραγματικό χρόνο. Στο κεφάλαιο αυτό περιγράφονται τα σύνολα δεδομένων που χρησιμοποιήθηκαν και για τα δυο προβλήματα, καθώς και δυο frameworks που χρησιμοποιήθηκαν για την εξαγωγή δισδιάστατων σκελετών και οπτικής ροής.

4.1 IsoGD Chalearn

Για την αναγνώριση ανθρώπινων δράσεων χρησιμοποιήθηκε το Isolated Gesture Recognition Dataset (IsoGD) [67], κυρίως λόγω του πλήθους των διαφορετικών χειρονομιών που περιέχει. Συγκεκριμένα το εν λόγω σύνολο δεδομένων αποτελείται από 47933 clips τα οποία περιλαμβάνουν μία χειρονομία έκαστο. Οι περίπου 48 χιλιάδες χειρονομίες χωρίζονται σε 249 κλάσεις και εκτελούνται από 21 διαφορετικούς ανθρώπους, εκ των οποίων οι 17 εμφανίζονται στο σύνολο εκπαίδευσης και από 2 εμφανίζονται στα σύνολα εκτίμησης και τεστ. Η κατανομή των χειρονομιών και των ανθρώπων που τις εκτελούν ανά σύνολο φαίνονται αναλυτικά στον επόμενο πίνακα:

Sets	**IsoGD dataset**			
	#labels	#gestures	#RGB-D videos	#performers
Train	249	35878	35878	17
Validation	249	5784	5784	2
Test	249	6271	6271	2
All	249	47933	47933	21

Πίνακας 4.1: IsoGD: Κατανομή δεδομένων στα training, validation και test σύνολα

Το συγκεκριμένο σύνολο δεδομένων εμφανίζει αρκετές προκλήσεις που οφείλονται στο γεγονός ότι αρκετοί δράστες εκτελούν πολλές διαφορετικές χειρονομίες σε ποικίλα backgrounds και θέσεις ως προς την κάμερα. Η IsoGD είναι μια RGB+D βάση, περιέχει δηλαδή για κάθε μια από τις 48 περίπου χιλιάδες χειρονομίες δύο βίντεο, ένα με την RGB πληροφορία και να με την πληροφορία βάθους (depth) που έχουν δημιουργηθεί μέσω του αισθητήρα kinect.

α β γ

Σχήμα 4.1: IsoGD: Χειρονομία 33 στα RGB: α) training, β) validation, γ) test sets

Ένας επιπλέον λόγος που μας οδήγησε στην επιλογή αυτού του συνόλου δεδομένων είναι ακριβώς η ύπαρξη αυτού του καναλιού βάθους. Συγκεκριμένα μέσω των αποχρώσεων των pixel δίνεται σε ένα σύστημα η δυνατότητα να κατανοήσει το βάθος, προσδιορίζοντας την απόσταση του εξεταζόμενου αντικειμένου από το επίπεδο της εικόνας. Η χρήση της πληροφορίας βάθους, εκτός του ότι προσθέτει μια ακόμη διάσταση στα δεδομένα που συλλέγουμε από το RGB κανάλι, μειώνει και την αρνητική επίδραση του background, στην αναγνώριση της χειρονομίας, κάτι το οποίο επιτυγχάνεται και με την χρήση του σκελετού, όπως θα δούμε σε επόμενη ενότητα.

α β γ

Σχήμα 4.2: IsoGD: Χειρονομία 33 στα Depth: α) training, β) validation, γ) test sets

4.2 THUMOS '14

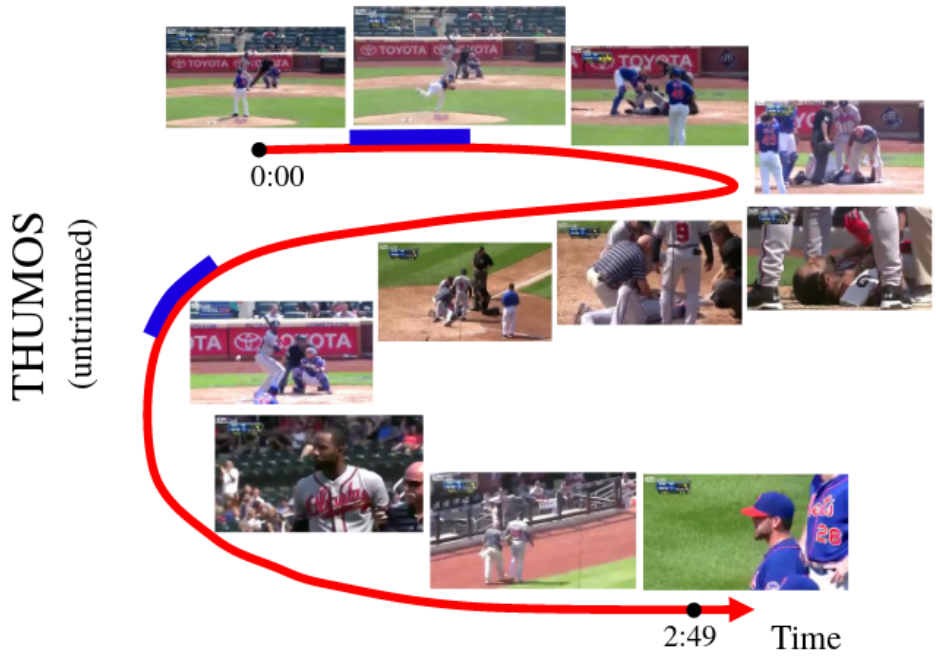
Τα χρονικά κατακερματισμένα κλιπ δεν αντικατοπτρίζουν τον πραγματικό κόσμο όπου οι ενέργειες είναι συνήθως ενσωματωμένες σε σύνθετες δυναμικές σκηνές με πλούσιες αιτιώδεις και χωρικές σχέσεις μεταξύ ανθρώπων και αντικειμένων. Η βάση THUMOS'14 [28], την οποία και εμείς χρησιμοποιούμε στο δεύτερο πειραματικό μέρος αυτής της διπλωματικής, εισήγαγε χιλιάδες μη τριμαρισμένα βίντεο στα σύνολα εκτίμησης και δοκιμής καθώς και στα δεδομένα παρασκηνίου. Παρέχει με αυτόν τον τρόπο στην επιστημονική κοινότητα ένα πρώτης κλάσης σύνολο δεδομένων για τον χρονικό εντοπισμό και την αναγνώριση δράσεων σε ρεαλιστικές συνθήκες.

Οι κατηγορίες δράσης έχουν επιλεγεί από το UCF101 [61] σύνολο δεδομένων και χωρίζονται σε πέντε βασικές κατηγορίες: α) αλληλεπίδραση ανθρώπου-αντικειμένου, β) αποκλειστική κίνηση σώματος, γ) αλληλεπίδραση ανθρώπων, δ)παίξιμο μουσικών οργάνων και ε) αθλήματα. Όλα τα βίντεο είναι δημοσίως διαθέσιμα στο YouTube και έχουν επισημειωθεί στο χέρι τόσο για την αναγνώριση δράσεων όσο και για τον χρονικό εντοπισμό αυτών. Στον επόμενο πίνακα 4.2 φαίνεται το πλήθος των βίντεο, οι περιεχόμενες κλάσεις και το είδος των βίντεο σε κάθε σύνολο. Το εξεταζόμενο σύνολο δεδομένων περιέχει

set	THUMOS'14				
	#videos	#classes	#temp_video	#temp_classes	type of video
Train	13320	101	-	-	trimmed
Background	2500	-	-	-	-
Validation	1010	101	200	22	untrimmed
Test	1574	101	213	22	untrimmed

Πίνακας 4.2: THUMOS'14: Κατανομή δεδομένων στα training, validation, background και test σύνολα για τα task του action recognition και temporal segmentation

περισσότερες από 254 ώρες βίντεο και περισσότερα από 25 εκατομμύρια καρέ, παρόλα αυτά για την εργασία του χρονικού εντοπισμού δράσεων χρησιμοποιούνται 200 βίντεο από το σύνολο εκτίμησης για την εκπαίδευση των μοντέλων και 213 βίντεο το σύνολο δοκιμής για την εκτίμηση των μοντέλων. Τα βίντεο αυτά περιέχουν 22 συνολικά κλάσεις, 20 εκ



Σχήμα 4.3: Μη τριμαρισμένο βίντεο από την βάση THUMOS'14, όπου με μπλε χρώμα σημειώνονται τα χρονικά διαστήματα της δράσης BaseballPitch ενώ τα υπόλοιπα είναι background. Παράλληλα μέσω αυτού του σχήματος οπτικοποιείται ο τρόπος επισημείωσης των δεδομένων. [25]

των προαναφερθέντων 101 κλάσεων του συνολικού dataset, μια διφορούμενη και μια υποβάθρου. Διφορούμενη θεωρείται μια δράση σε περιπτώσεις μερικής ορατότητας, ελλιπούς εκτέλεσης ή ισχυρής απόκλισης στο στυλ. Τα βίντεο υποβάθρου μοιράζονται σκηνές και

αντικείμενα παρόμοια με αυτά των υπολοίπων δράσεων, αλλά δεν περιέχουν καμία από αυτές. Τέτοια είδους background μπορούν να υποβαθμίσουν τον ρόλο των appearance χαρακτηριστικών και γενικά των στατικών πληροφοριών. Για λόγους πληρότητας αναφέρουμε τις 22 κλάσεις δράσης αλφαβητικά: Ambiguous, Background, BaseballPitch, BasketballDunk, Billiards, CleanAndJerk, CliffDiving, CricketBowling, CricketShot, Diving, FrisbeeCatch, GolfSwing, HammerThrow, HighJump, JavelinThrow, LongJump, PoleVault, Shotput, SoccerPenalty, TennisSwing, ThrowDiscus, VolleyballSpiking.

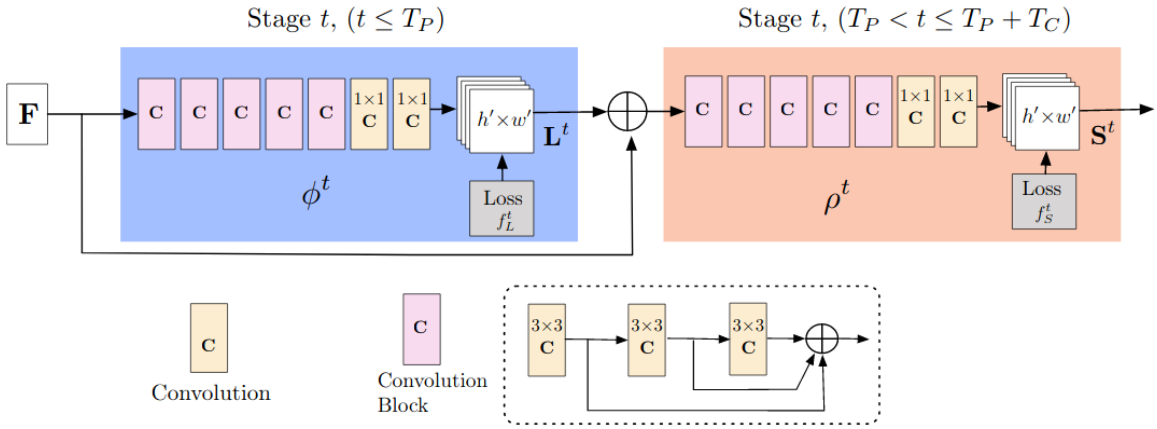
4.3 OpenPose

Το OpenPose [4] είναι ένα έργο που υλοποιήθηκε από τους Cao et al και επιτυγχάνει σε πραγματικό χρόνο να εκτιμά αποτελεσματικά τις δισδιάστατες θέσεις διαφόρων σημείων του ανθρώπινου σώματός προσδιορίζοντας με αυτόν τον τρόπο την ανθρώπινη πόζα σε εικόνες ακόμη και με περισσότερους από έναν ανθρώπους. Αξιοσημείωτη είναι και η επίδοση τους σε εφαρμογές πραγματικού χρόνου όπου η ταχύτητα ανίχνευσης χαρακτηριστικών τίθεται στα 25fps.

Σχήμα 4.4: Ταυτόχρονη εύρεση πολλαπλών σκελετών. [3]

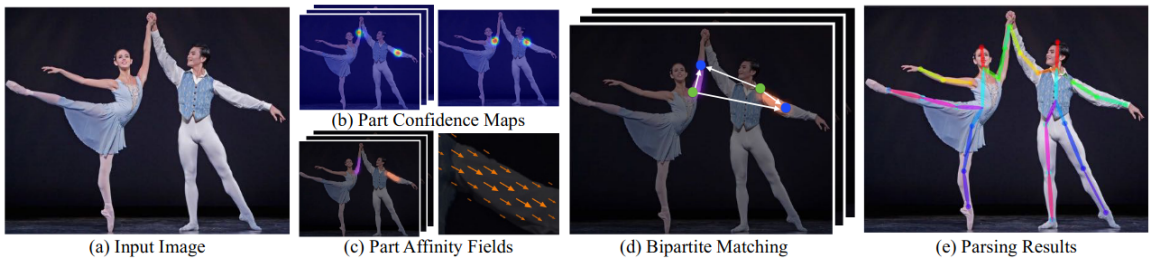
Τα πεδία μερικής συνάφειας (Part Affinity Fields - PAFs) είναι αναπαραστάσεις που αποτελούνται από ένα σύνολο πεδίων ροής, τα οποία κωδικοποιούν μη δομημένες σχέσεις ζεύγους μεταξύ σημείων του σώματος για ποικίλο αριθμό ανθρώπων. Για τον προσδιορισμό των σκελετών χρησιμοποιήθηκε για πρώτη φορά μια από κάτω προς τα πάνω αναπαράσταση των βαθμών συσχέτισης μέσω PAFs. Συγκεκριμένα ορίστηκε ένα σύνολο από δισδιάστατα διανύσματα που κωδικοποιούν τη θέση και τον προσανατολισμό των άκρων του σώματος.

Όσον αφορά την αρχιτεκτονική του μοντέλου αρχικά η εικόνα αναλύεται από ένα CNN που έχει αρχικοποιηθεί με τα πρώτα 10 στρώματα του VGG-19 και προκύπτει ένα σύνολο χαρτών χαρακτηριστικών που αποτελούν την είσοδο του συστήματος. Στη συνέχεια το μοντέλο με την χρήση συνελικτικών δικτύων παράγει ένα σύνολο από PAFs τα οποία σε κάθε στάδιο ανανεώνονται μέσω της σχέσης $L^t = \varphi^t(F, L^{t-1}), \forall 2 \leq t \leq T_p$, όπου L^t είναι το εκάστοτε πεδίο μερικής συνάφειας, φ^t το αντίστοιχο CNN και F τα χαρακτηριστικά της εικόνας. Μετά από T_p επαναλήψεις η διαδικασία επαναλαμβάνεται για τον εντοπισμό των χαρτών εμπιστοσύνης και δίνεται από την αναδρομική σχέση $S^t = \rho^t(F, L^{T^p}, S^{t-1}), \forall T_p < t \leq T_p + T_c$, όπου S^t είναι ο εκάστοτε χάρτης εμπιστοσύνης, ρ^t το αντίστοιχο CNN και T_c ο αριθμός των επαναλήψεων. Κάθε χάρτης εμπιστοσύνης είναι μια δισδιάστατη αναπαράσταση της πεποίθησης ότι κάθε συγκεκριμένο μέρος του σώματος μπορεί να βρίσκεται οπουδήποτε, σε οποιοδήποτε δοθέν pixel.



Σχήμα 4.5: Αρχιτεκτονική του OpenPose. [3]

Αρχικά δημιουργούνται ατομικοί χάρτες εμπιστοσύνης για κάθε εικονιζόμενο πρόσωπο και στη συνέχεια ο χάρτης εμπιστοσύνης του εκάστοτε σημείου προκύπτει από το μέγιστο των χαρτών εμπιστοσύνης στο σημείο αυτό. Έχοντας πλέον εντοπιστεί τα διάφορα μέρει του σώματος πρέπει αυτά να ενωθούν για να σχηματιστεί η ανθρώπινη πόζα. Ο πιο αποτελεσματικός τρόπος να γίνει αυτό είναι η εκ νέου χρήση των PAFs, που πρακτικά αποτελούν τα διανύσματα που ενώνουν δυο μέρη του σώματος, τα οποία μπορεί όμως να ανήκουν σε διαφορετικούς εικονιζόμενους. Η εύρεση όλων των πιθανών ακμών γίνεται μέσω ενός αναλυτή, ενώ στο τέλος το σύστημα εμφανίζει το πιο πιθανό αποτέλεσμα (πρόβλημα ταιριάσματος μέγιστου βάρους σε διμερή γράφο) δίνοντας χαρακτηριστικά σημεία εκφρασμένα ξεχωριστά για κάθε άνθρωπο.



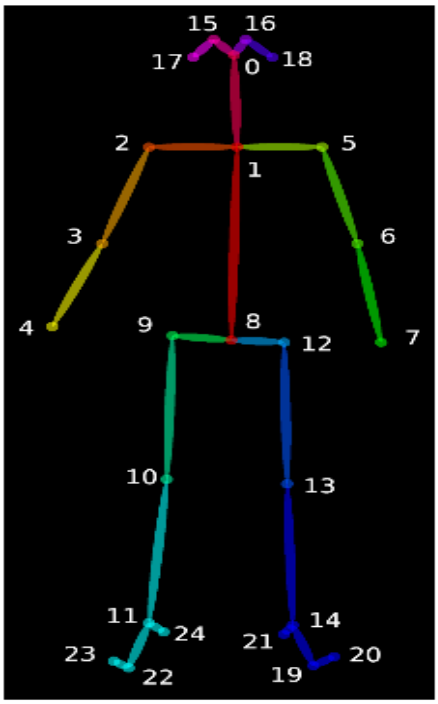
Σχήμα 4.6: OpenPose pipeline. [3]

Ένα ακόμα πλεονέκτημα του OpenPose πέραν της εύρεσης πολλαπλών σκελετών, είναι το γεγονός ότι παραμένει ανεπηρέαστο από το παρασκήνιο γεγονός πολύ προσοδοφόρο για την αναγνώριση χειρονομιών. Παράλληλα λόγω του ότι κάθε σημείο εντοπίζεται χωρικά ανεξάρτητα από τα υπόλοιπα το μοντέλο είναι αποτελεσματικό και σε περιπτώσεις όπου κάποιο μέρος του σώματος δεν είναι ορατό στην εικόνα. Τέλος το OpenPose χρησιμοποιεί και το χρονικό context ώστε να προβλέψει κάποια πληροφορία από προηγούμενα καρέ, αν αυτή λείπει στο εξεταζόμενο καρέ.

Έχοντας πλέον κατανοήσει την αρχιτεκτονική του, εξετάζουμε τα χαρακτηριστικά που μας δίνει ανάλογα με τα ορίσματα που του δίνουμε. Το OpenPose δίνει για το πρόσωπο, τα χέρια και τον κορμό σημεία-κλειδιά (keypoints) τα οποία προγράφονται από τρεις μεταβλητές, την x συνιστώσα, την y συνιστώσα και έναν συντελεστή εμπιστοσύνης c. Αναλυτικά έχουμε:

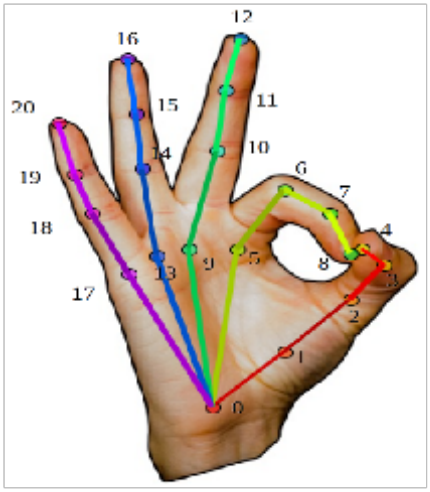
- **Κορμός:** Απαρτίζεται από 25 σημεία. Συγκεκριμένα 2 σημεία για τους καρπούς, 2 σημεία για τους αγκώνες, και 2 σημεία για τους ώμους. Επίσης ο βασικός κορμός ορίζεται από 3 σημεία, το πρόσωπο, το μέσο των ώμων και το μέσο της λεκάνης.

Για τα πόδια διαθέτει 2 σημεία στα ισχία, 2 στα γόνατα και 8 στα πέλματα. Τέλος υπάρχουν 4 σημεία που ορίζουν τα μάτια.



Σχήμα 4.7: 25 σημεία-κλειδιά για τον κορμό [45].

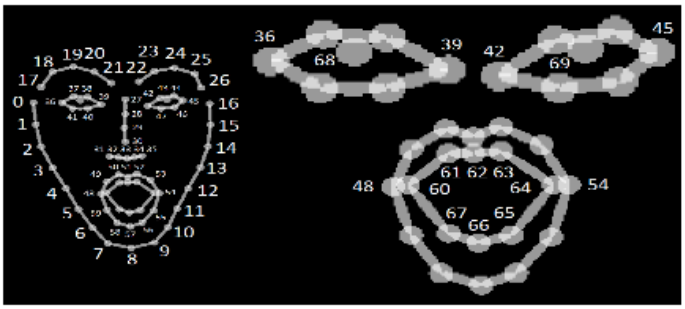
- **Χέρι:** Σχηματίζεται από 21 σημεία-κλειδιά. Συγκεκριμένα έχουμε 4 σημεία για κάθε δάχτυλο, 3 από τα οποία αντιπροσωπεύουν από μια άρθρωση και ένα σε κάθε απόληξη. Το 21 πρώτο σημείο αντιστοιχεί στον καρπό και πρακτικά ταυτίζεται με τα αντίστοιχα του κορμού.



Σχήμα 4.8: 21 σημεία-κλειδιά για το κάθε χέρι [45].

- **Πρόσωπο:** Χρησιμοποιούνται 70 σημεία ώστε να μπορέσουν να απεικονίσουν οποιαδήποτε μυική κίνηση μπορεί να γίνει από τα φρύδια, τα μάτια, τη μύτη, το στόμα και

το περίγραμμα του προσώπου.

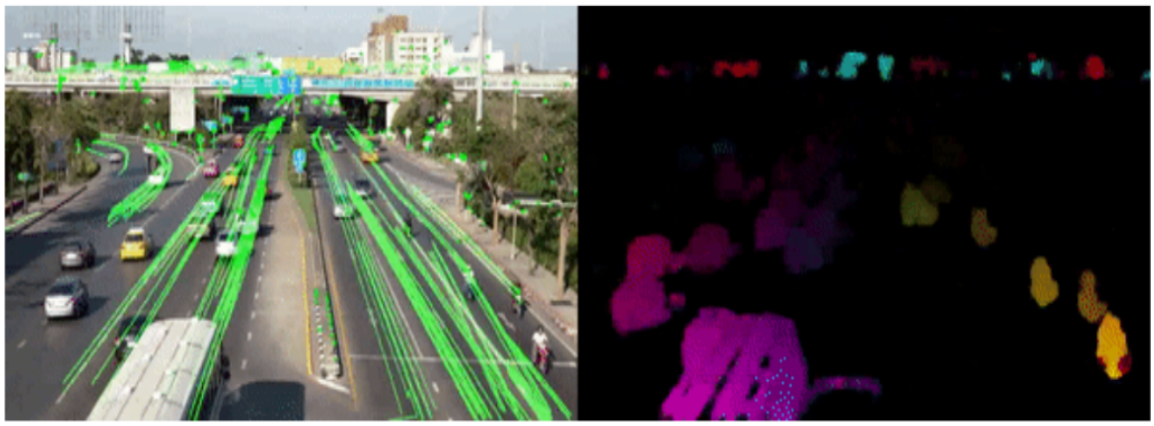


Σχήμα 4.9: 70 σημεία-κλειδιά για το κάθε χέρι [45].

4.4 Οπτική Ροή

Ως οπτική ροή (optical flow) ορίζουμε το μοτίβο της φαινομενικής κίνησης μεταξύ δύο διαδοχικών καρέ, λόγω της απόστασης του εξεταζόμενου σημείου από την κάμερα. Πρόκειται για μια κατανομή μεταβολών της φωτεινότητας (intensity) σε μια εικόνα. Συνεπώς η οπτική ροή μπορεί να διαχειριστεί ως δισδιάστατο διανυσματικό πεδίο όπου κάθε φορέας μετατόπισης αντιστοιχεί στην μετατόπιση των σημείων μεταξύ δύο διαδοχικών καρέ.

Διακρίνεται σε αραιή οπτική ροή (sparse) και σε πυκνή οπτική ροή (dense). Η αραιή οπτική ροή επιλέγει ένα σετ αραιών χαρακτηριστικών όπως ακμές και γωνίες. Έχουν αναπτυχθεί πολλοί αλγόριθμοι για τον υπολογισμό της με έναν από τους κυριότερους αυτόν των Lucas-Kanade. Αντίθετα η πυκνή οπτική ροή προσπαθεί να υπολογίσει ένα διάνυσμα οπ-



α **β**

Σχήμα 4.10: Διαφορές α) Αραιής και β) Πυκνής οπτικής ροής

τικής ροής για κάθε pixel του κάθε καρέ. Για τα πειράματά που ακολουθούν στο επόμενο κεφάλαιο επιλέξαμε την χρήση πυκνής οπτικής ροής λόγω μεγαλύτερης ευρωστίας που προσφέρει. Έχουν αναπτυχθεί αρκετοί αλγόριθμοι για τον υπολογισμό της πυκνής οπτικής ροής, με έναν από τους κυριότερους να είναι ο TV-L1. Η μέθοδος αυτή βασίζεται στην ολική διακύμανση (total variation - TV) και στην κανονικοποίηση και την ευρωστία που προσφέρει η L1 νόρμα. Αυτή η διαμόρφωση μπορεί να εμποδίσει τις ασυνέχειες στο πεδίο της ροής και

να προσφέρει αυξημένη ευρωστία έναντι στις αλλαγές του φωτισμού, τις εμφράξεις και το θόρυβο.

Από πλευράς λογισμικού χρησιμοποιήθηκε το dense-flow εργαλείο που περιέχει OpenCV υλοποιήσεις διαφόρων αλγορίθμων για τον υπολογισμό της οπτικής ροής και έχει δημιουργηθεί τους Wang et al. [73].

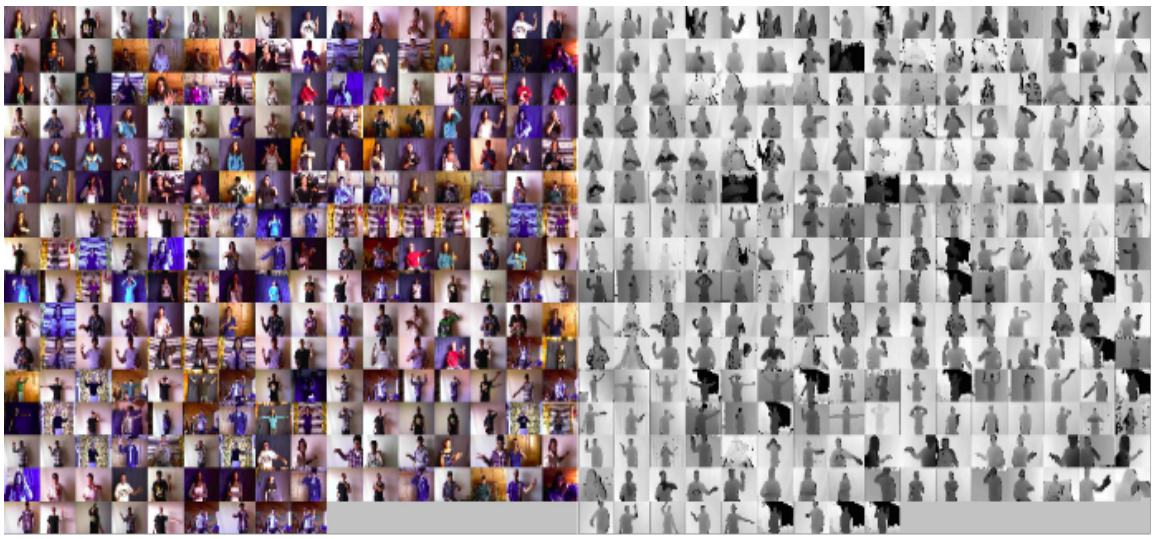
Κεφάλαιο 5

Αναγνώριση Χειρονομιών – Πειράματα, Αποτελέσματα και Συγκρίσεις

Στο κεφάλαιο αυτό παρουσιάζουμε τα πειραματικά μας αποτελέσματα τόσο για το πρόβλημα της αναγνώρισης ανθρώπινων χειρονομιών σε βίντεο και συγκεκριμένα ένα σύνολο από υλοποιήσεις. Για κάθε μία από τις υλοποιήσεις μας αρχικά περιγράφουμε την διαδικασία προ-επεξεργασίας των δεδομένων, τα βασικότερα στοιχεία της καθώς και τα παραγόμενα αποτελέσματα. Μετά το πέρας όλων τον υλοποιήσεων πραγματοποιούμε μια συνολική σύγκριση των αποτελεσμάτων όλων των μεθόδων.

5.1 Παράμετροι και Μετρικές

Για την αναγνώριση της ανθρώπινων χειρονομιών σε βίντεο χρησιμοποιήθηκε το IsoGD σύνολο δεδομένων, σχήμα 5.1, το οποίο περιγράψαμε και στο κεφάλαιο 4. Για την αντιμετώπιση αυτού του αντικειμένου έρευνας υλοποιήσαμε τρία βασικά μοντέλα καθώς και παραλλαγές αυτών (LSTMs, C3Ds, I3Ds).



Σχήμα 5.1: Το σύνολο των χειρονομιών του IsoGD για την καλύτερη εποπτεία των μεθόδων και την εξαγωγή συμπερασμάτων

Ως συνάρτηση σφάλματος (loss function) χρησιμοποιείται η Cross Entropy Loss για όλα τα πειράματα που εκτελέστηκαν. Κατά την χρήση αυτής της συνάρτησης σφάλματος κάθε προβλεπόμενη πιθανότητα κλάσης συγκρίνεται με την πραγματική επιθυμητή έξοδο κλάσης 0 ή 1 και υπολογίζεται μια βαθμολογία/σφάλμα που τιμωρεί την πιθανότητα με βάση το πόσο μακριά είναι από την πραγματική αναμενόμενη τιμή. Η ποινή είναι λογαριθμική στη φύση αποφέρει μια μεγάλη βαθμολογία για μεγάλες διαφορές κοντά στο 1 και μικρή βαθμολογία για μικρές διαφορές που τείνουν στο 0. Η Cross Entropy Loss δίνεται από την σχέση

$$L_{CE} = -\sum_{i=1}^{n} t_i \log(p_i)$$

όπου n ο αριθμός των κλάσεων, t_i η πραγματική τιμή και p_i η Softmax πιθανότητα για την i-οστή κλάση.

Για την προσαρμογή των βαρών του δικτύου σε κάθε εποχή χρησιμοποιούνται αλγόριθμοι βελτιστοποίησης (optimizers), οι οποίοι ανανεώνουν τα βάρη αυτά με βάση την παράγωγο της συνάρτησης κόστους. Βάση των αλγορίθμων αυτών αποτελεί ο αλγόριθμος κατάβασης κλίσης (gradient descent algorithm) και ο αλγόριθμος οπισθοδιάδοσης (back-propagation). Οι πιο εμφανιζόμενοι, τους οποίους χρησιμοποιούμε και εμείς στις υλοποιήσεις μας είναι οι Adam και SGD.

- **Stochastic Gradient Descent:** Υπολογίζει επαναληπτικά σημεία στην συνάρτηση κόστους προκειμένου να καταλήξει στο σημείο ελαχιστοποίησης της. Ο ρυθμός εκπαίδευσης αλλά και κάποιος βαθμός τυχαιότητας είναι η βάση για την επιλογή των σημείων, με τον δεύτερο να είναι ο κυριότερος παράγοντας περιορισμού της πολυπλοκότητας εκτέλεσης.

- **Adam:** Αποτελεί συνδυασμό των προσαρμοστικού αλγορίθμου και του αλγορίθμου ρίζας μέσης τετραγωνικής επέκτασης. Λόγω του προηγούμενου συνδυασμούς επιτυγχάνει πολύ καλά αποτελέσματα καθώς διαχειρίζεται αραιές παραγώγους και θορυβώδη δεδομένα.

Για την αξιολόγηση των δικτύων που υλοποιήθηκαν χρησιμοποιήθηκε η απλή μετρική ακρίβειας που ορίζεται ως το πηλίκο του αθροίσματος των σωστών προβλέψεων δια του συνόλου των εξετασθέντων δειγμάτων:

$$accuracy = \frac{\text{Πλήθος σωστών προβλέψεων}}{\text{Πλήθος εξεταζόμενων δειγμάτων}}$$

Οι βιβλιοθήκες που χρησιμοποιήθηκαν για την υλοποίηση των μοντέλων μας είναι η PyTorch και η TensorFlow, όπου ο τανυστής είναι η βασική δομή που χρησιμοποιούν. Ένας τανυστής είναι ένα αλγεβρικό αντικείμενο που περιγράφει μια bi-linear (πολυγραμμική) σχέση μεταξύ συνόλων αλγεβρικών αντικειμένων που σχετίζονται με τον διανυσματικό χώρο. Όσον αφορά το λογισμικό/υλικό, όλα τα πειράματα εκτελέσθηκαν σε GeForce RTX 1080 Ti GPU σε σύστημα με RAM 64MB και ubuntu 16.04.

5.2 LSTMs με Δισδιάστατους Σκελετούς

Αρχικά μετατρέψαμε τα βίντεο σε ακολουθίες από καρέ κρατώντας τον ρυθμό δειγματοληψίας της kinect μέσω τις συνάρτησης VideoCapture της βιβλιοθήκης cv2. Τα χαρακτηριστικά που θα τροφοδοτήσουν το LSTM δίκτυο είναι οι συνταγμένες των keypoints του δισδιάστατους σκελετού. Συνεπώς το επόμενο βήμα της εργασίας μας είναι η εξαγωγή των σκελετών μέσω του OpenPose. Παρατηρώντας το σύνολο των δεδομένων βλέπουμε ότι δεν

υπάρχει χρήσιμη πληροφορία στις εκφράσεις του προσώπου και για το λόγο αυτό τα keypoints από το πρόσωπο δεν συλλέγονται. Θεωρητικά θα αρκούσε να χρησιμοποιήσουμε μόνο τα keypoints του καρπού όμως παρατηρούμε ότι και η κίνηση του βραχίωνα περιέχει σημαντική πληροφορία για το μοντέλο μας. Εκτελώντας το OpenPose με ορίσματα τον κορμό και τα χέρια προκύπτει ένα json αρχείο για κάθε καρέ. Όπως έχουμε προαναφέρει στην περιγραφή του OpenPose το json αρχείο περιέχει 201 παραμέτρους για τον κάθε εικονιζόμενο. Υπάρχουν δείγματα στα οποία απεικονίζονται περισσότεροι του ενός ανθρώπου στο κάθε καρέ, όμως εμείς κρατάμε τον κυριότερο άνθρωπο, το νοηματιστή. Για να το επιτύχουμε αυτό υπολογίζουμε για κάθε άνθρωπο το μέσο όρο όλων των συντελεστών εμπιστοσύνης και κρατάμε τον μεγαλύτερο.

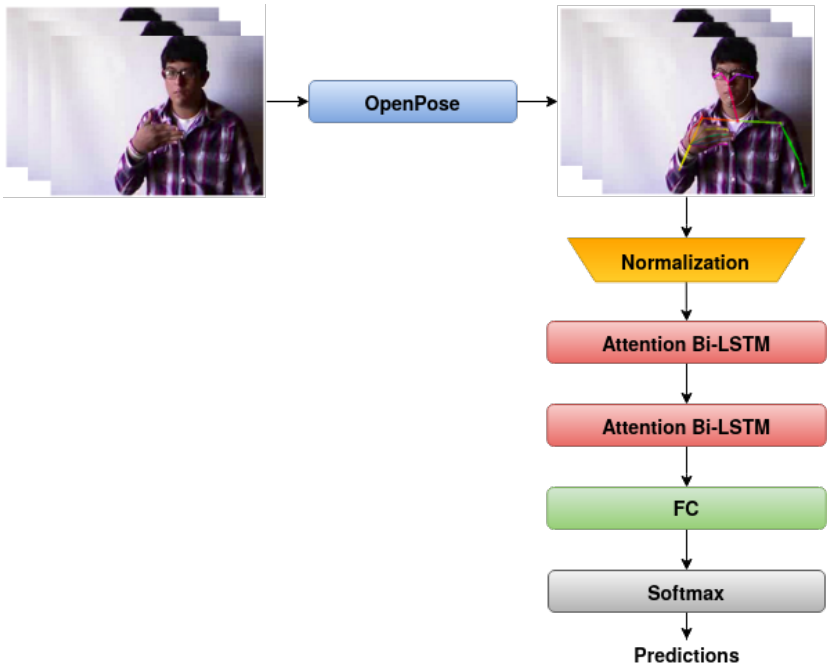
Δεδομένου του ότι οι τιμές των (x,y) στο json εκφράζουν συντεταγμένες πάνω στην εικόνα, με την αρχή των αξόνων να βρίσκεται στην άνω αριστερή γωνία, και του ότι οι νοηματιστές έχουν διαφορετικά φυσικά χαρακτηριστικά (π.χ. ύψος), βρίσκονται σε διαφορετικές θέσεις και αποστάσεις από την κάμερα, απαιτείται κανονικοποίηση των δεδομένων μας ώστε το μοντέλο να εκπαιδευτεί αποτελεσματικά. Έχοντας λοιπόν τις συντεταγμένες των 46 σημείων, εκφρασμένες ως προς το σύστημα συντεταγμένων της κάμερας, τις μεταφέρουμε στο σύστημα συντεταγμένων του ανθρώπινου κορμού, θέτοντας την αρχή των συντεταγμένων στο μέσο της ανθρώπινης λεκάνης, αν αυτό υπάρχει στο καρέ. Σε αντίθετη περίπτωση θα δημιουργήσουμε ένα πλασματικό σημείο μέσου λεκάνης μετά την κανονικοποίηση των υπόλοιπων δεδομένων μας. Η κανονικοποίηση γίνεται με βάση το ύψος του κορμού, την απόσταση δηλαδή του αυχένα από το μέσο της λεκάνης. Έχοντας πλέον κανονικοποιημένους τους σκελετούς, βρίσκουμε το μέσο ύψος των νοηματιστών και ανάλογα προσδιορίζουμε το πλασματικό μέσο της λεκάνης για να ολοκληρωθεί η προ-επεξεργασία και στους υπόλοιπους.

Στον επόμενο πίνακα φαίνονται οι παραλλαγές του LSTM που δοκιμάστηκαν. Συγκεκριμένα δοκιμάστηκαν μονοστρωματικά και διστρωματικά LSTM ακολουθούμενα από ένα γραμμικό πλήρως συνδεδεμένο επίπεδο (FC) και μια Softmax. Παρατηρούμε ότι η χρήση δύο στρωμάτων είναι σε όλες τις περιπτώσεις καλύτερη από την αντίστοιχη μονοστρωματική. Όπως ήταν αναμενόμενο τα Bidirrectional LSTMs δίνουν καλύτερη εκτίμηση από τα απλά

Network	Accuracy
1 LSTM	32.09
2 LSTM	33.87
1 Bi-LSTM	34.12
2 Bi-LSTM	36.22
1 Attention Bi-LSTM	39.39
2 Attention Bi-LSTM	**42.94**

Πίνακας 5.1: IsoGD: Συγκρίσεις Αναδρομικών Δικτύων σε Αναγνώριση Χειρονομιών με χρήση Δισδιάστατου σκελετού

καθώς ο συνδυασμός της παρελθοντικής, της τρέχουσας και της μελλοντικής πληροφορίας ενισχύει την ικανότητα του δικτύου να κατανοεί καλύτερα το context. Αξιοσημείωτη είναι και η συμβολή του Self Attention layer παρότι δεν έχει χρησιμοποιηθεί ιδιαίτερα στο task της αναγνώρισης ανθρώπινων δράσεων. Το καλύτερο μοντέλο μεταξύ των αναδρομικών δικτύων είναι το διστρωματικό Attention Bi-LSTM του οποίου η αρχιτεκτονική φαίνεται και στην επόμενη εικόνα:



Σχήμα 5.2: Η Αρχιτεκτονική του Καλύτερου Αναδρομικού Δικτύου, αποτελούμενη από δύο αμφίδρομα LSTM εμπλουτισμένα με έναν μηχανισμό προσοχής έκαστο.

Κατά την διάρκεια των πειραμάτων -διεξήχθησαν σε PyTorch - δοκιμάστηκαν αρκετές παράμετροι για το δίκτυο και επιλέχθηκαν οι εξής:

Παράμετροι του 2 Attention Bi-LSTM	
Batch Size	256
RNN Size	180
Loss Function	Cross Entropy Loss
Optimizer	Adam (lr=0.001, weight_decay= 0.001)

Πίνακας 5.2: Παράμετροι του Καλύτερου Αναδρομικού Δικτύου

5.3 C3Ds με RGB και πληροφορία Βάθους

Το συνελικτικό 3D δίκτυο, έχει την δυνατότητα να μαθαίνει χωροχρονική πληροφορία. Χρησιμοποιώντας μόνο το RGB κανάλι επιτυγχάνει αρκετά καλά αποτελέσματα, όπως έχουμε δει στην βιβλιογραφία και επαληθεύουμε στις επόμενες υλοποιήσεις. Παρόλα αυτά, σε σύγκριση με το RGB βίντεο, η ακολουθία βάθους είναι περισσότερο αναίσθητη στις αλλαγές φωτισμού και πιο λεπτομερειακή, λόγω της ικανότητάς της να πιάνει γεωμετρικές πληροφορίες αντικειμένων. Το C3D δίκτυο είναι προ-εκπαιδευμένο στο Sports 1M dataset, γεγονός που επιτρέπει να αποφεύγει το overfitting.

Στον επόμενο πίνακα 5.3 φαίνονται οι παραλλαγές του C3D που δοκιμάστηκαν. Αρχικά επιλέχθηκε να γίνει χρήση μόνο της RGB πληροφορίας. Στη συνέχεια υλοποιήθηκε ένα C3D με αποκλειστική είσοδο την πληροφορία από του κανάλι του βάθους. Παρατηρώντας τα αποτελέσματα από τα προηγούμενα δύο πειράματα, επιλέξαμε να προχωρήσουμε στο σχεδιασμό ενός μοντέλου C3D διπλής ροής. Συγκεκριμένα τα δύο υποδίκτυα που το αποτελούν

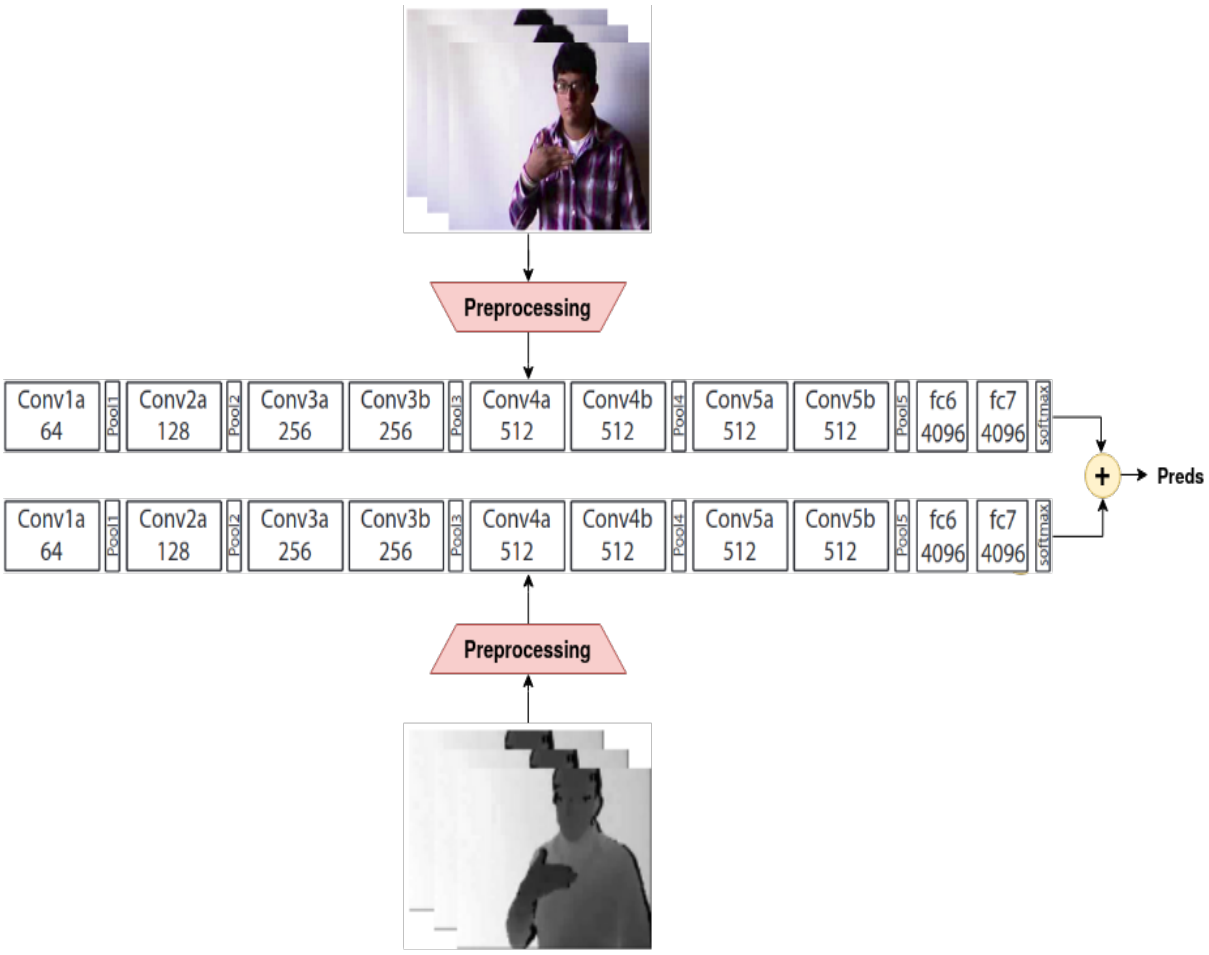
Network	Accuracy
RGB_C3D	36.94
Depth_C3D	46.71
Two-Stream_C3D	**48.90**

Πίνακας 5.3: IsoGD: Συγκρίσεις C3Ds σε Αναγνώριση Χειρονομιών με τη χρήση RGB πληροφορίας και πληροφορίας Βάθους

(RGB ροή, ροή Βάθους) εκπαιδεύονται ανεξάρτητα και εν συνεχεία εφαρμόζουμε late fusion, υπολογίζοντας τον μέσο όρων των δύο Softmax scores.

Και για τα τρία πειράματα εφαρμόζουμε την ίδια προ-επεξεργασία για τα δεδομένα μας, όπως αυτή περιγράφεται στο [63]. Συγκεκριμένα διαχωρίζουμε τα βίντεο σε clips μήκους 16 καρέ, μη επικαλυπτόμενα. Στη συνέχεια μεταβάλλουμε τις διαστάσεις του κάθε καρέ σε 128×171, και ακολούθως επιλέγουμε τυχαία ένα κομμάτι του καρέ διάστασης 112×112. Συνεπώς οι τελικές διαστάσεις των δεδομένων εισόδου ανά clip είναι $3 \times 16 \times 112 \times 112$.

Η αρχιτεκτονική του καλύτερου C3D μοντέλου φαίνεται στην επόμενη εικόνα:



Σχήμα 5.3: Η Αρχιτεκτονική του Καλύτερου C3D Δικτύου. Αποτελείται από δύο ροές, την RGB ροή και τη ροή Βάθους, οι οποίες εκπαιδεύονται ανεξάρτητα και επιδέχονται late fusion.

Κατά την διάρκεια των πειραματισμών - διεξήχθησαν σε PyTorch - δοκιμάστηκαν αρκετές παράμετροι για το δίκτυο και επιλέχθηκαν οι εξής:

Παράμετροι του Two Stream C3D	
Batch Size	20
Loss Function	Cross Entropy Loss
Optimizer	SGD(lr=0.001, momentum=0.9, weightdecay= $5e-4$)
Scheduler	scheduler.StepLR(optimizer, step_size=10, gamma=0.1

Πίνακας 5.4: Παράμετροι του Καλύτερου C3D Δικτύου

5.4 I3Ds με RGB, Flow και πληροφορία Βάθους

Θέλοντας να επιτύχουμε ακόμη μεγαλύτερη απόδοση, επιλέξαμε να ακολουθήσουμε την βιβλιογραφία και να επανα-υλοποιήσουμε το I3D μοντέλο. Αρχικά εκπαιδεύσαμε μόνο την RGB ροή του μοντέλου. Στην συνέχεια ακολούθησε η εκπαίδευση της ροής που δέχεται ως είσοδο την οπτική ροή. Εν τέλει εκπαιδεύσαμε και ολόκληρο το μοντέλο I3D, χρησιμοποιώντας και τις δύο ροές πληροφορίας. Ορμώμενοι από την χρήση της πληροφορίας βάθους στο C3D δίκτυο, επιλέξαμε να διενεργήσουμε δύο ακόμα πειραματισμούς, αντικαθιστώντας την RGB πληροφορία με την πληροφορία βάθους. Συγκεκριμένα αρχικά εκπαιδεύσαμε την appearance ροή χρησιμοποιώντας αποκλειστικά τα depth videos, ενώ στη συνέχεια υλοποιήσαμε ολοκληρωμένο το I3D δίκτυο, με τις δύο ροές να είναι το βάθος και η οπτική ροή. Τα αποτελέσματα που προέκυψαν από τους προηγούμενους πειραματισμούς φαίνονται στον επόμενο πίνακα:

Network	Accuracy
RGB_I3D	58.22
Depth_I3D	60.04
Flow_I3D	61.28
{RGB+Flow}_I3D	**67.09**
{Depth+Flow}_I3D	65.47

Πίνακας 5.5: IsoGD: Συγκρίσεις I3Ds σε Αναγνώριση Χειρονομιών με τη χρήση RGB, Βάθους και Οπτικής Ροής

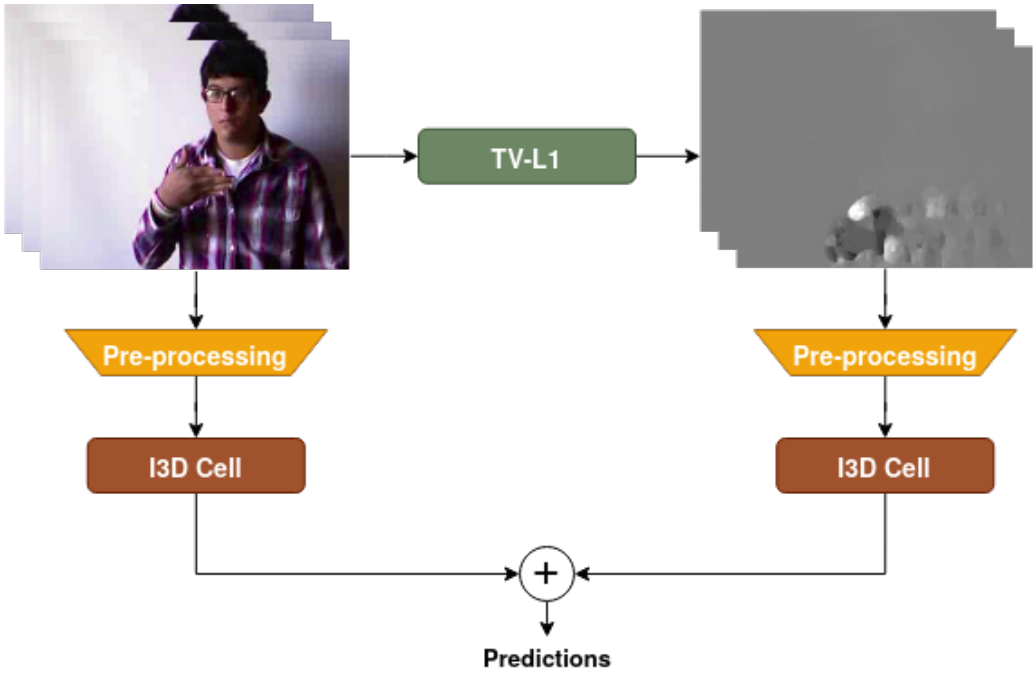
Όπως σε κάθε μοντέλο έτσι και στο I3D, πριν την εκπαίδευση του, είναι απαραίτητη η προ-επεξεργασία των δεδομένων που θα το τροφοδοτήσουν. Τα βίντεο τεμαχίζονται σε καρέ με ρυθμό δειγματοληψίας 25fps ανεξαρτήτου ροής.

Ξεκινώντας από τα RGB frames, μεταβάλουμε το μέγεθος τους, με προϋπόθεση την διατήρηση της αναλογίας, θέτοντας την μικρότερη διασταση στα 256 pixel. Η διαδικασία αυτή πραγματοποιείται χρησιμοποιώντας διγραμμική παρεμβολή (bilinear interpollation). Στην συνέχεια οι τιμές των pixel κανονικοποιούνται στο διάστημα $[-1, 1]$. Τέλος κατά την διάρκεια της εκπαίδευσης επιλέγεται ένα τυχαίο τμήμα του εκάστοτε καρέ διαστάσεων 224×224, σε αντίθεση με την φάση της δοκιμής, όπου επιλέγεται το κεντρικό κομμάτι του εκάστοτε καρέ, ίδιων διαστάσεων με την περίπτωση της εκπαίδευσης.

Όσον αφορά την οπτική ροή, αρχικά μετατρέπουμε τα καρέ σε ασπρόμαυρες εικόνες, τις οποίες τροφοδοτούμε στον TV-L1 αλγόριθμο, που περιγράψαμε στο προηγούμενο κεφάλαιο. Οι εξαγόμενες τιμές των pixel περιορίζονται στο διάστημα $[-20, 20]$ και στη συνέχεια κανο-

νικοποιούνται στο διάστημα $[-1, 1]$. Ακολούθως περικόπτουμε τα καρέ όπως ακριβώς κάναμε και στα RGB καρέ, ώστε να αποκτήσουν διαστάσεις 224×224.

Η δομή του καλύτερου I3D δικτύου από τα υλοποιηθέντα είναι η ακόλουθη:



Σχήμα 5.4: Η Δομή του Καλύτερου I3D Δικτύου

Κατά την διάρκεια των πειραματισμών - διεξήχθησαν σε TensorFlow - δοκιμάστηκαν αρκετές παράμετροι για το δίκτυο και επιλέχθηκαν οι εξής:

Παράμετροι του I3D	
Batch Size	6
Loss Function	Sparse softmax cross entropy with logits
Optimizer	Adam(lr=0.0001)

Πίνακας 5.6: Παράμετροι του Καλύτερου I3D Δικτύου

5.5 Σύγκριση Αποτελεσμάτων

Στον πίνακα 5.7 παρουσιάζονται συγκεντρωμένα όλα τα αποτελέσματα των πειραματισμών μας πάνω στο κομμάτι της αναγνώρισης χειρονομιών σε βίντεο. Όπως μπορούμε να παρατηρήσουμε, η χρήση του δισδιάστατου σκελετού, σε συνδυασμό με την δυνατότητα του LSTM να αναγνωρίζει μεγάλες ακολουθίες δίνει ιδιαίτερα ικανοποιητικά αποτελέσματα. Συγκεκριμένα τα δίκτυα 1 Bi-LSTM και 2 Bi-LSTM επιτυγχάνουν ακρίβεια 39,39% και 42,94% αντίστοιχα, ξεπερνώντας αρκετά την ακρίβεια του C3D μοντέλου που έχει εκπαιδευτεί αποκλειστικά σε RGB δεδομένα και φτάνει το 36,94%. Σημαντικός παράγοντας που επηρεάζει την απόδοση του C3D δικτύου είναι η ποικιλομορφία του background, του εξεταζόμενου dataset. Αντίθετα, αν χρησιμοποιηθεί η πληροφορία βάθους, που είναι λιγότερο ευαίσθητη σε τέτοιες μεταβολές, παρατηρούμε μια μεγάλη άνοδο της τάξης του 10% στην ακρίβεια του

Network	Accuracy
1 LSTM	32.09
2 LSTM	33.87
1 Bi-LSTM	34.12
2 Bi-LSTM	36.22
1 Attention Bi-LSTM	39.39
2 Attention Bi-LSTM	**42.94**
RGB_C3D	36.94
Depth_C3D	46.71
Two-Stream_C3D	**48.90**
RGB_I3D	58.22
Depth_I3D	60.04
Flow_I3D	61.28
{RGB+Flow}_I3D	**67.09**
{Depth+Flow}_I3D	65.47
State-of-the-art[1]	82.07

Πίνακας 5.7: IsoGD: Συγκεντρωτικός Πίνακας όλων των πειραμάτων

μοντέλου μας, η οποία φτάνει το 46.71%. Παρόλα αυτά το RGB κανάλι προσφέρει πληροφορία, η οποία δεν μπορεί να εντοπιστεί στο κανάλι βάθους, γεγονός που επαληθεύεται από το αποτέλεσμα του συνδυασμού των δύο καναλιών (RGB - Βάθος), όπου αγγίζει το 49%. Τέλος όσον αφορά το I3D μοντέλο, όπως ήταν αναμενόμενο και από την βιβλιογραφία, δίνει ακόμα μεγαλύτερη ακρίβεια, φτάνοντας έως και το 67%. Το γεγονός αυτό αποδεικνύει την σημαντικότητα της προ-εκπαίδευσης στις βάσεις Imagnet και Kinetics. Αξιοσημείωτο είναι επίσης το γεγονός ότι σε αυτό το δίκτυο η συμβολή της πληροφορίας βάθους φαίνεται να μην είναι ιδιαίτερα σημαντική.

Στη βιβλιογραφία έχουν αναπτυχθεί και αρκετές ακόμα μέθοδοι πάνω στο συγκεκριμένο σύνολο δεδομένων. Κάποιες από αυτές χρησιμοποιούν και άλλα κανάλια πληροφορίας, όπως το Saliency, ενώ άλλες συνδυάζουν και περισσότερα των δύο καναλιών, ως είσοδο στο μοντέλο τους (πχ RGB + Depth + FLow + Saliency). Υπάρχουν ακόμη μέθοδοι που προ-επεξεργάζονται με διαφορετικό τρόπο τα δεδομένα εισόδου, όπως είναι η απομάκρυνσή του background και η περικοπή του καρέ, ώστε να περιλαμβάνει μόνο τον καρπό. Παρόλα αυτά, στην περίπτωση μελέτης της διπλωματικής μας δεν επεκταθήκαμε σε τέτοιες μεθόδους, καθώς επιθυμούσαμε να χρησιμοποιήσουμε μεθόδους και δεδομένα, τα οποία δε θα προσθέσουν επιπλέον υπολογιστική επιβάρυνση στο σύστημα, σε περίπτωση μεταφοράς του σε online συνθήκες. Αυτό αποτελεί το δεύτερο κομμάτι της παρούσας εργασίας και παρουσιάζεται στο επόμενο κεφάλαιο.

[1] Το state-of-the-art μοντέλο χρησιμοποιεί ένα ResNet-50 δίκτυο που εφαρμόζεται μόνο στην περιοχή των χεριών. Στο μοντέλο αυτό εφαρμόζονται διαφορετικές ροές πληροφορίας και διαφορετικές προ-επεξεργασίες δεδομένων που δεν αποτελούν αντικείμενο έρευνας της παρούσας διπλωματικής εργασίας [67].

Κεφάλαιο 6

Χρονικός Εντοπισμός Δράσεων σε Πραγματικό Χρόνο – Πειράματα, Αποτελέσματα και Συγκρίσεις

Στο κεφάλαιο αυτό παρουσιάζουμε τα πειραματικά μας αποτελέσματα για το πρόβλημα του χρονικού εντοπισμού δράσεων σε πραγματικό χρόνο και συγκεκριμένα ένα σύνολο από υλοποιήσεις. Για κάθε μία από τις υλοποιήσεις μας αρχικά περιγράφουμε την διαδικασία προ-επεξεργασίας των δεδομένων, τα βασικότερα στοιχεία της καθώς και τα παραγόμενα αποτελέσματα. Μετά το πέρας όλων τον υλοποιήσεων πραγματοποιούμε μια συνολική σύγκριση των αποτελεσμάτων όλων των μεθόδων.

6.1 Παράμετροι και Μετρικές

Για το χρονικό εντοπισμό ανθρώπινων δράσεων σε πραγματικό χρόνο χρησιμοποιήθηκε το THUMOS'14 σύνολο δεδομένων. Κατά την φάση της δοκιμής δεν παρέχονταν στο σύστημα τα μελλοντικά καρέ, ώστε να μπορέσουμε να αξιολογήσουμε την online επίδοση του μοντέλου. Για τη μελέτη αυτού του αντικειμένου έρευνας τροποποιήθηκε το TRN [78] και στη συνέχεια ενισχύθηκε με διαφορετικά χωροχρονικά χαρακτηριστικά. Τα χαρακτηριστικά αυτά εξήχθησαν από τα μοντέλα που εξετάστηκαν στο πρώτο κομμάτι της διπλωματικής, δηλαδή C3D χαρακτηριστικά, I3D χαρακτηριστικά, ανθρώπινη πόζα.

Ακολουθώντας την προηγούμενες εργασίες, επιλέξαμε και εμείς να χρησιμοποιήσουμε την mAP (mean Average Precision) μετρική για την αξιολόγηση των μοντέλων μας. Για να μπορέσουμε να ορίσουμε την μετρική αυτή, χρειάζεται αρχικά να ορίσουμε δύο άλλες μετρικές την precision (ακρίβεια) και recall (ανάκληση). Οπότε έχουμε:

- Precision: Γνωστή και ως ακρίβεια της θετικής προβλεπόμενης τιμής και ορίζεται ως ο λόγος του πραγματικού θετικού (TP) προς το συνολικό αριθμό των προβλεπόμενων θετικών. Δίνεται από τον εξής τύπο:

$$Precision = \frac{\text{TP}}{\text{TP + FP}}$$

- Recall (TPR): Γνωστή και ως πραγματικός θετικός ρυθμός ή ευαισθησία και ορίζεται ως ο λόγος του πραγματικού θετικού (TP) προς το συνολικό αριθμό των θετικών δειγμάτων. Δίνεται από τον εξής τύπο:

$$Recall(TPR) = \frac{\text{TP}}{\text{TP + FN}}$$

- Average Precision (AP): Αποτελεί έναν συνδυασμό των προηγούμενων δύο μετρικών καθώς αποτελεί τον σταθμισμένο μέσο όρο των precisions, που επιτυγχάνονται σε κάθε κατώφλι. Τα αντίστοιχα βάρη είναι η αύξηση του recall από το προηγούμενο κατώφλι. Δίνεται από τον τύπο:

$$AP = \sum_n (R_n - R_{n-1}) P_n$$

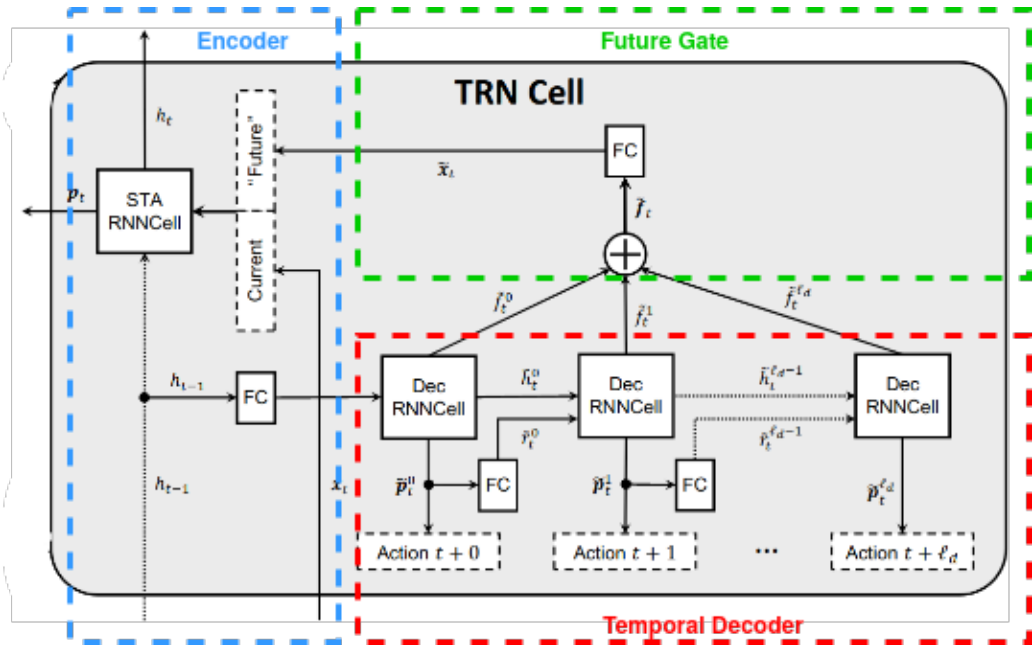
, όπου P_n και R_n τα precision και recall του n-ιοστού κατωφλιού.

- mean Average Precision (mAP): Είναι η μέση τιμή όλων των AP, για όλες τις κλάσεις του εκάστοτε προβλήματος. Δίνεται από τον τύπο:

$$mAP = \frac{\sum_n AP_n}{N}$$

Ως συνάρτηση σφάλματος χρησιμοποιείται η CrossEntropyLoss η οποία έχει οριστεί στο προηγούμενο κεφάλαιο (βλ. ενότητα 5.1), καθώς χρησιμοποιήθηκε και στο task της αναγνώρισης χειρονομιών σε βίντεο. Η διαφορά που υπάρχει μεταξύ της τρέχουσας συνάρτησης σφάλματος και αυτής του προηγούμενου κεφαλαίου είναι ότι στην τρέχουσα περίπτωση διαχειρίζεται πολλές κλάσεις σε κάθε βίντεο. Όπως έχουμε προαναφέρει και στο προηγούμενο κεφάλαιο, για την προσαρμογή των βαρών του δικτύου χρησιμοποιούνται κάποιοι optimizers. Στα πειράματα που ακολουθούν χρησιμοποιήθηκε ο Adam optimizer. Η βιβλιοθήκη που χρησιμοποιήθηκε για την υλοποίηση των μοντέλων είναι η Pytorch. Τέλος όσον αφορά το λογισμικό/υλικό, όλα τα πειράματα εκτελέστηκαν σε GeForce RTX 2080 Ti GPU σε σύστημα με RAM 64MB και ubuntu 16.04.

Η δομή του TRN cell έχει περιγραφεί αναλυτικά στο κεφάλαιο 3, αλλά για λόγους πληρότητας κάνουμε μια μικρή επανάληψη της δομής του. Το κύτταρο δέχεται ένα διάνυσμα

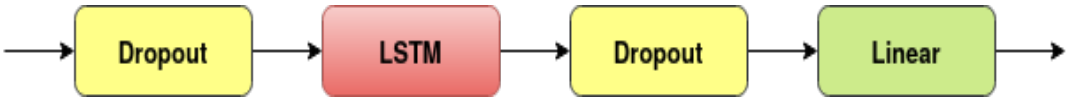


Σχήμα 6.1: Η αρχιτεκτονική του TRN κυττάρου [78]

χαρακτηριστικών, που εισέρχεται σε έναν χρονικό αποκωδικοποιητή, του οποίου οι κρυφές καταστάσεις στη συνέχεια τροφοδοτούν την πύλη χαρακτηριστικών. Από την πύλη αυτή

εξέρχεται η δημιουργούμενη μελλοντική πληροφορία, η οποία συνδυάζεται με την τρέχουσα και το τελικό διάνυσμα τροφοδοτεί τον κωδικοποιητή. Η έξοδος του κωδικοποιητή είναι αυτή που μας δίνει τις επιζητούμενες προβλέψεις. Στο σχήμα 6.1 παραθέτουμε ξανά την εικόνα του TRN κυττάρου ενώ στη συνέχεια προσδίδουμε επιπλέον λεπτομέρειες για την αρχιτεκτονική των τμημάτων του (Decoder, Feature Gate, Encoder).

Η αρχιτεκτονική του κωδικοποιητή και του αποκωδικοποιητή περιλαμβάνουν τα ίδια στρώματα, τα οποία φαίνονται στην επόμενη εικόνα, ενώ στον επόμενο πίνακα αναλύονται και οι παράμετροι του κάθε δικτύου.

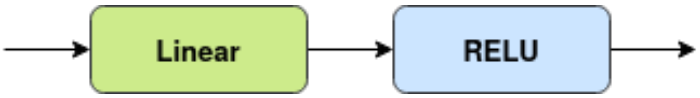


Σχήμα 6.2: Η αρχιτεκτονική των κωδικοποιητή και αποκωδικοποιητή

	Encoder	Decoder
Dropout	p=0.1	p=0.1
LSTM	fusion size, hidden size=4096	hidden size=4096, hidden size=4096
Dropout	p=0.1	p=0.1
Linear	hidden size=4096, num classes=22	hidden size=4096, num classes=22

Πίνακας 6.1: Παράμετροι του Κωδικοποιητή και του Αποκωδικοποιητή

Τα χαρακτηριστικά εισόδου είναι χαρακτηριστικά διπλής ροής. Συγκεκριμένα, η μία ροή είναι τα χαρακτηριστικά εμφάνισης (appearance features), ενώ η άλλη τα χαρακτηριστικά κίνησης (motion features), τα οποία συνενώνονται και τροφοδοτούνται σε ένα γραμμικό επίπεδο, με μία ReLU συνάρτηση ενεργοποίησης. Στην ίδια ακριβώς αρχιτεκτονική, που φαίνεται στην επόμενη εικόνα, έχει δομηθεί και η feature gate.



Σχήμα 6.3: Η αρχιτεκτονική της πύλης χαρακτηριστικών

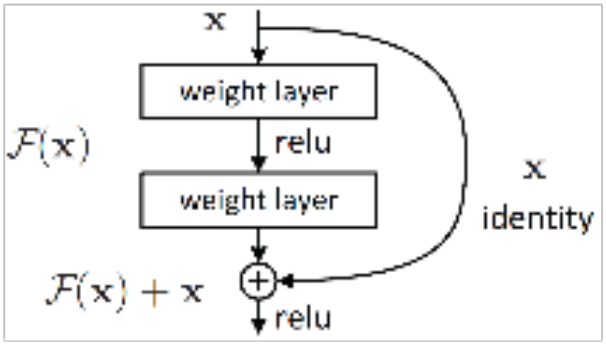
Λόγω του μεγάλου πλήθους παραμέτρων του μοντέλου, τόσο το απλό TRN, όσο και οι προτεινόμενες επεκτάσεις πραγματοποιήθηκαν με batch_size 2. Παράλληλα για όλα τα μοντέλα τα βίντεο έχουν δειγματοληφθεί σε συχνότητα δειγματοληψίας 30fps και έχουν διασπαστεί είτε σε κομμάτια των 6 είτε των 16 καρέ, ανάλογα με την υλοποίηση. Τα καρέ της οπτικής ροής έχουν εξαχθεί μέσω του εργαλείου dense_flow, που περιγράψαμε στα προηγούμενα κεφάλαια. Τέλος το πλήθος βημάτων του κωδικοποιητή παίρνει την τιμή 64, ενώ το πλήθος των βημάτων του αποκωδικοποιητή παίρνει την τιμή 8. Το τελικό διάνυσμα που προκύπτει τροφοδοτεί το TRN cell. Στο εξής δεν θα υπάρξει καμία αλλαγή στις προηγούμενες αρχιτεκτονικές, καθώς το αντικείμενο μελέτης μας είναι η εξερεύνηση της συνεισφοράς διαφόρων χαρακτηριστικών εισόδου, στην ακρίβεια του online μοντέλου.

6.2 Baseline & OpenPose TRN

Στο baseline δίκτυο χρησιμοποιήθηκαν χαρακτηριστικά διπλής ροής, όπως αυτά έχουν προταθεί στα [76][74]. Συγκεκριμένα, για την περίπτωση της appearance ροής, χρησιμοποιήθηκε το ResNet-200 δίκτυο, το οποίο εφαρμόστηκε πάνω στα RGB δεδομένα. Παράλληλα για την motion ροή χρησιμοποιήθηκε το Bn-Inception δίκτυο, το οποίο εφαρμόστηκε πάνω στα δεδομένα οπτικής ροής. Σύμφωνα με το [74], ο συνδυασμός τους είναι πολύ αποτελεσματικός. Γεγονός στο οποίο βασίστηκαν και οι δημιουργοί του baseline TRN. Στο σημείο αυτό αξίζει να τονίσουμε ότι πιο σύνθετοι μηχανισμοί συνδυασμού, όπως το IDT [72] και το TDD [71] μπορούν να εφαρμοστούν και να επιφέρουν καλύτερα αποτελέσματα. Οι αρχιτεκτονικές των ResNet-200 και Bn-Inception φαίνονται στις εικόνες 6.4 και 6.5 αντίστοιχα.

layer name	output size	200-layer
conv1	112×112	7×7, 64, stride 2
conv2_x	56×56	3×3 max pool, stride 2 $\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$
conv3_x	28×28	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 24$
conv4_x	14×14	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 36$
conv5_x	7×7	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$
	1×1	average pool, 1000-d fc, softmax
FLOPs		15×10^9

α

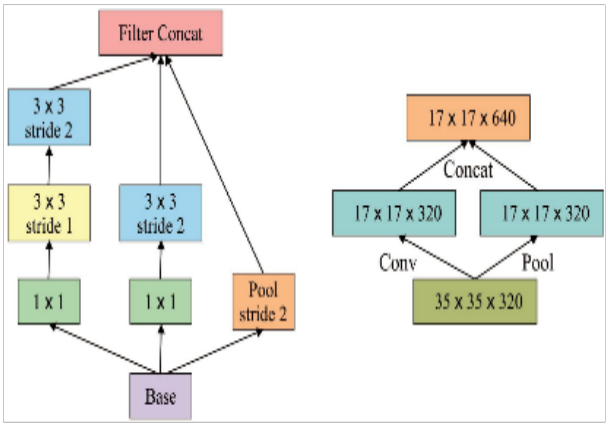


β

Σχήμα 6.4: α) Παράμετροι και β) αρχιτεκτονική του ResNet-200 δικτύου [22]

type	patch size/ stride	output size	depth	#1×1	#3×3 reduce	#3×3	double #3×3 reduce	double #3×3	Pool +proj
convolution*	7×7/2	112×112×64	1						
max pool	3×3/2	56×56×64	0						
convolution	3×3/1	56×56×192	1		64	192			
max pool	3×3/2	28×28×192	0						
inception (3a)		28×28×256	3	64	64	64	64	96	avg + 32
inception (3b)		28×28×320	3	64	64	96	64	96	avg + 64
inception (3c)	stride 2	28×28×576	3	0	128	160	64	96	max + pass through
inception (4a)		14×14×576	3	224	64	96	96	128	avg + 128
inception (4b)		14×14×576	3	192	96	128	96	128	avg + 128
inception (4c)		14×14×576	3	160	128	160	128	160	avg + 128
inception (4d)		14×14×576	3	96	128	192	160	192	avg + 128
inception (4e)	stride 2	14×14×1024	3	0	128	192	192	256	max + pass through
inception (5a)		7×7×1024	3	352	192	320	160	224	avg + 128
inception (5b)		7×7×1024	3	352	192	320	192	224	max + 128
avg pool	7×7/1	1×1×1024	0						

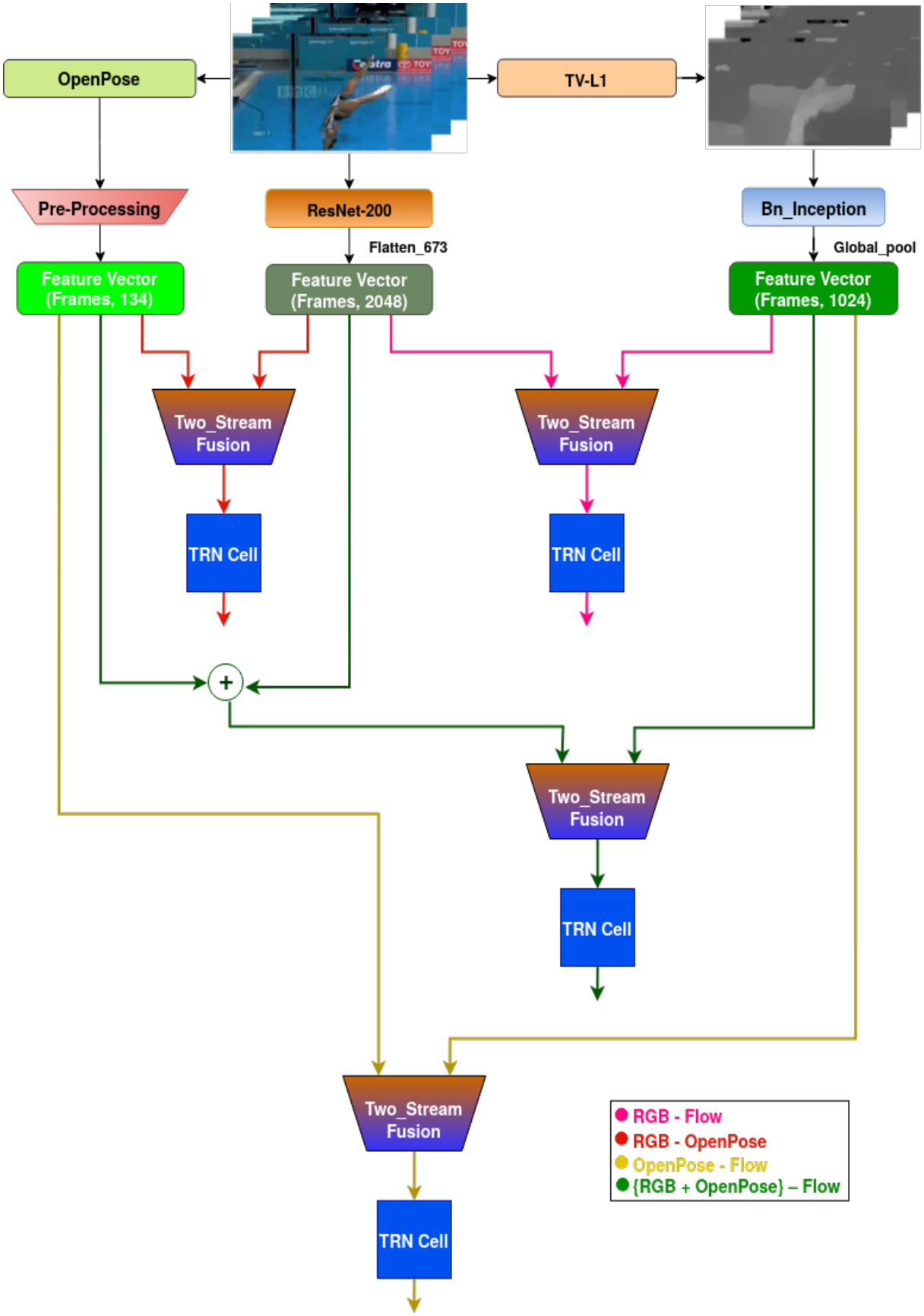
α



β

Σχήμα 6.5: α) Παράμετροι και β) αρχιτεκτονική του Bn_Inception δικτύου [26]

Οι συντεταγμένες του ανθρώπινου σκελετού είναι υψηλής ακρίβειας και μπορούν να αναπαραστήσουν την χρονική δυναμική των δράσεων με μεγάλη ακρίβεια [14]. Για το λόγο αυτό επιλέξαμε να χρησιμοποιήσουμε εκ νέου το OpenPose και να παράγουμε δισδιάστατους σκελετούς που θα ενισχύσουν τα baseline RGB και Flow χαρακτηριστικά. Από τα αρχεία json, που εξάγει το OpenPose, επιλέγουμε να κρατήσουμε τα σημεία κλειδιά του κορμού και των χεριών, τα οποία επιδέχονται την προ-επεξεργασία που έχουμε περιγράψει αναλυτικά στο προηγούμενο κεφάλαιο. Έχοντας πλέον κανονικοποιήσει τα δεδομένα μας δημιουργούμε ένα διάνυσμα 134 στοιχείων, το οποίο στη συνέχεια χρησιμοποιούμε είτε ως μοναδικά appearance χαρακτηριστικά είτε ως μοναδικά motion χαρακτηριστικά είτε ως συνιστώσα των appearance χαρακτηριστικών. Η επιλογή μας αυτή να χρησιμοποιήσουμε τα χαρακτηριστικά της ανθρώπινης πόζας είτε ως appearance είτε ως motion χαρακτηριστικά πηγάζει από τα ιδιαίτερα σημαντικά αποτελέσματα που πέτυχε το μοντέλο που τα χρησιμοποιούσε στο προηγούμενο σετ πειραματισμών μας (LSTM + OpenPose για gesture recognition). Παράλληλα η εκτεταμένη χρήση των σκελετών στην βιβλιογραφία, ως motion χαρακτηριστικά, ενδυνάμωσε ακόμη περισσότερο την επιλογή μας. Η δομή του δικτύου που χρησιμοποιήθηκε παρουσιάζεται στην επόμενη εικόνα:



Σχήμα 6.6: Δομή των τεσσάρων πειραμάτων με βάση το baseline και το OpenPose

Όπως φαίνεται και στο προηγούμενο σχήμα, τα χαρακτηριστικά του ResNet-200 εξάγονται από το Flatten_673 επίπεδο και έχουν διαστάσεις $\#Frames \times 2048$. Τα χαρακτηριστικά του Bn-Inception από το global_pool επίπεδο και έχουν διαστάσεις $\#Frames \times 1024$, ενώ τα χαρακτηριστικά του OpenPose υφίστανται την προ-επεξεργασία, που έχουμε αναφέρει αρκετές φορές στα προηγούμενα κεφάλαια. Το chuck_size σε αυτά τα πειράματα έχει οριστεί στην τιμή 6. Τα αποτελέσματα που προκύπτουν από τους πειραματισμούς αυτούς συγκεντρώνονται στον επόμενο πίνακα:

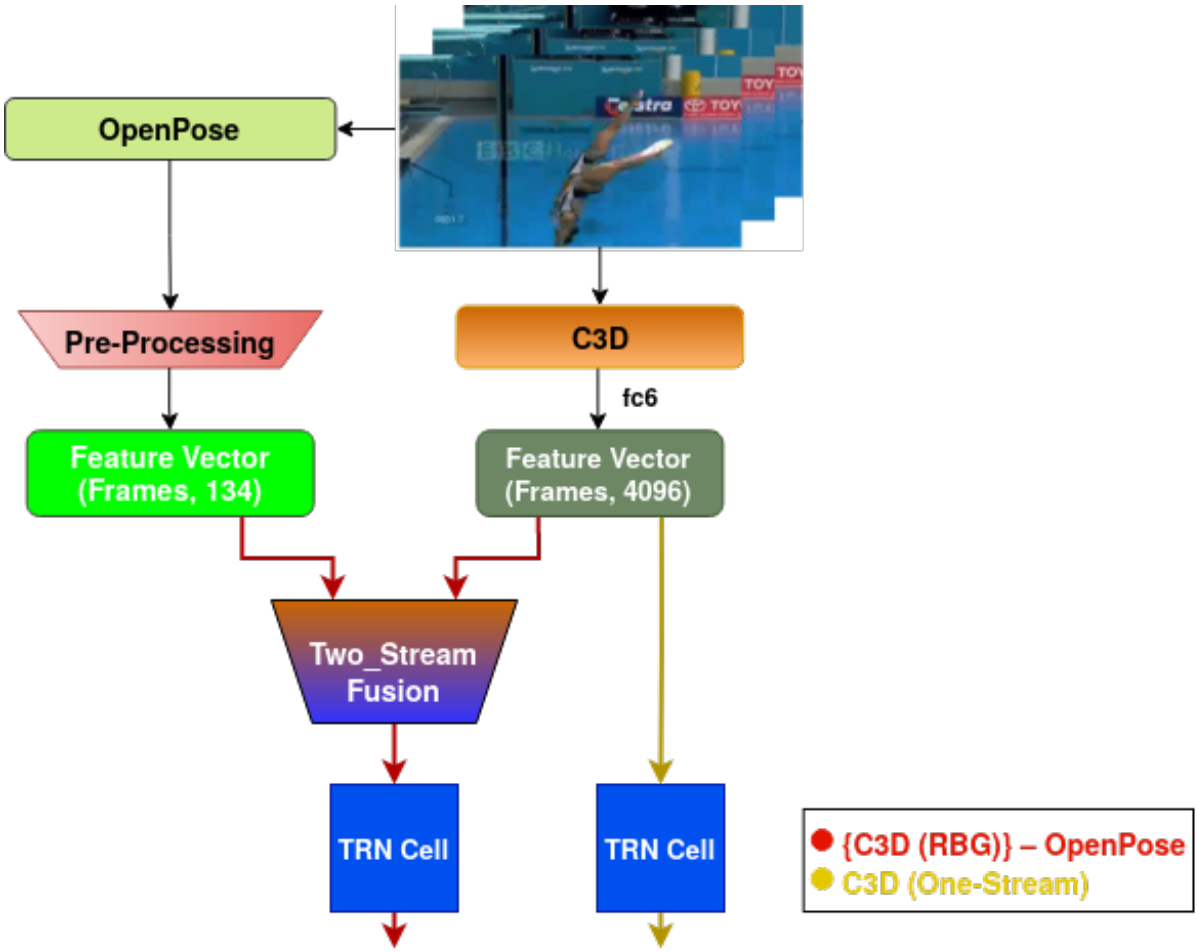
Method	Features Chunk size = 6 frames	Encoder	Decoder - Time predicted into the future (seconds)								
			0.25s	*0.50s*	*0.75s*	*1.00s*	*1.25s*	*1.50s*	*1.75s*	*2.00s*	*Avg*
Baseline	RGB – Flow	25.93	**26.15**	25.89	25.79	25.73	25.66	**25.68**	**25.66**	**25.57**	**25.77**
Ours	{RGB + OpenPose} – Flow	24.25	23.11	25.63	**26.72**	26.18	25.57	24.94	24.40	23.94	25.06
Ours	RGB – OpenPose	**37.57**	25.54	**25.93**	26.44	**26.60**	**26.28**	25.57	24.75	24.00	25.64
Ours	OpenPose – Flow	36.30	21.77	22.59	23.57	23.19	22.28	21.30	20.49	19.83	21.88

Πίνακας 6.2: THUMOS'14: Αποτελέσματα συνδυασμού Baseline και OpenPose

Από τον προηγούμενο πίνακα παρατηρούμε ότι η χρήση των χαρακτηριστικών του σκελετού, τόσο ως appearance χαρακτηριστικά όσο και ως motion χαρακτηριστικά αυξάνει αρκετά την επίδοση του μοντέλου. Στην περίπτωση χρήσης της πόζας ως motion χαρακτηριστικό η ακρίβεια του μοντέλου φτάνει το 37,57%. Όσον αφορά τον αποκωδικοποιητή, την δυνατότητα του μοντέλου να προβλέπει την πληροφορία κάποιων καρέ μπροστά, παρατηρούμε ότι το baseline παραμένει το καλύτερο μοντέλο. Αυτό εξηγείται από το γεγονός ότι στην περίπτωση του baseline τα RGB καρέ αποτελούν το ήμισυ της εισερχόμενης πληροφορίας, σε αντίθεση με τις υπόλοιπες περιπτώσεις, που είτε απουσιάζουν είτε μειοψηφούν. Η προγούμενη παρατήρηση επιβεβαιώνεται και από το γεγονός ότι η μέση ακρίβεια του anticipation στο τελευταίο πείραμα (OpenPose + Flow) έχει πέσει στην τιμή 21,88%.

6.3 C3D & OpenPose TRN

Η χρήση των C3D χαρακτηριστικών είναι ιδιαίτερα διαδεδομένη στις αναπαραστάσεις χαρακτηριστικών. Αυτό οφείλεται στο γεγονός ότι οι τρισδιάστατες συνελικτικές ενότητες μπορούν να εξάγουν χωροχρονικά χαρακτηριστικά ταυτόχρονα, σε αντίθεση με ResNet-200 που περιορίζονται στην εξαγωγή appearance χαρακτηριστικών. Παράλληλα το C3D δίκτυο είναι προ-εκπαιδευμένο στο Sports-1M σύνολο δεδομένων, κάτι που μειώνει την ανάγκη για fine-tunning. Καθώς το C3D δίκτυο δίνει χαρακτηριστικά μιας ροής, και ορμώμενοι και από την αποτελεσματικότητα των χαρακτηριστικών της πόζας στην προηγούμενη ενότητα, επιλέξαμε να δημιουργήσουμε ξανά χαρακτηριστικά διπλής ροής. Συγκεκριμένα επιλέξαμε να χρησιμοποιήσουμε τα C3D χαρακτηριστικά, ως appearance χαρακτηριστικά, και να τα συνδυάσουμε με τα OpenPose χαρακτηριστικά, τα οποία διαχειριστήκαμε ως motion χαρακτηριστικά. Η δομή του δικτύου μας φαίνεται στην επόμενη εικόνα:



Σχήμα 6.7: Δομή των δύο πειραμάτων με βάση το C3D και το OpenPose

Όπως φαίνεται και στο προηγούμενο σχήμα, τα C3D χαρακτηριστικά εξάγονται από το fc6 επίπεδο και έχουν διαστάσεις $\#Frames \times 4096$, ενώ για τα OpenPose χαρακτηριστικά ισχύει ότι και προηγουμένως. Το chunck_size σε αυτά τα πειράματα έχει οριστεί 16, λόγω του περιορισμού που θέτει το C3D δίκτυο. Τα αποτελέσματα που προκύπτουν από τους πειραματισμούς αυτούς φαίνονται στον επόμενο πίνακα:

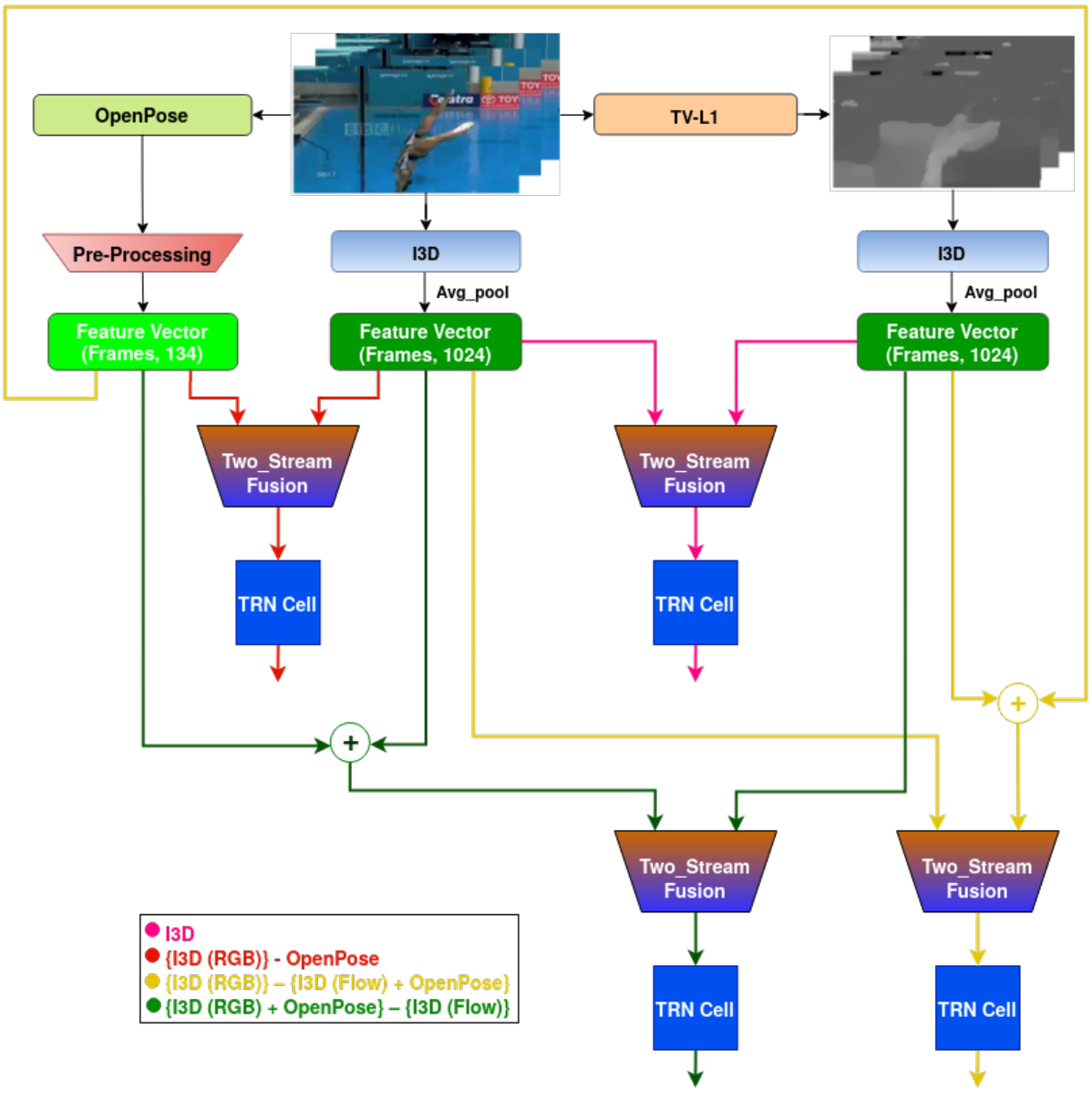
Method	Features Chunk size = 16 frames	Encoder	Decoder - Time predicted into the future (seconds)								
			0.25s	*0.50s*	*0.75s*	*1.00s*	*1.25s*	*1.50s*	*1.75s*	*2.00s*	*Avg*
Ours	C3D (One-Stream)	35.43	**34.34**	**31.05**	28.22	26.46	25.37	**24.75**	**24.39**	**24.22**	**27.35**
Ours	{C3D (RBG)} – OpenPose	**36.44**	32.98	30.56	**28.37**	**26.61**	**25.38**	24.54	23.78	23.22	26.93

Πίνακας 6.3: THUMOS'14: Αποτελέσματα συνδυασμού C3D και OpenPose

Από τον προηγούμενο πίνακα παίρνουμε τα αναμενόμενα αποτελέσματα. Συγκεκριμένα το απλό C3D μοντέλο δίνει καλύτερα χαρακτηριστικά σε σχέση με το baseline. Όπως και στην προηγούμενη περίπτωση, η χρήση του OpenPose ως motion ροή, ενισχύει την ακρίβεια του μοντέλου. Τέλος το anticipation, όπως και προηγουμένως, είναι καλύτερο στην περίπτωση που η RGB πληροφορία κυριαρχεί.

6.4 I3D & OpenPose TRN

Εμπνεόμενοι από την χρήση χαρακτηριστικών διπλής ροής σε όλους τους προηγούμενους πειραματισμούς, επιλέξαμε να χρησιμοποιήσουμε ένα μοντέλο, που είναι εξ' ορισμού διπλής ροής. Συγκεκριμένα χρησιμοποιήθηκε το I3D μοντέλο, το οποίο στη συνέχεια συνδυάστηκε και με τα OpenPose χαρακτηριστικά. Οι συνδυασμοί που έγιναν είναι αντίστοιχοι αυτών που έγιναν στο πρώτο σετ πειραμάτων. Αναλυτικότερα, εκτός του απλού I3D μοντέλου, τα OpenPose χαρακτηριστικά είτε συνδυάστηκαν με RGB χαρακτηριστικά του I3D ώστε μαζί να αποτελέσουν τα appearance χαρακτηριστικά, είτε συνδυάστηκαν με τα Flow χαρακτηριστικά του I3D ώστε στο σύνολο τους να σχηματίσουν τα motion χαρακτηριστικά, είτε χρησιμοποιήθηκαν ως αυτοτελή motion χαρακτηριστικά. Η προ-εκπαίδευση του I3D στο kinetics σύνολο δεδομένων εξελίσσει την ικανότητα του μοντέλου να γενικεύεται και αποφεύγεται το overfitting. Η δομή του δικτύου μας φαίνεται αναλυτικά στην επόμενη εικόνα:



Σχήμα 6.8: Δομή των δύο πειραμάτων με βάση το I3D και το OpenPose

Όπως φαίνεται και από το προηγούμενο σχήμα, τα I3D χαρακτηριστικά εξάγονται από το avg_pool επίπεδο και έχουν διαστάσεις $\#Frames \times 1024$, ενώ για τα OpenPose χαρακτηριστικά, ισχύει ό,τι και προηγουμένως. Η οπτική ροή που τροφοδοτεί το I3D δίκτυο έχει εξαχθεί και πάλι μέσω του TV-L1 αλγόριθμου. Όπως και στα πειράματα με βάση το C3D, έτσι και εδώ το chunk_size παίρνει την τιμή 16 λόγω των περιορισμών που θέτει το I3D δίκτυο. Τα αποτελέσματα των προηγούμενων πειραματισμών φαίνονται αναλυτικά στον επόμενο πίνακα:

Method	Features Chunk size = 16 frames	Encoder	Decoder - Time predicted into the future (seconds)								
			0.25s	*0.50s*	*0.75s*	*1.00s*	*1.25s*	*1.50s*	*1.75s*	*2.00s*	*Avg*
Ours	**I3D**	**55.25**	**52.57**	**46.69**	**41.94**	**38.39**	**35.90**	**34.22**	**33.00**	**32.08**	**39.35**
Ours	{I3D (RGB) + OpenPose} - {I3D (Flow)}	49.21	46.65	40.78	36.42	33.19	30.90	29.42	28.43	27.71	34.19
Ours	{I3D (RGB)} - OpenPose	47.43	44.59	40.08	36.77	34.24	32.37	31.29	30.56	30.06	35.00
Ours	{I3D (RGB)} - {I3D (Flow) + OpenPose}	44.47	29.55	31.92	29.62	27.21	25.63	24.78	24.20	23.68	27.07

Πίνακας 6.4: THUMOS'14: Αποτελέσματα συνδυασμού I3D και OpenPose

Από τον προηγούμενο πίνακα συμπεραίνουμε την αποτελεσματικότητα του I3D δικτύου στην εξαγωγή χωροχρονικών χαρακτηριστικών, καθώς επιτυγχάνει τα μέγιστα αποτελέσματα, τόσο στην πρόβλεψη των μελλοντικών καρέ (39.35%), όσο και στον χρονικό εντοπισμό και αναγνώριση των δράσεων (52.57%). Παρότι το απλό I3D μοντέλο είναι το καλύτερο, συγκρίνοντας τα υπόλοιπα μοντέλα μεταξύ τους, όπως και στα δύο προηγούμενα σετ πειραμάτων, η ακρίβεια του anticipation είναι μεγαλύτερη, όταν η RGB πληροφορία κυριαρχεί.

6.5 Σύγκριση Αποτελεσμάτων

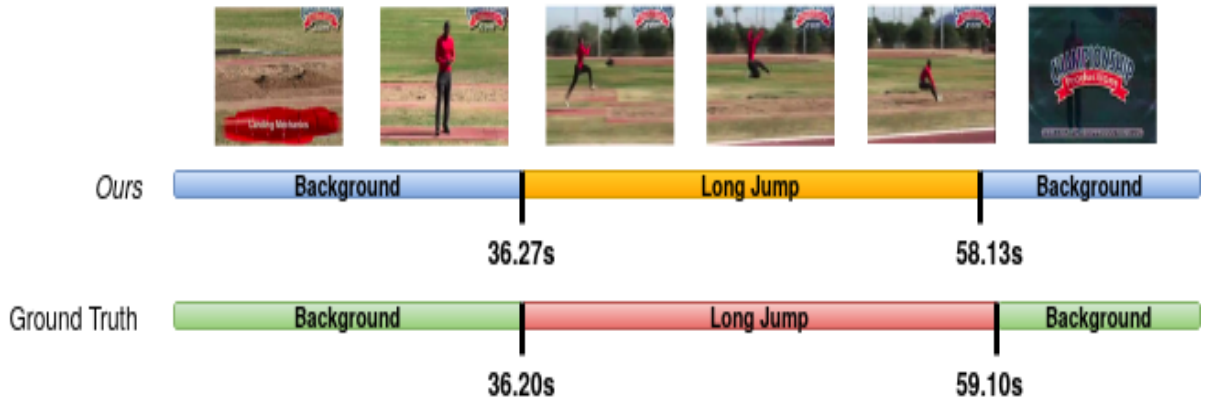
Στον επόμενο πίνακα [1] φαίνονται συγκεντρωμένα τα αποτελέσματα όλων των μοντέλων

Method	Features	Encoder	Decoder - Time predicted into the future (seconds)								
			0.25s	0.50s	0.75s	1.00s	1.25s	1.50s	1.75s	2.00s	Avg
Baseline[78]	RGB – Flow	25.93	**26.15**	25.89	25.79	25.73	25.66	**25.68**	**25.66**	**25.57**	**25.77**
Ours	{RGB + OpenPose} – Flow	24.25	23.11	25.63	**26.72**	26.18	25.57	24.94	24.40	23.94	25.06
Ours	RGB – OpenPose	**37.57**	25.54	**25.93**	26.44	**26.60**	**26.28**	25.57	24.75	24.00	25.64
Ours	OpenPose – Flow	36.30	21.77	22.59	23.57	23.19	22.28	21.30	20.49	19.83	21.88
Ours	C3D (One-Stream)	35.43	**34.34**	**31.05**	28.22	26.46	25.37	**24.75**	**24.39**	**24.22**	**27.35**
Ours	{C3D (RBG)} – OpenPose	**36.44**	32.98	30.56	**28.37**	**26.61**	**25.38**	24.54	23.78	23.22	26.93
Ours	**I3D**	**55.25**	**52.57**	**46.69**	**41.94**	**38.39**	**35.90**	**34.22**	**33.00**	**32.08**	**39.35**
Ours	{I3D (RGB) + OpenPose} – {I3D (Flow)}	49.21	46.65	40.78	36.42	33.19	30.90	29.42	28.43	27.71	34.19
Ours	{I3D (RGB)} – OpenPose	47.43	44.59	40.08	36.77	34.24	32.37	31.29	30.56	30.06	35.00
Ours	{I3D (RGB)} – {I3D (Flow) + OpenPose}	44.47	29.55	31.92	29.62	27.21	25.63	24.78	24.20	23.68	27.07

Πίνακας 6.5: THUMOS'14: Συγκριτικός πίνακας αποτελεσμάτων TRN

[1] Το baseline μοντέλο ήταν το προηγούμενο state-of-the-art με ακρίβεια 47.2%. Επανυλοποιήθηκε με batch_size 2 ώστε να έχουμε βάση σύγκρισης, γεγονός που έριξε την ακρίβεια στο 25.93%

ώστε να μπορέσουμε να έχουμε μια πληρέστερη εικόνα και να βγάλουμε ασφαλή συμπεράσματα. Η χρήση του OpenPose ως μοναδική motion πληροφορία στα δύο πρώτα μοντέλα (Baseline, RGB), δίνει την μεγαλύτερη ακρίβεια στο task του χρονικού εντοπισμού και αναγνώρισης δράσεων. Η παραπάνω θέση αποδεικνύει ότι η δυναμική της ανθρώπινης κίνησης είναι ιδιαίτερα σημαντική για την online αναγνώριση δράσεων. Από την άλλη πλευρά δεν ισχύει το ίδιο για το τρίτο σύνολο πειραμάτων. Όπως βλέπουμε, η χρήση των OpenPose χαρακτηριστικών αντί των χαρακτηριστικών οπτικής ροής από το I3D δίκτυο ρίχνουν την ακρίβεια του μοντέλου από το 55.25% στο 47.43%. Συμπεραίνουμε λοιπόν, ότι τα χαρακτηριστικά της πόζας κωδικοποιούν πολύ καλά την πληροφορία της κίνησης, χωρίς όμως να ξεπερνούν τις αντίστοιχες δυνατότητες του I3D μοντέλου.



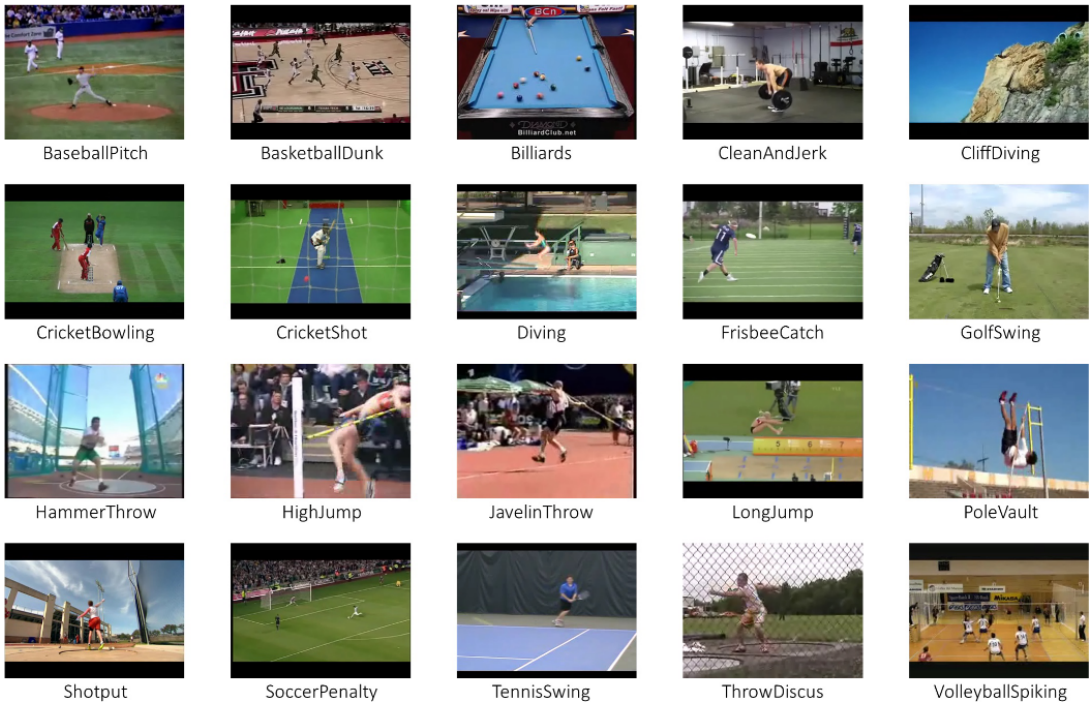
Σχήμα 6.9: THUMOS'14: Οπτικοποίηση των αποτελεσμάτων του καλύτερου δικτύου - I3D-TRN. Η εκτίμηση του δικτύου γίνεται σε πραγματικό χρόνο καθώς δεν γνωρίζει την πληροφορία από τα μελλοντικά καρέ. Επιτυγχάνει χρονικό εντοπισμό πολύ κοντά στα δεδομένα παρατήρησης.

Όπως είδαμε, η χρήση των κατάλληλων motion χαρακτηριστικών επηρεάζει αρκετά την ακρίβεια του μοντέλου, στο χρονικό εντοπισμό και την αναγνώριση δράσεων. Από την άλλη όμως πλευρά, σε όλα τα σύνολα πειραμάτων βλέπουμε, ότι, όσο μικρότερο είναι το ποσοστό της appearance πληροφορίας, τόσο πέφτει η ακρίβεια του anticipation. Η περίπτωση χρήσης μόνο των OpenPose και των flow χαρακτηριστικών από το Bn_Inception αποδεικνύει τον προηγούμενο συλλογισμό. Όπως έχουμε αναφέρει τα χαρακτηριστικά του σκελετού είναι κυρίως motion χαρακτηριστικά. Συνεπώς η παρουσία appearance χαρακτηριστικών στο μοντέλο αυτό είναι μηδαμινή και η ακρίβεια του έχει πέσει στο 21.88% από το 25.77%, που εμφανίζει το baseline μοντέλο. Παράλληλα το απλό C3D δίκτυο, που χρησιμοποιεί πληροφορία μόνο από την appearance ροή, στο task του anticipation, εμφανίζει την μεγαλύτερη ακρίβεια μεταξύ όσων δικτύων δεν χρησιμοποιούν καθόλου I3D χαρακτηριστικά.

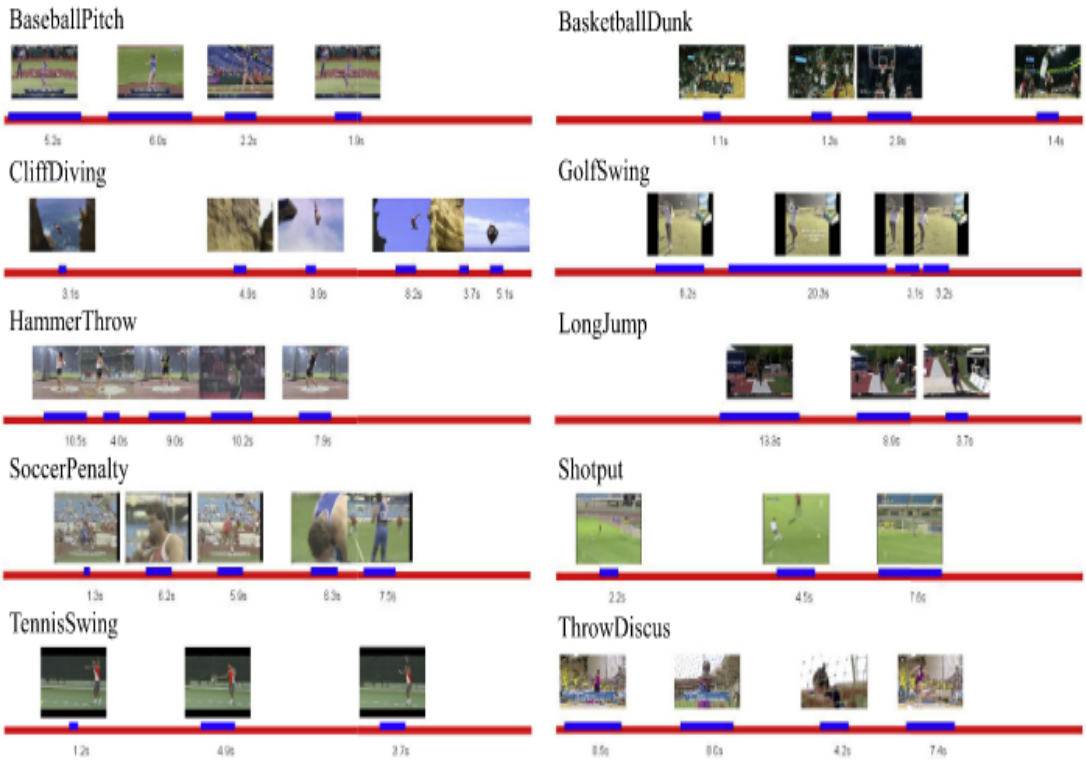
Τέλος, στις περιπτώσεις που το OpenPose χρησιμοποιείται ως μοναδική motion πληροφορία, παρατηρείται μια κορύφωση της ακρίβειας του anticipation στο διάστημα $0.75s - 1.25s$ μπροστά από την χρονική στιγμή της αναγνώρισης. Η παρατηρούμενη αυτή διακύμανση δεν έχει μελετηθεί στην βιβλιογραφία και αποτελεί τροφή για σκέψη και έρευνα.

Συγκρίνοντας τα τρία σύνολα πειραματισμών, συμπεραίνουμε ότι τα τρισδιάστατα συνελικτικά δίκτυα (C3D, I3D) εμφανίζουν μεγαλύτερη ακρίβεια από τα υπόλοιπα δίκτυα, με το I3D να είναι το πιο εύρωστο και το πιο εύκολα γενικεύσιμο.

Το μοντέλο μας επιτυγχάνει state-of-the-art αποτελέσματα, τόσο στο task του χρονικού εντοπισμού και αναγνώρισης δράσεων σε πραγματικό χρόνο, όσο και στο task της πρόβλεψης των μελλοντικών καρέ. Επιτύχαμε ακρίβεια 55.25% για το πρώτο task 39.35% για το δεύτερο task, ξεπερνώντας αρκετά τα εως τώρα state-of-the-art, που είχαν τιμές 47.20% και 38.90% αντίστοιχα.



Σχήμα 6.10: Οι 20 κλάσεις του Thumos'14 συνόλου δεδομένων



Σχήμα 6.11: Παραδείγματα ground truth της THUMOS'14 [25]

Κεφάλαιο 7

Επίλογος και Επεκτάσεις

7.1 Επίλογος

Στη παρούσα διπλωματική εργασία μελετήσαμε σε βάθος το πρόβλημα της αναγνώρισης χειρονομιών και ανθρώπινων δράσεων, τόσο σε offline, όσο και σε online συνθήκες. Το πρόβλημα αυτό αποτελεί ένα από τους κυριότερους τομείς έρευνας της όρασης υπολογιστών και έχει πολλαπλές και πολύ ενδιαφέρουσες κατευθύνσεις.

Η μελέτη μας ξεκίνησε από τη διερεύνηση της βιβλιογραφίας, τόσο πάνω στην αναγνώριση χειρονομιών σε βίντεο, όσο και πάνω στον χρονικό εντοπισμό και την αναγνώριση ανθρωπίνων δράσεων σε πραγματικό χρόνο. Έχοντας αποκτήσει τις απαραίτητες γνώσεις, ξεκινήσαμε τους πειραματισμούς μας για το task της αναγνώρισης χειρονομιών.

Για την αναγνώριση χειρονομιών χρησιμοποιήθηκε το IsoGD σύνολο δεδομένων, που περιέχει τόσο RGB βίντεο, όσο και βίντεο βάθους. Χρησιμοποιήσαμε το OpenPose για να εξάγουμε δισδιάστατους σκελετούς και τον αλγόριθμο TV-L1, για να εξάγουμε οπτική ροή από τα βίντεο μας. Στη συνέχεια πειραματιστήκαμε με αρχιτεκτονικές βαθιών νευρωνικών δικτύων (π.χ. LSTM, C3D, I3D) και με διαφορετικές πληροφορίες εισόδου ώστε να δημιουργήσουμε ακριβή μοντέλα. Όλες οι υλοποιήσεις έγιναν χρησιμοποιώντας είτε την βιβλιοθήκη PyTorch ειτε την βιβλιοθήκη TensorFlow της Python. Τέλος συγκεντρώσαμε τα αποτελέσματα των πειραματισμών μας και συγκρίναμε τα μοντέλα μας. Έχοντας πλέον αποκτήσει την προηγούμενη γνώση μεταβήκαμε στο πρόβλημα της online αναγνώρισης δράσεων.

Για την αναγνώριση δράσεων σε πραγματικό χρόνο, χρησιμοποιήθηκε ένα ιδιαίτερα γνωστό σύνολο δεδομένων, το THUMOS'14. Παράλληλα έχοντας μελετήσει την κατάλληλη βιβλιογραφία επιλέξαμε να διερευνήσουμε την επιρροή του χρονικού context και της δυναμικής της ανθρώπινης κίνησης στον online χρονικό εντοπισμό και την αναγνώριση δράσεων. Πειραματιστήκαμε λοιπόν πάνω στο TRN δίκτυο, που πρότειναν οι Xu et al. [78], τροφοδοτώντας το με ποικίλα appearance και motion χαρακτηριστικά. Τα χαρακτηριστικά αυτά εξήχθησαν από τα μοντέλα που ερευνήσαμε στο πρώτο μέρος αυτής της διπλωματικής (OpenPose, C3D, I3D) αλλά και από τα δύο δίκτυα που προτείνουν οι δημιουργοί του (ResNet-200, Bn_Inception). Με την εργασία μας αυτή επιτύχαμε state-of-the-art, τα οποία υποβάλαμε προς δημοσίευση στο European Signal Processing Conference (EUSIPCO 2021).

Στην επόμενη ενότητα παρουσιάζουμε κάποιες προτάσεις για βελτίωση των παραπάνω μεθόδων και επίτευξη ακόμη καλύτερων αποτελεσμάτων. Ολοκληρώνοντας αυτή την εργασία προσδοκούμε να βοηθήσει έστω και στο ελάχιστο τον αναγνώστη να κατανοήσει αυτή την περιοχή έρευνας και να αποτελέσει τροφή για περαιτέρω εξέλιξη.

7.2 Μελλοντικές Επεκτάσεις

Παρά την αποτελεσματικότητα των μεθόδων μας, οι απαιτήσεις της καθημερινότητας επιζητούν όλο και ακριβέστερα μοντέλα. Συνεπώς η προσπάθεια που κάναμε στην διπλωματική αυτή τόσο στο task του offline gesture recognition όσο και στο task του online action detection and recognition θα μπορούσε ενδεικτικά να επεκταθεί προς τις ακόλουθες κατευθύνσεις:

[Α] Για το task του offline gesture recognition:

- Εφαρμογή των μοντέλων σε πιο μεγάλες βάσεις δεδομένων, ώστε να μπορούν τα αποτελέσματα να γενικεύονται.

[Β] Για το task του online action detection:

- Χρήση διαφόρων μηχανισμών σύνθεσης των appearance και motion χαρακτηριστικών, όπως τα Improved Dense Trajectories (IDT) και Trajectory-Pooled Deep-Convolutional Descriptors (TDD).
- Χρήση διαφορετικών χαρακτηριστικών για την πρόβλεψη της μελλοντικής πληροφορίας και διαφορετικών για τον χρονικό εντοπισμό και αναγνώριση δράσεων.
- Ενσωμάτωση της μεθόδου σε ρομποτική εφαρμογή, ώστε να αξιολογηθεί και στην πράξη η αποδοτικότητα της.

[Γ] Και για τα δύο tasks:

- Χρήση και επιπλέον καναλιών πληροφορίας. Μπορεί να δοκιμαστεί η χρήση του αρνητικού RGB, του saliency και του depth (όπου δεν χρησιμοποιείται).
- Επιπλέον προ-επεξεργασία των δεδομένων. Μπορεί να χρησιμοποιηθεί το RCNN-Masking, που χρησιμοποιείται ιδιαίτερα στα task του object detection. Ενώ παράλληλα πιθανόν θετική να είναι και η επίδραση της αφαίρεσης των background πληροφοριών και της περικοπής των καρέ με τέτοιο τρόπο, ώστε να περιέχουν μόνο τα σημεία ενδιαφέροντος (π.χ. χέρια).
- Χρήση τρισδιάστατων σκελετών, που προσφέρουν πιο ολοκληρωμένη πληροφορία για την κίνηση. Με αυτόν τον τρόπο πιθανόν η δυναμική της ανθρώπινης κίνησης να προσφέρει ακόμα περισσότερα στα tasks της αναγνώρισης δράσεων. Σημαντική επίσης μπορεί να αποβεί η χρήση της ταχύτητας των σκελετών.

Βιβλιογραφία

[1] N. Burrus. Kinect Calibration. http://nicolas.burrus.name/index.php/Research/KinectCalibration#tocLink7S

[2] C. Camacho. Convolutional Neural Networks. https://cezannec.github.io/Convolutional_Neural_Networks/.

[3] Z. Cao, T. Simon, S. Wei, and Y. Sheikh, "OpenPose: realtime multi-person 2D pose estimation using part affinity fields". In IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 1, pp. 172-186, 1 Jan. 2021, doi: 10.1109/TPAMI.2019.2929257.

[4] Z. Cao, T. Simon, S. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In Proc. CVPR, 2017.

[5] J. Carreira and A. Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In Proc. CVPR, 2017.

[6] G. Chalvatzaki, Petros Koutras, A. Tsiami, C. Tzafestas and P. Maragos. i-Walk Intelligent Assessment System: Activity, Mobility, Intention, Communication. In Proc. ECCV Workshops, 2020.

[7] D. Chen, G. Li, Y. Sun, J. Kong, G. Jiang, H. Tang, Z. Ju, H. Yu and H. Liu. An Interactive Image Segmentation Method in Hand Gesture Recognition. In Sensors (Basel, Switzerland) vol. 17, no. 2, pp. 253, 27 Jan. 2017, doi:10.3390/s17020253.

[8] S. Cho, M. H. Maqbool, F. Liu and H. Foroosh. Self-Attention Network for Skeleton-based Human Action Recognition. In Proc. WACV, 2020.

[9] C. Cortes and V. Vapnik. Support-vector networks. In Mach. Learn. vol. 20, no. 3, pp. 273–297 Sept. 1995, doi:https://doi.org/10.1023/A:1022627411411

[10] R. De Geest and T. Tuytelaars. Modeling temporal structure with lstm for online action detection. In Proc. WACV, 2018.

[11] R. De Geest, E. Gavves, A. Ghodrati, Z. Li, C. Snoek, and T. Tuytelaars. Online action detection. In Proc. ECCV, 2016.

[12] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparsespatio-temporal features. In Proc. VS-PETS, 2005.

[13] J. Donahue et al. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. In IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 677-691, 1 April 2017, doi: 10.1109/TPAMI.2016.2599174.

[14] Y. Du, Y. Fu and L. Wang. Representation Learning of Temporal Dynamics for Skeleton-Based Action Recognition. In IEEE Transactions on Image Processing, vol. 25, no. 7, pp. 3010-3022, July 2016, doi: 10.1109/TIP.2016.2552404.

[15] Y. Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. In Proc. CVPR, 2015.

[16] V. Escorcia, F. C. Heilbron, J. C. Niebles, and B. Ghanem. Daps: Deep action proposals for action understanding. In Proc. ECCV, 2016.

[17] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In Proc. CVPR, 2016.

[18] J. Gao, Z. Yang, and R. Nevatia. RED: Reinforced encoder-decoder networks for action anticipation. In Proc. BMVC, 2017.

[19] J. Gao, Z. Yang, C. Sun, K. Chen, and R. Nevatia. TURN TAP: Temporal unit regression network for temporal action proposals. In Proc. ICCV, 2017.

[20] J. Hadfield, G. Chalvatzaki, P. Koutras, M. Khamassi, C. S. Tzafestas and P. Maragos. A Deep Learning Approach for Multi-View Engagement Estimation of Children in a Child-Robot Joint Attention Task. In Proc. IROS, 2019.

[21] C. Harris and M.J. Stephens. A combined corner and edge detector. In Proc. Alvey Vision Conference, 1988.

[22] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proc. CVPR, 2016.

[23] K. He, G. Gkioxari, P. Dollár and R. Girshick. Mask R-CNN. In Proc. ICCV, 2017.

[24] S. Hochreiter and J. Schmidhuber. Long short-term memory. In Neural Comput. vol. 9, no. 8, Nov. 1997, pp. 1735–1780. doi:https://doi.org/10.1162/neco.1997.9.8.1735

[25] H. Idrees, A.R. Zamir, Y. Jiang, A. Gorban, I. Laptev, R. Sukthankar and M. Shah. The THUMOS challenge on action recognition for videos "in the wild". In Computer Vision and Image Understanding, vol. 155, 2017, pp. 1-23, doi:https://doi.org/10.1016/j.cviu.2016.10.018.

[26] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift.In arXiv:1502.03167, 2015.

[27] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. In IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 1, pp. 221-231, Jan. 2013, doi: 10.1109/TPAMI.2012.59.

[28] Y.-G. Jiang, J. Liu, A. R. Zamir, G. Toderici, I. Laptev, M. Shah, R. Sukthankar. THUMOS'14: ECCV Workshop on Action Recognition with a Large Number of Classes. In Proc. ECCV, 2014.

[29] S. Karaman, L. Seidenari, and A. D. Bimbo. Fast saliency based pooling of fisher encoded dense trajectories. In Proc. ECCV, 2014.

[30] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-Scale Video Classification with Convolutional Neural Networks. In Proc. CVPR, 2014.

[31] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, A. Natsev, M. Suleyman, and A. Zisserman. The Kinetics Human Action Video Dataset. In ArXiv, abs/1705.06950, 2017.

[32] Q. Ke, M. Bennamoun, S. An, F. Sohel, F. Boussaid: A New Representation of Skeleton Sequences for 3D Action Recognition. In Proc. CVPR, 2017.

[33] V. Khong and T. Tran. Improving Human Action Recognition with Two-Stream 3D Convolutional Neural Network. In 1st International Conference on Multimedia Analysis and Pattern Recognition (MAPR), 2018.

[34] A. Kläser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3D-gradients. In Proc. BMVC, 2008.

[35] O. Köpüklü, X. Wei and G. Rigoll. You Only Watch Once: A Unified CNN Architecture for Real-Time Spatiotemporal Action Localization. In ArXiv, 2019.

[36] K. Lai and S. N. Yanushkevich. CNN+RNN Depth and Skeleton based Dynamic Hand Gesture Recognition. In Proc. ICPR, 2018.

[37] I. Laptev and T. Lindeberg. Space-time interest points. In Proc. ICCV, 2003.

[38] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In Proc. CVPR, 2008.

[39] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner. Gradient-based learning applied to document recognition. In Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, Nov. 1998, doi: 10.1109/5.726791.

[40] T. Lin, X. Zhao and Z. Fan. Temporal action localization with two-stream segment-based RNN. In Proc. ICIP, 2017.

[41] D. Liu, T. Jiang and Y. Wang. Completeness Modeling and Context Separation for Weakly Supervised Temporal Action Localization. In Proc. CVPR 2019.

[42] F. Long, T. Yao, Z. Qiu, X. Tian, J. Luo and T. Mei. Gaussian Temporal Awareness Networks for Action Localization. In Proc. CVPR, 2019.

[43] S. Ma, L. Sigal, and S. Sclaroff. Learning activity progression in lstms for activity detection and early detection. In Proc. CVPR, 2016.

[44] A. Mittal. Understanding RNN and LSTM. https://aditi-mittal.medium.com/understanding-rnn-and-lstm-f7cdf6dfc14e

[45] F. M. Noori, B. Wallace, Md. Z. Uddin,and J. Torresen: A Robust Human Activity Recognition Approach Using OpenPose, Motion Features, and Deep Recurrent Neural Network. In Proc. SCIA, 2019.

[46] D. Oneata, J. Verbeek, and C. Schmid. Action and Event Recognition with Fisher Vectors on a Compact Feature Set. In Proc. ICCV, 2013.

[47] D. Oneata, J. Verbeek, and C. Schmid. 2014. The LEAR submission at Thumos 2014. In ECCV THUMOS Workshop, 2014.

[48] M. Ramezani and F. Yaghmaee. A review on human action analysis in videos for retrieval applications. Artif. Intell. Rev. 46, 4, 2016, 485–514.

[49] B. Ramsundar, R.B. Zadeh. Chapter 4 Fully Connected Deep Networks. https://www.oreilly.com/library/view/tensorflow-for-deep/9781491980446/ch04.html.

[50] W. Rawat and Z. Wang. Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. In Neur.Comp., 2017, pp. 1-98.

[51] J. Sánchez, E. Meinhardt-Llopis and G. Facciolo. TV-L1 optical flow estimation. In Proc. IPOL, 2013.

[52] G. Serpen and R. H. Khan. Real-time Detection of Human Falls in Progress: Machine Learning Approach. In Proc. CASE, 2018.

[53] A. Shahroudy, J. Liu, T. Ng and G. Wang. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. In Proc. CVPR, 2016.

[54] S. Sharma. Activation Functions in Neural Networks. https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6

[55] Z. Shou, J. Chan, A. Zareian, K. Miyazawa, and S.-F. Chang. CDC: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In Proc. CVPR, 2017.

[56] Z. Shou, J. Pan, J. Chan, K. Miyazawa, H. Mansour, A. Vetro, X. Giroi Nieto, and S.-F. Chang. Online action detection in untrimmed, streaming videos-modeling and evaluation. In Proc. ECCV, 2018.

[57] Z. Shou, D. Wang, S.F. Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In Proc. CVPR, 2016.

[58] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In Proc. NIPS, 2014.

[59] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In Proc. ICLR, 2015.

[60] G. Singh, S. Saha, M. Sapienza, P. Torr and F. Cuzzolin. Online Real-Time Multiple Spatiotemporal Action Localisation and Prediction. In Proc ICCV, 2017.

[61] K. Soomro, A. R. Zamir, M. Shah. UCF101: A Dataset of 101 Human Action Classes from Videos in the Wild, Tech. In Rep. CRCV-TR-12-01, 2012. URL: http://arxiv.org/abs/1212.0402

[62] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. In J. Mach. Learn. Res. vol. 15, 2014, pp. 1929-1958.

[63] D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri. Learning Spatiotemporal Features with 3D Convolutional Networks. In Proc. ICCV, 2015.

[64] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad and S. W. Baik. Action Recognition in Video Sequences using Deep Bi-Directional LSTM With CNN Features. In IEEE Access, vol. 6, pp. 1155-1166, 2018, doi: 10.1109/ACCESS.2017.2778011.

[65] V. I. Vasileiou, N. Kardaris, P.Maragos. Exploring Temporal Context and Human Movement Dynamics for Online Action Detection in Videos. Submitted in Proc. EUSIPCO, 2021.

[66] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones,A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is allyou need. In Proc. NIPS, 2017.

[67] J. Wan, Y. Zhao, S. Zhou, I. Guyon, S. Escalera and S.Z. Li. ChaLearn Looking at People RGB-D Isolated and Continuous Datasets for Gesture Recognition. In Proc. CVPR, 2016.

[68] H. Wang, A. Kläser, C. Schmid et al. Dense Trajectories and Motion Boundary Descriptors for Action Recognition. Int J Comput Vis 103, 60–79 (2013).

[69] H. Wang, M.M. MUllah, A. Kläser, I. Laptev and C. Schmid. Evaluation of Local Spatio-temporal Features for Action Recognition. In Proc. BMVC, 2009.

[70] L. Wang, Y. Qiao, and X. Tang. Action recognition and detection by combining motion and appearance features. In ECCV THUMOS Workshop, 2014

[71] L. Wang, Y. Qiao, and X. Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In Proc. CVPR, 2015.

[72] H. Wang and C. Schmid. Action recognition with improved trajectories. In Proc. ICCV, 2013.

[73] L. Wang and Y.X. Wang. Dense_Flow tool. In GitHub, 2016. https://github.com/yjxiong/dense_flow.

[74] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: towards good practices for deep action recognition. In Proc. ECCV, 2016.

[75] G. Willems, T. Tuytelaars, and L. Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In Proc. ECCV, 2008.

[76] Y. Xiong, L. Wang, Z. Wang, B. Zhang, H. Song, W. Li, D. Lin, Y. Qiao, L. Van Gool and X. Tang. CUHK & ETHZ & SIAT submission to activitynet challenge 2016. In arXiv:1608.00797, 2016.

[77] H. Xu, A. Das, and K. Saenko. R-C3D: Region convolutional 3d network for temporal activity detection. In Proc. ICCV, 2017.

[78] M. Xu and M. Gao and Y. Chen and L. S. Davis and D. J. Crandall: Temporal Recurrent Networks for Online Action Detection. In Proc. ICCV, 2019.

[79] H. Yakura et al. . Malware Analysis of Imaged Binary Samples by Convolutional Neural Network with Attention Mechanism. In Mar, 2018, pp. 127–134.

[80] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime TV-L1 optical flow. In Trans. Pattern Recognition, 2007, pp. 214–223.

[81] L. Zhang, G. Zhu, L. Mei, P. Shen, S. Shah and M. Bennamoun. Attention in Convolutional LSTM for Gesture Recognition. In Proc. NeurIPS, 2018.

[82] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie. Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In Proc. AAAI, 2016.

[83] C. Zimmermann, T. Welschehold, C. Dornhege, W. Burgard and T. Brox. 3D Human Pose Estimation in RGBD Images for Robotic Task Learning. In Proc. ICRA, 2018.

[84] A. Zlatintsi, A.C. Dometios, N. Kardaris, I. Rodomagoulakis, P. Koutras, X. Papageorgiou, P. Maragos, C.S. Tzafestas, P. Vartholomeos, K. Hauer, C. Werner, R. Annicchiarico, M.G. Lombardi, F. Adriano, T. Asfour, A.M. Sabatini, C. Laschi, M. Cianchetti, A. Güler, I. Kokkinos, B. Klein, and R. López. 2020. I-Support: A robotic platform of an assistive bathing robot for the elderly population. In Robot. Auton. Syst., 2020.