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

Αυτόματη Εκτίμηση **Engagement** κατά την αλληλεπίδραση παιδιών με ρομπότ με τη χρήση πόζας

Διπλωματική Εργασία
της
Δάφνης Αναγνωστοπούλου

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Δεκέμβριος, 2020

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

Αυτόματη Εκτίμηση Engagement κατά την αλληλεπίδραση παιδιών με ρομπότ με τη χρήση πόζας

Διπλωματική Εργασία
της
Δάφνης Αναγνωστοπούλου

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 7η Δεκεμβρίου 2020.

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

Δεκέμβριος, 2020

.....
Δάφνη Γ. Αναγνωστοπούλου
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Copyright © Δάφνη Γ. Αναγνωστοπούλου, 2020
Με επιφύλαξη παντός δικαιώματος. All rights reserved.

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

Περίληψη

Σκοπός αυτής της διπλωματικής εργασίας είναι η αυτόματη εκτίμηση του engagement κατά την αλληλεπίδραση παιδιών με ρομπότ με την αξιοποίηση της πόζας του ανθρώπινου σώματος. Η εκτίμηση του engagement αποτελεί απαραίτητη προϋπόθεση για την οικοδόμηση φυσικής αλληλεπίδρασης μεταξύ παιδιού και ρομπότ. Ειδικά στην περίπτωση παιδιών με διαταραχές αυτιστικού φάσματος, η δυνατότητα των ρομπότ να εκτιμούν αυτόματα το engagement των παιδιών τους επιτρέπει να προσαρμόζουν τη συμπεριφορά τους ανάλογα με τους εκάστοτε εκπαιδευτικούς και θεραπευτικούς σκοπούς. Πραγματοποιούμε εκτίμηση του engagement με τη χρήση της πόζας τόσο για παιδιά με διαταραχές αυτιστικού φάσματος όσο και για τυπικώς αναπτυσσόμενα παιδιά. Δουλέψαμε με αρκετές διαφορετικές βαθιές αρχιτεκτονικές και καταλήξαμε σε συνελικτικές αρχιτεκτονικές που υπερτερούν σε σχέση με προηγούμενες προτεινόμενες μεθόδους, είτε χρησιμοποιούν τρισδιάστατες είτε δισδιάστατες συντεταγμένες των σημείων του σκελετού των παιδιών. Ακόμη, για να αξιολογήσουμε τη δυνατότητα γενίκευσης των μοντέλων μας, τα έχουμε εφαρμόσει σε διαφορετικά σύνολα δεδομένων με διαφορετικές καταστάσεις και συμμετέχοντες κατά τις αλληλεπιδράσεις. Τα μοντέλα μας επιτυγχάνουν σημαντικά αποτελέσματα στην εκτίμηση του engagement των παιδιών σε ποικιλία συνθηκών, καταστάσεων και αλληλεπιδράσεων.

Λέξεις Κλειδιά

Engagement, Αλληλεπίδραση Παιδιού-Ρομπότ, Διαταραχές Αυτιστικού Φάσματος, Κοινωνικά ρομπότ, Πόζα, OpenPose, Βαθιά Συνελικτικά Δίκτυα.

Abstract

In this work we estimate children engagement during Child Robot Interaction using pose data. Estimating the engagement of children is an essential prerequisite for constructing natural Child-Robot Interaction. Especially in the case of children with Autism Spectrum Disorder, monitoring the engagement of the other party allows robots to adjust their actions according to the educational and therapeutic goals in hand. We delve into engagement estimation using pose data for both children with autism spectrum disorders and typically developing children. We have been working with several different deep neural network architectures and have concluded to the ones that outperform previous methods when they use either 3D or 2D coordinates of pose keypoints. and explore their performance under variable conditions, in different databases depicting ASD and TD children interacting with robots or humans. Moreover, in order to evaluate the generalization of our models, we test them on different data sets with different situations and participants during the interaction. Our resulting models achieve important success in engagement estimation for children in a variety of conditions, situations and interactions.

Key Words

Engagement, Child-Robot Interaction, Autism Spectrum Disorders, Social Robots, Pose, OpenPose, Deep Convolutional Networks.

Ευχαριστίες

Αρχικά για την ανάθεση αυτής της διπλωματικής εργασίας, την καθοδήγηση και τις συμβουλές του θα ήθελα να ευχαριστήσω τον καθηγητή μου κ. Πέτρο Μαραγκό. Ακόμη, για τη σημαντική της συμβολή και τις συμβουλές της θα ήθελα να ευχαριστήσω την καθηγήτρια κα. Χριστίνα Παπαηλιού. Τέλος, θα ήθελα να ευχαριστήσω τη Νίκη Ευθυμίου για την πολύτιμη στήριξη και ακούραστη καθοδήγησή της καθόλη την εκπόνηση αυτής της διπλωματικής εργασίας.

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Σχημάτων

Acronyms

ASD Autism Spectrum Disorder. 33

AU Action Units. 40

CNN Convolutional Neural Networks. 17

CRI Child Robot Interaction. 12

csv comma-separated values. 41

CVAE Conditional Variational Autoencoder. 59

FACS Facial Action Coding System. 40

FC Fully Connected. 66

fps frame per second. 68

GCN Graph Convolutional Network. 56

GRU Gated Recurrent Unit. 63

GTAP Global Temporal Average Pooling. 53

HMM Hidden Markov Model. 59

HRI Human Robot Interaction. 12

ICC intraclass correlation. 21

ICP Iterative Closest Point. 28

LDA Linear Discriminant Analysis. 15

LSTM Long-Short Term Memory. 17

PAFs Part Affinity Fields. 48

PSMs Pictorial Structures Models. 45

ReLU Rectified Linear Unit. 40

RGB Red-Green-Blue. 22

RGB-D Red-Green-Blue-Depth. 37

RNN Recurrent Neural Network. 53

ST-GCN Spatial Temporal Graph Convolutional Network. 58

SVM Support Vector Machines. 15

TCN Temporal Convolutional Network. 53

Κεφάλαιο 1

Εισαγωγή

1.1 Περιγραφή του προβλήματος

Τα κοινωνικά ρομπότ (social robots) εντάσσονται ολοένα και περισσότερο στην καθημερινότητά μας. Μία από τις πτυχές της κοινωνικής ζωής στις οποίες αξιοποιούνται είναι η εκπαίδευση των παιδιών, και ειδικότερα των παιδιών με διαταραχές αυτιστικού φάσματος [16, 17]. Για να επιτευχθεί όμως ποιοτική αλληλεπίδραση μεταξύ παιδιών και ρομπότ είναι απαραίτητο τα ρομπότ να έχουν τη δυνατότητα να εκτιμούν κάθε στιγμή το βαθμό στον οποίο ανταποκρίνονται τα παιδιά στο περιεχόμενο της μεταξύ τους αλληλεπίδρασης [18]. Ένα βασικό χαρακτηριστικό που μαρτυρά την ανταπόκριση ενός ατόμου σε μία αλληλεπίδραση στην οποία συμμετέχει είναι το engagement. Στη βιβλιογραφία συναντάμε, όπως αναφέρουμε και στο επόμενο κεφάλαιο, πλήθος διαφορετικών ορισμών για το engagement. Εισαγωγικά, μπορούμε να πούμε ότι το επίπεδο του engagement εκφράζει το βαθμό στον οποίο ένα άτομο προσέχει - παρακολουθεί τον παρτενέρ του, ενώ ταυτόχρονα συνεργάζεται ενεργά με αυτόν για την επίτευξη ενός κοινού στόχου στα πλαίσια της αλληλεπίδρασής τους. Όταν το ρομπότ έχει τη δυνατότητα να εκτιμά αυτόματα το επίπεδο του engagement τότε μπορεί να προσαρμόζει τη συμπεριφορά του ώστε να βελτιώνεται η ποιότητα της αλληλεπίδρασής του με τα παιδιά και τελικά να επιτυγχάνονται οι εκπαιδευτικοί και θεραπευτικοί στόχοι της διαδικασίας.

Έτσι, λοιπόν, το πρόβλημα που μας απασχολεί στην παρούσα διπλωματική εργασία είναι η αυτόματη εκτίμηση του engagement κατά τις αλληλεπιδράσεις παιδιών με ρομπότ. Ο ερευνητικός στόχος που θέτουμε είναι η σχεδίαση και υλοποίηση μίας μεθόδου που να εκτιμά το επίπεδο engagement των παιδιών χρησιμοποιώντας δεδομένα video από τις αλληλεπιδράσεις τους με κοινωνικά ρομπότ. Στόχος μας είναι να εφαρμόσουμε τη μέθοδό μας σε διαφορετικά είδη δεδομένων, σε αλληλεπιδράσεις τυπικώς αναπτυσσόμενων παιδιών με ρομπότ, ποικίλες αλληλεπιδράσεις παιδιών με αυτισμό με ρομπότ, αλληλεπιδράσεις παιδιών με αυτισμό με τις μητέρες τους στο περιβάλλον του σπιτιού τους κ.α. Κεντρικός άξονας στην παρούσα εργασία είναι η αξιοποίηση της πόζας ώστε να επιτευχθεί επιτυχημένη εκτίμηση του engagement.

1.2 Εφαρμογή

Τα κοινωνικά ρομπότ μπορούν να βοηθήσουν παιδιά με αυτισμό να καλλιεργήσουν δεξιότητες και να αποκτήσουν γνώσεις καθώς έχουν δείξει σημαντικά πλεονεκτήματα κατά τη διάρκεια αλληλεπιδράσεων με εκπαιδευτικούς και θεραπευτικούς σκοπούς [19, 20, 21, 22, 23]. Κατά την αλληλεπίδρασή τους με ρομπότ, τα παιδιά με αυτισμό δείχνουν μεγαλύτερο

ενδιαφέρον και είναι περισσότερο συγκεντρωμένα, σε σχέση με αντίστοιχες αλληλεπιδράσεις με ανθρώπους. Επιπλέον, είναι πιθανότερο να διατηρήσουν πιο ήρεμη διάθεση, αισθάνονται πιο άνετα να εκφραστούν συναισθηματικά και εμπλέκονται ενεργότερα στην αλληλεπίδραση [24, 25]. Τα ευρήματα αυτά δείχνουν ότι τα παιδιά μπορούν να ωφεληθούν σημαντικά από την αλληλεπίδρασή τους με κοινωνικά ρομπότ. Μία από τις σημαντικότερες προκλήσεις στις αλληλεπιδράσεις αυτές είναι η διατήρηση του engagement των παιδιών. Συνεπώς, η δυνατότητα αυτόματης εκτίμησης του είναι ζωτικής σημασίας για να μπορέσουν σε μεγάλη κλίμακα τα κοινωνικά ρομπότ να βοηθήσουν τα παιδιά με αυτισμό στην αντιμετώπιση των συνεπειών της διαταραχής αυτής.

Γενικότερα, σε όλες τις εφαρμογές που περιλαμβάνουν μεγάλης διάρκειας αλληλεπιδράσεις ανθρώπου-ρομπότ (HRI) ή και παιδιού-ρομπότ (CRI), η δυνατότητα να αναγνωρίζει το ρομπότ το επίπεδο του engagement που παρουσιάζουν οι άνθρωποι που λαμβάνουν μέρος στη διαδικασία αλληλεπίδρασης είναι πολύ χρήσιμη. Το επίπεδο του engagement μπορεί κατ΄ αρχάς να αξιοποιηθεί ως ένα μέτρο αξιολόγησης της συμπεριφοράς του ρομπότ. Χρησιμοποιώντας την πληροφορία αυτή είναι δυνατό να βελτιωθεί η συμπεριφορά του ρομπότ ώστε να μεγιστοποιηθεί το engagement του χρήστη και κατ΄ επέκταση και τα οφέλη που αποκομίζει από τη διαδικασία, συνολικά να βελτιστοποιηθεί η εμπειρία του.

Κατά τη διάρκεια εκπαιδευτικών διαδικασιών, όπως αυτές που λαμβάνουν χώρα σε ένα σχολείο ή σε ένα μουσείο, η δυνατότητα να προσελκύεται το ενδιαφέρον των συμμετεχόντων με τρόπο τέτοιο ώστε να αφοσιώνονται στην αλληλεπίδραση είναι τεράστιας σημασίας. Είναι γνωστό ότι υψηλότερα επίπεδα engagement οδηγούν σε καλύτερα αποτελέσματα εκμάθησης, ενώ παρόμοιο αποτέλεσμα φαίνεται να έχει και το engagement των παιδιών στα πλαίσια μιας εκπαιδευτικής δραστηριότητας που πραγματοποιείται με τη συμμετοχή ενός ρομπότ. [1].

1.3 Προκλήσεις

Μία από τις βασικότερες δυσκολίες που αντιμετωπίζουμε για την επίτευξη του στόχου μας, προκύπτει από την ίδια τη φύση του χαρακτηριστικού που προσπαθούμε να εκτιμήσουμε. Είναι δύσκολο να περιγράψουμε τι είναι εκείνο που καθιστά ένα παιδί engaged στην αλληλεπίδραση στην οποία συμμετέχει παρ΄ όλο που το χαρακτηριστικό αυτό είναι ζωτικής σημασίας για να αποκομίσει το παιδί τα μέγιστα οφέλη από την αλληλεπίδραση αυτή. Ενώ το engagement είναι μία εσωτερική ψυχική - νοητική κατάσταση των παιδιών που αλληλεπιδρούν με το ρομπότ, το ίδιο το ρομπότ (όπως και κάποιος άνθρωπος παρατηρητής) είναι υποχρεωμένο να περιοριστεί στην ανάλυση εξωτερικών χαρακτηριστικών (όψη, λόγος, ήχος) για να εξάγει συμπέρασμα για το επίπεδό του [26]. Η δυσκολία αυτή αναδεικνύεται και από το γεγονός ότι ο ίδιος ο ορισμός του engagement διαφέρει σημαντικά από πηγή σε πηγή στη βιβλιογραφία. Τελικά, αποτελεί σημαντική πρόκληση να επιλέξουμε το συνδυασμό εκείνο εξωτερικών χαρακτηριστικών και τη μέθοδο με την οποία μπορούμε να τα επεξεργαστούμε ώστε να μας δώσουν μία επιτυχημένη εκτίμηση του επιπέδου του engagement.

Επιπλέον, για την εκτίμηση του engagement σε παιδιά που εμφανίζουν διαταραχές του αυτιστικού φάσματος η δυσκολία πολλαπλασιάζεται. Αποτελέσματα πειραμάτων μας δείχνουν ότι είναι δυσκολότερη η εκτίμηση του engagement σε αυτή την περίπτωση [27], καθώς για παράδειγμα ενδείξεις που είναι ιδιαίτερα βοηθητικές για τον υπολογισμό του engagement όπως το σταθερό βλέμμα, το άνοιγμα και το κλείσιμο των ματιών, η θέση του κεφαλιού, αλλά και άλλες, δεν εμφανίζονται σε τόσο άμεση σύνδεση με την προσήλωση σε παιδιά με διαταραχές αυτιστικού φάσματος [28].

Επιπρόσθετα, στις περιπτώσεις αλληλεπιδράσεων στις οποίες τα παιδιά είναι καθισμένα και στατικά δημιουργείται μία ακόμη δυσκολία καθώς οι πόζες τους, στις οποίες όπως θα δείξουμε στη συνέχεια βασιζόμαστε σε μεγάλο βαθμό για να εκτιμήσουμε το engagement, παρουσιάζουν πολύ μικρή ποικιλία δυσκολεύοντας την εκτίμηση.

Τέλος, στα σύνολα δεδομένων που περιλαμβάνουν αλληλεπιδράσεις παιδιών με ρομπότ τα στιγμιότυπα στα οποία τα παιδιά είναι πλήρως engaged στην αλληλεπίδραση, καθώς και τα στιγμιότυπα στα οποία τα παιδιά είναι πλήρως disengaged είναι πολλές φορές συντριπτικά λιγότερα από εκείνα τα στιγμιότυπα στα οποία τα παιδιά παρουσιάζουν ενδιάμεσο επίπεδο engagement. Ωστόσο, η σωστή εκτίμηση και των οριακών καταστάσεων είναι ιδιαίτερα σημαντική, καθώς για παράδειγμα όταν τα παιδιά είναι πλήρως disengaged το ρομπότ πρέπει δραστικά να προσαρμόσει τη συμπεριφορά του για να κερδίσει το ενδιαφέρον και την προσοχή τους. Το γεγονός ότι οι οριακές καταστάσεις δεν είναι τόσο συχνές στα δεδομένα αυξάνει τη δυσκολία να εκτιμώνται σωστά.

1.4 Δομή της διπλωματικής

Η παρούσα διπλωματική αποτελείται από έξι κεφάλαια, το πρώτο από τα οποία είναι εισαγωγικό.

Στο δεύτερο κεφάλαιο ορίζουμε πληρέστερα το πρόβλημα εκτίμησης του engagement και παρουσιάζουμε βασικές στρατηγικές αντιμετώπισης του εν λόγω προβλήματος που εφαρμόζονται τα τελευταία χρόνια. Εξετάζουμε διάφορες επιλογές προσέγγισης της λύσης του προβλήματος.

Στο τρίτο κεφάλαιο παρουσιάζουμε αναλυτικά τα σύνολα δεδομένων με τα οποία δουλέψαμε (BabyRobot, BabyAffect, Pinsoro), καθώς και τις διαδικασίες με τις οποίες επεξεργαστήκαμε τα δεδομένα ώστε να μπορούμε να τα χρησιμοποιήσουμε στην εκτίμηση.

Στο τέταρτο κεφάλαιο αρχικά παρουσιάζουμε μία ανασκόπηση των μεθόδων εκτίμησης της ανθρώπινης πόζας καθώς και της εξέλιξής τους τα προηγούμενα χρόνια. Στη συνέχεια, παρουσιάζουμε αναλυτικά τον αλγόριθμο OpenPose. Πρόκειται για το εργαλείο που χρησιμοποιούμε στην παρούσα διπλωματική για την εξαγωγή των σημείων του ανθρώπινου σκελετού από τα δεδομένα. Τέλος, συμπεριλαμβάνουμε διάφορες εφαρμογές και μεθόδους αναγνώρισης ανθρώπινης δράσης στις οποίες αξιοποιούνται ως είσοδοι τα σημεία του σκελετού των ανθρώπων. Ορισμένα στοιχεία από τις μεθόδους αυτές έχουμε υλοποιήσει και προσαρμόσει και στην παρούσα διπλωματική.

Στο πέμπτο κεφάλαιο παρουσιάζουμε τις μεθόδους που υλοποιήσαμε για την εκτίμηση του engagement καθώς και τα σημαντικότερα αποτελέσματα που πήραμε από τα πειράματά μας στα διάφορα σύνολα δεδομένων. Ξεκινώντας από αλληλεπιδράσεις του συνόλου δεδομένων BabyRobot στις οποίες συμμετείχαν τυπικώς αναπτυσσόμενα παιδιά αρχικά υλοποιήσαμε αναδρομικά δίκτυα για την εκτίμηση του engagement. Ύστερα, προχωρήσαμε στην υλοποίηση διάφορων συνελικτικών δικτύων. Στη συνέχεια, αξιοποιήσαμε τα καλύτερα από τα δίκτυά μας ώστε να επεκτείνουμε την εκτίμηση του engagement σε παιδιά με διαταραχές αυτιστικού φάσματος που αλληλεπιδρούσαν τόσο στην ίδια συνθήκη με τα τυπικώς αναπτυσσόμενα παιδιά όσο και σε διαφορετικά παιχνίδια (BabyRobot). Ακόμη επεκτείναμε την εκτίμηση του επιπέδου engagement των παιδιών στο σύνολο δεδομένων BabyAffect, σε πειράματα από σχολείο στον Πειραιά (BabyRobot) αλλά και στο σύνολο δεδομένων Pinsoro.

Τέλος, στο έκτο κεφάλαιο προχωράμε στη σύνοψη των συμπερασμάτων που προέκυψαν από τη μελέτη και τα πειράματα της παρούσας διπλωματικής εργασίας, ενώ προτείνουμε επίσης ορισμένες πιθανές κατευθύνσεις για τη συνέχεια.

Κεφάλαιο 2

Εκτίμηση του engagement στη βιβλιογραφία

Στο Κεφάλαιο αυτό παραθέτουμε χαρακτηριστικές προσεγγίσεις αντιμετώπισης του προβλήματος εκτίμησης του engagement. Παρατηρούμε ότι η επιλογή της προσέγγισης του προβλήματος εξαρτάται σημαντικά από τον τρόπο που ορίζεται προηγουμένως το engagement. Οι διαφορετικές προσεγγίσεις που συναντούμε στη βιβλιογραφία και επεκτείνουν τη συνολική μας γνώση για την καλύτερη αντιμετώπισή του προβλήματος εμφανίζουν μεγάλη ποικιλομορφία. Παρακάτω, έχουμε οργανώσει τα περιεχόμενα του Κεφαλαίου με τέτοιο τρόπο ώστε να αναδεικνύονται οι πιο σημαντικές κατευθύνσεις σε διάφορες πτυχές των μεθόδων που προτείνονται.

2.1 Ορισμός engagement

Μία από τις θεμελιώδεις προκλήσεις στην εκτίμηση του engagement αποτελεί το ευρύ φάσμα των τρόπων με των οποίων το ορίζουμε, καθώς επίσης και των τρόπων με τους οποίους το αναπαριστούμε υπολογιστικά στα μοντέλα μας. Συμπεριλαμβάνουμε κάποιους από τους ορισμούς που συναντάμε στη βιβλιογραφία και που εισάγουν ορισμένα σημαντικά στοιχεία για τους τρόπους προσέγγισης του προβλήματος: Στη δουλειά των Chorianopoulou et al. [27] το engagement περιγράφεται ως μία πτυχή της εμπειρίας του χρήστη η οποία χαρακτηρίζεται από ιδιότητες όπως η πρόκληση, η θετική επίδραση, η διάρκεια, το αισθητικό ενδιαφέρον, η συγκέντρωση των αισθήσεων, η προσοχή, η ανάδραση, η ποικιλία/ καινοτομία, η διαδραστικότητα καθώς και η αίσθηση ελέγχου από το χρήστη. Στο The Role of Dialogue in Human Robot Interaction [29] βρίσκουμε ότι το engagement μπορεί να οριστεί ως μία διαδικασία που εμπλέκει δύο ή περισσότερους εταίρους, οι οποίοι αλληλεπιδρούν μεταξύ τους εντός ενός πλαισίου και βασίζεται σε κοινές συγκεκριμένες πτυχές της κατάστασης τους όπως η αντίληψη του περιβάλλοντος, ένας κοινός στόχος κλπ. Ακόμη στους Khamassi et al. [30] ορίζεται ως η διαδικασία με την οποία οι αλληλεπιδρώντες ξεκινούν, διατηρούν, και τελειώνουν την κοινώς αντιληπτή μεταξύ τους σύνδεση κατά τη διάρκεια μίας κατάστασης αλληλεπίδρασης. Τέλος, βρίσκουμε στο [31] ότι το engagement είναι το επίπεδο στο οποίο ένας συμμετέχων συμβάλλει στο στόχο του να βρίσκεται μαζί με άλλους συμμετέχοντες κατά τη διάρκεια μίας κοινωνικής αλληλεπίδρασης και το κατά πόσο διατηρεί αυτή την αλληλεπίδραση, προτείνεται δηλαδή ότι είναι η ποσοτικοποίηση του πόσο «μαζί» βρίσκονται οι αλληλεπιδρώντες. Ο τρόπος με τον οποίο ορίζουμε το engagement μπορεί να επηρεάσει πολλές πτυχές της αντιμετώπισης του προβλήματος, όπως η επιλογή του περιβάλλοντος στο οποίο λαμβάνει χώρα, τα

χαρακτηριστικά που αξιοποιούνται για τη λύση του, τις κλάσεις που επιλέγονται για την αξιολόγηση, τις μετρικές αξιολόγησης της λύσης κ.α. Το engagement συνδέεται άμεσα και με την έννοια της «κοινής προσοχής» (joint attention), τη διαδικασία δηλαδή κατά την οποία το άτομο προσανατολίζει μαζί με άλλους και επικεντρώνει την προσοχή του ώστε να μάθει επιτυχώς από το περιβάλλον του [32].

2.2 Στρατηγικές Αντιμετώπισης του Προβλήματος

Με δεδομένες τις παραπάνω προσπάθειες ορισμού του engagement παρατηρούμε πως στις διάφορες υπάρχουσες προσεγγίσεις επίλυσης του προβλήματος βρίσκουμε διαφορετικές επιλογές ως προς τις κλάσεις του προβλήματος, τα χαρακτηριστικά με τα οποία εργάζονται αλλά και τον τρόπο εξαγωγής τους, την αρχιτεκτονική του νευρωνικού δικτύου που χρησιμοποιείται και τον τρόπο εκπαίδευσής του, την έκταση χρήσης επισημειώσεων των δεδομένων, την εξατομίκευση ή όχι του μοντέλου, τις μετρικές αξιολόγησης του παραγόμενου μοντέλου. Στη βιβλιογραφία ξεχωρίζουμε δύο βασικούς άξονες προσέγγισης του προβλήματος.

2.2.1 Εκτίμηση του engagement με αξιοποίηση δευτερογενών χαρακτηριστικών όπως η πόζα

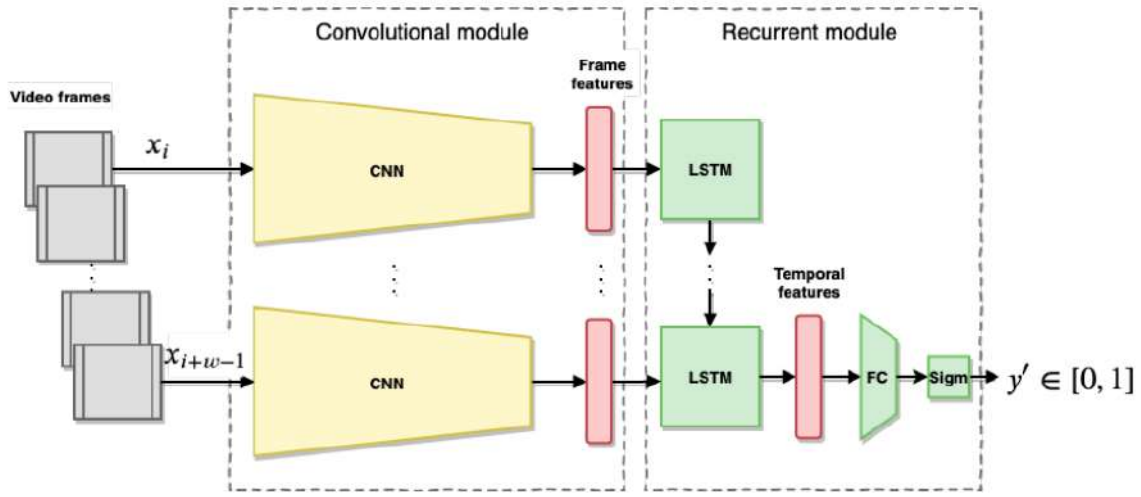
Ο πρώτος άξονας προσέγγισης του προβλήματος βασίζεται στην ανίχνευση της παρουσίας συγκεκριμένων ενδείξεων για το engagement με τη βοήθεια διάφορων μεθόδων όρασης υπολογιστών (ή/και επεξεργασίας φωνής). Ο δεύτερος άξονας βασίζεται στη χρήση ισχυρών ταξινομητών με επίβλεψη που εκπαιδεύονται με κατάλληλα χαρακτηριστικά και δεδομένα στην επίλυση του συγκεκριμένου προβλήματος. Ξεκινώντας από τον πρώτο άξονα, οι πρώτες προσπάθειες εκτίμησης του engagement βασίζονταν κυρίως στην αξιολόγηση του βλέμματος και της ομιλίας των συμμετεχόντων. Υπάρχουν αρκετές προσπάθειες που βασίζονται σε μη λεκτικές ενδείξεις για το engagement όπως η κατεύθυνση του βλέμματος, η στάση του σώματος, οι εκφράσεις του προσώπου, καθώς και συμπεριφορές σχετικές με την παρατηρούμενη εργασία. Επίσης έχουν γίνει προσπάθειες που βασίζονται αποκλειστικά στην κατεύθυνση του βλέμματος καθώς και προσπάθειες που βασίζονται στην ανθρώπινη πόζα αναφορικά με τη θέση του ρομπότ. Για την εξαγωγή των επιθυμητών χαρακτηριστικών από τα video έχουν χρησιμοποιηθεί διάφορα εργαλεία αυτόματης εύρεσης ανθρώπινης πόζας, όπως τα OpenFace, OpenPose, OpenSMILE, αλλά και άλλοι αλγόριθμοι [33]. Στη συνέχεια, για την εξαγωγή συμπερασμάτων από τα δεδομένα αυτά έχουν χρησιμοποιηθεί διάφορες αρχιτεκτονικές όπως Support Vector Machines (SVM), decision trees, Linear Discriminant Analysis (LDA), βαθιά νευρωνικά δίκτυα κ.α.

Η δουλειά των Hadfield et al. [2] αποτελεί ένα παράδειγμα στο οποίο ακολουθείται τέτοια στρατηγική. Εδώ η εκτίμηση του engagement γίνεται σε αλληλεπιδράσεις από το σύνολο δεδομένων BabyRobot, το οποίο αξιοποιούμε και στην παρούσα διπλωματική εργασία και για το οποίο δίνουμε περισσότερες λεπτομέρειες αργότερα. Συνοπτικά, σε αυτή την προσέγγιση, μέσω κατάλληλης σύνθεσης των εικόνων ενός δικτύου καμερών επιτυγχάνεται μία καλή εκτίμηση της πόζας των παιδιών κατά την οποία αντιμετωπίζονται και περιπτώσεις στις οποίες δε φαίνονται ορισμένα μέρη του σώματος, ενώ ταυτόχρονα εξασφαλίζεται η ελεύθερη κίνηση των παιδιών στο δωμάτιο (σε αντίθεση με περιπτώσεις αναγκαστικής παραμονής μπροστά από μία κάμερα). Η ανθρώπινη πόζα θεωρείται ένα

από τα πιο σημαντικά στοιχεία για την εκτίμηση του engagement. Ωστόσο ο εντοπισμός της αποτελεί ένα ιδιαίτερα δύσκολο πρόβλημα. Με τη βοήθεια του λογισμικού OpenPose παράγεται ένας ακριβής δισδιάστατος προσδιορισμός της από δεδομένα video. Στη συνέχεια, παράγεται μία 3D αναπαράσταση με τη βοήθεια του δισδιάστατου χάρτη δεδομένων από τις κάμερες βάθους και ενός βαθιού νευρωνικού δικτύου. Τελικά, έχοντας αυτή την πληροφορία καθώς και πληροφορία για τη θέση του ρομπότ πρέπει να γίνει επιλογή των χαρακτηριστικών εκείνων που θα βοηθήσουν στην εκτίμηση του engagement. Τα χαρακτηριστικά που επιλέγονται και παράγονται από τα σημεία του σκελετού στη συγκεκριμένη περίπτωση είναι η γωνία ανάμεσα στο βλέμμα του παιδιού και το ρομπότ, η γωνία ανάμεσα στον προσανατολισμό του κορμιού του παιδιού και το ρομπότ, η απόσταση των χεριών από τους αντίστοιχους ώμους.

2.2.2 Εκτίμηση του engagement απευθείας από τις εικόνες

Στο δεύτερο άξονα, δεν παράγονται από τις εικόνες χαρακτηριστικά υψηλότερου επιπέδου που μπορεί να σχετίζονται με το engagement, όπως η πόζα, η κατεύθυνση του βλέμματος κλπ. Αντίθετα, χρησιμοποιούνται απευθείας ως χαρακτηριστικά οι τιμές των pixel των frame των video, με τις οποίες εκπαιδεύονται κατάλληλα διαμορφωμένα βαθιά νευρωνικά δίκτυα. Για παράδειγμα, στη δουλειά των Rudovic et al. [34] έχει χρησιμοποιηθεί η αρχιτεκτονική ResNet-50, εκπαιδευμένη προηγουμένως στο dataset ImageNet. Με αυτόν τον τρόπο αποκτώνται χαρακτηριστικά όπως ακμές, γωνίες, σχήματα και άλλες αναπαραστάσεις των δεδομένων που λειτουργούν καλά ως χαρακτηριστικά γενικού σκοπού για εφαρμογές ταξινόμησης εικόνων. Για την εξειδίκευση του προ-εκπαιδευμένου ResNet σε χαρακτηριστικά προσώπων και τη ρύθμιση των βαρών του έχει επίσης χρησιμοποιηθεί το AffectNet (με πάνω από 500.000 επισημειωμένες εικόνες όσον αφορά την ευχαρίστηση/δυσαρέσκεια και την αδιαφορία/ενθουσιασμό). Ένα από τα επιχειρήματα για τη χρήση βαθιών νευρωνικών στην εφαρμογή εκτίμησης του engagement σε παιδιά με διαταραχές αυτιστικού φάσματος είναι ότι οι ενδείξεις που είναι ιδιαίτερα βοηθητικές για τον υπολογισμό του engagement τυπικώς αναπτυσσόμενων παιδιών όπως το βλέμμα, το άνοιγμα και το κλείσιμο των ματιών, η πόζα-θέση του κεφαλιού, αλλά και άλλες, δεν είναι συνήθως τόσο ξεκάθαρες σε παιδιά με διαταραχές αυτιστικού φάσματος.



Σχήμα 2.1: Δίκτυο για την εκτίμηση του engagement που αποτελείται από ένα συνελικτικό και ένα αναδρομικό τμήμα [1].

Στο "Are you still with me?" των Baxter et al. [1] βρίσκουμε μία πρόσφατη προσέγγιση του προβλήματος που ακολουθεί αυτή τη στρατηγική. Εδώ ο στόχος είναι ο αυτόματος υπολογισμός του engagement από ένα ρομπότ-ξεναγό. Ως είσοδος στο δίκτυο εκτίμησης χρησιμοποιούνται οι εικόνες των ανθρώπων των οποίων το engagement πρέπει να εκτιμηθεί, όπως αυτές αποκτώνται από την οπτική του ρομπότ-ξεναγού. Η αρχιτεκτονική του δικτύου που προτείνεται φαίνεται στο Σχήμα 2.1. Το δίκτυο αποτελείται από δύο βασικά τμήματα: ένα συνελικτικό, το οποίο εξάγει από τα καρέ εισόδου χαρακτηριστικά των εικόνων και ένα αναδρομικό, το οποίο συγκεντρώνει τα χαρακτηριστικά αυτά σε ένα χρονικό διάστημα ώστε να παράγει ένα διάνυσμα χρονικών χαρακτηριστικών για τη σκηνή που έχει καταγραφεί.

Το συνελικτικό κομμάτι είναι ένα ResNetXt-50 Convolutional Neural Network, το οποίο είναι προ - εκπαιδευμένο στο σύνολο δεδομένων ImageNet. Τα χαρακτηριστικά των εικόνων αποκτώνται με την ενεργοποίηση της τελευταίας πλήρους στιβάδας του Convolutional Neural Networks (CNN) (με διάσταση 2048) πριν τη Softmax στιβάδα. Το αναδρομικό κομμάτι είναι μία μοναδική στιβάδα Long-Short Term Memory (LSTM) (με 2048 μονάδες) που ακολουθείται από μία πλήρη στιβάδα. Η στιβάδα LSTM λαμβάνει ως είσοδο μία ακολουθία από σύνολα χαρακτηριστικών που αντιστοιχούν σε N καρέ από το συνελικτικό τμήμα του δικτύου και παράγει με τη σειρά της ένα διάνυσμα χαρακτηριστικών που αντιπροσωπεύει ολόκληρη την ακολουθία των N καρέ. Με αυτό τον τρόπο συλλαμβάνει την συμπεριφορά των ανθρώπων μέσα στο χρονικό παράθυρο N. Τα χρονικά χαρακτηριστικά οδηγούνται στην τελική πλήρη στιβάδα, η οποία διαθέτει σιγμοειδή συνάρτηση ενεργοποίησης και παράγει τιμές που ανήκουν στο διάστημα $[0, 1]$. Το αναδρομικό δίκτυο εκπαιδεύεται με τη χρήση ενός πλήρως επισημειωμένου για το engagement συνόλου δεδομένων με καταγραφές ξεναγήσεων από το ρομπότ-ξεναγό, ενώ το συνελικτικό δίκτυο είναι σταθερό.

Ένα βασικό συμπέρασμα που διαπερνά τη βιβλιογραφία είναι ότι καθώς το engagement εξαρτάται σημαντικά από χρονικές πληροφορίες, παρατηρείται σημαντική βελτίωση με την εισαγωγή LSTM στιβάδων στα χρησιμοποιούμενα μοντέλα. Ένας λόγος για αυτό είναι ότι το παιδί εμφανίζει συνήθως το σταθερό επίπεδο engagement για κάποια συνεχόμενα δευτερόλεπτα. Για παράδειγμα, αν η κατεύθυνση του βλέμματος του παιδιού αλλάζει συνεχώς, αυτό καταδεικνύει ότι δεν είναι στοχοπροσηλωμένο, σε αντίθεση με ένα σταθερό βλέμμα [2]. Οι LSTM στιβάδες επιτρέπουν την εκμάθηση χρονικών εξαρτήσεων σε μεγάλη εμβέλεια και έχουν συμβάλει σε επιτυχημένες εφαρμογές αναγνώρισης δράσης και ανάλυσης φωνής. Έτσι, σε όλες τις πρόσφατες προσεγγίσεις του προβλήματος χρησιμοποιούνται τέτοιου τύπου στιβάδες στα νευρωνικά δίκτυα που εκπαιδεύονται.

2.3 Επιλογή Χαρακτηριστικών

Ήδη έχουμε αναφερθεί στην επιλογή ορισμένων χαρακτηριστικών που αξιοποιούνται για τον υπολογισμό του engagement. Μπορούμε εύκολα να συμπεράνουμε από τη βιβλιογραφία ότι η κατεύθυνση του βλέμματος αποτελεί σίγουρα ένα από τα πιο σημαντικά χαρακτηριστικά, αφού η χρήση της είναι πολύ διαδεδομένη στις προσπάθειες υπολογισμού του engagement. Έχουν ωστόσο εξεταστεί πολλοί διαφορετικοί συνδυασμοί τέτοιων χαρακτηριστικών στις διάφορες προσπάθειες προσέγγισης του προβλήματος.

Για παράδειγμα, στη δουλειά των Khamassi et al. [35] το engagement αντιμετωπίζεται απλά ως μέτρο της προσοχής που δείχνει το παιδί σε ένα συγκεκριμένο αντικείμενο που υποδεικνύεται από το ρομπότ. Υπό αυτό το πρίσμα, ο υπολογισμός του γίνεται ως εξής: μετρώνται με μεθόδους όρασης υπολογιστών οι γωνίες κλίσης και περιστροφής του κεφα-

λιού του παιδιού ως προς το συγκεκριμένο αντικείμενο και το engagement υπολογίζεται ως αντίστροφο της διακύμανσης των μεγεθών αυτών.

Στη δουλειά των Feng et al. [18] τα χαρακτηριστικά που επιλέγονται είναι ο προσανατολισμός του προσώπου (και η διάρκειά του σε κάθε θέση), η απόσταση μεταξύ παιδιού και ρομπότ (και η τάση για αλλαγή αυτής) και τέλος η παραγωγή ήχου όπως το κλάμα, η κραυγή και το γέλιο (και η διάρκειά τους). Τα χαρακτηριστικά αυτά δίνονται ως είσοδοι σε ένα μπεϋζιανό δυναμικό δίκτυο για να προκύψει συμπέρασμα για το engagement. Η παραμετροποίηση του δικτύου πραγματοποιείται από ειδικούς, που περιγράφουν ποιοτικά το βαθμό της στοχοπροσήλωσης των παιδιών (πολύ υψηλός, υψηλός, σχετικά υψηλός, ουδέτερος, σχετικά χαμηλός, χαμηλός και πολύ χαμηλός) και στη συνέχεια η περιγραφή αυτή μετατρέπεται σε ποσοτική με τη βοήθεια triangular fuzzy number. Το μοντέλο που προκύπτει με αυτό τον τρόπο έχει σταθερές παραμέτρους, γεγονός που αποτελεί περιορισμό σε περίπτωση που αλλάζει η διαδικασία της αλληλεπίδρασης του παιδιού και του ρομπότ.

Στο "Recognizing engagement in human-robot interaction" [36] χρησιμοποιούνται λίγο διαφορετικά χαρακτηριστικά για την εκτίμηση του engagement από τις προηγούμενες προσεγγίσεις. Εδώ η εκπαίδευση βασίζεται στο άμεσο βλέμμα σε ένα αντικείμενο που υποδεικνύεται από το ρομπότ, στην ανταλλαγή βλέμματος μεταξύ ανθρώπου-ρομπότ, στα ζεύγη αλληλοσύνδεσης. Με τον όρο αυτό περιγράφουμε για παράδειγμα την άμεση απόκριση του ανθρώπου (λεκτική ή μη) σε κάτι που του είπε το ρομπότ ή ορισμένα πιο έμμεσα σημάδια ανταπόκρισης του ανθρώπου σε όσα λέει το ρομπότ, όπως το γνέψιμο, μία παραμόρφωση του προσώπου ή ορισμένοι ήχοι, ή και τους χρόνους καθυστέρησης κατά τη διάρκεια των ενδείξεων αυτών.

Στη δουλειά των Castellano et al. [37] χρησιμοποιούνται ως χαρακτηριστικά το βλέμμα του παιδιού (συγκεντρωμένο ή όχι στο ρομπότ) και το χαμόγελό του. Ωστόσο, στην προσέγγιση αυτή με δεδομένο ότι η στοχοπροσήλωση του παιδιού εξαρτάται άμεσα και από την έκβαση της εργασίας που πραγματοποιούν μαζί με το ρομπότ κατά την αλληλεπίδρασή τους (στη συγκεκριμένη εργασία πρόκειται για παρτίδα σκακιού), όπως και από τις διάφορες εκφράσεις του ρομπότ χρησιμοποιούνται και αυτά ως χαρακτηριστικά για τον υπολογισμό του engagement. Συμπερασματικά, της εργασίας αυτής, τα τελευταία αυτά χαρακτηριστικά παρ' ότι δίνουν χειρότερα αποτελέσματα από τη χρήση των υπολοίπων μπορούν να αξιοποιηθούν συμπληρωματικά όταν είναι δύσκολη η εξαγωγή χαρακτηριστικών για τη συμπεριφορά του χρήστη σε περίπτωση θορυβωδών δεδομένων ή μερικής έλλειψης ορισμένων οπτικών δεδομένων.

Στο "Explorations in engagement for humans and robots" [38], στο οποίο το ζητούμενο είναι ο υπολογισμός του joint attention, δίνεται μεγάλη σημασία στον εντοπισμό της κατεύθυνσης του βλέμματος, ο οποίος γίνεται σε τρία βήματα: στο πρώτο βήμα με τη βοήθεια τεσσάρων καμερών επιτυγχάνεται η εύρεση του προσανατολισμού του προσώπου ενός παιδιού που βρίσκεται καθισμένο απέναντι από το ρομπότ ως προς τις συντεταγμένες μίας από τις τέσσερις κάμερες που σίγουρα θα εντοπίζει το πρόσωπο του παιδιού. Στο δεύτερο βήμα, ο προσανατολισμός του προσώπου μετατρέπεται με κατάλληλο μετασχηματισμό σε πραγματικές συντεταγμένες. Στο τρίτο βήμα με τη χρήση ενός κατάλληλου αλγορίθμου αντιστοιχίζεται ο προσανατολισμός του προσώπου στην κατεύθυνση του βλέμματος.

Αυτά είναι τα βασικότερα από τα χαρακτηριστικά που αξιοποιούνται στη βιβλιογραφία για την εκτίμηση του engagement. Παρατηρούμε ότι η πόζα σίγουρα μπορεί να παίξει πρωτεύοντα ρόλο στην διαδικασία αυτή. Πρώτον, επειδή τα σημεία του σκελετού και του προσώπου μπορούν να αποτελέσουν και από μόνα τους χαρακτηριστικά που δίνουν

πληροφορία για το engagement του παιδιού και άρα που μπορούν να αξιοποιηθούν στην εκπαίδευση σχετικών μοντέλων. Δεύτερον, επειδή τα περισσότερα από τα χαρακτηριστικά υψηλότερου επιπέδου που χρησιμοποιούνται για την εκτίμηση του engagement σε διάφορες προσεγγίσεις, με συνηθέστερο από αυτά την κατεύθυνση του βλέμματος, μπορούν να παραχθούν από τα σημεία της πόζας. Επομένως, στην προσέγγισή μας στην παρούσα διπλωματική θα χρησιμοποιήσουμε εκτεταμένα την πόζα στην προσπάθειά μας να εκτιμήσουμε το engagement.

2.4 Εξατομίκευση των Μοντέλων

Σύμφωνα με πιο πρόσφατες προσπάθειες ένα από τα μειονεκτήματα των κλασικότερων προσεγγίσεων εκτίμησης συμπεριφορών είναι ότι χρησιμοποιούν μοντέλα τα οποία εκπαιδεύονται σε συγκεκριμένα υποκείμενα και τα οποία συνήθως δεν αποδίδουν καλά όταν δοκιμάζονται σε υποκείμενα που δεν έχουν ξανασυναντήσει. Οι προσεγγίσεις "one size fits all" μπορεί να μην είναι τόσο επιτυχημένες εδώ καθώς υπάρχουν σημαντικές διαφορές στην έκφραση του engagement ανάλογα με το άτομο, οι οποίες γίνονται ακόμη πιο έντονες όταν πρόκειται για παιδιά με διαταραχές αυτιστικού φάσματος. Επιπρόσθετα, έχει προταθεί ότι υπάρχουν σημαντικές διαφορές στην έκφραση του engagement που πηγάζουν από το πολιτισμικό υπόβαθρο των παιδιών και εμποδίζουν τα μοντέλα να κάνουν σωστές εκτιμήσεις [39].

Έτσι, σε κάποιες προσπάθειες εκτίμησης του engagement υιοθετείται εξατομικευμένη στρατηγική πρόβλεψης. Στο CultureNet [28] προτείνεται ένα δίκτυο που εκπαιδεύεται αρχικά ως ένα σημείο του με κάποια γενικά χαρακτηριστικά, παγώνει ως το σημείο αυτό, και στη συνέχεια εξειδικεύεται για συγκεκριμένα πολιτισμικά υπόβαθρα. Παρόμοια προσπάθεια θα μπορούσε να γίνει με εξειδίκευση του δικτύου ανάλογα με την ηλικία ή το φύλο του παιδιού, αν παρατηρούνται σταθερές διαφορές στην εκδήλωση του engagement ανάλογα με αυτά τα χαρακτηριστικά των παιδιών.

Ένας άλλος τρόπος εκμάθησης που προτείνεται από τους Rudovic et al. [34] με τη χρήση active learning και reinforcement learning είναι ο εξής: το μοντέλο αποφασίζει είτε να εκτιμήσει το επίπεδο engagement του παιδιού είτε να ζητήσει επισημείωση όταν δεν είναι βέβαιο. Η διαδικασία αυτή πραγματοποιείται με τη βοήθεια αλγορίθμου Q-learning με θετικές ή αρνητικές ανταμοιβές ανάλογα με την απόφαση του μοντέλου. Οι καταγραφές για τις οποίες έχει ζητηθεί επισημείωση αργότερα επισημειώνονται χειροκίνητα και χρησιμοποιούνται για να εξατομικευθεί η στρατηγική εκτίμησης του μοντέλου στο συγκεκριμένο παιδί. Στη συγκεκριμένη προσπάθεια αυτόματου υπολογισμού υπάρχει σημαντική βελτίωση των αποτελεσμάτων για την πλειοψηφία των παιδιών, υπάρχουν όμως και ορισμένες περιπτώσεις παιδιών στις οποίες το εξατομικευμένο μοντέλο αποτυγχάνει σε σχέση με το ομαδικό λόγω overfitting, και οι οποίες επομένως θα πρέπει να αντιμετωπιστούν. Σε μελλοντικές προσεγγίσεις, η εξατομίκευση θα μπορούσε να πραγματοποιείται αυτόνομα από το ίδιο το ρομπότ, ρωτώντας ερωτήσεις όπως: "Θα ήθελες να συνεχίσεις το παιχνίδι;" όταν εκτιμούσε το engagement του παιδιού ως χαμηλό ή όταν δεν ήταν βέβαιο για την εκτίμησή του [34].

2.5 Κλάσεις Ταξινόμησης

Όσον αφορά στις επιλεγόμενες κλάσεις ταξινόμησης για το πρόβλημα αυτό παρατηρούμε τα εξής: Στις παλαιότερες προσπάθειες εκτίμησης του engagement [36, 37] χρησιμοποιο-

ύνταν δύο μόνο κλάσεις: engaged, όταν το παιδί ήταν στοχοπροσηλωμένο, αφοσιωμένο, εμπλεκόταν με διάφορους τρόπους στην αλληλεπίδραση και disengaged, όταν δεν παρατηρούνταν τα προαναφερθέντα. Σε πιο πρόσφατες προσπάθειες υπολογισμού η αντιμετώπιση αυτή θεωρήθηκε ανεπαρκής για να περιγράψει τις πραγματικές διαφορές στη συμπεριφορά των παιδιών και έτσι εισάγονται και άλλες κλάσεις.

Στο [2] των Hadfield et al. χρησιμοποιούνται τρεις κλάσεις που αντιστοιχούν η πρώτη στις περιπτώσεις που το παιδί προσέχει ελάχιστα ή και καθόλου το ρομπότ, η δεύτερη στις περιπτώσεις που το παιδί προσέχει το ρομπότ αλλά δε συνεργάζεται μαζί του και η τρίτη σε περιπτώσεις που το παιδί συνεργάζεται ενεργά με το ρομπότ με στόχο την εκπλήρωση ενός ορισμένου σκοπού. Παρατηρούμε πως σε αυτή την περίπτωση χρησιμοποιείται πληρέστερα ο ορισμός για το engagement καθώς υπάρχει διάκριση ανάμεσα στην απλή ένδειξη προσοχής και στην διάθεση συμμετοχής εκ μέρους του παιδιού και άρα με τη σωστή εκτίμηση θα μπορέσει να υπάρξει πιο στοχευμένη αντίδραση από το ρομπότ. Ακόμη, στο [1] των Baxter et al. το engagement αντιμετωπίζεται ως ένα βαθμωτό συνεχές μέγεθος που λαμβάνει τιμές από 0% έως 100% για κάθε frame.

Μία άλλη πρόταση στο CultureNet [28] για τις κλάσεις που μπορούν να χρησιμοποιηθούν είναι η χρήση πέντε κλάσεων με διαβαθμίσεις ως προς την ενέργειες που κάνει το παιδί σε σχέση με έναν κοινό σκοπό που τίθεται και ως προς το βαθμό στον οποίο αντιλαμβάνεται και προσέχει το συνεργάτη-ρομπότ κατά τη διαδικασία αυτή.

2.6 Τρόποι Επισημείωσης των Δεδομένων

Η επισημείωση των δεδομένων video ως προς το επίπεδο του engagement που παρουσιάζουν οι συμμετέχοντες μπορεί να γίνεται βασικά με δύο τρόπους. Είτε επισημειώνονται ολοκληρωμένα διαστήματα στα οποία το επίπεδο του engagement διατηρείται σταθερό είτε επισημειώνονται καθορισμένα σύντομα χρονικά διαστήματα.

Στη δουλειά των Rudovic et al. (A cross cultural study) [39] επισημειώνονται διαστήματα στα οποία παρατηρείται από την αρχή έως το τέλος σταθερό επίπεδο engagement. Η επιλογή αυτή γίνεται καθώς διατυπώνεται το επιχείρημα ότι με τον τρόπο αυτό η επισημείωση προσαρμόζεται στα ιδιαίτερα χαρακτηριστικά της εκάστοτε αλληλεπίδρασης, με αποτέλεσμα να διατηρείται το πλαίσιο στο οποίο παρατηρείται το επίπεδο του engagement, ώστε να μην είναι πιθανό να χαθεί η αρχή και το τέλος της δραστηριότητας που μας ενδιαφέρει.

Από την άλλη, έχουν παρατηρηθεί συγκεκριμένα μικρά χρονικά διαστήματα στα οποία το engagement και συνολικά η έκφραση της εσωτερικής ψυχικής κατάστασης των παιδιών παραμένουν σταθερά. Για παράδειγμα, στη δουλειά των Kim et al. [40] το engagement επισημειώνεται για σταθερά διαστήματα 5 δευτερολέπτων. Στο Κεφάλαιο 5 θα αναφερθούμε εκτενέστερα στο χρονικό διάστημα που απαιτείται για την παρατήρηση του engagement των παιδιών. Από τα πειράματά μας καταλήγουμε στο συμπέρασμα πως το χρονικό αυτό διάστημα είναι περίπου 3 δευτερόλεπτα. Το συμπέρασμα αυτό επιβεβαιώνεται και από τη βιβλιογραφία.

Τέλος, μπορεί να είναι βοηθητική η επισημείωση και άλλων ειδικότερων πληροφοριών, πέρα από την επισημείωση του επιπέδου engagement για το συγκεκριμένο απόσπασμα. Για παράδειγμα, έχουν χρησιμοποιηθεί επισημειώσεις των δεδομένων σε συνεχή κλίμακα όσον αφορά την ευχαρίστηση ή δυσαρέσκεια που παρουσιάζουν τα παιδιά καθώς επίσης και την αδιαφορία ή τον ενθουσιασμό που δείχνουν, όπως επίσης και η εκφραστικότητα του προσώπου [40]. Η επιλογή των χαρακτηριστικών αυτών γίνεται επειδή είναι ανιχνεύσιμα από αλγορίθμους όρασης υπολογιστών.

2.7 Μετρικές Αξιολόγησης

Για την αξιολόγηση των αποτελεσμάτων χρησιμοποιούνται μετρικές όπως τα standard accuracy, F1 score, balanced accuracy, precision, Pearson correlation, concordance correlation coefficient και intraclass correlation coefficient [2].

Το απλό standard accuracy ή accuracy (2.1) εκφράζει το ποσοστό των δειγμάτων που ταξινομήθηκαν στη σωστή κλάση. Στα περισσότερα από τα σύνολα δεδομένων που χρησιμοποιούνται για την εκτίμηση του engagement τα δεδομένα δεν είναι ίσα κατανεμημένα στις κλάσεις. Το standard accuracy δεν είναι απόλυτα ενδεικτικό της επιτυχίας του εκπαιδευμένου μοντέλου, καθώς όπως φαίνεται χαρακτηριστικά και στους πίνακες αξιολόγησης των αποτελεσμάτων στο πέμπτο κεφάλαιο της παρούσας εργασίας δε λαμβάνει καθόλου υπόψιν την ανισοκατανομή των κλάσεων.

$$\text{accuracy} = \frac{\text{correct samples}}{\text{all samples}} \tag{2.1}$$

Για αυτό το λόγο χρησιμοποιούνται και άλλες καταλληλότερες μετρικές. Δύο από αυτές είναι το Fscore (2.4) και το balanced accuracy (2.5). Για να είναι υψηλό το Fscore πρέπει ταυτόχρονα να είναι υψηλό και το recall (2.2) και το precision (2.3), γεγονός που θέλουμε να πετύχουμε στις περιπτώσεις μεγάλης ανισοκατανομής των κλάσεων. Ειδικότερα, το weighted Fscore υπολογίζει τα επιμέρους Fscore όλων των κλάσεων και στη συνέχεια χρησιμοποιεί το διάνυσμα βαρών των κλάσεων για να υπολογίσει το σταθμισμένο μέσο όρο τους.

$$\text{recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \tag{2.2}$$

$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \tag{2.3}$$

$$\text{Fscore} = 2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{2.4}$$

Το balanced accuracy (2.5) υπολογίζεται ως ο μέσος όρος των recall όλων των κλάσεων και επομένως λαμβάνει επίσης υπόψιν την ανισοκατανομή των κλάσεων, αφού για να πάρει υψηλές τιμές πρέπει να πετυχαίνουμε καλές τιμές recall σε όλες τις κλάσεις (πλήθος κλάσεων k).

$$\text{balanced accuracy} = \frac{1}{k}\sum_{n=1}^{N}\text{recall}(k_n) \tag{2.5}$$

Όσον αφορά την αξιολόγηση των αποτελεσμάτων ένας επιπρόσθετος παράγοντας που είναι χρήσιμο να συνυπολογίζουμε μαζί με τις μετρικές των νευρωνικών δικτύων από τα οποία αποτελούνται τα μοντέλα μας είναι και το ποσοστό των επισημειώσεων με τις οποίες επιτυγχάνονται τα αποτελέσματα των μεθόδων μας. Κάνοντας σύγκριση των αποτελεσμάτων των διάφορων προτεινόμενων μεθόδων μπορούμε να εξετάζουμε τις τιμές των διάφορων μετρικών και υπό αυτό το πρίσμα, καθώς η επισημείωση των δεδομένων είναι μία ακριβή και χρονοβόρα διαδικασία.

Τέλος, η intraclass correlation (ICC) είναι μία μετρική που χρησιμοποιείται συχνά στις συμπεριφορικές επιστήμες για την αξιολόγηση της συμφωνίας των επισημειώσεων [33]. Η intraclass correlation κωδικοποιεί τη συνέπεια μεταξύ των προβλέψεων του μοντέλου και των ετικετών ($y' = y+b$), η μετρική correlation coefficient μετρά τη συμφωνία με αφετηρία

την τέλεια γραμμικότητα ($y' = y$) και η Pearson correlation συλλαμβάνει τη γενικότερη γραμμική σχέση ($y' = ay + b$), όπου τα a και b είναι η κλίμακα και το bias αντίστοιχα, y η ετικέτα και y' η πρόβλεψη του μοντέλου) [28].

2.8 Αντιμετώπιση του engagement στην παρούσα διπλωματική εργασία

Έπειτα από τη μελέτη της βιβλιογραφίας για τις μεθόδους εκτίμησης του engagement καταλήξαμε σε ορισμένες αποφάσεις για την προσέγγιση που θα ακολουθούσαμε στην παρούσα διπλωματική εργασία.

Ορίζουμε το engagement ως το το βαθμό στον οποίο ένα άτομο προσέχει - παρακολουθεί τον παρτενέρ του, ενώ ταυτόχρονα συνεργάζεται ενεργά με αυτόν για την επίτευξη ενός κοινού στόχου στα πλαίσια της αλληλεπίδρασής τους.

Ειδικότερα, **αξιοποιούμε** για να εκτιμήσουμε το engagement κυρίως δευτερεύοντα χαρακτηριστικά που εξάγαμε από τα Red-Green-Blue (RGB) δεδομένα. Σε μικρότερο βαθμό συγκρίναμε ορισμένα από τα αποτελέσματά μας με εκείνα που πετύχαμε με εκτίμηση απευθείας από τα RGB δεδομένα. Τα δευτερεύοντα χαρακτηριστικά, στα οποία επικεντρωνόμαστε, είναι η πόζα του ανθρώπινου σώματος καθώς και χαρακτηριστικά όπως η κατεύθυνση σώματος και κεφαλιού. Σε ορισμένα από τα σύνολα δεδομένων, στα οποία η πόζα του σώματος δεν επαρκούσε, δοκιμάσαμε να εντάξουμε και τα σημεία του προσώπου.

Λόγω της ποσότητας των δεδομένων ανά διαφορετικό περιβάλλον και τύπο αλληλεπίδρασης, αλλά και ανά παιδί που συμμετείχε στα πειράματα **δεν εφαρμόζουμε εξατομικευμένη προσέγγιση** σε κάθε παιδί. Κάτι τέτοιο επιχειρήσαμε δοκιμαστικά σε περιορισμένη κλίμακα, σε λίγα μόνο από τα δεδομένα μας, χωρίς τελικά να βελτιώσουμε σημαντικά τα αποτελέσματα της εκτίμησης του engagement κάθε παιδιού. Όπως θα δείξουμε και θα σχολιάσουμε και παρακάτω (Κεφάλαια 5 και 6) αποδεικνύεται πολύ μεγάλης σημασίας η επιτυχημένη και εύκολη προσαρμογή - εξειδίκευση των μοντέλων μας στα διαφορετικά περιβάλλοντα και στις διαφορετικές συνθήκες των αλληλεπιδράσεων.

Όσον αφορά τις **κλάσεις ταξινόμησης** που επιλέγουμε αυτές αντιπροσωπεύουν τέσσερα διακριτά επίπεδα engagement: 0,1,2 και 3. Πιο συγκεκριμένα, στα στιγμιότυπα που κατηγοριοποιούνται στην κλάση 0 δεν υπάρχει οποιαδήποτε απόκριση από το παιδί, ενώ η δράση του και η έκφρασή του δεν σχετίζονται με τον κοινό σκοπό της αλληλεπίδρασης. Στα στιγμιότυπα της κλάσης 1 δεν υπάρχει αναγνώριση του παρτενέρ από την πλευρά του παιδιού, ωστόσο η δράση ή η έκφρασή του σχετίζεται με τον κοινό στόχο της αλληλεπίδρασής του με τον παρτενέρ. Στην κλάση 2 το παιδί αναγνωρίζει τον παρτενέρ του στη συγκεκριμένη αλληλεπίδραση, ωστόσο δεν δρα ή εκφράζεται ανάλογα με τον κοινό τους στόχο. Τέλος, στα στιγμιότυπα που κατατάσσονται στην κλάση 3, δηλαδή υψηλό επίπεδο engagement, το παιδί ταυτόχρονα αναγνωρίζει τον παρτενέρ του στην αλληλεπίδραση και δρα ή εκφράζεται κατάλληλα για την εκπλήρωση του κοινού τους στόχου.

Οι κλάσεις αυτές κατηγοριοποιούν επιτυχώς το engagement των παιδιών. Επίσης, καθεμία συνδέεται με συγκεκριμένα οπτικά χαρακτηριστικά. Είναι δηλαδή δυνατόν να ξεχωρίσουμε τα διαφορετικά επίπεδα κοιτώντας ένα στιγμιότυπο. Για παράδειγμα, όταν το παιδί κοιτά τριγύρω του ή επεξεργάζεται ένα αντικείμενο που δε σχετίζεται με το στόχο της αλληλεπίδρασης και δε μιλά τότε κατατάσσουμε το στιγμιότυπο στην κλάση 0. Αντίστοιχα, όταν το παιδί δεν κοιτά τον παρτενέρ αλλά επιλέγει για παράδειγμα μία από τις εικόνες που εμφανίζονται σε οθόνη σχετική με το μεταξύ τους παιχνίδι, κατατάσσουμε το στιγμιότυπο στην κλάση 1. Ένα παράδειγμα στιγμιότυπου που κατατάσσουμε στην

κλάση 2 είναι όταν το παιδί κοιτά σταθερά τον παρτενέρ αλλά δεν πραγματοποιεί καμία ενέργεια. Τέλος, κατατάσσουμε ένα στιγμιότυπο στο οποίο το παιδί κοιτά σταθερά τον παρτενέρ και εκτελεί μία κίνηση, δράση ή χειρονομία που εκείνος του υπέδειξε στην κλάση 3. Σε όσα από τα δεδομένα δεν ήταν επισημειωμένα προχωράμε στην επισημείωσή τους ακολουθώντας αυτό το σχήμα κωδικοποίησης. Ακολουθούμε το σχήμα αυτό υπό την καθοδήγηση της καθηγήτριας ψυχολόγου κα. Χριστίνας Παπαηλιού.

Ωστόσο, για να υπάρχει συνοχή στα αποτελέσματά μας και να ταυτίζονται οι κλάσεις σε όλα τα δεδομένα μας, πραγματοποιούμε συγχώνευση των κλάσεων 1 και 2 όπου έχουμε τέσσερις κλάσεις. Έτσι, τα μοντέλα μας, προβλέπουν τρία επίπεδα engagement, με το μεσαίο επίπεδο να περιλαμβάνει στιγμιότυπα στα οποία τα παιδιά αναγνωρίζουν τον παρτενέρ αλλά δε δρουν για τον κοινό τους στόχο, αλλά και στιγμιότυπα στα οποία τα παιδιά δεν αναγνωρίζουν τον παρτενέρ, αλλά δρουν σχετικά με τον κοινό στόχο.

Αρχικά, **επισημειώναμε τα δεδομένα** σύμφωνα με το παραπάνω σχήμα για διαστήματα στα οποία διατηρούνταν σταθερό το επίπεδο του engagement. Αυτός ο τρόπος επισημείωσης των δεδομένων είναι σημαντικά πιο χρονοβόρος σε σχέση με την επισημείωση καθορισμένων σύντομων χρονικών διαστημάτων. Εξετάζοντας ενδελεχώς το χρονικό διάστημα το οποίο έπαιρναν υπόψιν τα δίκτυα για την εκτίμηση του engagement, παρατηρήσαμε πως ο ελάχιστος χρόνος για αποδοτική εκτίμηση ήταν τα 3 δευτερόλεπτα, πράγμα που συνάδει με τη σχετική βιβλιογραφία από πλευράς ανάλυσης των ειδικών ψυχολόγων. Έτσι, σε συνεννόηση με την ψυχολόγο καταλήξαμε πως η επισημείωση των δεδομένων για σταθερά διαστήματα των 3 δευτερολέπτων είναι ικανή να αποδώσει το επίπεδο του engagement και φυσικά η επισημείωσή του στα δεδομένα γίνεται λιγότερο χρονοβόρα.

Για να **αξιολογήσουμε** τα μοντέλα μας χρησιμοποιήσαμε εκτός από το απλό accuracy, το Fscore και το balanced accuracy. Ακόμη, στα σημαντικά συγκεντρωτικά αποτελέσματά μας χρησιμοποιούμε και το weighted precision δηλαδή ένα σταθμισμένο ανά κλάση μέσο όρο των precision. Χρησιμοποιούμε αυτές τις μετρικές για να λαμβάνουμε υπόψιν στην αξιολόγησή μας το γεγονός ότι τα δεδομένα μας είναι ιδιαίτερα ανισοκατανεμημένα στις κλάσεις που κωδικοποιούν τα επίπεδα του engagement.

Κεφάλαιο 3

Σύνολα δεδομένων και χαρακτηριστικά εκτίμησης

Στο Κεφάλαιο αυτό πραγματοποιούμε μία παρουσίαση των συνόλων δεδομένων που χρησιμοποιήθηκαν για την εκτίμηση του engagement στην παρούσα διπλωματική. Ακόμη, παρουσιάζουμε τα χαρακτηριστικά που χρησιμοποιούμε για την εκτίμηση του engagement και την επεξεργασία των RGB ή RGB-D video για την εξαγωγή των χαρακτηριστικών αυτών και την παραγωγή των τελικών διανυσμάτων που αξιοποιούνται από τα δίκτυά μας (Κεφάλαιο 5) για την εκτίμηση.

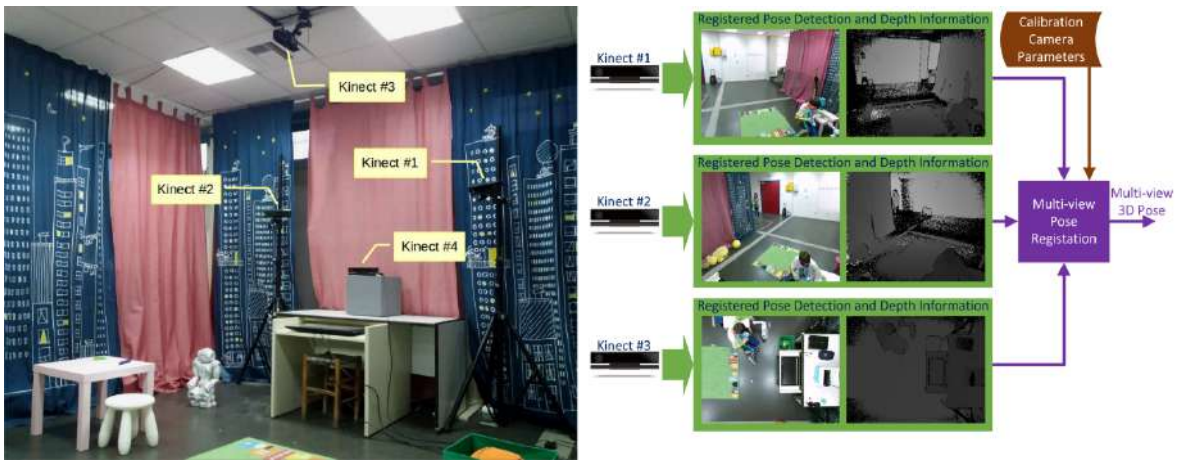
Τα σύνολα δεδομένων που χρησιμοποιούμε έχουν προκύψει από πειράματα που έγιναν στο πλαίσιο των προθεστ BabyRobot, BabyAffect και της PInSoRo βάσης δεδομένων. Χρησιμοποιούμε δεδομένα από διαφορετικές πηγές καθώς για τη μέθοδο που θα σχεδιάσουμε είναι τεράστιας σημασίας η δυνατότητα εκτίμησης του engagement σε πολύ διαφορετικές συνθήκες (ελεύθερες ή καθορισμένες, σε εργαστηριακό ή οικιακό περιβάλλον, με τυπικώς αναπτυσσόμενα παιδιά ή παιδιά με αυτισμό, αλληλεπιδράσεις με ρομπότ ή με ανθρώπους). Ο κύριος όγκος των πειραμάτων που χρησιμοποιούμε προέρχεται από το πρόγραμμα BabyRobot και γι΄ αυτό χωρίζουμε τα δεδομένα αυτά σε υποκατηγορίες ώστε να μπορούμε να πραγματοποιήσουμε στοχευμένες συγκρίσεις μεταξύ τους και τελικά να εξάγουμε συμπεράσματα τόσο για το engagement όσο και για την εκτίμησή του. Τα κριτήρια για το χωρισμό των δεδομένων αυτών σε υποκατηγορίες είναι το αν στα πειράματα συμμετέχουν τυπικώς αναπτυσσόμενα παιδιά ή παιδιά με αυτισμό, ποια παιχνίδια παίζουν τα παιδιά, καθώς και ο τόπος που λαμβάνουν χώρα οι αλληλεπιδράσεις. Σύμφωνα με τα κριτήρια αυτά έχουμε χωρίσει τα δεδομένα BabyRobot που αξιοποιήσαμε στις κατηγορίες TD-Joint Attention, ASD-Joint Attention, ASD-Games και ASD-School. Στη συνέχεια, παρουσιάζουμε στοιχεία για τις διατάξεις των πειραμάτων - αλληλεπιδράσεων από τα οποία προέκυψαν τα δεδομένα, για τις κλάσεις που χρησιμοποιήθηκαν για την επισημείωσή τους -όπου δεν πραγματοποιήσαμε εμείς την επισημείωση σύμφωνα με το σχήμα που περιγράφηκε στην Ενότητα 2.8-, για τα εργαλεία, τον τρόπο εξαγωγής και την επεξεργασία των επιλεγμένων χαρακτηριστικών από τα video που συγκεντρώθηκαν.

Η επιτυχής εξαγωγή της πόζας αποτελεί ένα πολύ σημαντικό βήμα για την αναγνώριση ανθρώπινης δράσης και για πολλές ακόμη εφαρμογές όρασης υπολογιστών. Στο Κεφάλαιο 4 αναφερόμαστε αναλυτικά στις διάφορες μεθόδους εξαγωγής πόζας, στη βιβλιοθήκη OpenPose με την οποία πραγματοποιούμε εξαγωγή της πόζας στην παρούσα διπλωματική, καθώς και σε αρκετές εφαρμογές αναγνώρισης ανθρώπινης δράσης και συναισθήματος που αξιοποιούν την πόζα και το OpenPose ειδικότερα. Σε αυτό το Κεφάλαιο θεωρούμε δεδομένη τη δυνατότητα εξαγωγής της πόζας, δηλαδή της θέσης σημείων του ανθρώπινου

σκελετού με τα οποία μπορούμε να περιγράψουμε τη στάση του ανθρώπινου σώματος. Τα σύνολα αυτά θέσεων σημείων εξάγουμε σε όλα τα σύνολα δεδομένων μας και τα αξιοποιούμε ως χαρακτηριστικά εκτίμησης του engagement.

Συνοπτικά, στα δεδομένα TD και ASD-Joint Attention χρησιμοποιήσαμε ως διάνυσμα χαρακτηριστικών τα τρισδιάστατα σημεία του σκελετού, εμπλουτισμένα με τέσσερα ακόμη χαρακτηριστικά (κατεύθυνση κεφαλιού, κατεύθυνση σώματος, απόσταση χεριών από αντίστοιχους ώμους) τα οποία περιγράφονται αναλυτικά παρακάτω. Στα δεδομένα BabyAffect χρησιμοποιήσαμε τα δισδιάστατα σημεία του σκελετού, εμπλουτισμένα με τα τέσσερα χαρακτηριστικά. Στα δεδομένα ASD-Games και ASD-School το διάνυσμα χαρακτηριστικών μας αποτελείται από τα δισδιάστατα σημεία του σκελετού εμπλουτισμένα με τα δισδιάστατα σημεία του προσώπου. Τέλος, στα δεδομένα PInSoRo αξιοποιήσαμε πολλά ως διάνυσμα χαρακτηριστικών πολλά από τα χαρακτηριστικά που είναι διαθέσιμα, όπως κομμάτι των δισδιάστατων σημείων της πόζας και action units.

3.1 Το σύνολο δεδομένων BabyRobot



Σχήμα 3.1: Το περιβάλλον των καταγραφών της BabyRobot και πολλαπλές όψεις για την εκτίμηση της πόζας [2].

Στο σύνολο δεδομένων BabyRobot περιλαμβάνονται πολυάριθμες καταγραφές αλληλεπιδράσεων παιδιών με ενήλικες, ρομπότ καθώς και με άλλα παιδιά. Στα πειράματα συμμετέχουν τόσο τυπικά αναπτυσσόμενα παιδιά όσο και παιδιά με διαταραχές του αυτιστικού φάσματος. Οι καταγραφές περιλαμβάνουν διάφορους τύπους αλληλεπιδράσεων - παιχνιδιών οι οποίες πραγματοποιήθηκαν σε ειδικά διαμορφωμένο εργαστήριο (Σχήμα 3.1) ή σε αντίστοιχα διαμορφωμένη αίθουσα σε σχολείο του Πειραιά (Σχήμα 3.7).

Όπως αναφέραμε και παραπάνω έχουμε χωρίσει τα δεδομένα BabyRobot που αξιοποιήσαμε στις κατηγορίες TD-Joint Attention, ASD-Joint Attention, ASD-Games και ASD-School. Σε όλες αυτές τις κατηγορίες δεδομένων τα παιδιά συμμετέχουν σε συγκεκριμένες αλληλεπιδράσεις με ρομπότ. Πολλές από τις αλληλεπιδράσεις αυτές πραγματοποιήθηκαν και με παρτενέρ ψυχολόγο, έχουν καταγραφεί και περιλαμβάνονται στα δεδομένα BabyRobot. Οι αλληλεπιδράσεις BabyRobot με τις οποίες έχουμε δουλέψει στην παρούσα διπλωματική εργασία έχουν πραγματοποιηθεί στο ειδικά διαμορφωμένο εργαστήριο εκτός από τα δεδομένα ASD-School που έχουν πραγματοποιηθεί σε σχολείο στον Πειραιά.

Στο χώρο του εργαστηρίου τα παιδιά μπορούσαν να κινούνται τελείως ελεύθερα. Η καταγραφή των πειραμάτων έγινε από τέσσερις συσκευές Kinect που ήταν τοποθετημένες

(α′) Furhat robot[42]

(β′) Nao robot[41]

Σχήμα 3.2: Τα ρομπότ που χρησιμοποιήθηκαν στο BabyRobot.

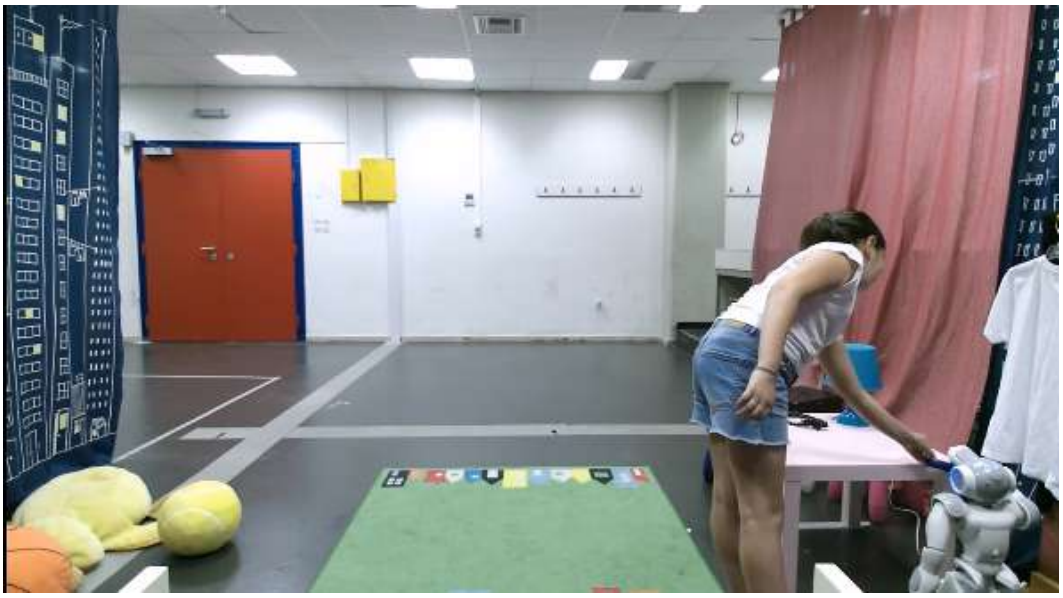
σε διαφορετικά σημεία ώστε να καταγράφονται από πολλαπλές όψεις οι αλληλεπιδράσεις. Στις αλληλεπιδράσεις χρησιμοποιήθηκαν το ρομπότ Nao [41], το οποίο βρισκόταν στο πάτωμα του δωματίου και κινούνταν ελεύθερα καθώς και το ρομπότ Furhat [42] (Σχήμα 3.2), το οποίο βρισκόταν τοποθετημένο σε ένα τραπέζι στη μία πλευρά του δωματίου πάνω από μία οθόνη αφής.

3.1.1 Αλληλεπιδράσεις TD και ASD-Joint Attention

Αρχικά, χρησιμοποιήσαμε τα πειράματα Joint Attention του BabyRobot. Σε αυτά έχουν καταγραφεί 25 αλληλεπιδράσεις τυπικώς αναπτυσσόμενων παιδιών και 15 αλληλεπιδράσεις παιδιών με διαταραχές αυτιστικού φάσματος. Από τα 15 αυτά παιδιά τα 8 δεν ανταποκρίθηκαν με κανένα τρόπο στην αλληλεπίδραση. Γι΄ αυτό το λόγο χρησιμοποιήσαμε μόνο τα videos των υπόλοιπων 7 παιδιών, με αποτέλεσμα να έχουμε τα σύνολα δεδομένων TD-Joint Attention με video από 25 TD παιδιά και ASD-Joint Attention με video από 7 ASD παιδιά. Στην αλληλεπίδραση Joint Attention τα παιδιά είναι ελεύθερα να κινούνται στο δωμάτιο όπως επιθυμούν. Κατά τη διάρκεια του παιχνιδιού το ρομπότ Nao πλησιάζει ένα τουβλάκι και στη συνέχεια με μία σειρά κινήσεων προσπαθεί να δείξει στο παιδί ότι επιθυμεί να πιάσει το τουβλάκι (πλησιάζοντας προς αυτό, ανοιγοκλείνοντας το χέρι του, γέρνοντας προς αυτό). Αν δεν καταφέρνει να τραβήξει την προσοχή του παιδιού με τις κινήσεις αυτές, ζητά το τουβλάκι και λεκτικά από το παιδί (Σχήμα 3.3, Σχήμα 3.4). Εδώ στόχος της αλληλεπίδρασης είναι να αντιληφθεί το παιδί την πρόθεση του ρομπότ και να το βοηθήσει να πραγματοποιήσει την επιθυμία του, δίνοντας του το τουβλάκι.

Στα πειράματα Joint Attention η επισημείωση του engagement πραγματοποιήθηκε με τη χρήση τριών διακριτών επιπέδων engagement, όπως περιγράφουμε στην Ενότητα 2.8. Το πρώτο (κλάση 0) δηλώνει ότι το παιδί δίνει ελάχιστη ή και καθόλου προσοχή στο ρομπότ, το δεύτερο (κλάση 1) δηλώνει ότι το παιδί παρακολουθεί το ρομπότ αλλά δε συνεργάζεται με αυτό και τέλος το τρίτο (κλάση 2) δηλώνει ότι το παιδί συνεργάζεται ενεργά με το ρομπότ για να ολοκληρώσει τη δραστηριότητα.

Για την εξαγωγή των σημείων του σκελετού από τις καταγραφές ακολουθήθηκε η παρακάτω διαδικασία [2]. Η βιβλιοθήκη του OpenPose χρησιμοποιήθηκε για να ανιχνευθούν από τις έγχρωμες εικόνες δισδιάστατοι χάρτες σημείων σκελετού: ένα σημείο για κάθε καρπό, αγκώνα, ώμο, γοφό, γόνατο και αστράγαλο, ένα για το λαιμό και πέντε σημεία προσώπου, δύο για τα αυτιά, δύο για τα μάτια και ένα για τη μύτη. Παρουσιάζουμε αναλυτικά την εκτίμηση της ανθρώπινης πόζας και πιο συγκεκριμένα από το OpenPose



Σχήμα 3.3: Στιγμιότυπο από αλληλεπίδραση TD-Joint Attention, οπτική από μία κάμερα.

Σχήμα 3.4: Στιγμιότυπο από αλληλεπίδραση ASD-Joint Attention, οπτικές από όλες τις κάμερες.

στο Κεφάλαιο 4. Στη συνέχεια οι χάρτες αυτοί τροφοδοτήθηκαν σε ένα βαθύ νευρωνικό δίκτυο μαζί με πλέγματα πληρότητας στοιχειακών όγκων που εξήχθησαν από τις εικόνες βάθους. Το δίκτυο εκπαιδεύτηκε ώστε να παράγει εκτιμήσεις για τις θέσεις των 18 σημείων στον τρισδιάστατο χώρο.

Όταν χρησιμοποιούνται πολλές κάμερες, τα σημεία του σκελετού μπορούν να εξαχθούν για κάθε μία όψη και να συγχωνευθούν για να παράγουν τις τελικές εκτιμήσεις. Το πρώτο βήμα για να επιτευχθεί αυτό ήταν να εισαχθούν τα σημεία από το καρέ αναφοράς κάθε κάμερας σε ένα κοινό καρέ αναφοράς. Οι αντίστοιχες παράμετροι υπολογίστηκαν με τη χρήση του αλγορίθμου Iterative Closest Point (ICP) [43], ο οποίος παρέχει το μετασχηματισμό που ταιριάζει καλύτερα το νέφος σημείων της μίας κάμερας στην άλλη, με δεδομένο έναν αρχικό μετασχηματισμό που του δίνεται. Μετά από τη μετατροπή των συντεταγμένων των σημείων του σκελετού από όλες τις κάμερες σε ένα κοινό καρέ αναφοράς, το επόμενο βήμα ήταν να καθοριστεί ποια από τα σημεία κάθε όψης ήταν έγκυρα.

Ο αλγόριθμος εκτίμησης πόζας αποτυγχάνει σε ορισμένες περιπτώσεις, για παράδειγμα όταν κάποιο από τα μέλη του παιδιού είναι κρυμμένο στην εικόνα. Σε αυτές τις περιπτώσεις, παράγει θορυβώδεις εκτιμήσεις είτε για μερικά είτε για όλα τα σημεία, ενώ μπορεί να αποτύχει στην παραγωγή οποιασδήποτε εκτίμησης. Ένα άλλο πρόβλημα είναι ότι ο αλγόριθμος μερικές φορές δίνει ως έξοδο πολλαπλές πόζες, όταν υπάρχει και άλλο άτομο στην εικόνα ή όταν περιστασιακά το δίκτυο μπερδεύεται από αντικείμενα του περιβάλλοντος. Για να αντιμετωπιστούν αυτά τα προβλήματα, λαμβανόταν ο μέσος όρος μόνο των σημείων εκείνων που ήταν αρκετά κοντά στα σημεία του προηγούμενου στιγμιότυπου. Εάν δεν υπήρχαν τέτοιες εκτιμήσεις για κάποιο σημείο, αυτό θεωρούνταν πως έλειπε στο συγκεκριμένο στιγμιότυπο. Στη συνέχεια, πραγματοποιήθηκε γραμμική παρεμβολή των τιμών που έλειπαν. Τα σημεία τέλος περιστράφηκαν κατάλληλα, ώστε οι άξονες των συντεταγμένων τους να συμπίπτουν με τις πλευρές του δωματίου.

Για το συγκεκριμένο πρόβλημα η πόζα του παιδιού μας ενδιαφέρει κυρίως σε σχέση με τη δραστηριότητα που λαμβάνει χώρα, δηλαδή σε σχέση με το ρομπότ Nao. Συγκεκριμένα, έπρεπε να υπολογιστούν τα τρισδιάστατα σημεία της πόζας σε σχέση με τη θέση του ρομπότ. Έτσι, ανιχνεύθηκε η θέση του Nao σε όλα τα στιγμιότυπα των καταγραφών. Στη συνέχεια, αφαιρέθηκαν οι αντίστοιχες τιμές από τις εκτιμήσεις για κάθε σημείο της πόζας του παιδιού. Με αυτό τον τρόπο τα χαρακτηριστικά είναι εκφρασμένα ως προς το ρομπότ.

Μετέπειτα, δημιουργήθηκε ένα σύνολο από χαρακτηριστικά ώστε να βοηθήσουν τη διαδικασία ταξινόμησης. Αυτά είναι η γωνία μεταξύ του ρομπότ και του βλέμματος του παιδιού, η γωνία μεταξύ του ρομπότ και της κατεύθυνσης που κοιτάει το σώμα του παιδιού και οι αποστάσεις των χεριών από τους αντίστοιχους ώμους. Η κατεύθυνση του βλέμματος υπολογίστηκε από τα σημεία του προσώπου που είχαν ανιχνευθεί, λαμβάνοντας το διάνυσμα από το ένα αυτί στο άλλο στο δισδιάστατο επίπεδο και περιστρέφοντάς το κατά 90 μοίρες. Με παρόμοιο τρόπο υπολογίστηκε η κατεύθυνση του σώματος του παιδιού από τα σημεία των ώμων. Από τις δύο γωνίες αφαιρέθηκε η γωνία μεταξύ παιδιού και ρομπότ, η οποία υπολογίστηκε με τη χρήση του κέντρου μάζας και της θέσης του ρομπότ που έχει ανιχνευθεί. Τα τέσσερα αυτά χαρακτηριστικά τοποθετήθηκαν μαζί με τις τιμές των τριών συντεταγμένων των σημείων του σκελετού και προέκυψαν τα διανύσματα εκτίμησης του engagement.

Σχήμα 3.5: Στιγμιότυπο από αλληλεπίδραση Show me the Gesture.

Σχήμα 3.6: Στιγμιότυπο από αλληλεπίδραση Pantomime.

3.1.2 Το σύνολο δεδομένων ASD-Games

Σε δεύτερη φάση χρησιμοποιήσαμε ορισμένα ακόμη από τα πειράματα του Baby Robot, στα οποία συμμετέχουν τα επτά παιδιά με διαταραχές αυτιστικού φάσματος που συμμετέχουν και στα παιχνίδια Joint Attention. Συγκεκριμένα, για κάθε παιδί χρησιμοποιήσαμε καταγραφές τεσσάρων επιπλέον παιχνιδιών, τα οποία λαμβάνουν χώρα στο ίδιο περιβάλλον με τα παιχνίδια Joint Attention. Τα παιχνίδια αυτά είναι τα ακόλουθα:

- Show me the Gesture: Το ρομπότ Furhat που βρίσκεται τοποθετημένο πάνω από την οθόνη αφής ζητά από το παιδί να κάνει ορισμένες χειρονομίες. Στην αρχή του ζητά να το χαιρετήσει, έπειτα να το καλέσει κοντά του και να του δείξει να σταματήσει αυτό που κάνει. Τέλος, το ρομπότ ζητά από το παιδί να του δείξει το τραπέζι και να του δείξει να κάτσει. Αν το παιδί δεν ανταποκρίνεται το ρομπότ επαναλαμβάνει δύο φορές την οδηγία πριν περάσει σε επόμενη χειρονομία. (Σχήμα 3.5).

- Express the Feeling: Το ρομπότ Furhat ζητά από το παιδί να μιμηθεί, παίρνοντας την κατάλληλη έκφραση στο πρόσωπό του, τα συναισθήματα που βλέπει σε εικόνες τις οποίες διαλέγει από την οθόνη αφής μπροστά του. Τα συναισθήματα αυτά είναι χαρά, λύπη, αηδία, φόβος και θυμός. Αφού το παιδί έχει χρόνο να εκφράσει το συναίσθημα, παίρνει και το ρομπότ την αντίστοιχη έκφραση.

- Guess the Object: Το ρομπότ Furhat ζητά από το παιδί να εντοπίσει ορισμένα αντικείμενα (μπάλα, βιβλίο, κουτί παπουτσιών, μολύβι και κουτάλι) που βρίσκονται γύρω του από την περιγραφή τους και να τα τοποθετήσει σε συγκεκριμένη θέση. Για παράδειγμα, το ρομπότ περιγράφει την μπάλα ως ένα πράγμα πορτοκαλί, στρογγυλό που με αυτό παίζουμε. Επαναλαμβάνει μερικές φορές την περιγραφή όσο το παιδί δεν ανταποκρίνεται πριν προχωρήσει σε επόμενο αντικείμενο. Όταν το παιδί βρει την μπάλα, το ρομπότ του ζητά να την τοποθετήσει στο μεγάλο πράσινο κουτί και συνεχίζει λέγοντας στο παιδί ότι έχει διαλέξει επόμενο αντικείμενο.

- Pantomime: Το ρομπότ Nao ζητά από το παιδί να μιμηθεί, κάνοντας κινήσεις με το σώμα του, τις δραστηριότητες που βλέπει σε εικόνες τις οποίες διαλέγει από την οθόνη αφής. Οι δραστηριότητες αυτές είναι κολύμβηση, ανάγνωση βιβλίου, γυμναστική, χτένισμα, χορός και φαγητό. Αφού το παιδί έχει χρόνο να κάνει την κίνηση που βλέπει στην κάρτα το ρομπότ μαντεύει την κίνηση που έκανε και στη συνέχεια μιμείται και αυτό τη δραστηριότητα (Σχήμα 3.6).

Για τα δεδομένα αυτά πραγματοποιήσαμε την επισημείωση των δεδομένων όπως αναφέρουμε στην Ενότητα 2.8, δηλαδή χρησιμοποιώντας τέσσερα διακριτά επίπεδα engagement: το πρώτο (κλάση 0), δηλώνει ότι το παιδί δεν δίνει προσοχή στο ρομπότ και δεν ασχολείται με κανένα τρόπο με τη μεταξύ τους αλληλεπίδραση-δραστηριότητα, το δεύτερο (κλάση 1) δηλώνει ότι το παιδί δεν δίνει προσοχή στο ρομπότ, αλλά δρα στα πλαίσια της κοινής δραστηριότητάς τους, το τρίτο επίπεδο (κλάση 2) δηλώνει ότι το παιδί προσέχει το ρομπότ, ενώ δε συμμετέχει συμβάλλοντας στην αλληλεπίδρασή τους και τέλος το τέταρτο επίπεδο (κλάση 3) δηλώνει ότι το παιδί όχι μόνο προσέχει το ρομπότ, αλλά συμμετέχει και ενεργά στην μεταξύ τους δραστηριότητα. Χρησιμοποιήσαμε αυτή την κωδικοποίηση του engagement καθώς είναι ακριβέστερη και πληρέστερη, περιγράφει καλύτερα τη συμπεριφορά των παιδιών. Ωστόσο για λόγους συνοχής με τα επισημειωμένα δεδομένα Joint Attention στο πλαίσιο της παρούσας εργασίας συγχωνεύσαμε για τα πειράματα μας τις κλάσεις 1 και 2.

Σχήμα 3.7: Στιγμιότυπο από τα δεδομένα από το σχολείο (ASD-School).

Για τα παιχνίδια αυτά (με εξαίρεση το Pantomime Game, για το οποίο δουλέψαμε όπως και με τα Joint Attention) δεν μας ενδιέφερε η απόσταση από το Nao, καθώς σε αυτά ο παρτενέρ είναι το σταθερό ρομπότ Furhat, στην πλευρά του οποίου βρίσκονται τοποθετημένες οι κάμερες καταγραφής. Εδώ, επίσης, δεν ήταν διαθέσιμα τα δεδομένα του βάθους και επομένως καταλήξαμε σε δισδιάστατες συντεταγμένες για τα σημεία της πόζας. Ακόμη, εκτός από τα σημεία του σκελετού αξιοποιήσαμε και τα 70 σημεία του προσώπου που δίνει το OpenPose, καθώς δεν πρόκειται για τόσο απλές συνθήκες όσο η συνθήκη Joint Attention. Τέλος, παράγαμε τα 4 επιπλέον χαρακτηριστικά που παράχθηκαν και για τα δεδομένα Joint Attention (κατεύθυνση κεφαλιού, κατεύθυνση σώματος, αποστάσεις χεριών από ώμους).

3.1.3 Το σύνολο δεδομένων ASD-School

Τέλος, από το BabyRobot αξιοποιήσαμε και ορισμένες από τις αλληλεπιδράσεις που έλαβαν χώρα σε σχολείο του Πειραιά, ώστε να εμπλουτίσουμε και άλλο τα δεδομένα που χρησιμοποιούμε και να εφαρμόσουμε τις μεθόδους μας και σε αρκετά διαφορετικές συνθήκες αλληλεπίδρασης. Χρησιμοποιήσαμε καταγραφές στις οποίες συμμετέχουν τρία παιδιά με διαταραχές αυτιστικού φάσματος. Στις καταγραφές αυτές τα παιδιά υπό την καθοδήγηση ενός σταθερού ρομπότ που βρίσκεται τοποθετημένο απέναντί τους παίζουν δύο παιχνίδια (Emotions και Sums). Τα πειράματα αυτά γίνονταν επαναλαμβανόμενα 2 φορές την εβδομάδα για 3 μήνες. Στο χρονικό αυτό διάστημα καταγράφηκαν για τα τρία παιδιά 25 video με το παιχνίδι Sums και 16 video με το παιχνίδι Emotions.

Στο πρώτο παιχνίδι (Emotions) καλούνται να μιμηθούν με την έκφραση του προσώπου τους συναισθήματα που τους ζητά το ρομπότ. Τα συναισθήματα αυτά είναι χαρά, λύπη, φόβος και θυμός. Το ρομπότ μιμείται επίσης με το πρόσωπό του τα συναισθήματα και ζητά από τα παιδιά να κάνουν το ίδιο. Στο δεύτερο παιχνίδι (Sums) τα παιδιά καλούνται να βοηθήσουν το ρομπότ να βρει τα σωστά τα αποτελέσματα μερικών απλών πράξεων. Εδώ, το ρομπότ στην αρχή κάνει λάθος στις πράξεις και ζητά τη βοήθεια του παιδιού για να βρει το σωστό αποτέλεσμα. Καθώς εξελίσσεται η διαδικασία το ρομπότ μαθαίνει σταδιακά τα αποτελέσματα που του δείχνει το παιδί και τα βρίσκει σωστά. Στόχος της

συγκεκριμένης αλληλεπίδρασης δεν είναι να βρίσκουν τα παιδιά τα αποτελέσματα των πολύ απλών πράξεων, αλλά να μπουν επιτυχημένα στο ρόλο του βοηθού του ρομπότ. Για παράδειγμα, όταν το ρομπότ λέει στο παιδί Ώραία, θα δοκιμάσω εγώ τώρα, το παιδί είναι πλήρως engaged αν περιμένει το ρομπότ να διαλέξει αποτέλεσμα και όχι αν διαλέξει μόνο του το σωστό αποτέλεσμα.

Και οι δύο αλληλεπιδράσεις πραγματοποιούνται με τη βοήθεια μίας οθόνης αφής, η οποία βρίσκεται τοποθετημένη στο τραπέζι ανάμεσα στα παιδιά και το ρομπότ. Η καταγραφή των πειραμάτων γίνεται από μία κάμερα που βρίσκεται απέναντι από τα παιδιά, πάνω από το ρομπότ. Είναι χαρακτηριστικό το γεγονός ότι σε αυτή την περίπτωση τα παιδιά βρίσκονται καθισμένα μπροστά από την οθόνη και το ρομπότ σε όλη τη διάρκεια των αλληλεπιδράσεων (Σχήμα 3.7) με αποτέλεσμα να είναι πολύ περιορισμένη η ποικιλία των κινήσεών τους και άρα και της πόζας τους.

Για την επεξεργασία των παραπάνω δεδομένων ακολουθήσαμε την ίδια διαδικασία με τα δεδομένα ASD-Games και καταλήξαμε στο ίδιο διάνυσμα χαρακτηριστικών.

3.1.4 Επεξεργασία των διανυσμάτων χαρακτηριστικών

Για τα χαρακτηριστικά που αντιπροσωπεύουν τις συντεταγμένες των σημείων του σκελετού των παιδιών, επιλέγουμε να αφαιρέσουμε από τις τιμές τους τις αντίστοιχες τιμές του σημείου που αντιπροσωπεύει το δεξιό γοφό ανά στιγμιότυπο. Με αυτό τον τρόπο όλα τα σημεία του σκελετού πλέον δίνονται στα δίκτυα που θα εκπαιδεύσουμε ως προς το δεξιό γοφό. Με αυτόν τον τρόπο οι τιμές αντιπροσωπεύουν απευθείας τις σχέσεις και τις αποστάσεις μεταξύ των διάφορων μελών του σώματος, περιγράφουν δηλαδή αμεσότερα την πόζα του παιδιού. Ο δεξιός γοφός επιλέγεται καθώς είναι σχετικά σταθερό σημείο του σώματος. Αντίστοιχα στο ASD-School επιλέχθηκε ο λαιμός καθώς τα παιδιά βρίσκονται καθισμένα με αποτέλεσμα να μην ανιχνεύονται τιμές για τις θέσεις των σημείων του σκελετού από τη μέση και κάτω.

Στη συνέχεια, πραγματοποιούμε κανονικοποίηση για όλα τα χαρακτηριστικά στο διάστημα [0,1] και χωρίζουμε τα σύνολα των δεδομένων σε train και test δεδομένα. Κάθε φορά τα test δεδομένα αποτελούν το 30% των συνολικών δεδομένων. Με στόχο να εμπλουτίσουμε τα train δεδομένα, ώστε να προσφέρουν στο δίκτυο μεγαλύτερες δυνατότητες γενίκευσης και να αποφευχθεί το overfitting εφαρμόζουμε δύο μεθόδους data augmentation, που συνηθίζονται σε τέτοιου τύπου προβλήματα και δεδομένα. Προσθέτουμε με πιθανότητα 25% ένα μικρό γκαουσιανό θόρυβο σε όλα τα χαρακτηριστικά του διανύσματος εισόδου και με πιθανότητα 10% παράγουμε το καθρέπτισμα της πόζας.

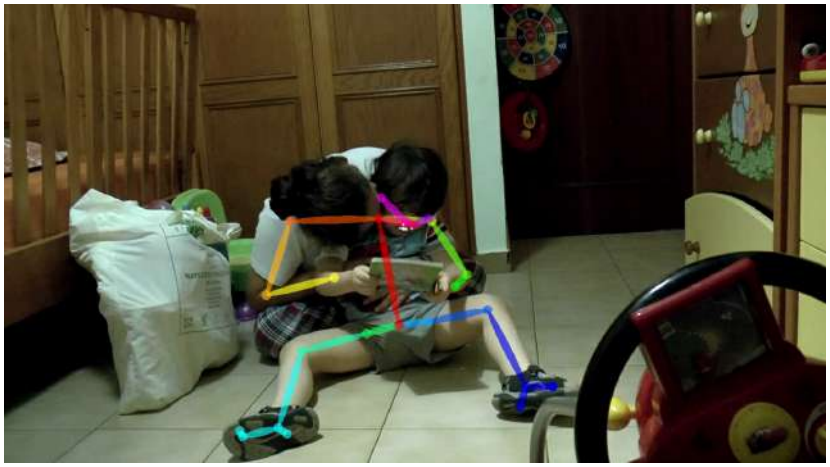
Επιπρόσθετα, κατασκευάζουμε το διάνυσμα βαρών των κλάσεων των συνόλων δεδομένων, ώστε να το αξιοποιήσουμε στην εκπαίδευσή. Για κάθε κλάση διαιρούμε το πλήθος των δειγμάτων που ανήκουν στη συχνότερη κλάση με το πλήθος των δειγμάτων που ανήκουν στη συγκεκριμένη κλάση. Παράγουμε το διάνυσμα βαρών των κλάσεων μόνο με βάση τα train δεδομένα, ώστε να μην εισάγουμε ουσιαστικά στο δίκτυο κατά την εκπαίδευσή του πληροφορία για τα test δεδομένα. Για παράδειγμα, το διάνυσμα βαρών των κλάσεων για το συγκεκριμένο χωρισμό train-test δεδομένων στα τυπικώς αναπτυσσόμενα παιδιά Joint Attention είναι [11.1095, 1.0000, 2.9630]. Χρησιμοποιούμε το διάνυσμα βαρών στην εκπαίδευση για να εκπαιδεύουμε το δίκτυο λαμβάνοντας υπ΄ όψιν τη σημαντική ανισοκατανομή των κλάσεων.

Τα δεδομένα είναι πλέον έτοιμα να χρησιμοποιηθούν ως είσοδοι σε κατάλληλα δίκτυα με στόχο την εκτίμηση του engagement. Παρουσιάζουμε τα αποτελέσματα των μεθόδων που εφαρμόσαμε για την εκτίμηση στο Κεφάλαιο 5.

(α′) Στιγμιότυπο από το παιχνίδι του εξερευνητή του BabyAffect.

(β′) Στιγμιότυπο του BabyAffect στο οποίο η μητέρα προσποιείται πως έχει χτυπήσει.

Σχήμα 3.8: Στιγμιότυπα του συνόλου δεδομένων BabyAffect.



Σχήμα 3.9: Στιγμιότυπο BabyAffect στο οποίο μητέρα και παιδί εντοπίζονται ως ένας άνθρωπος.

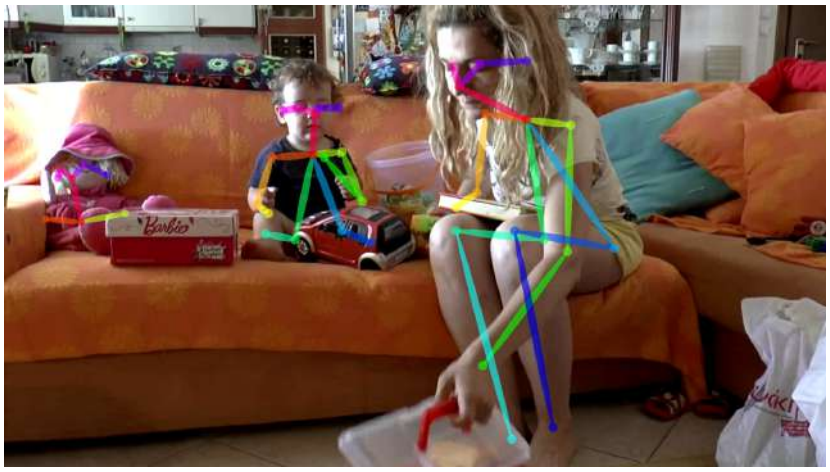
3.2 Το σύνολο δεδομένων BabyAffect

Επιπρόσθετα, χρησιμοποιήσαμε τα δεδομένα BabyAffect. Πρόκειται για καταγραφές παιχνιδιών ανάμεσα στις μητέρες και τα παιδιά τους. Υπάρχουν καταγραφές για 15 τυπικώς αναπτυσσόμενα παιδιά και 25 παιδιά στο αυτιστικό φάσμα. Οι καταγραφές διαρκούν 20 περίπου λεπτά.

Από τα παραπάνω δεδομένα διαθέτουμε επισημειώσεις για το engagement για έξι από τις Autism Spectrum Disorder (ASD) καταγραφές και συγκεκριμένα για τα διαστήματα από το ένατο έως το δέκατο τέταρτο λεπτό και από το εικοστό έως το εικοστό πρώτο. Στο πρώτο διάστημα οι μητέρες με τα παιδιά παίζουν το παιχνίδι του¨εξερευνητή' (¨εξερευνούν' μαζί ένα κουτί με διάφορα πλαστικά παιχνίδια: φρούτα, οικιακά σκεύη κλπ.)(Σχήμα 3.8), ενώ στο δεύτερο διάστημα η μητέρα προσποιείται πως έχει χτυπήσει και πως κλαίει, ώστε να καταγραφεί η αντίδραση του παιδιού σε μία τέτοια κατάσταση (Σχήμα 3.8). Οι επισημειώσεις έχουν γίνει και εδώ σε τέσσερα διακριτά επίπεδα engagement σύμφωνα με τον τρόπο που περιγράφεται παραπάνω.

3.2.1 Επεξεργασία των δεδομένων BabyAffect

Από τις έξι καταγραφές για τις οποίες υπάρχουν προς το παρόν επισημειώσεις, δεν ήταν δυνατόν να αξιοποιήσουμε τις δύο καθώς οι μητέρες είχαν για το μεγαλύτερο διάστημα



Σχήμα 3.10: Στιγμιότυπο BabyAffect στο οποίο εντοπίζεται και ο σκελετός μίας κούκλας.

των αλληλεπιδράσεων τα παιδιά στην αγκαλιά τους, με αποτέλεσμα ο αλγόριθμος του OpenPose σε σημαντικό κομμάτι των καταγραφών να μη καταφέρνει να εντοπίσει σωστά το σκελετό του παιδιού, να τον ταυτίζει με αυτόν της μητέρας (Σχήμα 3.9). Συνεπώς, τέσσερις μόνο από τις καταγραφές μπορούσαν να αξιοποιηθούν από τη μέθοδό μας. Από αυτές στις τρεις συμμετέχουν ASD παιδιά και στην τέταρτη ένα τυπικώς αναπτυσσόμενο παιδί. Για αυτό το λόγο χρησιμοποιήσαμε τελικά τις τρεις από τις τέσσερις καταγραφές, ώστε να αποτελείται όλο το σύνολο από δεδομένα ASD παιδιών.

Στα videos παρήγαμε όπως προηγουμένως με τη βοήθεια του OpenPose τους σκελετούς. Στη συνέχεια, έπρεπε σε κάθε στιγμιότυπο να επιλέξουμε τους σκελετούς του παιδιού και της μητέρας. Σε ορισμένα στιγμιότυπα η μητέρα ή το παιδί απομακρύνονται από το κάδρο ή δεν καταφέρνουν να εντοπιστούν (Σχήμα 3.11), σε άλλα εντοπίζονται εκτός από τους σκελετούς της μητέρας και του παιδιού και άλλοι σκελετοί (άλλα παιδιά ή παιχνίδια - κούκλες) που δε μας ενδιαφέρουν (Σχήμα 3.10), ενώ σε κάποια στιγμιότυπα μητέρα και παιδί διασταυρώνονται καθώς αλλάζουν θέσεις μεταξύ τους, καθώς αγκαλιάζονται κλπ. Συνεπώς η διαδικασία της αντιστοίχισης των σκελετών που εντοπίζει το OpenPose στα πρόσωπα που μας ενδιαφέρουν παρουσιάζει ορισμένες δυσκολίες.

Έτσι, ακολουθήσαμε την παρακάτω μέθοδο: Πρώτα, αρχικοποιούμε χειροκίνητα τους σκελετούς του πρώτου στιγμιότυπου κάθε καταγραφής. Υπολογίσαμε το μήκος κορμού του παιδιού και της μητέρας, δηλαδή τις αποστάσεις μεταξύ λαιμού και δεξιού γοφού. Καθώς τα παιδιά που συμμετέχουν είναι μικρής ηλικίας, τα μήκη αυτά έχουν σημαντική διαφορά, ενώ ταυτόχρονα έχουν σημαντική διαφορά και από τα μήκη κορμών των σκελετών των παιχνιδιών που εντοπίζονται λανθασμένα και είναι πολύ μικροί.

Στη συνέχεια για κάθε στιγμιότυπο υπολογίσαμε τις αποστάσεις των πιθανών σκελετών από τους σκελετούς του παιδιού και της μητέρας στο προηγούμενο στιγμιότυπο. Για να υπολογίσουμε τις αποστάσεις αυτές χρησιμοποιούμε το λαιμό και τους γοφούς, που αποτελούν σχετικά σταθερά σημεία του σώματος. Έτσι, από τους πιθανούς σκελετούς βρίσκουμε εκείνους με τη μικρότερη απόσταση από τους προηγούμενους γνωστούς πλέον σκελετούς. Αν η απόσταση αυτή είναι μικρότερη από ένα κατάλληλα επιλεγμένο κατώφλι τότε ο σκελετός αυτός αντιστοιχίζεται στο άτομο. Ταυτόχρονα ελέγχεται και το μήκος κορμού αυτού του σκελετού ώστε να μη διαφέρει περισσότερο από 10% από τα μήκη κορμού του παιδιού και της μητέρας αντίστοιχα. Αν σε κάποιο στιγμιότυπο κανένας σκελετός δεν ικανοποιεί τις παραπάνω προϋποθέσεις τότε θέτουμε τις τιμές των σημείων του σκελετού μηδενικές στο συγκεκριμένο στιγμιότυπο.

Σχήμα 3.11: Στιγμιότυπο BabyAffect στο οποίο η μητέρα έχει φύγει προσωρινά από το κάδρο.

Σε περίπτωση που κάποιος από τους σκελετούς είναι μηδενικός στο προηγούμενο στιγμιότυπο προκειμένου να εντοπιστεί εκ νέου η μητέρα ή το παιδί για κάθε πιθανό σκελετό υπολογίζεται το μήκος κορμού. Αν το μήκος αυτό διαφέρει λιγότερο από 5% (αυστηρότερο κριτήριο από προηγουμένως καθώς εδώ δεν έχουμε το κριτήριο της χρονικής συνέχειας) από το μήκος κορμού που έχει υπολογιστεί για το παιδί ή τη μητέρα τότε ο σκελετός αντιστοιχίζεται στο αντίστοιχο πρόσωπο.

Στη συνέχεια, πραγματοποιούμε γραμμική παρεμβολή για σύντομα διαστήματα στα οποία οι τιμές των σκελετών είναι μηδενικές (λιγότερο από 2 sec) και απορρίπτουμε τα μεγαλύτερα διαστήματα στα οποία δεν έχουμε καταφέρει να εντοπίσουμε κάποιον από τους δύο σκελετούς. Μας ενδιαφέρει εξίσου και ο σκελετός της μητέρας καθώς το engagement εδώ ορίζεται ως προς μία από κοινού δραστηριότητα των παιδιών με τις μητέρες τους. Επομένως, αν για παράδειγμα η μητέρα αποχωρήσει για λίγο από τη σκηνή δεν έχει ουσιαστικά νόημα η εκτίμηση του engagement.

Τέλος, μετασχηματίζουμε τις τιμές των σημείων του σκελετού των παιδιών ώστε να εκφράζουν τις διαφορές από το σημείο της μύτης του σκελετού της μητέρας τους, από το κέντρο του προσώπου της δηλαδή και ακολουθούμε την υπόλοιπη επεξεργασία που έγινε και για τα δεδομένα BabyRobot.

3.3 Σύγκριση των συνόλων δεδομένων και συμπεράσματα

Παρακάτω πραγματοποιούμε μία σύντομη σύγκριση των δεδομένων των παραπάνω ομάδων, ώστε να προκύψουν ορισμένα συμπεράσματα για το engagement των παιδιών. Για την εξαγωγή των συμπερασμάτων αυτών συμβουλευτήκαμε την καθηγήτρια ψυχολόγο κα. Χριστίνα Παπαηλιού. Η σύγκριση αυτή καθώς και τα συμπεράσματα που απορρέουν από αυτή θα μας βοηθήσουν αργότερα και στην εξήγηση των αποτελεσμάτων των πειραμάτων μας. Στον πίνακα 3.1 παρουσιάζουμε την κατανομή των στιγμιοτύπων σε τρεις κλάσεις (ενωμένες οι μεσαίες κλάσεις) για τα σύνολα δεδομένων TD-Joint Attention, ASD-Joint Attention, ASD Other Games, ASD-School και BabyAffect. Ακόμη, στον πίνακα συμπεριλαμβάνουμε αντίστοιχα στοιχεία για καταγραφές παρόμοιες με τις καταγραφές ASD-Joint Attention και ASD Other Games αλλά με άνθρωπο-ψυχολόγο στη

θέση του ρομπότ (ASD-Joint Attention Human και ASD Games Human). Συμπεριλαμβάνουμε επίσης και στοιχεία για καταγραφές παρόμοιες με τις καταγραφές BabyAffect στις οποίες όμως συμμετέχουν τυπικώς αναπτυσσόμενα παιδιά (TD-BabyAffect).

	Distribution (%)			Total #
Data	Class0	Class1	Class2	Frames
TD-Joint Attention-Robot	7.80	83.42	8.78	108,408
ASD-Joint Attention-Robot	17.68	66.07	16.22	50,869
ASD-Games-Robot	19.35	70.09	10.56	109,381
ASD-Joint Attention-Human	34.62	50.00	15.38	4,680
ASD-Games-Human	31.98	52.69	15.33	75,150
ASD-BabyAffect	14.72	71.48	13.80	27,207
TD-BabyAffect	31.00	48.78	20.22	26,830

Table 3.1: Κατανομή κλάσεων για τα διάφορα σύνολα δεδομένων.

Κατ΄ αρχάς, από την κατανομή στον Πίνακα φαίνεται η σημαντική ανισοκατανομή των στιγμιοτύπων, στην οποία έχουμε αναφερθεί και προηγουμένως, και η οποία ισχύει για όλα τα σύνολα δεδομένων. Εκτός όμως από την παρατήρηση αυτή μεγάλη χρησιμότητα έχει η σύγκριση των δεδομένων ως προς τις ακραίες κλάσεις, δηλαδή ως προς τα χρονικά διαστήματα στα οποία τα παιδιά ήταν είτε πλήρως engaged, είτε πλήρως disengaged. Συγκρίνουμε τα χρονικά διαστήματα στα οποία τα TD και τα ASD παιδιά είναι engaged ή disengaged όταν αλληλεπιδρούν με ανθρώπους ή με ρομπότ, όταν αλληλεπιδρούν σε ελεύθερες ή σε περισσότερο καθορισμένες συνθήκες.

Από την κατανομή στον Πίνακα παρατηρούμε ότι τα ASD παιδιά είναι για πολύ περισσότερη ώρα disengaged όταν αλληλεπιδρούν με ανθρώπους σε συνθήκες εργαστηρίου (πάνω από το 30% του χρόνου) από ότι όταν αλληλεπιδρούν με ρομπότ στις αντίστοιχες συνθήκες (λιγότερο από 20% του χρόνου). Η παρουσία του ρομπότ και οι καθορισμένες συνθήκες στο πλαίσιο των οποίων λαμβάνει χώρα η αλληλεπίδρασή τους τα παρακινεί να επικοινωνήσουν για περισσότερο χρόνο. Η παρατήρηση αυτή ενισχύει αντίστοιχες παρατηρήσεις που συναντάμε στη βιβλιογραφία για το πως τα παιδιά με διαταραχές αυτιστικού φάσματος μπορούν να επωφεληθούν από την αλληλεπίδρασή τους με κοινωνικά ρομπότ.

Ενισχυτικά σε αυτό, ενώ, τα τυπικώς αναπτυσσόμενα (TD) παιδιά μένουν για περισσότερη ώρα πλήρως engaged όταν αλληλεπιδρούν στο σπίτι με τις μητέρες τους από ότι στη συνθήκη Joint Attention με το ρομπότ (20.22% αντί 8.78%), τα ASD παιδιά μένουν για μεγαλύτερα χρονικά διαστήματα πλήρως engaged σε μία πιο καθορισμένη συνθήκη (όπως η συνθήκη Joint Attention) αλληλεπιδρώντας με ένα ρομπότ (16.22% αντί 13.80%).

Προς επίρρωση των παραπάνω παρατηρήσεων είναι αξιοσημείωτο ότι κατά τις αλληλεπιδράσεις με το ρομπότ στην κοινή συνθήκη Joint Attention τα παιδιά με διαταραχές αυτιστικού φάσματος παραμένουν πλήρως engaged το διπλάσιο χρόνο από ότι τα τυπικώς αναπτυσσόμενα παιδιά (16.22% και 8.78% αντίστοιχα). Πρέπει να σημειώσουμε ότι όπως αναφέραμε και προηγουμένως τόσο στα πειράματά μας όσο και στις κατανομές του επιπέδου του engagement σε κλάσεις που παρουσιάζουμε στην Ενότητα αυτή δεν έχουμε συμπεριλάβει τα 8 από τα 15 παιδιά με αυτιστικές διαταραχές που συμμετείχαν στο πρόγραμμα BabyRobot. Τα παιδιά αυτά αποτελούν σοβαρές περιπτώσεις στο φάσμα του αυτισμού και δεν αποκρίθηκαν καθόλου σε κανένα από τα παιχνίδια είτε συμμετείχε ρομπότ είτε συμμετείχε άνθρωπος. Για το λόγο αυτό εξαιρέθηκαν από τα πειράματά μας καθώς δεν ήταν δυνατό να βοηθήσουν στην αξιολόγηση των αποτελεσμάτων των αλληλεπιδράσεων ή στην προσπάθεια εκτίμησης του engagement.

(α′) Κλάση 0

(β′) Κλάση 1

(γ′) Κλάση 2

Σχήμα 3.12: Στιγμιότυπα από τις τρεις διαφορετικές κλάσεις engagement.

3.4 Το σύνολο δεδομένων της PInSoRo

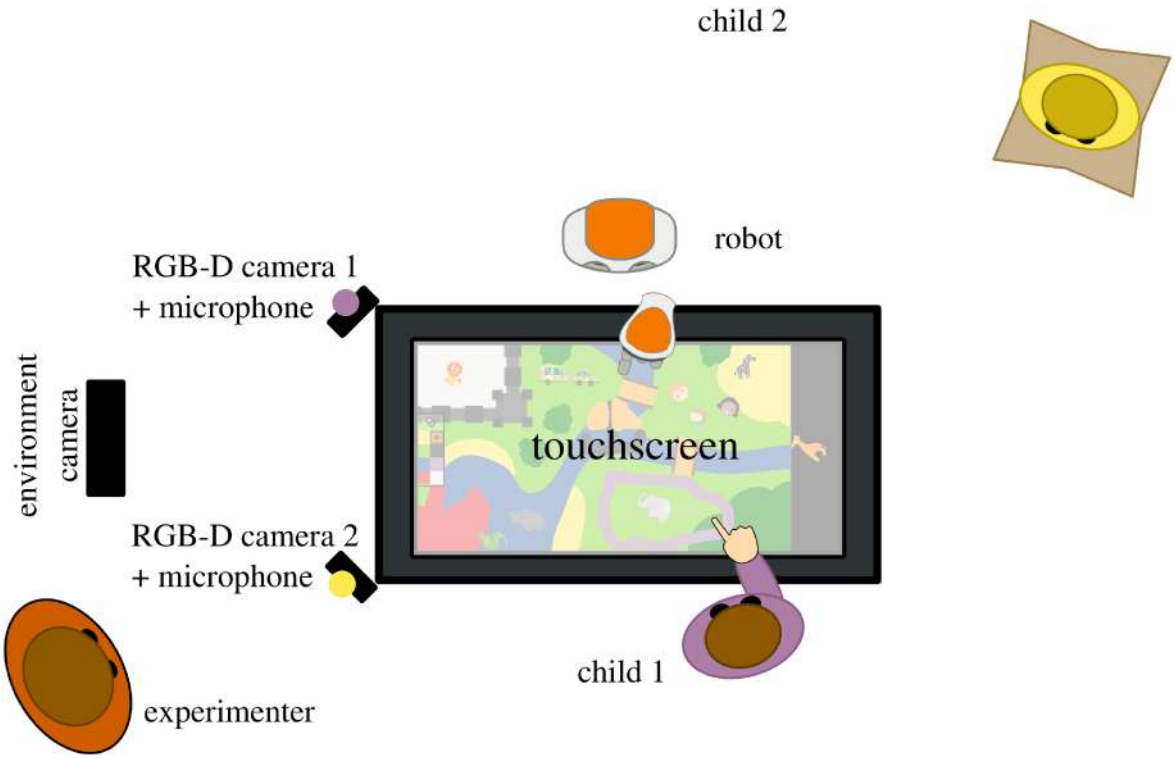
Το σύνολο δεδομένων της PInSoRo παρουσιάζεται αναλυτικά στο "The PInSoRo dataset" των Lemaignan et al. [3]. Πρόκειται για ένα ελεύθερο σύνολο δεδομένων από κοινωνικές συναναστροφές παιδιών, που σχεδιάστηκε ώστε να ταιριάζει σε ερευνητικές μεθόδους βασιζόμενες στα δεδομένα. Σε αυτό περιλαμβάνονται αλληλεπιδράσεις ελεύθερου παιχνιδιού, εσκεμμένα μη πλήρως καθορισμένες, αλλά μεθοδολογικά συνεπείς. Με αυτό τον τρόπο συλλαμβάνεται ένα πλούσιο σύνολο μοτίβων συμπεριφοράς που συναντώνται σε αλληλεπιδράσεις μεταξύ παιδιών. Το τελικό σύνολο δεδομένων αποτελείται από περισσότερες από 45 ώρες καταγραφών κοινωνικών αλληλεπιδράσεων μεταξύ 45 ζευγαριών παιδιών και 30 ζευγαριών παιδιών-ρομπότ. Τα παιδιά ήταν σε ηλικίες από 4 έως 8 ετών. Όλες οι καταγραφές είναι επισημειωμένες από ερευνητές. Το σύνολο των δεδομένων περιλαμβάνει εκτός από επισημειώσεις κοινωνικών συμπεριφορών, πλήρως βαθμονομημένες καταγραφές video, 3D καταγραφές των προσώπων, πληροφορίες για το σκελετό, πλήρεις καταγραφές ήχου καθώς και καταγραφές των παιχνιδιών.

3.4.1 Καταγραφή των αλληλεπιδράσεων

Οι καταγραφές της PInSoRo πραγματοποιούνται γύρω από το free-play sandbox task, το οποίο βασίζεται σε αλληλεπιδράσεις ελεύθερου παιχνιδιού πρόσωπο με πρόσωπο στις οποίες παρεμβάλλεται μία μεγάλη, οριζόντια οθόνη αφής. Ζευγάρια παιδιών (ή αντίστοιχα, ένα παιδί και ένα ρομπότ) καλούνται να σχεδιάσουν ελεύθερα και να αλληλεπιδράσουν με εικονιζόμενα αντικείμενα σε ένα διαδραστικό τραπέζι, χωρίς να τίθενται συγκεκριμένοι στόχοι από τον πειραματιστή. Η δραστηριότητα έχει σχεδιαστεί ώστε τα παιδιά να μπορούν εμπλακούν σε παιχνίδι μη κατευθυντικό, χωρίς ορισμένο τέλος. Παρ΄ όλα αυτά, είναι επαρκώς περιορισμένο ώστε να είναι κατάλληλο για καταγραφή.

Τα πρόσωπα των παιδιών καταγράφονταν με δύο Red-Green-Blue-Depth (RGB-D) κάμερες μικρού εύρους (0.2μ. έως 1.2μ.), που βρίσκονταν τοποθετημένες στις γωνίες της

οθόνης αφής και κεκλιμένες ώστε να κοιτούν τα πρόσωπα των παιδιών. Οι κάμερες ήταν σταθερά τοποθετημένες σε ειδικά διαμορφωμένους βραχίονες, ώστε να καθίσταται δυνατή η ακριβής μέτρηση της 3D πόζας ως προς την οθόνη αφής (εξωτερική βαθμονόμηση). Ακόμη, μία τρίτη RGB κάμερα κατέγραφε όλο το περιβάλλον της αλληλεπίδρασης. Η καταγραφή αυτή προοριζόταν για να χρησιμεύσει στους ερευνητές που πραγματοποίησαν τις επισημειώσεις και δεν ήταν βαθμονομημένη με ακρίβεια. Στην περίπτωση ζεύγους παιδιού-ρομπότ, χρησιμοποιήθηκε ένα ρομπότ Nao. Το ρομπότ παρέμενε σε όρθια θέση καθ΄ όλη τη διάρκεια των αλληλεπιδράσεων.

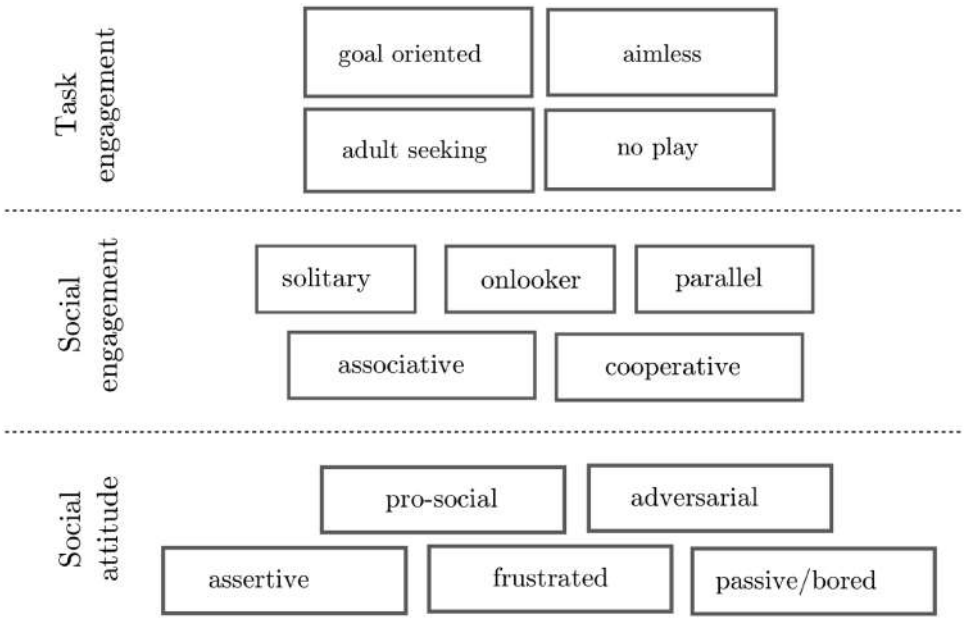


Σχήμα 3.13: Το πειραματικό περιβάλλον της PInSoRo [3].

3.4.2 Επισημειώσεις των αλληλεπιδράσεων

Το σύνολο των δεδομένων επισημειώθηκε με τη χρήση ενός συνδυασμού τριών σχημάτων κωδικοποίησης κοινωνικών συναναστροφών. Αυτά τα σχήματα κωδικοποίησης έχουν προκύψει από προσαρμογή καθιερωμένων κλιμάκων κωδικοποίησης κοινωνικής συμπεριφοράς. Το τελικό σχήμα κωδικοποίησης περιείχε τρεις συγκεκριμένους άξονες: το επίπεδο του task engagement (που διακρίνει μεταξύ συγκεντρωμένων, προσανατολισμένων στο παιχνίδι συμπεριφορών και μη στοχοπροσηλωμένων συμπεριφορών, που ωστόσο μπορεί να είναι σε ορισμένες περιπτώσεις ιδιαίτερα κοινωνικές), το επίπεδο του social engagement (που βασίζεται στα στάδια του Parten's για το παιχνίδι και περιγράφει περισσότερο τη σχέση μεταξύ των παιδιών) και το social attitude (που κωδικοποιεί συμπεριφορές όπως υποστήριξη, επιθετικότητα, κυριαρχία, ενόχληση, κ.α.) [3]. Από τους τρεις αυτούς άξονες εργαστήκαμε με το task engagement.

Συγκεκριμένα, το task engagement στοχεύει στο να επιτύχει μία ευρεία διάκριση μεταξύ συμπεριφορών που σχετίζονται με το παιχνίδι-'on-task' και που δε σχετίζονται με το παιχνίδι-'off-task' (παρ΄ όλο που το free-play sandbox δεν απαιτούσε σαφώς από τα παιδιά



Σχήμα 3.14: Οι άξονες επισημειώσεων της PInSoRo και οι επιμέρους κλάσεις που χρησιμοποιούνται [3].

να προσπαθήσουν για ένα συγκεκριμένο στόχο, υπήρχε ένας υποκείμενος στόχος: να παίξουν με το παιχνίδι). Οι συμπεριφορές που σχετίζονται με το παιχνίδι χαρακτηρίζονται ως goal oriented και περιλαμβάνουν συγκεκριμένες, σχεδιασμένες ενέργειες (που μπορεί να είναι ή να μην είναι κοινωνικές). Συμπεριφορές που δε σχετίζονται με το παιχνίδι, όπως όταν τα παιδιά φέρονται ανόητα, συζητούν για διαφορετικά πράγματα ή γελούν μεταξύ τους κλπ χαρακτηρίζονται ως aimless. Οι συμπεριφορές αυτής της κατηγορίας ήταν ορισμένες φορές ιδιαίτερα κοινωνικές και συνέβαλαν σημαντικά στην εδραίωση της εμπιστοσύνης και της συνεργασίας μεταξύ των παιδιών. Από αυτή την άποψη, θεωρούνται εξίσου σημαντικές με τις συμπεριφορές που ανήκουν στην κατηγορία goal oriented. Στη συνέχεια, ακολουθούν οι συμπεριφορές που κατηγοριοποιούνται ως adult seeking και αφορούν στιγμές που τα παιδιά αναζητούν την έγκριση του επιβλέποντα ενήλικα ή ρωτούν ερωτήσεις κλπ. Τελευταία βρίσκεται η κατηγορία no play στην οποία τα παιδιά δεν πραγματοποιούν καμία προφανή δραστηριότητα.

Παρ΄ όλο που όπως αναφέρεται και παρακάτω οι κατηγορίες aimless και adult seeking ήταν σημαντικά περιορισμένες σε σχέση με τις άλλες δύο και κυρίως σε σχέση με την κατηγορία goal oriented δεν δοκίμασα να ενώσω τις κατηγορίες αυτές, λόγω της μεγάλης σημασίας που αποδίδουν οι δημιουργοί της PInSoRo στην κατηγορία aimless.

Η επισημείωση του συνόλου των δεδομένων πραγματοποιήθηκε από πέντε ανθρώπους με πείρα στην επισημείωση δεδομένων. Συνολικά, προέκυψαν 13289 επισημειώσεις και η μέση διάρκεια των επισημειωμένων επεισοδίων είναι 48.8 δευτερόλεπτα.

Καθώς κατατάσσονται οι κοινωνικές συμπεριφορές ανάλογα με την ηλικία των παιδιών, παρατηρούνται οι αναμενόμενες τάσεις: καθώς τα παιδιά μεγαλώνουν μειώνονται οι συμπεριφορές adult seeking, μεταξύ μεγαλύτερων παιδιών παρατηρείται συχνότερα cooperative play, ενώ στις μικρότερες ηλικίες κυριαρχεί η κατηγορία parallel play. Αντίθετα, για τον άξονα social attitudes οι κλάσεις είναι ομοίως κατανεμημένες στις διαφορετικές ηλικίες.

3.4.3 Επεξεργασία των καταγραφών

Όπως και προηγουμένως με τη βοήθεια του OpenPose εξήχθησαν για κάθε παιδί ο σκελετός (18 σημεία), 70 σημεία του προσώπου, συμπεριλαμβανομένης της θέσης της κόρης του ματιού καθώς και ο σκελετός των χεριών. Επιπλέον, για κάθε καρέ εξήχθησαν 17 Action Units (AU) με τα αντίστοιχα επίπεδα βεβαιότητας για το καθένα, με τη χρήση του OpenFace. Τα action units που αναγνωρίστηκαν από το OpenFace και παρέχονται μαζί με τα υπόλοιπα δεδομένα είναι AU01, AU02, AU04, AU05, AU06, AU07, AU09, AU10, AU12, AU14, AU15, AU17, AU20, AU23, AU25, AU26, AU28 και AU45.

Ακόμη, υπολογίστηκε η κατεύθυνση του βλέμματος με τη χρήση δύο τεχνικών. Πρώτα, πραγματοποιήθηκε εκτίμηση της πόζας του κεφαλιού και χρησιμοποιήθηκε για την εκτίμηση της κατεύθυνσης του βλέμματος. Παρ΄ όλο που η τεχνική αυτή είναι αποτελεσματική για να γνωρίζουμε την κατεύθυνση του βλέμματος σε ένα γενικό επίπεδο (π.χ. βλέμμα στο τραπέζι του παιχνιδιού ή βλέμμα στο άλλο παιδί ή βλέμμα στον ερευνητή), προσφέρει περιορισμένη ακρίβεια όταν θέλουμε να προσδιορίσουμε για παράδειγμα το ακριβές σημείο του τραπεζιού του παιχνιδιού στο οποίο κατευθύνεται το βλέμμα (αφού δεν προσδιορίζεται η θέση της κόρης του ματιού). Η εκτίμηση της πόζας του κεφαλιού πραγματοποιήθηκε από ένα νευρωνικό δίκτυο (ένα απλό perceptron με επτά πλήρως συνδεδεμένες στιβάδες, ενεργοποιήσεις Rectified Linear Unit (ReLU) και 64 μονάδες ανά στιβάδα, που εφαρμόστηκε με τη χρήση του Caffe.

Για τη δεύτερη τεχνική, εκπαιδεύτηκε ένα δίκτυο από μία πραγματική αντιστοιχία των προσώπων των παιδιών και των 2D συντεταγμένων του βλέμματός τους. Τα δεδομένα εκπαίδευσης αποκτήθηκαν ζητώντας από τα παιδιά να ακολουθήσουν ένα στόχο στην οθόνη για σύντομο χρονικό διάστημα πριν ξεκινήσουν το παιχνίδι. Η θέση του στόχου αντιπροσωπεύει τος πραγματικές συντεταγμένες του βλέμματος στην οθόνη. Για κάθε καρέ, το δίκτυο λαμβάνει ως είσοδο ένα διάνυσμα χαρακτηριστικών που αποτελείται από 32 σημεία προσώπου και σκελετού (x, y) σχετικά με την εκτίμηση του βλέμματος (θέση της κόρης κάθε ματιού, περιγράμματα ματιών, φρύδια, μύτη, λαιμός, ώμοι και αυτιά). Με τη χρήση της τεχνικής αυτής μετρήθηκε στα δεδομένα ελέγχου σφάλμα 12.8% μεταξύ της πραγματικής θέσης του στόχου στην οθόνη και της εκτιμώμενης κατεύθυνσης του βλέμματος (δηλ. 9εκ σε μία οθόνη πλάτους 70εκ). Το ίδιο προεκπαιδευμένο δίκτυο χρησιμοποιείται στη συνέχεια ώστε να παρέχει εκτιμήσεις της κατεύθυνσης του βλέμματος για το υπόλοιπο του παιχνιδιού.

Τα Action Units που περιλαμβάνονται στα χαρακτηριστικά που παρέχει η PInSoRo για τις καταγραφές της ανήκουν στο Facial Action Coding System (FACS). Πρόκειται για ένα σύνολο από κινήσεις των μυών του προσώπου (Action Units) που αντιστοιχούν σε κάποιο εκπεμπόμενο συναίσθημα. Με τη βοήθεια του FACS, μπορούμε να καθορίσουμε το συναίσθημα ενός ανθρώπου η εικόνα του οποίου καταγράφεται. Η ανάλυση αυτή των εκφράσεων του προσώπου είναι μία από τις πολύ λίγες διαθέσιμες τεχνικές για την εκτίμηση συναισθημάτων σε πραγματικό χρόνο.

Για πολλά χρόνια οι ερευνητές βασιζόμενοι στα action units που περιγράφονται από το FACS επισημείωναν χειροκίνητα καταγραφές ανθρώπων, γεγονός ιδιαίτερα χρονοβόρο. Η διαδικασία αυτή μπορεί πλέον να ολοκληρωθεί με αυτόματη ανάλυση έκφρασης του προσώπου με τη βοήθεια τεχνικών όρασης υπολογιστών.

Το OpenFace [44] είναι μία βιβλιοθήκη που επιτρέπει σε μεγάλο βαθμό τον εντοπισμό βασικών σημείων του προσώπου, την εκτίμηση της πόζας του κεφαλιού, την αναγνώριση action units στο πρόσωπο και την εκτίμηση της κατεύθυνσης του βλέμματος ανθρώπων από εικόνες και video.

3.4.4 Επεξεργασία των διανυσμάτων χαρακτηριστικών

Δουλέψαμε με το πλήρως διαθέσιμο κομμάτι της PInSoRo. Σε αυτό για κάθε μία καταγραφή είναι διαθέσιμο ένα αρχείο comma-separated values (csv), το οποίο περιλαμβάνει όλα τα χαρακτηριστικά του κυρίως συνόλου δεδομένων, δειγματοληπτημένα στα 30Hz.

Τα χαρακτηριστικά που περιλαμβάνονται σε κάθε σειρά του αρχείου αυτού είναι τα εξής:

- 'timestamp': το UNIX timestamp της σειράς. Το πρώτο timestamp κάθε καταγραφής είναι είναι το timestamp του πρώτου καταγεγραμμένου καρέ (από οποιαδήποτε από τις δύο κάμερες)(0).

- 'id': η ταυτότητα αυτής της καταγραφής (απλώς η ημερομηνία και η ώρα έναρξης του πειράματος - συνήθως λίγα λεπτά πριν το timestamp του πρώτου καρέ του video(1).

- 'condition': child-child αν πρόκειται για καταγραφή στην οποία συμμετέχουν δύο παιδιά (ένα με μοβ και ένα με κίτρινο γιλέκο- το μοβ και κίτρινο παιδί αντίστοιχα) ή αν συμμετέχει ένα μόνο παιδί με το ρομπότ (το ρομπότ παίρνει σε όλες τις καταγραφές τη θέση του κίτρινου παιδιού και όλα τα αντίστοιχα πεδία είναι κενά)(2).

- 'annotators': τα ονόματα των ανθρώπων που πραγματοποίησαν τις επισημειώσεις αυτής της αλληλεπίδρασης. Αν δεν υπάρχει επισημείωση το πεδίο αυτό είναι κενό, ενώ αν έκαναν τις επισημειώσεις περισσότεροι από ένας τα ονόματα χωρίζονται από ένα ′+′(3).

- 'complete': είναι αληθές αν όλα τα δεδομένα είναι διαθέσιμα για αυτό το timestamp(4).

- 'purple_child_age', 'purple_child_gender', 'yellow_child_age', 'yellow_child_gender': οι ηλικίες και τα φύλα των παιδιών που συμμετέχουν (5-8).

- 'purple_frame_idx': ο αριθμός του καρέ στη μετάδοση video της μοβ κάμερας (9).

- 'purple_child_face00..69_x,y': οι 2D συντεταγμένες 70 σημείων προσώπου (μεταξύ των οποίων και οι κόρες των ματιών), κανονικοποιημένες στο διάστημα [0.0, 0.1], όπως εξάγονται από το OpenPose (10-149).

- 'purple_child_skel00..17_x,y': οι 2D συντεταγμένες 18 σημείων σκελετού, κανονικοποιημένες στο διάστημα [0.0, 0.1], όπως εξάγονται από το OpenPose (150-185).

- 'purple_child_head_x,y,z,rx,ry,rz': η εκτίμηση της πόζας του κεφαλιού σε m και rad, αναφορικά με το κέντρο του τραπεζιού. Έχει υπολογιστεί με τη χρήση του OpenFace (186-191).

- 'purple_child_gaze_x,y,z': το διάνυσμα του βλέμματος, σταθμισμένο και για τα δύο μάτια, αναφορικά με το κέντρο του τραπεζιού. Έχει υπολογιστεί με τη χρήση του OpenFace (192-194).

- 'purple_child_au01,02,04,05,06,07,09,10,12,14,15,17,20,23,25,26,28,45': Η ένταση 18 action units του προσώπου. Έχουν υπολογιστεί με τη χρήση του OpenFace (195-212).

- 'purple_child_motion_intensity_avg,stdev,max': ο μέσος όρος, η τυπική απόκλιση και το μέγιστο του πλάτους της κίνησης που παρατηρείται στο καρέ. Αυτά υπολογίζονται εφαρμόζοντας υπολογισμό οπτικής ροής με τη χρήση του αλγορίθμου Dual TVL1 και παίρνοντας το μέσο όρο των τιμών σε ολόκληρο το καρέ (213-215).

- 'purple_child_motion_direction_avg,stdev': ο μέσος όρος και η τυπική απόκλιση της κατεύθυνσης της κίνησης που παρατηρείται στο καρέ. Αυτά υπολογίζονται εφαρμόζοντας υπολογισμό οπτικής ροής με τη χρήση του αλγορίθμου Dual TVL1 και παίρνοντας το μέσο όρο των τιμών σε ολόκληρο το καρέ (216-217).

- Τα ίδια χαρακτηριστικά και για την κίτρινη κάμερα-κίτρινο παιδί. Τα χαρακτηριστικά αυτά λείπουν στις περιπτώσεις αλληλεπίδρασης παιδιού-ρομπότ (218,219-358,359-394,395-400,401-403,404-421,422-424,425-426).

- 'audio_00..15': χαρακτηριστικά ήχου που προς το παρόν λείπουν (427-442).

- 'purple,yellow_child_task_engagement,social_engagement,social_attitude': οι επισημειώσεις για το συγκεκριμένο στιγμιότυπο. Αν έκαναν τις επισημειώσεις περισσότεροι από ένας και υπάρχει διαφωνία τότε οι διαφορετικές αυτές τιμές χωρίζονται από ένα ′+′ (443-448).

Για την προεπεξεργασία των δεδομένων ακολουθήσαμε την εξής διαδικασία. Αρχικά, διαβάσαμε τα αρχεία csv με τη βοήθεια της βιβλιοθήκης Pandas. Ήμαστε υποχρεωμένοι να κρατήσουμε μόνο τα στιγμιότυπα εκείνα για τα οποία η τιμή 'complete' είναι αληθής για να εξασφαλίσουμε ότι δε θα λείπουν από τα στιγμιότυπα της εκπαίδευσής οι επισημειώσεις ή ολόκληρα κανάλια πληροφορίας. Τα δεδομένα από τα csv αρχεία είναι complete σε ποσοστό περίπου 56,6%. Παρατηρήσαμε ότι σε πολλά από τα video πολύ σημαντικό ποσοστό των timestamps δεν ήταν complete. Συνεπώς, αφαιρώντας όλα αυτά τα timestamps τα δεδομένα που προκύπτουν σίγουρα δε διατηρούσαν την χρονική συνέχεια, που είναι πολύ σημαντική για την εκμάθηση εκτίμησης του engagement.

Επεξεργαστήκαμε ένα σύνολο δεδομένων που περιλαμβάνει μόνο 20 καταγραφές στις οποίες περισσότερο από 85% των timestamps είναι πλήρη. Μάλιστα παρατηρήσαμε ότι στις καταγραφές αυτές τα μη πλήρη timestamps είναι συγκεντρωμένα στην πλειοψηφία τους στην αρχή και στο τέλος των καταγραφών (σε αντίθεση με τις υπόλοιπες καταγραφές που βρίσκονται διασπαρμένα σε όλη τη διάρκειά τους). Αυτό το υποσύνολο της PInSoRo περιλαμβάνει 573.889 timestamps, έναντι 1.589.903 timestamps. Ακόμη, αποτελείται κυρίως από καταγραφές στις οποίες ένα παιδί αλληλεπιδρά με το ρομπότ και όχι καταγραφές με ζεύγη παιδιών.

Στη συνέχεια, επιλέξαμε χαρακτηριστικά που θα χρησιμοποιούνταν για την εκπαίδευσή των δικτύων. Για ορισμένα από τα χαρακτηριστικά που δίνονται λείπουν οι τιμές σε σημαντική έκταση των δεδομένων. Για τα σημεία του σκελετού των παιδιών οι θέσεις των γοφών, των γονάτων και των ποδιών των παιδιών (8-13) δεν υπάρχουν για τη συντριπτική πλειοψηφία των στιγμιότυπων. Το γεγονός αυτό είναι απολύτως λογικό καθώς κατά τη διάρκεια του παιχνιδιού τα παιδιά βρίσκονται καθισμένα μπροστά στο τραπέζι/οθόνη αφής. Ακόμη, οι θέσεις και των δύο χεριών των παιδιών (4,7) λείπουν στα μισά περίπου στιγμιότυπα και επομένως παραλείψαμε και αυτά τα χαρακτηριστικά. Τέλος, λείπει στα περισσότερα στιγμιότυπα η θέση του ενός από τα δύο αυτιά των παιδιών (16/17, ανάλογα από ποια πλευρά του παιδιού είναι τοποθετημένη η κάμερα που το καταγράφει) και έτσι παραλείψαμε και το αντίστοιχο χαρακτηριστικό για κάθε παιδί. Τα υπόλοιπα σημεία του

Χαρακτηριστικά	
0-15	purple_child_skel_{0,1,2,3,5,14,15,17}_{x,y}
16-21	purple_child_head_{x,y,z,rx,ry,rz}
22-24	purple_child_gaze_{x,y,z}
25-41	purple_child_au{01,02,04,05,06,07,09,10,12,14,15,17,20,23,25,26,45}
42-44	purple_child_motion_intensity_{avg,stdev,max}
45-46	purple_child_motion_direction_{avg,stdev}
47	purple_task_engagement

Πίνακας 3.2: Το τελικό διάνυσμα χαρακτηριστικών για τα δεδομένα της PInSoRo

Κλάση		Μοβ παιδί	Κίτρινο παιδί
0	noplay	17.44%	28.48%
1	adultseeking	5.78%	5.34%
2	aimless	16.26%	14.23%
3	goal-oriented	60.50%	51.93%

Πίνακας 3.3: Κατανομή των κλάσεων για το task engagement της PInSoRo.

σκελετού: μύτη, μάτια,αυτί (από τη μεριά της κάμερας), λαιμός, ώμοι, αλλά και οι αγκώνες (0-3,5,6,14,15,16/17) υπάρχουν για σημαντικό κομμάτι του συνόλου των στιγμιότυπων.

Τέλος, για να μειώσουμε τον όγκο των δεδομένων παραλείψαμε τα χαρακτηριστικά του προσώπου καθώς κατά βάση η πληροφορία που εμπεριέχουν, μπορεί να εξαχθεί και από τον συνδυασμό των υπολοίπων χαρακτηριστικών (αντίστοιχα σημεία σκελετού, θέση και προσανατολισμός κεφαλιού, προσανατολισμός βλέμματος, action units). Παραλείποντας τα χαρακτηριστικά αυτά το διάνυσμα χαρακτηριστικών γίνεται τέσσερις φορές μικρότερο και άρα σημαντικά ταχύτερη και ευκολότερη η εκπαίδευση. Συνεπώς, δοκιμάσαμε να εκπαιδεύσουμε τα δίκτυα εκτίμησης με το διάνυσμα χαρακτηριστικών που παρουσιάζεται στον Πίνακα 3.2.

Στον πίνακα 3.3 φαίνεται η κατανομή των κλάσεων. Παρατηρούμε μεγάλη ανισοκατανομή στις κλάσεις, ωστόσο δοκιμάζουμε να μην ενοποιήσουμε την κλάση adultseeking, που εμφανίζεται δώδεκα φορές λιγότερο από την πολυπληθέστερη κλάση (goal-oriented) με την κλάση aimless, παρ΄ όλο που τέτοιου είδους ενοποιήσεις κλάσεων συνηθίζονται σε προσπάθειες εκτίμησης του engagement, καθώς η συμπεριφορές που χαρακτηρίζονται ως aimless έχουν ιδιαίτερη σημασία όπως αναφέρθηκε και παραπάνω. Παρ΄ όλο που δε σχετίζονται άμεσα με το παιχνίδι-στόχο είναι σημαντικές ώστε να χτιστεί εμπιστοσύνη και συνεργασία μεταξύ των συμμετεχόντων στο παιχνίδι.

Κεφάλαιο 4

Εκτίμηση Ανθρώπινης Πόζας και Αναγνώριση Δράσης με τη Βοήθεια της Πόζας

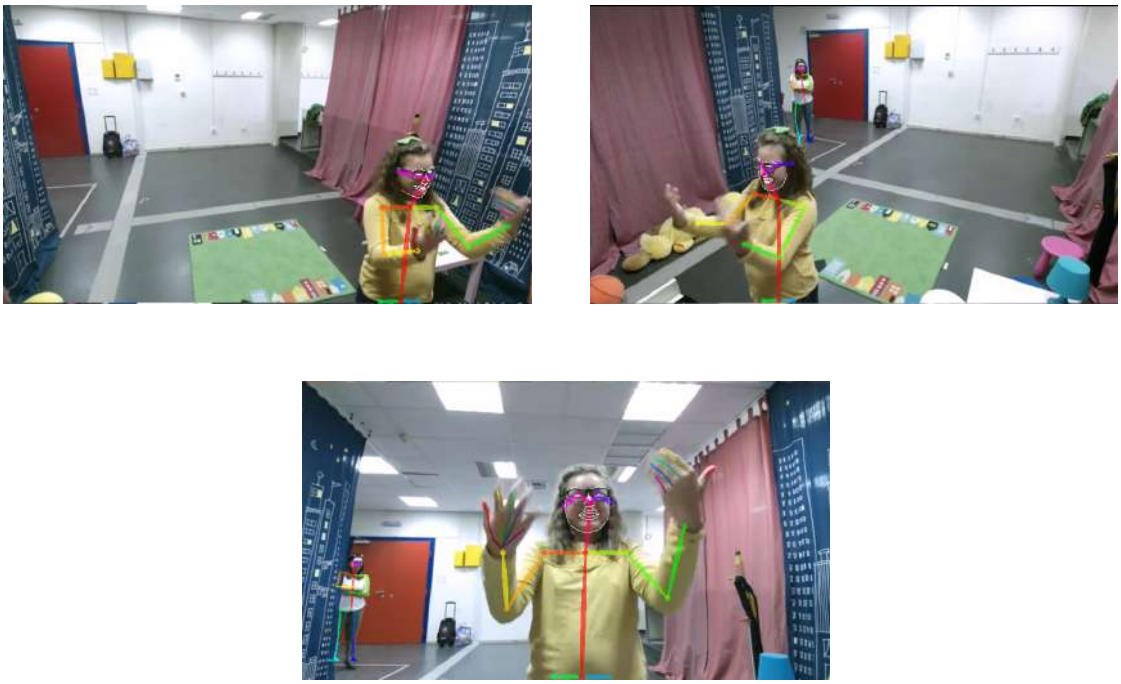
Η εκτίμηση ανθρώπινης πόζας είναι η διαδικασία αυτόματου προσδιορισμού της διαμόρφωσης του σώματος από μία μόνο εικόνα. Αποτελεί ένα από τα βασικά προβλήματα όρασης υπολογιστών και έχει μελετηθεί για περισσότερα από 25 χρόνια. Η μεγάλη σημασία της έγκειται στο γεγονός ότι επιτρέπει ή βελτιώνει σημαντικά μία τεράστια ποικιλία εφαρμογών. Με την εκτίμηση της ανθρώπινης πόζας επιτυγχάνεται ένα ανώτερο επίπεδο αλληλεπίδρασης ανθρώπου υπολογιστή (Human Computer Interaction) καθώς και αναγνώρισης δράσης (Action Recognition) [45].

Στις πιο σημαντικές προκλήσεις για την επιτυχή αντιμετώπιση αυτού του προβλήματος συγκαταλέγονται: (1) η ποικιλία στους τρόπους με τους οποίους εμφανίζονται οι άνθρωποι στις διάφορες εικόνες, (2) η μεταβλητότητα των συνθηκών φωτισμού, (3) η ποικιλία της ανθρώπινης σωματικής διάπλασης, (4) η κάλυψη μεταξύ μερών του ανθρώπινου σώματος λόγω της άρθρωσης του και των διάφορων στρωμάτων αντικειμένων στη σκηνή, (5) η πολυπλοκότητα της ανθρώπινης σκελετικής δομής, (6) το πλήθος των βαθμών ελευθερίας της πόζας και (7) η απώλεια τρισδιάστατων πληροφοριών που προκύπτει από την παρατήρηση της στάσης μέσω δισδιάστατων εικονικών προβολών. [45]. Επιπροσθέτως, στις εικόνες στις οποίες υπάρχουν πολλοί άνθρωποι το πρόβλημα γίνεται δυσκολότερο καθώς είναι άγνωστος ο αριθμός των ανθρώπων που μπορεί να βρίσκονται σε οποιαδήποτε θέση και κλίμακα της εικόνας, η αλληλεπίδραση μεταξύ των ανθρώπων εμποδίζει την αντιστοίχιση μερών του σώματος με τα σωστά άτομα και τέλος, η πολυπλοκότητα του προβλήματος τείνει να αυξάνεται μαζί με τον αριθμό των ανθρώπων στην εικόνα [4].

4.1 Εκτίμηση ανθρώπινης πόζας για μεμονωμένα άτομα

Για πολλά χρόνια η προσπάθεια εστιαζόταν στην εκτίμηση της πόζας για μεμονωμένα άτομα. Περιγράφουμε σύντομα ορισμένες προσεγγίσεις του προβλήματος, η εξέλιξη των οποίων δείχνει την τεράστια πρόοδο που έχει επιτευχθεί στην εκτίμηση της πόζας.

Στο "Strike a Pose" των Ramanan et al. [46] το 2005 ακολουθήθηκε η εξής στρατηγική που αποτελείται από δύο στάδια: Αρχικά σε ένα σύνολο διαδοχικών εικόνων εφαρμόζεται ένας ανιχνευτής ανθρώπου που περπατάει κατά μήκος της εικόνας. Κάτι τέτοιο



Σχήμα 4.1: Παράδειγμα στιγμιότυπου BabyRobot για το οποίο έχουμε εξάγει την πόζα με τη χρήση του OpenPose σε τρεις διαφορετικές όψεις.

επιλέχθηκε ώστε να γίνεται με μεγαλύτερη επιτυχία η ανίχνευση και ώστε να εκπαιδεύεται ο ανιχνευτής στην εμφάνιση των άκρων του ανθρώπινου σώματος (πόδια, χέρια, κορμός, κεφάλι) του συγκεκριμένου ατόμου στις συγκεκριμένες εικόνες. Στις εικόνες εφαρμοζόταν ανιχνευτής ακμών και στη συνέχεια βρίσκονταν κατάλληλα ορθογώνια για κάθε άκρο, ώστε να πληρούνται προϋποθέσεις θέσης, αποστάσεων, γωνιών μεταξύ τους για τη συγκεκριμένη πόζα του περπατήματος και να σχηματίζεται ένας ανθρώπινος σκελετός. Με τη βοήθεια των ορθογωνίων που προέκυπταν ο ανιχνευτής μάθαινε την εμφάνιση κάθε άκρου. Στο δεύτερο στάδιο, ο ανιχνευτής εμφάνισης άκρων που προέκυψε εφαρμοζόταν στη συνέχεια στο σύνολο των εικόνων, στις οποίες το άτομο μπορεί να βρισκόταν σε άλλη κλίμακα ή σε μία περίεργη πόζα και κατάφερνε σε αρκετές περιπτώσεις να εκτιμήσει με επιτυχία την πόζα αυτή.

Αρκετές προσεγγίσεις βασίζονται σε Pictorial Structures Models (PSMs), με τα οποία γίνεται αναγνώριση αντικειμένων μέσω ενός συνόλου επιμέρους τμημάτων τους, οργανωμένων σε ευέλικτες διατάξεις. Το βέλτιστο ταίριασμα επιλέγεται με την ελαχιστοποίηση μίας συνάρτησης που μετρά τόσο το κόστος της θέσης κάθε μεμονωμένου τμήματος όσο και το κόστος των συνδέσεών μεταξύ των τμημάτων. Ουσιαστικά, η διάταξη του ανθρώπινου σκελετού μοντελοποιείται από ένα σύνολο αυστηρά καθορισμένων μελών και από ένα σύνολο ζευγαρωτών πιθανοτήτων που αντιπροσωπεύουν τις συνδέσεις μεταξύ τους. Οι συνδέσεις μεταξύ των μελών θεωρείται ότι σχηματίζουν μία δενδροειδή μορφή ώστε να επιτρέπεται αποδοτική εξαγωγή συμπερασμάτων σε εύλογο χρόνο.

Για παράδειγμα, στη δουλειά των Johnson et al. [47] το μοντέλο που προτείνεται αποτελείται από δύο δομικά στοιχεία, το πρώτο (appearance) μοντελοποιεί την πιθανότητα ενός μέρους του ανθρώπινου σώματος να βρίσκεται σε μία συγκεκριμένη θέση και να έχει συγκεκριμένο προσανατολισμό με δεδομένη τη δοσμένη εικόνα, ενώ το δεύτερο (prior) μοντελοποιεί την κατανομή της πιθανότητας της πόζας, περιορίζοντας την εκτιμώμενη πόζα ώστε να ταιριάζει σε ανθρώπινο σώμα. Πηγαίνοντας ένα βήμα παραπέρα, η κατανομή πιθανότητας της πόζας δεν ξεκινά από μία μόνο απλή γκαουσιανή αλλά από ένα σύνολο

Σχήμα 4.2: Παράδειγμα στιγμιότυπων ASD-School για το οποία έχουμε εξάγει την πόζα με τη χρήση του OpenPose.

δενδροειδών μοντέλων που κωδικοποιούν ομάδες από παρόμοιες πόζες, ώστε να μοντελοποιούνται οι δυνατές πόζες με μεγαλύτερη πιστότητα. Στα πλαίσια κάθε τέτοιας ομάδας μοντελοποιείται η εμφάνιση των διάφορων σημείων του σώματος (ώστε να συμπεριλαμβάνεται η συσχέτιση μεταξύ τους με δεδομένη την πόζα) με τη βοήθεια μη γραμμικών SVM ταξινομητών.

Οι προσεγγίσεις που χρησιμοποιούν Pictorial Structure Models οδηγούν σε αποδοτικά και ακριβή συμπεράσματα, ωστόσο αποτυγχάνουν να συλλάβουν τις εξαρτήσεις μεταξύ των μη γειτονικών μελών του σώματος. Το πρόβλημα αυτό επιχειρεί να αντιμετωπίσει η παραπάνω προσέγγιση εισάγοντας ένα σύνολο αρχικών δενδροειδών μοντέλων. Στο "Poselet Conditioned Pictorial Structures" [48] για να αντιμετωπιστεί το πρόβλημα αυτό εισάγεται η ακόλουθη μέθοδος. Ο ανθρώπινος σκελετός χωρίζεται σε μικρότερα απλούστερα κομμάτια (poselets) για καθένα από τα οποία εκπαιδεύεται ένας ανιχνευτής (AdaBoost) ώστε να τα εντοπίζει. Στη συνέχεια με τη βοήθεια αυτών βελτιώνονται οι όροι του Pictorial Structure Models που χρησιμοποιείται για την εκτίμηση της πόζας στη συγκεκριμένη εικόνα. Το αποτέλεσμα είναι τα PSMs να αρχικοποιούνται σε κάθε περίπτωση σε μία πόζα που είναι σχετικά κοντά στην πραγματική και άρα να βελτιώνεται η ακρίβεια της τελικής εκτίμησης.

Αργότερα, μετά την επιτυχία των συνελικτικών νευρωνικών δικτύων (CNNs) σε αρκετά προβλήματα όρασης υπολογιστών, τέτοιου είδους δίκτυα αξιοποιήθηκαν και για την εκτίμηση της ανθρώπινης πόζας, ιδιαίτερα αφού μπορούσαν να εκμεταλλευτούν τα πολύ μεγάλα επισημειωμένα σύνολα δεδομένων που ήταν πλέον διαθέσιμα. Και αυτές οι προσεγγίσεις εξακολουθούν να αντιμετωπίζουν αρκετά από τα προβλήματα που παρουσιάζονται στην προσπάθεια εκτίμησης της ανθρώπινης πόζας, όπως το πρόβλημα των μελών του σώματος που δε φαίνονται στις εικόνες (αλληλοκαλύπτονται με άλλα μέλη του σώματος ή με άλλα αντικείμενα). Παρ΄ όλα αυτά τα συνελικτικά νευρωνικά δίκτυα οδήγησαν σε πολλές νέες προσεγγίσεις για την εκτίμηση της ανθρώπινης πόζας και έδωσαν ώθηση για την ολοκληρωτική επίλυση του προβλήματος αυτού.

Για παράδειγμα, στο [49] (Bulat et al.) για την εκτίμηση της πόζας χρησιμοποιείται μία αλυσίδα CNN, η οποία αποτελείται από δύο βαθιά υποδίκτυα. Το πρώτο βασίζεται στην αρχιτεκτονική VGG-16 και εκπαιδεύεται ώστε να παράγει για κάθε μέρος του σώματος έναν χάρτη πιθανότητας θέσης με τη βοήθεια μίας σιγμοειδούς συνάρτησης σφάλματος. Στη συνέχεια, το δεύτερο δίκτυο λαμβάνει ως είσοδο τους χάρτες που έχουν παραχθεί από το πρώτο καθώς και την αρχική εικόνα και πραγματοποιεί παλινδρόμηση ώστε να καταλήξει στις θέσεις των μελών του σώματος.

Σχήμα 4.3: Παράδειγμα στιγμιότυπου BabyAffect στο οποίο έχουμε εξάγει την πόζα με τη χρήση του OpenPose. Εδώ ανιχνεύονται οι πόζες για παραπάνω από έναν ανθρώπους (μητέρα και παιδί). Ακόμη, παρατηρούμε τη δυσκολία που περιγράψαμε στο Κεφάλαιο 3, καθώς εκτός από τους ανθρώπους ανιχνεύεται πόζα και για δύο κούκλες.

4.2 Εκτίμηση ανθρώπινης πόζας για πολλά άτομα ταυτόχρονα

Για την επίλυση του προβλήματος εκτίμησης πόζας για εικόνες με πολλά άτομα η συνηθέστερη προσέγγιση βασίζεται στις παραπάνω προσπάθειες. Με τη βοήθεια ενός ανιχνευτή ανθρώπων για κάθε άτομο που ανιχνεύεται πραγματοποιείται εκτίμηση πόζας για ένα μόνο άτομο. Η από τα πάνω προς τα κάτω προσέγγιση αυτή αξιοποιεί όλες τις προηγούμενες επιτυχημένες προσπάθειες εκτίμησης πόζας, έχει όμως ορισμένα σημαντικά μειονεκτήματα. Αν ο ανιχνευτής αποτύχει, γεγονός που είναι πολύ πιθανό να συμβεί όταν οι άνθρωποι στην εικόνα αλληλεπιδρούν σε κοντινή απόσταση, δεν υπάρχει τρόπος να διορθωθεί το λάθος του ανιχνευτή αλλά θα υπάρχει στην εκτίμηση μέχρι το τέλος της διαδικασίας. Επιπρόσθετα, ο χρόνος πραγματοποίησης της εκτίμησης αυξάνεται μαζί με το πλήθος των ατόμων στην εικόνα, καθώς για κάθε άτομο εκτελείται ξεχωριστά εκτίμηση πόζας.

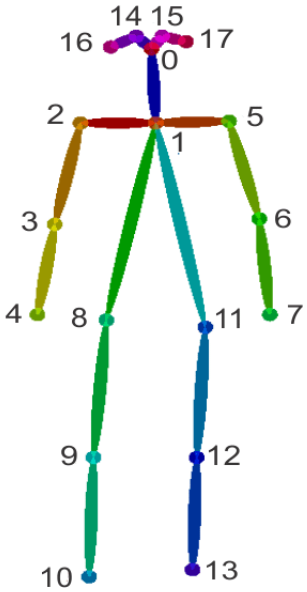
Στον αντίποδα βρίσκεται η από κάτω προς τα πάνω προσέγγιση που μπορεί να υπερκεράσει τα προαναφερθέντα μειονεκτήματα. Τέτοιες προσπάθειες ανιχνεύουν ταυτόχρονα τα υποψήφια μέρη του σώματος και στη συνέχεια τα συνδέουν με συγκεκριμένα άτομα. Ωστόσο η αντιστοίχιση αυτή ισοδυναμεί με τη λύση ενός προβλήματος ακέραιου γραμμικού προγραμματισμού σε ένα πλήρως συνδεδεμένο γράφο, το οποίο αποτελεί ένα NP-δύσκολο πρόβλημα, που χρειάζεται ώρες για να επιλυθεί. Ο χρόνος αυτός, όπως θα δείξουμε και παρακάτω μπορεί να μειωθεί με τη χρήση συντελεστών για τα διάφορα ζεύγη μερών των σωμάτων της εικόνας.

Το OpenPose αποτελεί μία από τις πιο πρόσφατες προσεγγίσεις του προβλήματος, που ξεπερνά σε μεγάλο βαθμό τις προαναφερθείσες δυσκολίες. Επιτυγχάνει σε πραγματικό χρόνο τον ακριβή προσδιορισμό των 2D σημείων του σκελετού του ανθρώπινου σώματος ακόμη για εικόνες με πολλούς ανθρώπους. Παρουσιάζει πολλά πλεονεκτήματα και χρησι-

μοποιείται πλέον ευρέως σε εφαρμογές όρασης υπολογιστών και μηχανικής μάθησης που απαιτούν την εκτίμηση ανθρώπινης πόζας. Όπως έχουμε αναφέρει και προηγουμένως και στην παρούσα διπλωματική πραγματοποιούμε την εκτίμηση της πόζας με τη χρήση του OpenPose [4, 50, 51, 52] (π.χ. Σχήμα 4.1 και Σχήμα 4.2 και Σχήμα 4.3).

4.3 OpenPose

Για την εκτίμηση της ανθρώπινης πόζας το OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields (Wei et al.) επιστρατεύει τη χρήση δισδιάστατων διανυσματικών πεδίων που ονομάζονται Part Affinity Fields (PAFs) και κωδικοποιούν την πληροφορία για το βαθμό συσχέτισης μεταξύ των σημείων, για τη θέση και τον προσανατολισμό των άκρων του σώματος στην εικόνα. Η χρήση τέτοιων διανυσματικών πεδίων επιτρέπει τη συσχέτιση κάθε μέρους σώματος στην εικόνα με το σωστό άτομο και μάλιστα σε ελάχιστο χρόνο σε σχέση με προηγούμενες προσπάθειες με τη χρήση άπληστου αλγορίθμου που μετατρέπει το πρόβλημα από NP Hard σε πρόβλημα διμερών ταιριασμάτων. Το OpenPose ανιχνεύει συνολικά 135 σημεία ανθρώπινου σκελετού, τα οποία αντιστοιχούν στο σώμα, τα πόδια, τα χέρια και το πρόσωπο των ανθρώπων. Στη συνέχεια περιγράφουμε τα σημαντικότερα βήματα της διαδικασίας που ακολουθείται [4].



Σχήμα 4.4: Απεικόνιση των δεκαοκτώ σημείων του σκελετού που εξάγονται από το OpenPose. [4]

Αρχικά η εικόνα αναλύεται από ένα convolutional δίκτυο το οποίο αρχικοποιείται με τις πρώτες 10 στιβάδες του VGG-19. Με τον τρόπο αυτό παράγεται ένα σύνολο χαρτών χαρακτηριστικών, που αποτελεί την είσοδο του συστήματος. Το σύστημα που χρησιμοποιείται αποτελείται από δύο διαδοχικά CNN δίκτυα. Το πρώτο δίκτυο προβλέπει ένα σύνολο από PAFs (4.3-4.4) το οποίο βελτιώνεται για αρκετά διαδοχικά στάδια με τη βοήθεια των αρχικών χαρτών χαρακτηριστικών και της προηγούμενης πρόβλεψης (4.5-4.6). Στη συνέχεια, η διαδικασία επαναλαμβάνεται για το δεύτερο δίκτυο το οποίο καταλήγει επίσης μετά από αρκετά διαδοχικά στάδια σε ένα σύνολο χαρτών εμπιστοσύνης (4.1-4.2), με τη βοήθεια των αρχικών χαρτών χαρακτηριστικών, του τελευταίου συνόλου PAFs καθώς και των προηγούμενων προβλέψεων για τους χάρτες εμπιστοσύνης (4.7-4.8).

$$L = (L_1, L_2, ..., L_C) \tag{4.1}$$

$$L_c \, \epsilon \, R^{m \times h \times 2}, \; c \, \epsilon \, \{1...C\} \tag{4.2}$$

$$S = (S_1, S_2, ..., S_J) \tag{4.3}$$

$$S_j \, \epsilon \, R^{m \times h}, \; j \, \epsilon \, \{1...J\} \tag{4.4}$$

$$L^1 = \phi^1(F) \tag{4.5}$$

$$L^t = \phi^t(F, L^{t-1}), \; \forall \; 2 \leq t \leq T_P \tag{4.6}$$

$$S^{T_P} = \rho^{T_P}(F, L^{T_P}) \tag{4.7}$$

$$S^t = \rho^t(F, L^{T_P}, S^{t-1}), \; \forall \; T_P \leq t \leq T_P + T_C \tag{4.8}$$

$m \times h$: size of image

F : feature maps

S : 2D confidence maps

L : 2D part affinity fields

J : number of parts

C : number of limbs

T_P : number of PAFs CNN stages

T_C : number of confidence maps CNN stages

$\phi, \; \rho$: CNNs

Για να οδηγηθεί το δίκτυο στην πρόβλεψη σωστών χαρτών εμπιστοσύνης και διανυσματικών πεδίων χρησιμοποιείται L2 συνάρτηση σφάλματος μεταξύ των προβλέψεων και των πραγματικών χαρτών και πεδίων αντίστοιχα (4.9-4.12).

$$f_L^{t_i} = \sum_{c=1}^{C} \sum_{p} W(p) \| L_c^{t_i}(p) - L_c^*(p) \|_2^2 \tag{4.9}$$

$$f_S^{t_k} = \sum_{j=1}^{J} \sum_{p} W(p) \| L_c^{t_k}(p) - L_c^*(p) \|_2^2 \tag{4.10}$$

$$f = \sum_{t=1}^{T_P} f_L^t + \sum_{t=T_P+1}^{T_P+T_C} f_S^t \tag{4.11}$$

$W(p) = 0$ when annotation is missing (4.12)

S^* : ground truth confidence maps

L^* : ground truth part affinity fields

f : error

p : image points

Ουσιαστικά οι χάρτες εμπιστοσύνης περιλαμβάνουν τις θέσεις των διάφορων σημείων του σώματος και παράγεται ένας τέτοιος χάρτης για κάθε σημείο του σκελετού. Αντίστοιχα, τα PAFs κωδικοποιούν την πληροφορία για το βαθμό συσχέτισης μεταξύ των σημείων αυτών και παράγεται ένα διανυσματικό πεδίο για κάθε τύπο άκρου (ζεύγος σημείων σκελετού).

Κάθε χάρτης εμπιστοσύνης είναι μία δισδιάστατη αναπαράσταση της πεποίθησης ότι ένα συγκεκριμένο σημείο του σώματος μπορεί να βρίσκεται σε ένα δοσμένο pixel. Αρχικά για κάθε σημείο του σώματος παράγονται τόσοι χάρτες εμπιστοσύνης όσοι οι άνθρωποι στην εικόνα. Η τιμή σε κάθε σημείο του χάρτη είναι η τιμή μίας γκαουσιανής κατανομής με κέντρο την πραγματική θέση του σημείου του σώματος (4.13). Στη συνέχεια, λαμβάνοντας το μέγιστο των χαρτών εμπιστοσύνης του συγκεκριμένου σημείου του σώματος όλων των ανθρώπων προκύπτει ο χάρτης εμπιστοσύνης του σημείου αυτού (4.14).

$$S^*_{j,k}(p) = \exp(-\frac{\|p - x_{j,k}\|_2^2}{\sigma^2}) \tag{4.13}$$

$$S^*_j(p) = \max_k S^*_{j,k}(p) \tag{4.14}$$

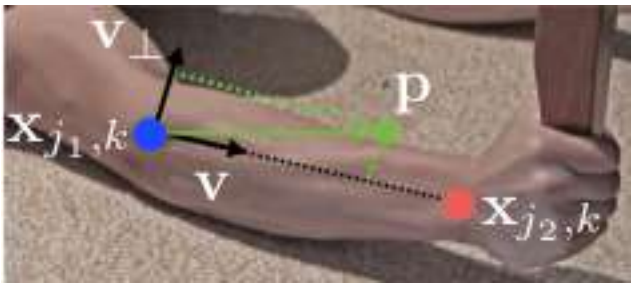
$x_{j,k}$: ground truth

k : person

Σχήμα 4.5: Στην εικόνα a φαίνεται ένα σύνολο από σημεία ανθρώπινων σκελετών που έχουν ανιχνευθεί. Στην εικόνα b φαίνονται οι δυνατές συνδέσεις μεταξύ των σημείων αυτών χρησιμοποιώντας τη μέθοδο του ενδιάμεσου σημείου. Με τη μέθοδο αυτή εκτός από τις ορθές (μαύρες) συνδέσεις μπορούν να προκύψουν και οι λανθασμένες (πράσινες) συνδέσεις. Τέλος, στην εικόνα c φαίνονται οι συνδέσεις μεταξύ των ίδιων σημείων χρησιμοποιώντας Part Affinity Fields [4].

Στη συνέχεια, υπάρχει ένα σύνολο ανιχνευμένων σημείων του σώματος, τα οποία πρέπει να ενωθούν σχηματίζοντας τον ανθρώπινο σκελετό. Πρέπει, επομένως, να βρεθεί ένα μέτρο της συσχέτισης μεταξύ των διάφορων σημείων. Ένα τέτοιο μέτρο συσχέτισης θα μπορούσε να είναι η ύπαρξη ή μη ανάμεσα τους ενός ενδιάμεσου σημείου στο εν λόγω άκρο (ζεύγος σημείων), το οποίο ενδιάμεσο σημείο επίσης θα ανιχνεύαμε. Ωστόσο, με τον τρόπο αυτό κωδικοποιείται μονάχα η θέση και όχι ο προσανατολισμός κάθε άκρου και επιπλέον η περιοχή στην οποία εκτείνεται ένα άκρο ελαχιστοποιείται σε ένα και μόνο

σημείο. Εξαιτίας αυτών η μέθοδος αυτή οδηγεί σε σφάλματα, όταν οι άνθρωποι βρίσκονται ο ένας κοντά στον άλλο (Σχήμα 4.5). Για το λόγο αυτό εισάγονται τα PAFs. Σε κάθε τύπο άκρου και για κάθε άνθρωπο αντιστοιχεί ένα διανυσματικό πεδίο που ενώνει τα δύο συσχετιζόμενα σημεία του σώματος.



Σχήμα 4.6: Απεικόνιση των PAFs που χρησιμοποιεί το OpenPose. Με x συμβολίζονται τα δύο άκρα του χεριού, p είναι ένα σημείο που ανήκει στο χέρι και v είναι η τιμή-διάνυσμα που λαμβάνουν όλα τα σημεία που ανήκουν στο χέρι [4].

Η τιμή του διανυσματικού αυτού πεδίου σε κάθε σημείο του είναι μηδέν αν το σημείο αυτό δεν ανήκει στο εν λόγω άκρο ή ίση με ένα μοναδιαίο διάνυσμα (v) με την κατεύθυνση του άκρου (από το ένα σημείο του στο άλλο) (4.15,4.17). Το αν ένα σημείο ανήκει ή όχι σε ένα τύπο άκρου υπολογίζεται με βάση ορισμένα για κάθε τύπο άκρου κατώφλια (4.18-4.19). Στο Σχήμα 4.6 παρουσιάζεται μία αναπαράσταση των PAFs. Για να προκύψει το διανυσματικό πεδίο κάθε άκρου λαμβάνεται ο μέσος όρος των επιμέρους διανυσματικών πεδίων όλων των ανθρώπων (4.16).

$$L^*_{c,k}(p) = v,\ if\ p\ on\ limb\ c,\ k \tag{4.15}$$

$$L^*_c(p) = \frac{1}{n_c(p)} \sum_k L^*_{c,k}(p) \tag{4.16}$$

$$v = \frac{x_{j2,k} - x_{j1,k}}{\|x_{j2,k} - x_{j1,k}\|_2} \tag{4.17}$$

$$0 \leq v(p - x_{j,k}) \leq l_{c,k} \tag{4.18}$$

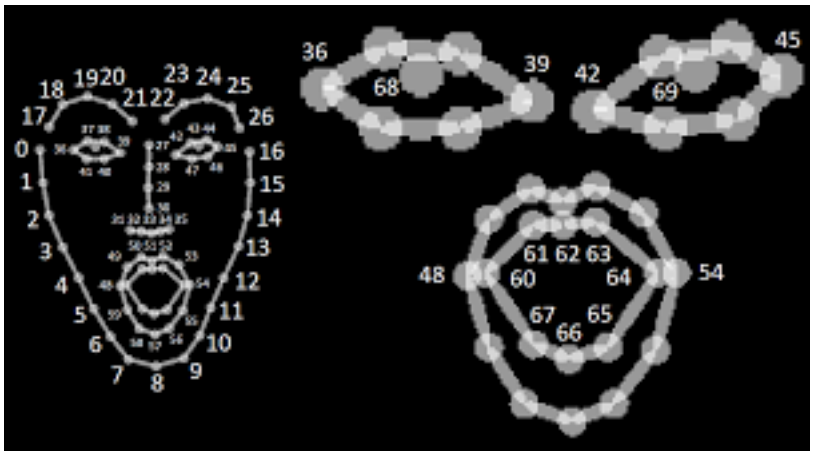
$$v_\perp(p - x_{j,k}) \leq \sigma_l \tag{4.19}$$

$l_{c,k}$: length of limb

σ_l : width of limb

$n_c(p)$: number of non zero vectors at p

Χρησιμοποιώντας ως βάρη αυτά τα διανυσματικά πεδία υπολογίζεται ένα γραμμικό ολοκλήρωμα (E) για κάθε υποψήφιο άκρο που προκύπτει από τα σημεία που έχουν ανιχνευθεί με τη βοήθεια των χαρτών εμπιστοσύνης (4.20-4.21). Έτσι, το πρόβλημα της εύρεσης της βέλτιστης λύσης αντιστοιχεί σε ένα πρόβλημα ταιριάσματος τόσων διαστάσεων όσα τα διαφορετικά άτομα στην εικόνα, το οποίο είναι NP-Hard. Πρέπει να βρεθούν τα τα ζεύγη ανιχνευμένων σημείων που αποτελούν πραγματικά ανθρώπινα άκρα. Έτσι, φτιάχνεται ένα σύνολο στο οποίο αντιπροσωπεύονται με μία δυαδική μεταβλητή όλες οι πιθανές συνδέσεις μεταξύ ανιχνευμένων σημείων. Για να απλοποιηθεί το πρόβλημα του ταιριάσματος γίνονται ορισμένες παραδοχές. Έτσι, εξετάζονται μόνο συνδέσεις μεταξύ τύπων σημείων του



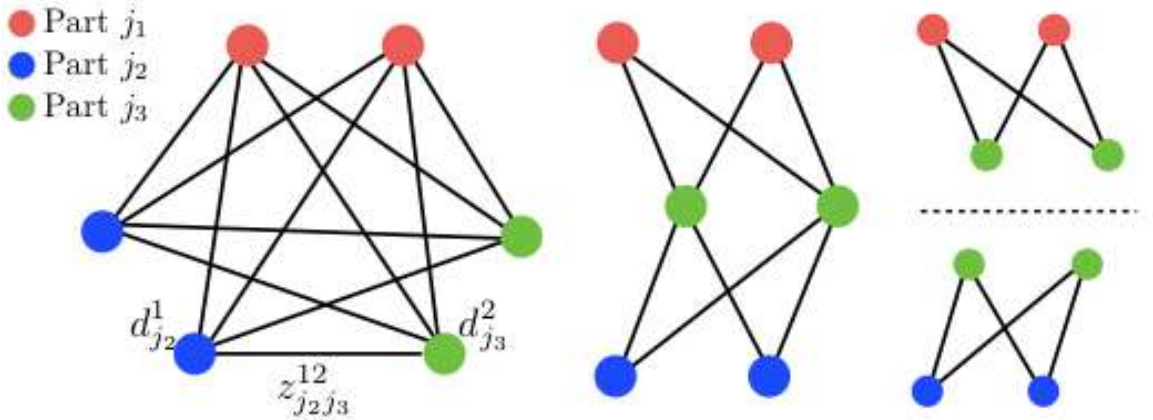
Σχήμα 4.7: Απεικόνιση των εβδομήντα σημείων του προσώπου που εξάγονται από το OpenPose [4].

σώματος που πράγματι συνδέονται μεταξύ τους (για παράδειγμα για τα σημεία των δεξιών αγκώνων εξετάζεται η σύνδεσή τους μόνο με τα σημεία των δεξιών χεριών και των δεξιών ώμων) και ακόμη κάθε σύνδεση εξετάζεται ανεξάρτητα από όλες τις άλλες. Οι παραδοχές αυτές και η απλοποίηση που φέρουν στο πρόβλημα παρουσιάζεται στο Σχήμα 4.8.

$$E = \int_{u=0}^{u=1} L_c(p(u)) \frac{d_j 2 - d_j 1}{\|d_j 2 - d_j 1\|_2} \tag{4.20}$$

$$p(u) = (1-u)d_{j1} + u d_{j2} \tag{4.21}$$

$$d_{j1}, d_{j2} : \text{two body parts}$$



Σχήμα 4.8: Παραδοχές κατά το ταίριασμα των σημείων του σώματος που έχουν ανιχνευθεί για το σχηματισμό του ανθρώπινου σκελετού από το OpenPose. Γίνονται αποδεκτά μόνο τα ζεύγη που αναπαριστούν πραγματικά άκρα και κάθε τύπος άκρου επιλύεται ξεχωριστά [4].

Έτσι, για δύο συγκεκριμένους τύπους σημείων, η εύρεση της βέλτιστης συσχέτισης αντιστοιχεί σε ένα πρόβλημα ταιριάσματος μέγιστου βάρους σε διμερή γράφο. Ως βάρη χρησιμοποιούνται τα γραμμικά ολοκληρώματα που έχουν υπολογιστεί. Στη συνέχεια

χρησιμοποιείται ένας άπληστος αλγόριθμος, δηλαδή για την επίλυση του ταιριάσματος επιλέγεται σε κάθε βήμα η τοπικά βέλτιστη λύση. Στόχος είναι η μεγιστοποίηση του αθροίσματος αυτών των ολοκληρωμάτων για να σχηματιστεί ο σκελετός των ανθρώπων που απεικονίζονται.

4.4 Αναγνώριση Δράσης με τη Βοήθεια της Πόζας

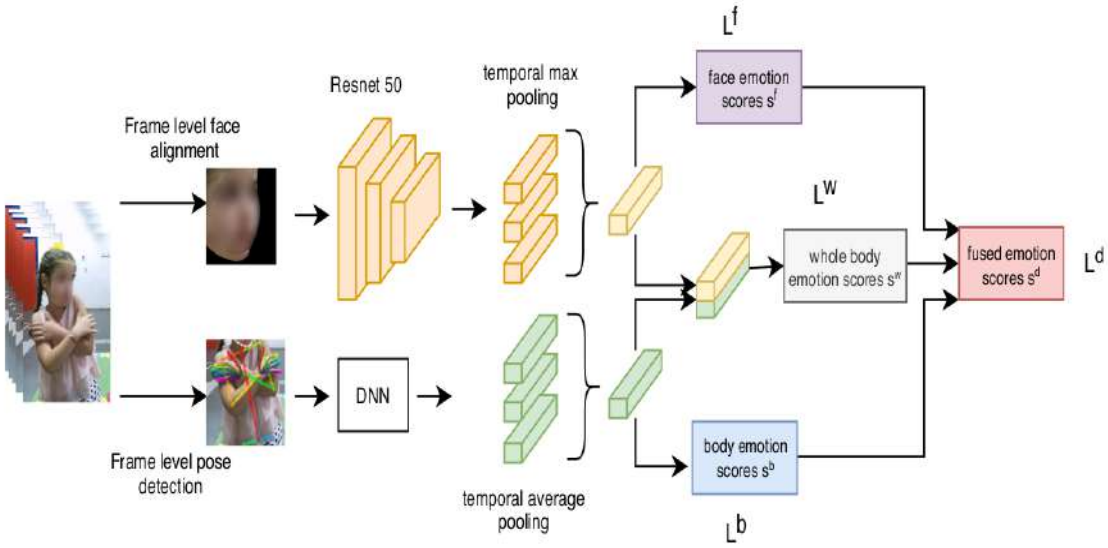
Όπως αναφέραμε και προηγουμένως, η εκτίμηση της ανθρώπινης πόζας μπορεί να χρησιμοποιηθεί από πληθώρα εφαρμογών αλληλεπίδρασης ανθρώπου υπολογιστή αλλά και αναγνώρισης δράσης. Και στην παρούσα διπλωματική αξιοποιούμε την πόζα που έχει εξαχθεί από καταγραφές αλληλεπίδρασης παιδιών με ρομπότ για να εκτιμήσουμε το engagement των παιδιών. Στη βιβλιογραφία συναντάμε πολλές ενδιαφέρουσες μεθόδους με τις οποίες γίνεται αναγνώριση ανθρώπινης δράσης ή και εκτίμηση συναισθήματος, συγκέντρωσης, engagement κλπ ανθρώπων μέσω και της πόζας. Οι περισσότερες από αυτές τις μεθόδους στηρίζονται είτε σε αναδρομικά νευρωνικά δίκτυα (Recurrent Neural Network (RNN)) είτε σε συνελικτικά νευρωνικά δίκτυα (CNNs) είτε σε συνελικτικά νευρωνικά δίκτυα σε γράφους (graph-based CNNs). Παρακάτω περιγράφουμε μερικά παραδείγματα τέτοιων μεθόδων με στόχο να βρούμε μεθόδους που θα μπορούσαν -πιθανώς παραλλαγμένες- να εφαρμοστούν με επιτυχία και στο δικό μας πρόβλημα, με βάση το στόχο μας, που είναι η εκτίμηση του engagement, αλλά και με βάση τα δεδομένα που διαθέτουμε.

4.4.1 Αναγνώριση δράσης/συναισθήματος αξιοποιώντας την πόζα με τη χρήση αναδρομικών - LSTM και συνελικτικών δικτύων - CNN

Στη δουλειά των Filntisis et al. [5] αποδεικνύεται ότι αξιοποιώντας τις πληροφορίες για την ανθρώπινη πόζα μπορούμε να καταλήξουμε σε σωστότερες εκτιμήσεις για τη συναισθηματική κατάσταση από ότι αν αξιοποιούμε μόνο πληροφορίες για τις εκφράσεις του προσώπου.

Έτσι, χρησιμοποιούνται δύο νευρωνικά δίκτυα, ένα το οποίο εκπαιδεύεται με εισόδους τις δισδιάστατες θέσεις των σημείων του σκελετού και των χεριών και ένα που εκπαιδεύεται με εισόδους τα σημεία του προσώπου. Πραγματοποιείται εκτίμηση συναισθήματος με καθένα ξεχωριστά, αλλά και με το συνδυασμό τους, που είναι και σημαντικά περισσότερο επιτυχής. Τα σημεία του σκελετού και των χεριών εξάγονται με τη χρήση του OpenPose. Για το σκέλος του σκελετού προτείνονται τρεις διαφορετικές αρχιτεκτονικές. Στην πρώτη γίνεται Global Temporal Average Pooling (GTAP), με τη χρήση μίας κρυφής πλήρους στιβάδας και ενεργοποίησης ReLU. Η δεύτερη αποτελείται από ένα χρονικό συνελικτικό δίκτυο (Temporal Convolutional Network (TCN)). Η τρίτη αποτελείται από ένα αμφίδρομο LSTM δύο στιβάδων, που ακολουθείται από μία πλήρη στιβάδα και ενεργοποίηση ReLU. Στις δύο τελευταίες αρχιτεκτονικές λαμβάνεται ο μέσος όρος των εξόδων για όλες τς χρονικές στιγμές ώστε να προκύψει αποτέλεσμα. Στη συγκεκριμένη μελέτη η πρώτη μέθοδος έχει καλύτερα αποτελέσματα, παρ΄ ότι είναι αρκετά απλούστερη. Το γεγονός αυτό αποδίδεται στο γεγονός ότι το πλήθος των δεδομένων ήταν αρκετά μικρό, με αποτέλεσμα η μέθοδος να επικεντρώνεται σε ορισμένες χαρακτηριστικές στάσεις του σώματος που συνηθίζονται κατά την έκφραση των συναισθημάτων, ενώ αγνοούσε την πληροφορία από τη χρονική συνέχεια.

Στη δουλειά των Marinou et al.([53]) ο στόχος είναι η αυτόματη εκτίμηση δράσης αλλά



Σχήμα 4.9: Δίκτυο για την εκτίμηση του engagement που αξιοποιεί τόσο τα σημεία του σκελετού όσο και τα RGB δεδομένα της περιοχής του προσώπου [5].

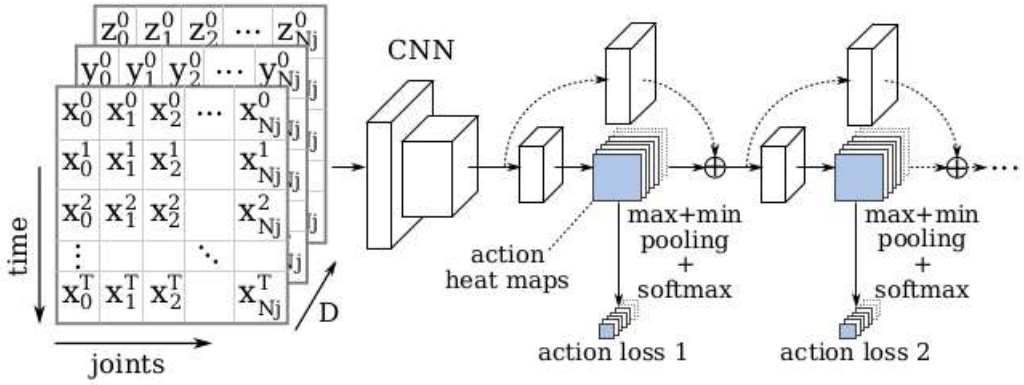
και συναισθήματος κατά την αλληλεπίδραση παιδιών με διαταραχές αυτιστικού φάσματος με ρομπότ. Και εδώ εξάγεται η πόζα των παιδιών με τη χρήση κατάλληλου αλγορίθμου. Είναι ενδιαφέρον το γεγονός ότι η συγκεκριμένη εργασία αξιοποιεί και την πόζα των βοηθών-ψυχολόγων που συμμετέχουν στις αλληλεπιδράσεις μαζί με τις πόζες των παιδιών για να εξαχθούν συμπεράσματα.

Για την εκτίμηση εκπαιδεύεται ένα σχετικά απλό συνελικτικό δίκτυο αποτελούμενο από τέσσερις συνελικτικές στιβάδες, με ενεργοποιήσεις ReLU και Max-Pooling ακολουθούμενες από μία στιβάδα Dropout καθώς και δύο πλήρεις στιβάδες. Το δίκτυο αυτό λαμβάνει ως είσοδο μία χρονική ακολουθία σκελετών. Ακόμη, εκπαιδεύεται και ένα αναδρομικό δίκτυο που αποτελείται από 5 bidirectional αναδρομικά υπό-δίκτυα, καθένα από τα οποία λαμβάνει ως είσοδο τα σημεία ενός από πέντε υπό-μέρη του ανθρώπινου σκελετού. Αυτά είναι ο κορμός, το αριστερό χει, το δεξί χέρι, το αριστερό πόδι και το δεξί πόδι. Έπειτα, οι αναπαραστάσεις που σχηματίζονται από τα υπό-δίκτυα συγχωνεύονται και δίνονται ως είσοδοι σε επόμενες στιβάδες.

4.4.2 Αναγνώριση δράσης/συναισθήματος αξιοποιώντας την πόζα με τη χρήση συνελικτικών δικτύων - CNN

Στο "2D/3D Pose Estimation and Action Recognition using Multitask Deep Learning" [6], γίνεται μία προσπάθεια ταυτόχρονης εκτίμησης της ανθρώπινης πόζας και αναγνώρισης της ανθρώπινης δράσης. Με στόχο να αξιοποιηθεί η υψηλού επιπέδου πληροφορία που εμπεριέχεται στις θέσεις των μελών του σώματος, μετατρέπονται τα σύνολα των μελών που αποτελούν τις πόζες σε αναπαραστάσεις που μοιάζουν με εικόνες. Ο κάθετος άξονας αντιπροσωπεύει το χρόνο, ενώ ο οριζόντιος αντιπροσωπεύει τα σημεία του σκελετού. Τέλος, οι συντεταγμένες κάθε σημείου είτε είναι 2D - (x,y) είτε είναι 3D - (x, y, z) αποτελούν τα κανάλια. Με αυτό τον τρόπο είναι δυνατό να χρησιμοποιηθούν κλασικές δισδιάστες συνελίξεις για να εξαχθούν μοτίβα απευθείας από την χρονική αλληλουχία των σημείων του σκελετού. Έτσι, χρησιμοποιείται ένα βαθύ συνελικτικό νευρωνικό δίκτυο για

να εξαχθούν χαρακτηριστικά από τις πόζες εισόδου και να παραχθούν heatmaps δράσης, όπως φαίνεται στο Σχήμα 4.10. Για να παραχθεί η πιθανότητα εξόδου για κάθε δράση, πραγματοποιείται max plus min pooling και softmax activation στους χάρτες δράσης.



Σχήμα 4.10: Δισδιάστατο συνελικτικό δίκτυο για εκτίμηση δράσης με είσοδο τα 2D ή 3D σημεία του σκελετού [6].

Ακόμη, στη δουλειά των Javed et al. ([7]), ο στόχος είναι η εκτίμηση του engagement. Εδώ με τη βοήθεια του OpenPose εξάγονται τα σημεία του σκελετού και του προσώπου. Στη συνέχεια, από αυτά εξάγονται τρία καινούρια χαρακτηριστικά υψηλότερου επιπέδου, που ορίζονται από την μέθοδο Laban Movement Analysis([54]). Πρόκειται για μία μέθοδο με στόχο την περιγραφή και την ερμηνεία όλων των τύπων ανθρώπινης κίνησης, που χρησιμοποιείται συχνά σε μία ποικιλία πεδίων όπως ο χορός, η υποκριτική, η μουσική και η φυσικοθεραπεία. Η μέθοδος αυτή κατηγοριοποιεί τις κινήσεις του σώματος με βάση κριτήρια ένα από τα οποία είναι το body effort, το οποίο αντιπροσωπεύει τη δυναμική της ανθρώπινης κίνησης και παρέχει χρήσιμη πληροφορία για τα δυσδιάκριτα χαρακτηριστικά των κινήσεων σε σχέση με τις εσωτερικές προθέσεις. Αυτό το καθιστά χρήσιμο για την εκτίμηση χαρακτηριστικών όπως το engagement. Το effort περιγράφεται από το χώρο (space), το βάρος (weight) και το χρόνο (time), τα οποία είναι τα τρία χαρακτηριστικά που παράγονται από τα σημεία του σκελετού σύμφωνα τις εξισώσεις του Σχήματος 4.11 και σε διαστήματα του ενός δευτερολέπτου και χρησιμοποιούνται εδώ για την εκτίμηση του engagement. Έτσι, τα χαρακτηριστικά που χρησιμοποιούνται τελικά είναι οι συντεταγμένες των 34 σημείων του προσώπου που αντιστοιχούν στα χείλη και στα μάτια καθώς και τα τρία χαρακτηριστικά του effort.

Στο Σχήμα 4.12 παρουσιάζεται το δίκτυο που χρησιμοποιήθηκε. Πρόκειται για ένα συνελικτικό νευρωνικό δίκτυο με πολλαπλά κανάλια και πολλαπλές στιβάδες καθώς αντιμετωπίζει ένα πρόβλημα ταξινόμησης πολλαπλών κλάσεων. Αποτελείται από δυο στιβάδες Conv1D, για να αναγνωρίζουν τα χρονικά μοτίβα στα δεδομένα (με 5 κανάλια με 64 και 128 φίλτρα αντίστοιχα και kernel size 3 με 20% dropout) και τρεις πλήρεις στιβάδες για την ταξινόμηση (με kernel size 256, 256 και 7(αριθμός των κλάσεων) αντίστοιχα). Η εκπαίδευση έγινε για κάθε παιδί ξεχωριστά. Οι Conv1D στιβάδες πρέπει να εξάγουν υψηλού επιπέδου χαρακτηριστικά από τα χρονικά δεδομένα εφόσον το σύνολο δεδομένων που χρησιμοποιείται έχει μεγάλη διάσταση εισόδου και σχετικά μικρό αριθμό σημείων. Καθώς τα δεδομένα δεν έχουν γραμμική δομή, οι πρώτες δύο πλήρεις στιβάδες χρησιμοποιούνται για να διασπείρουν την διάσταση των χαρακτηριστικών, ενώ η τελευταία παράγει την διάσταση εξόδου. Για να αποφευχθεί το overfitting χρησιμοποιούνται στιβάδες dropout.

Feature	Equation	
Space		$space = \left(0.5\lvert\vec{a}\rvert \left\lvert\vec{d}\right\rvert \sin(\theta_1)\right) + \left(0.5\lvert\vec{c}\rvert \left\lvert\vec{b}\right\rvert \sin(\theta_2)\right)$
	where	$\vec{a}$ = Position vector from left shoulder to left hand
		$\vec{b}$ = Position vector from right shoulder to left shoulder
		$\vec{c}$ = Position vector from right hand to right shoulder
		$\vec{d}$ = Position vector from left hand to right hand
		θ_1 = Angle between $\vec{a}$& $\vec{d}$
		θ_2 = Angle between $\vec{c}$& $\vec{b}$
Weight		$Weight = \sum_i \tau_i(t)$
	where	$\tau_i = L^2\omega_i{}^2\sin(\theta)$ *mass
		$\omega_i = \frac{d\theta}{dt}$
		L = Distance between joints
		i = Joint number
		$\dot{\omega}_i$ = Angular velocity for joint i
Time		$Time_i = \sum_i \dot{\omega}_i(t)$
	where	i = Joint number
		$\dot{\omega}_i$= Angular velocity for joint i

Σχήμα 4.11: Εξισώσεις για την εξαγωγή των χαρακτηριστικών Laban [7].

4.4.3 Αναγνώριση δράσης/συναισθήματος αξιοποιώντας την πόζα με τη χρήση GCN

Όπως αναφέρουμε παραπάνω για την εκτίμηση ανθρώπινης δράσης από στοιχεία της πόζας πρόσφατα έχουν αξιοποιηθεί graph based CNNs, πετυχαίνοντας αξιοσημείωτα αποτελέσματα. Τα δίκτυα αυτά μοντελοποιούν τους σκελετούς του ανθρώπινου σώματος χρησιμοποιώντας δομικά στοιχεία γράφων, όπως οι κόμβοι και οι ακμές.

Τα Graph Convolutional Network (GCN) αποτελούν ένα τύπο νευρωνικών δικτύων που εφαρμόζονται απευθείας σε γράφους, ώστε να εξάγουν συμπεράσματα από πληροφορίες που δίνει η δομή τους. Ως τέτοιοι γράφοι μπορούν να αντιμετωπιστούν τα σύνολα των σημείων των ανθρώπινων σκελετών, δηλαδή οι πόζες. Με δεδομένο ένα γράφο $G = (V, E)$ ένα GCN δίκτυο λαμβάνει ως είσοδο έναν πίνακα χαρακτηριστικών X: $N \times F^0$, όπου N το πλήθος των κόμβων και F^0 είναι το πλήθος των χαρακτηριστικών εισόδου για κάθε κόμβο, καθώς και έναν πίνακα A: $N \times N$ που περιγράφει τη δομή του γράφου. Έτσι, η κρυφή στιβάδα του GCN μπορεί να γραφτεί ως $H^1 = f(H^{(i-1)}, A))$, όπου $H^0 = X$ και f είναι μία συνάρτηση προώθησης. Κάθε στιβάδα H^i αντιστοιχεί σε ένα πίνακα χαρακτηρι-

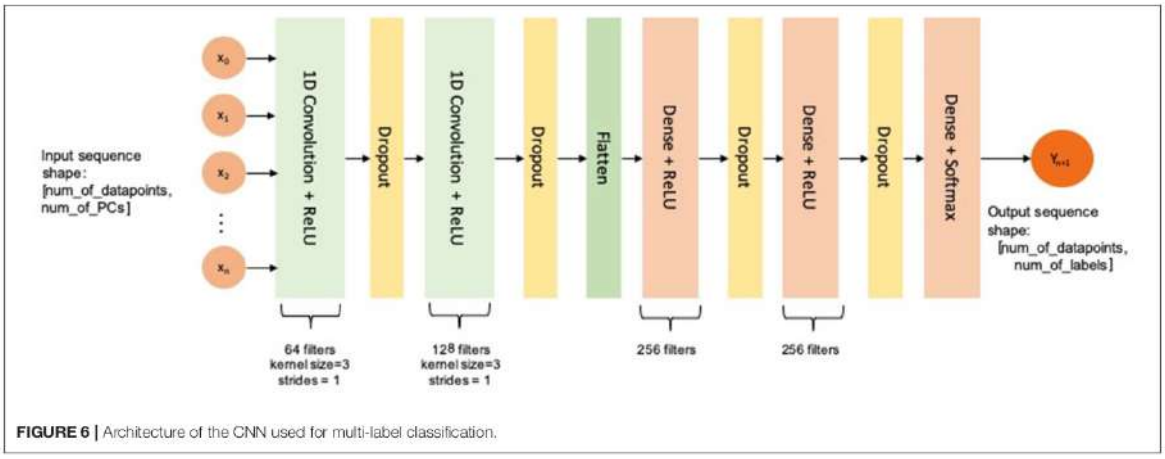


FIGURE 6 | Architecture of the CNN used for multi-label classification.

Σχήμα 4.12: Μονοδιάστατο συνελικτικό δίκτυο που χρησιμοποιείται για την εκτίμηση του engagement από σημεία σκελετού [7].

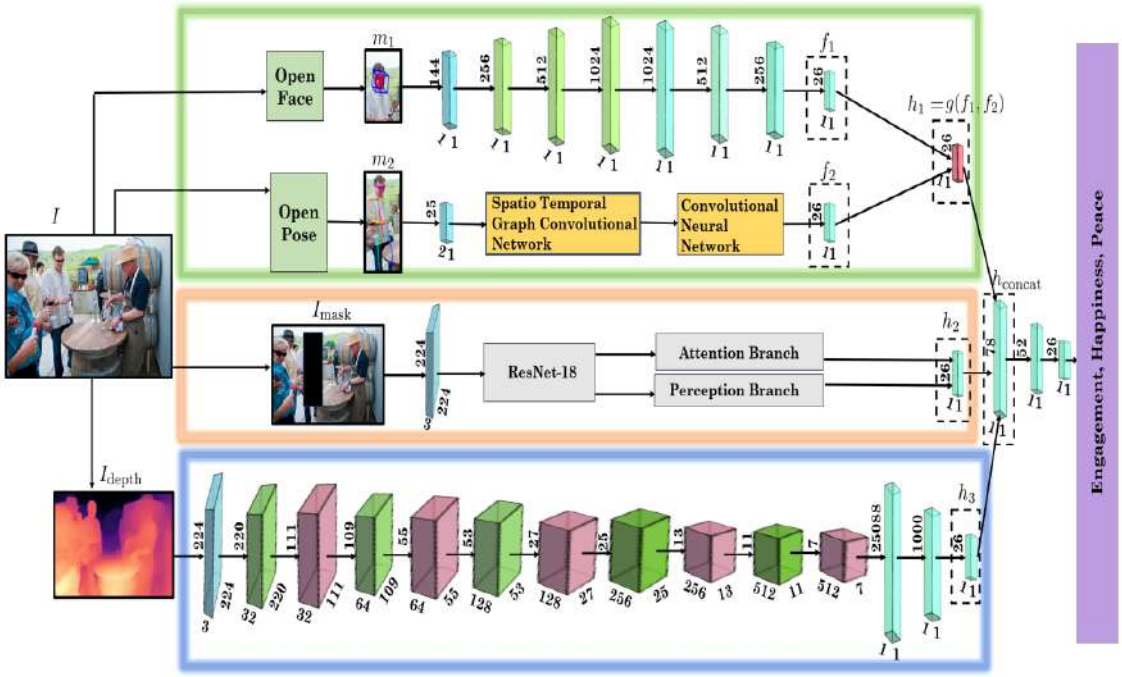
στικών $N \times F^i$, στον οποίο κάθε σειρά αποτελεί περιγραφή χαρακτηριστικών ενός κόμβου. Σε κάθε στιβάδα αυτά τα χαρακτηριστικά συνυπολογίζονται ώστε να σχηματίσουν τα χαρακτηριστικά της επόμενης στιβάδας, χρησιμοποιώντας τον κανόνα προώθησης f. Με αυτό τον τρόπο τα χαρακτηριστικά γίνονται ολοένα και πιο αφηρημένα σε κάθε διαδοχική στιβάδα. Βασικά, οι διάφοροι τύποι GCN διαφέρουν στον κανόνα προώθησης f που εφαρμόζουν.

Παραδείγματα εκτίμησης ανθρώπινης δράσης χρησιμοποιώντας ως είσοδο την πόζα και θεωρώντας τη γράφο βρίσκουμε στα [8], [9], [55], και [56].

Ειδικότερα, στο EmotiCon [8] παρουσιάζεται, ένας αλγόριθμος εκμάθησης για αναγνώριση ανθρώπινου συναισθήματος από εικόνες και videos. Η προσέγγιση αυτή συνδυάζει τρεις ερμηνείες του περιβάλλοντος πλαισίου για αναγνώριση συναισθήματος (βασιζόμενη στην Frege's Context Principle από την ψυχολογία). Η πρώτη ερμηνεία βασίζεται στη χρήση πολλαπλών λειτουργικοτήτων (για παράδειγμα πρόσωπα και βάδισμα) για την αναγνώριση του συναισθήματος. Για τη δεύτερη ερμηνεία, συγκεντρώνονται ουσιώδεις πληροφορίες από την εικόνα εισόδου και χρησιμοποιείται ένα δίκτυο CNN για να κωδικοποιηθεί το περιεχόμενο αυτό. Τέλος, χρησιμοποιούνται χάρτες βάθους για να μοντελοποιηθεί η τρίτη ερμηνεία που σχετίζεται με τις κοινωνικές-δυναμικές αλληλεπιδράσεις και την εγγύτητα μεταξύ των συμμετεχόντων.

Η κατάληξη σε τελικά συμπεράσματα για τα συναισθήματα των συμμετεχόντων προκύπτει από το συνδυασμό των συμπερασμάτων των τριών διαύλων. Ωστόσο, στην παρούσα εργασία μας απασχολεί κυρίως το πρώτο σκέλος της προτεινόμενης μεθόδου, δηλαδή η αξιοποίηση των σημείων του προσώπου και του σώματος. Με τη χρήση του OpenFace εξάγεται ένα διάνυσμα χαρακτηριστικών με τα σημεία του προσώπου. Επίσης, με τη χρήση του OpenPose εξάγεται ένα διάνυσμα χαρακτηριστικών με τα σημεία του σκελετού. Για κάθε σημείο, υπάρχουν οι τιμές των συντεταγμένων x και y. Τα διανύσματα αυτά χρησιμοποιούνται ως είσοδοι σε αντίστοιχα νευρωνικά δίκτυα. Τα αποτελέσματα των δικτύων αυτών συμπτύσσονται με multiplicative fusion, ώστε να γίνει ο αλγόριθμος ευσταθής σε παρεμβολές θορύβου.

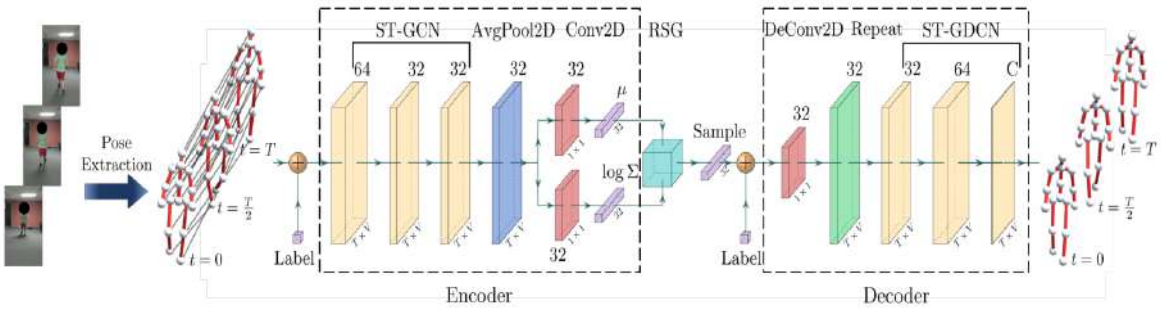
Για το διάνυσμα του προσώπου, χρησιμοποιούνται τρεις μονοδιάστατες συνελίξεις (που απεικονίζονται στο Σχήμα 4.13 με ανοιχτό πράσινο χρώμα) με batch normalization και ReLU activation. Στη συνέχεια, ακολουθούν στιβάδα max pooling καθώς και τρεις πλήρεις στιβάδες (με γαλάζιο χρώμα στο Σχήμα 4.13), επίσης με με batch normalization



Σχήμα 4.13: Δίκτυο αναγνώρισης συναισθήματος τριών τμημάτων με τη χρήση σκελετού, χαρτών βάθους και RGB δεδομένων στο [8].

και ReLU activation. Για το διάνυσμα του σκελετού του σώματος χρησιμοποιείται αρχιτεκτονική Spatial Temporal Graph Convolutional Network (ST-GCN), που προτείνεται στο STEP([9]) και βασίζεται στην εκμάθηση γράφων.

Έτσι, στο STEP προτείνεται μία αρχιτεκτονική ST-GCN για την ταξινόμηση ανθρώπινων συναισθημάτων με τη χρήση βηματισμού-πόζας. Το δίκτυο STEP φαίνεται στο Σχήμα 4.14.



Σχήμα 4.14: Το δίκτυο STEP που χρησιμοποιείται για την αναγνώριση ανθρώπινων συναισθημάτων από την πόζα - το βηματισμό [9].

Με είσοδο ένα RGB video ενός ανθρώπου που περπατά, το δίκτυο χρησιμοποιεί τα χαρακτηριστικά του βηματισμού για να ταξινόμηση τη συναισθηματική κατάσταση του ανθρώπου σε ένα από τέσσερα συναισθήματα: χαρά, λύπη, θυμός ή ουδέτερο συναίσθημα. Εκτός από εκατοντάδες επισημειωμένα με το χέρι πραγματικά video με ανθρώπους που

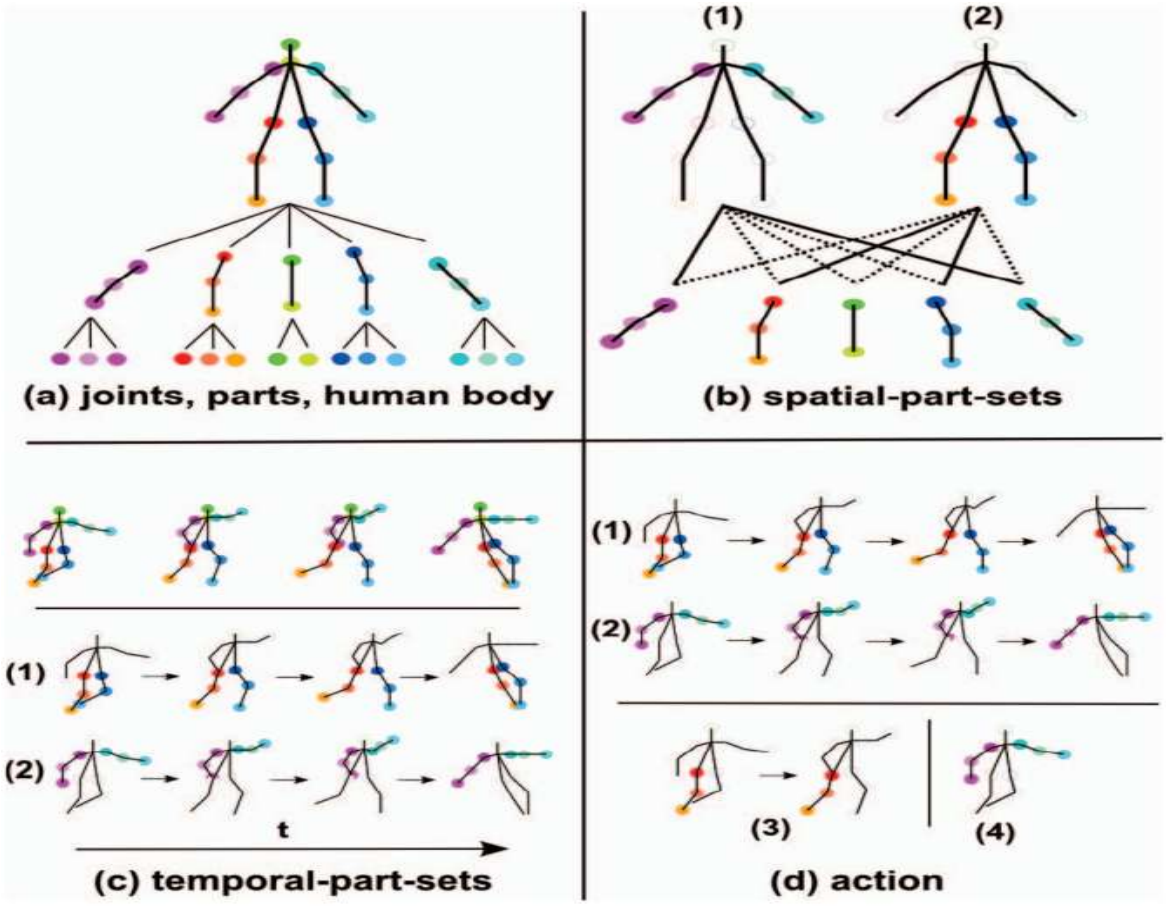
περπατούν, παράγονται χιλιάδες επισημειωμένοι συνθετικοί βηματισμοί με τη χρήση ενός επίσης ST-GCN δικτύου, χτισμένου πάνω σε έναν Conditional Variational Autoencoder (CVAE), ώστε να εμπλουτιστούν τα δεδομένα εκπαίδευσης του STEP.

Στο κυρίως δίκτυο κάθε σκελετός περνά από ένα σύνολο τριών στιβάδων ST-GCN, που έχουν 32, 64 και 64 πυρήνες αντίστοιχα. Η έξοδος της τελευταίας στιβάδες περνά από μία στιβάδα average pooling, τόσο στη χρονική όσο και στις χωρικές διαστάσεις και έπειτα από μία συνελικτική στιβάδα 1×1. Τέλος, το αποτέλεσμα περνά από μία πλήρη στιβάδα ακολουθούμενη από στιβάδα Softmax από την οποία προκύπτει η τελική κλάση της συναισθηματικής κατάστασης. Επίσης όλες οι στιβάδες, εκτός βέβαια από την τελευταία ακολουθούνται από batch normalization και ReLU activation.

4.4.4 Αναγνώριση δράσης/συναισθήματος αξιοποιώντας την πόζα με τη χρήση άλλων μεθόδων

Στο [10] ο στόχος είναι η αναγνώριση δράσης με δεδομένη την πόζα. Με τη βοήθεια ενός νευρωνικού δικτύου αποκτάται μία αναπαράσταση των χωρικών και διαδοχικών δομών σκελετού που συναντώνται σε πολλές ανθρώπινες δραστηριότητες. Η αναπαράσταση αυτή συλλαμβάνει τις διατάξεις των μελών του σώματος σε ένα καρέ (spatial-part-sets) καθώς και τις κινήσεις των μελών του σώματος σε μία αλληλουχία καρέ (temporal-part-sets), που χαρακτηρίζουν μία ανθρώπινη δράση. Για τις χαρακτηριστικές αυτές επαναλαμβανόμενες δομές του σώματος παράγονται σύνολα διατάξεων επιμέρους μελών του σώματος που συμβαίνουν ταυτόχρονα. Ομοίως παράγονται σύνολα διαδοχικών διατάξεων επιμέρους μελών του σώματος που συμβαίνουν ταυτόχρονα για τις χαρακτηριστικές επαναλαμβανόμενες κινήσεις του σώματος. Έτσι είναι δυνατή η αναπαράσταση των ανθρώπινων δράσεων ως ιστογράμματα των παραπάνω συνόλων χαρακτηριστικών διατάξεων. Τέλος, χρησιμοποιείται ένα μοντέλο bag of words, το οποίο εκμεταλλεύεται τα ιστογράμματα για να πραγματοποιήσει αναγνώριση της ανθρώπινης δράσης. Παραδείγματα των διατάξεων που παράγονται και αξιοποιούνται για την αναγνώριση της δράσης φαίνονται στο Σχήμα 4.15.

Στη δουλειά των Gaschler et al. ([57]) γίνεται προσπάθεια αναγνώρισης τυπικών ανθρώπινων συμπεριφορών σε ένα μπαρ με τη βοήθεια της πόζας. Η προσέγγιση αυτή χρησιμοποιεί συγκεκριμένα τα σημεία του ανθρώπινου σκελετού, τον προσανατολισμό του κεφαλιού καθώς και ορισμένα χαρακτηριστικά που εξάγονται με βάση τις σχέσεις μεταξύ των προσανατολισμών των σωμάτων των ανθρώπων που αλληλεπιδρούν. Για να μοντελοποιηθεί η χρονική αλληλουχία των καταστάσεων συμπεριφοράς των ανθρώπων χρησιμοποιούνται σε αυτή την προσέγγιση Hidden Markov Model (HMM). Τα HMMs είναι μία ισχυρή προσέγγιση μοντελοποίησης διαδοχικών και στατιστικών διαδικασιών που επιτρέπουν τις έμμεσες παρατηρήσεις. Στη συγκεκριμένη περίπτωση οι κοινωνικές συμπεριφορές μοντελοποιούνται ως οι κρυφές καταστάσεις του HMM μοντέλου. Αντίστοιχα η πόζα και τα υπόλοιπα χαρακτηριστικά αντιστοιχούν στις παρατηρήσεις των κρυφών καταστάσεων του μοντέλου. Έχει ενδιαφέρον ότι αξιολογήθηκαν διάφορα υποσύνολα του διανύσματος χαρακτηριστικών, με αποτέλεσμα σημαντικές διαφορές στην επιτυχία αναγνώρισης. Για το συγκεκριμένο πρόβλημα παρατηρήθηκε ότι το βέλτιστο σύνολο χαρακτηριστικών περιλάμβανε τον κορμό, τα χέρια (χωρίς τους ώμους και τους αγκώνες) και το κεφάλι.



Σχήμα 4.15: Προτεινόμενη αναπαράσταση των ανθρωπίνων δράσεων μέσω συνδυασμών των σημείων των σκελετών [10].

Κεφάλαιο 5

Μέθοδοι εκτίμησης

Στο κεφάλαιο αυτό παρουσιάζουμε τις μεθόδους που υλοποιήσαμε για την εκτίμηση του engagement στα σύνολα δεδομένων μας καθώς και τα αποτελέσματα που πετυχαίνει η κάθε μέθοδος. Αρχικά, περιγράφουμε τα διάφορα δίκτυα που σχεδιάσαμε για την εκτίμηση του engagement δοκιμάζοντας την απόδοση τους στο σύνολο TD-Joint Attention. Πρόκειται για αναδρομικά δίκτυα βασισμένα σε LSTM στιβάδες, μονοδιάστατα συνελικτικά δίκτυα και δισδιάστατα συνελικτικά δίκτυα. Παρουσιάζουμε αναλυτικά τη δομή και τα χαρακτηριστικά των αρχιτεκτονικών που χρησιμοποιούμε και στεκόμαστε ιδιαίτερα στο μέγεθος του χρονικού παραθύρου που επεξεργάζονται τα δίκτυα κάθε φορά.

Στη συνέχεια, έχοντας καταλήξει σε ορισμένα από αυτά με τα οποία πετυχαίνουμε καλά αποτελέσματα εκτίμησης τα εφαρμόζουμε και στα υπόλοιπα σύνολα δεδομένων και σε συνδυασμό δεδομένων, παρουσιάζοντας και εξηγώντας τα αποτελέσματα που λαμβάνουμε σε κάθε περίπτωση. Τέλος, εξετάζουμε την περίπτωση εξατομικευμένης εκτίμησης του engagement και συγκρίνουμε τα αποτελέσματα της εκτίμησής με τη χρήση απευθείας των RGB δεδομένων.

5.1 Υλοποίηση, Εκπαίδευση και Αξιολόγηση των Μεθόδων

Εισαγωγικά, πριν παρουσιάσουμε τις μεθόδους και τα αποτελέσματά τους, παραθέτουμε ορισμένα γενικά στοιχεία για τον τρόπο που υλοποιούνται, εκπαιδεύονται και αξιολογούνται όλα τα μοντέλα. Όλες οι προτεινόμενες μέθοδοι και τα αντίστοιχα μοντέλα έχουν υλοποιηθεί χρησιμοποιώντας τις βιβλιοθήκες του PyTorch [58].

Ως συνάρτηση σφάλματος (loss function) για την εκπαίδευση των δικτύων έχουμε χρησιμοποιήσει σε όλες τις περιπτώσεις την Cross Entropy Loss Function (Εξίσωση 5.1), ώστε να λαμβάνει το δίκτυο υπόψιν του την ανισοκατανομή των κλάσεων και να δίνει μεγαλύτερη βαρύτητα στα δείγματα που προέρχονται από τις πιο σπάνιες κλάσεις. Στην Cross Entropy Loss Function περνάμε ως όρισμα το διάνυσμα βαρών που προκύπτει για τα εκάστοτε train δεδομένα.

$$D = \frac{1}{N} \sum_{i=1}^{N} y_i * \log(\widetilde{y_i}) \tag{5.1}$$

$$D : \text{error}$$
$$y : \text{ground truth}$$
$$\widetilde{y} : \text{prediction}$$
$$N : \text{number of points}$$

Για να αξιολογήσουμε τα αποτελέσματα όλων των δικτύων χρησιμοποιούμε τέσσερις μετρικές. Οι τιμές όλων των μετρικών είναι εκφρασμένες σε ποσοστά με βέλτιστες τιμές το 100%. Η πρώτη είναι το απλό accuracy:

$$\text{accuracy} = \frac{\text{correct samples}}{\text{all samples}} \tag{5.2}$$

Επιλέξαμε και άλλες μετρικές ώστε να αξιολογούνται τα αποτελέσματα των μεθόδων μας με βάση τη μεγάλη ανισοκατανομή των κλάσεων, όπως περιγράφεται στην Ενότητα 2.7 (Μετρικές Αξιολόγησης). Συγκεκριμένα, χρησιμοποιούμε το weighted Fscore, δίνοντάς του ως όρισμα το διάνυσμα βαρών των κλάσεων για τα εκάστοτε train δεδομένα και το αναγράφουμε ως W. Fscore.

$$\text{weighted Fscore} = \sum_{n=1}^{N} \frac{\text{samples}_{k_n}}{\text{all samples}} \text{Fscore}(k_n) \tag{5.3}$$

Επίσης, χρησιμοποιούμε το balanced accuracy, το οποίο αναγράφουμε στους πίνακες παρουσίασης των αποτελεσμάτων ως Balanced Acc. Η μετρική αυτή είναι ένας απλός και όχι ένας σταθμισμένος μέσος όρος των recall όλων των κλάσεων. Παρ΄ όλα αυτά την αξιοποιούμε για την αξιολόγηση των αποτελεσμάτων μας καθώς είναι αρκετά σημαντικό το δίκτυό μας να εκτιμά σωστά και τις λιγότερο συνήθεις κλάσεις. Για παράδειγμα, τα στιγμιότυπα στα οποία τα παιδιά είναι πλήρως disengaged είναι λιγότερα από τα μισά στιγμιότυπα στα οποία τα παιδιά παρουσιάζουν ένα μέσο επίπεδο engagement σε όλα τα σύνολα δεδομένων. Παρ΄ όλα αυτά, είναι πολύ χρήσιμο να γνωρίζει το ρομπότ πότε το παιδί είναι πλήρως disengaged ώστε να αλλάξει τη συμπεριφορά του για να καταφέρει να το εμπλέξει στην εκάστοτε δραστηριότητα.

$$\text{balanced accuracy} = \frac{1}{k} \sum_{n=1}^{N} \text{recall}(k_n) \tag{5.4}$$

Στα τελικά και πιο σημαντικά αποτελέσματα για τα διάφορα σύνολα δεδομένων συμπεριλαμβάνουμε και το weighted precision για την πληρέστερη παρουσίαση και αξιολόγηση της απόδοσης των μεθόδων μας. Πρόκειται για το σταθμισμένο μέσο όρο των precision όλων των κλάσεων και το αναγράφουμε στους πίνακες των αποτελεσμάτων ως W. Precision.

$$\text{weighted precision} = \sum_{n=1}^{N} \frac{\text{samples}_{k_n}}{\text{all samples}} \text{precision}(k_n) \tag{5.5}$$

Τα μοντέλα εκπαιδεύονται για 100 περίπου εποχές. Διατηρείται κάθε φορά το μοντέλο εκείνο που πετυχαίνει το βέλτιστο W. Fscore.

Κατά την παρουσίαση των αποτελεσμάτων σε όλες τις μεθόδους για να αξιολογείται πληρέστερα η ποιότητα τους, παραθέτουμε αρχικά τις τιμές που θα είχαν οι μετρικές

αξιολόγησης αν όλα τα δείγματα κατατάσσονταν πάντα στην πολυπληθέστερη κλάση κάθε συνόλου δεδομένων. Τις τιμές αυτές αναγράφουμε κάθε φορά δίπλα από την ετικέτα common class.

5.2 Εκτίμηση με αναδρομικό δίκτυο

Όπως έχουμε αναφέρει και παραπάνω συχνά χρησιμοποιούνται αναδρομικά δίκτυα (RNN) για την αναγνώριση ανθρώπινης δράσης. Τα δίκτυα αυτά, και ειδικότερα οι μονάδες LSTM που χρησιμοποιούμε εδώ, επιτρέπουν την εκμάθηση χρονικών εξαρτήσεων σε μεγάλη εμβέλεια. Έχει παρατηρηθεί από πολλές εργασίες πάνω στην εκτίμηση του engagement ότι το επίπεδο του engagement των παιδιών διατηρείται μέσα σε ένα χρονικό διάστημα λίγων δευτερολέπτων και συνολικά παρουσιάζει χρονική συνέχεια. Επομένως, με τη χρήση LSTM στιβάδων μπορούμε να εξάγουμε ακριβέστερα συμπεράσματα για το engagement αξιοποιώντας και μαθαίνοντας από αυτή τη χρονική συνέχεια.

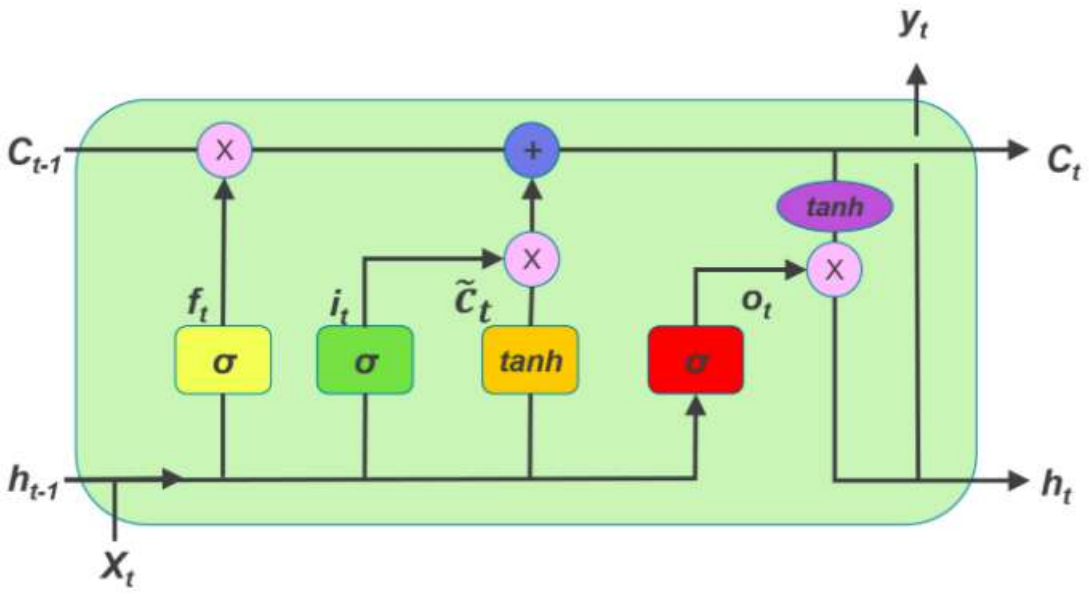
5.2.1 LSTM

Τα Long Short Term Memory (LSTM) αποτελούν υποκατηγορία των αναδρομικών δικτύων RNNs. Τα κλασικά RNN δεν έχουν τη δυνατότητα να μεταφέρουν σημαντική πληροφορία ανάμεσα σε βήματα που απέχουν σημαντικά μεταξύ τους. Αντιμετωπίζουν το πρόβλημα vanishing gradient. Ουσιαστικά οι παράγωγοι που χρησιμοποιούνται για την ανανέωση των βαρών του δικτύου τείνουν να μειώνονται ολοένα και περισσότερο. Έτσι, οι αρχικές στιβάδες, για τις οποίες οι παράγωγοι είναι εξαιρετικά μικρές σταματούν να μαθαίνουν με αποτέλεσμα το δίκτυο να ξεχνά σημαντικά συμπεράσματα που έχει δει στα δεδομένα.

Τα LSTM (όπως και τα Gated Recurrent Unit (GRU)) σχεδιάστηκαν για να αντιμετωπίσουν το πρόβλημα των vanishing gradients. Στο εσωτερικό τους υπάρχουν πύλες που ελέγχουν κατάλληλα τη ροή της πληροφορίας. Οι πύλες αυτές μπορούν να μάθουν ποια δεδομένα είναι σημαντικό να κρατήσουν και ποια να αποβάλλουν. Με τον τρόπο αυτό, το LSTM μπορεί να περάσει τη χρήσιμη πληροφορία κατά μήκος μίας μεγάλης αλυσίδας από αλληλουχίες για να καταλήξει σε προβλέψεις. Το σύνολο σχεδόν των state of the art αποτελεσμάτων που βασίζονται σε RNNs έχει επιτευχθεί με τη χρήση LSTM (και GRU). Τα δίκτυα αυτά χρησιμοποιούνται ευρέως στην αναγνώριση και σύνθεση φωνής, στη σύνθεση κειμένου, στην αναγνώριση δράσης και πολλές άλλες εφαρμογές. Στο Σχήμα 5.1 παρουσιάζεται μία μονάδα LSTM.

Το cell state του LSTM εμπεριέχει χρήσιμες πληροφορίες καθ' όλη τη διάρκεια επεξεργασίας μίας αλληλουχίας δεδομένων. Καθώς προχωρά η επεξεργασία της αλληλουχίας, πληροφορία προστίθεται και αφαιρείται από το cell state μέσω των πυλών. Οι πύλες αποτελούν διαφορετικά νευρωνικά δίκτυα που αποφασίζουν ποια πληροφορία θα υπάρχει στο cell state. Οι πύλες μαθαίνουν ποιες πληροφορίες είναι χρήσιμες κατά τη διάρκεια της εκπαίδευσης [59].

Οι πύλες χρησιμοποιούν σιγμοειδείς συναρτήσεις ενεργοποίησης, οι οποίες κανονικοποιούν όλες τις τιμές εισόδου στο διάστημα [0,1]. Το γεγονός αυτό διευκολύνει την ανανέωση ή την απόρριψη των δεδομένων καθώς κάθε αριθμός που πολλαπλασιάζεται με το μηδέν μηδενίζεται και άρα η πληροφορία του απορρίπτεται (ξεχνιέται). Αντίστοιχα, κάθε αριθμός που πολλαπλασιάζεται με το ένα παραμένει ο ίδιος και άρα η πληροφορία του διατηρείται. Μία μονάδα LSTM διαθέτει τρεις πύλες: την forget gate, την input gate και την output gate.



Σχήμα 5.1: Μονάδα LSTM [11].

Αρχικά, βρίσκεται η πύλη απόρριψης (forget gate). Η πύλη αυτή αποφασίζει ποια πληροφορία πρέπει να απορριφθεί και ποια να διατηρηθεί. Η πληροφορία από την προηγούμενη κρυφή κατάσταση και η πληροφορία εισόδου περνούν από μία σιγμοειδή συνάρτηση. Αν το αποτέλεσμα βρίσκεται πιο κοντά στο μηδέν σημαίνει ότι απορρίπτεται, ενώ αν βρίσκεται πιο κοντά στο ένα σημαίνει ότι διατηρείται (Εξίσωση 5.6).

$$f_t = \sigma(w_f[h_{t-1}, x_t] + b_f) \tag{5.6}$$

x_t : input
f_t : forget gate
w_f : forget gate weights
b_f : forget gate bias
h_{t-1} : previous hidden state
σ : sigmoid function

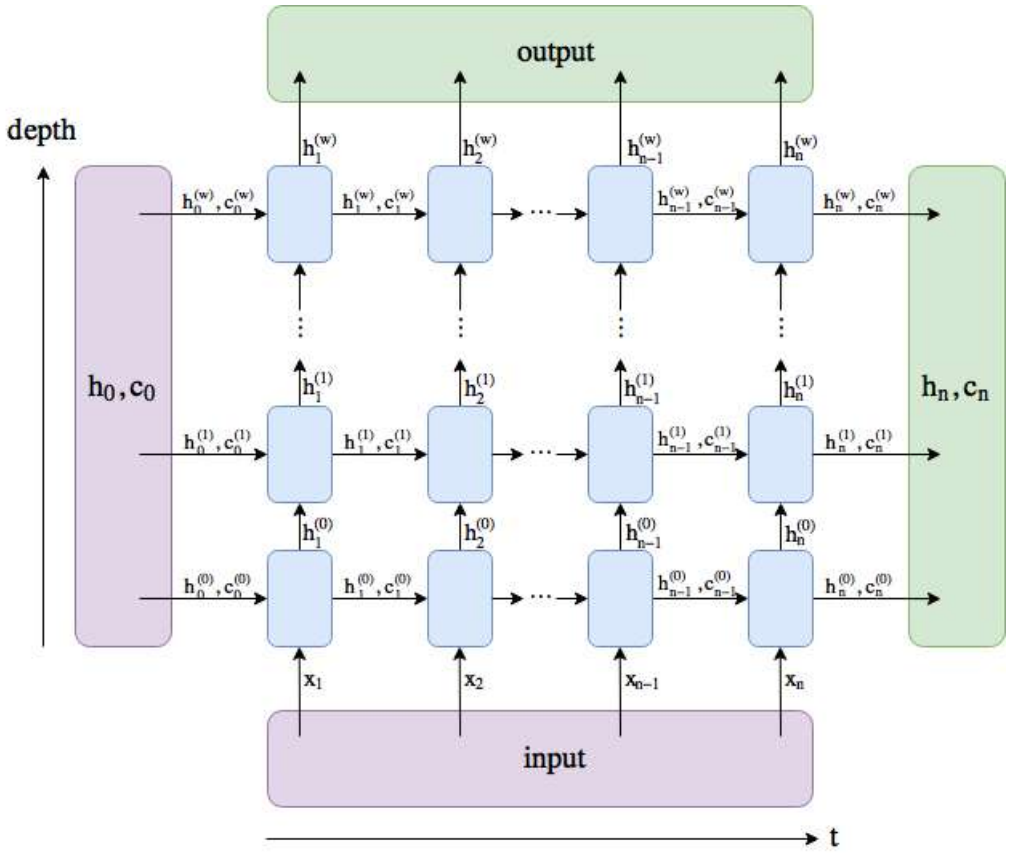
Για να ανανεωθεί το cell state, χρησιμοποιείται η πύλη εισόδου (input gate). Αρχικά, η προηγούμενη κρυφή κατάσταση και η είσοδος περνούν και πάλι από μία σιγμοειδή συνάρτηση. Αυτή αποφασίζει ποιες τιμές θα ανανεωθούν κανονικοποιώντας τις τιμές στο διάστημα [0,1]. Για άλλη μία φορά, το μηδέν σημαίνει μη σημαντική πληροφορία και το ένα σημαντική (Εξίσωση 5.7). Επιπρόσθετα, η προηγούμενη κρυφή κατάσταση και η είσοδος περνούν και από μία συνάρτηση tanh και οι τιμές τους κανονικοποιούνται στο [-1, 1] δίνοντας το υποψήφιο καινούριο cell state, (Εξίσωση 5.8). Οι έξοδοι της σιγμοειδούς και της tanh πολλαπλασιάζονται. Η έξοδος της σιγμοειδούς καθορίζει ποια πληροφορία πρέπει να διατηρηθεί από την έξοδο της tanh.

$$i_t = \sigma(w_f[h_{t-1}, x_t] + b_i) \tag{5.7}$$

i_t : input gate
i_f : input gate weights
b_i : input gate bias

$$\widetilde{c_t} = \tanh(w_c[h_{t-1}, x_t] + b_c) \tag{5.8}$$

$\widetilde{c_t}$: candidate cell state
w_c : candidate cell state weights
b_c : candidate cell state bias



Σχήμα 5.2: Απεικόνιση LSTM δικτύου [11]

Σε αυτό το σημείο υπάρχει επαρκής πληροφορία για να υπολογιστεί το επόμενο cell state. Αρχικά, το cell state πολλαπλασιάζεται σημείο προς σημείο με το διάνυσμα που έχει προκύψει από την πύλη απόρριψης. Επομένως, υπάρχει περίπτωση να απορριφθούν ορισμένες τιμές στο cell state, αν πολλαπλασιαστούν με τιμές κοντά στο μηδέν λόγω της εξόδου της forget gate. Στη συνέχεια, η έξοδος της πύλης εισόδου προστίθεται σημείο προς σημείο με το αλλαγμένο cell state, γεγονός που ανανεώνει το cell state με τιμές που το νευρωνικό δίκτυο θεωρεί χρήσιμες. Έτσι προκύπτει το νέο cell state(Εξίσωση 5.9).

$$c_t = i_t * \widetilde{c_t} + f_t * c_{t-1} \tag{5.9}$$

c_t : current cell state

c_{t-1} : previous cell state

Τέλος, η πύλη εξόδου (output gate) αποφασίζει ποια θα πρέπει να είναι η επόμενη κρυφή κατάσταση. Για άλλη μία φορά η προηγούμενη κρυφή κατάσταση και η είσοδος περνούν από μία σιγμοειδή συνάρτηση. Επίσης, το καινούριο cell state περνά από μία συνάρτηση tanh (Εξίσωση 5.10). Οι έξοδοι της σιγμοειδούς και της tanh πολλαπλασιάζονται ώστε να αποφασιστεί ποια πληροφορία θα πρέπει να μεταφέρει η καινούρια κρυφή κατάσταση (Εξίσωση 5.11). Η έξοδος (y_t) είναι η κρυφή κατάσταση. Τέλος, η καινούρια κρυφή κατάσταση και το νέο cell state μεταφέρονται στο επόμενο βήμα, όπως φαίνεται και στο Σχήμα 5.2.

$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o) \tag{5.10}$$

o_t : output gate

w_o : output gate weights

b_o : output gate bias

$$h_t = o_t * \tanh(c_t) \tag{5.11}$$

h_t : new hidden state (output)

5.2.2 Υλοποίηση της μεθόδου

Υλοποιήσαμε και εκπαιδεύσαμε το δίκτυο εκτίμησης. Ξεκινήσαμε χρησιμοποιώντας τα δεδομένα TD-Joint Attention του BabyRobot. Επιλέξαμε τη δομή του δικτύου έπειτα από ορισμένες δοκιμές εναλλακτικών συνδυασμών Fully Connected (FC) και LSTM στιβάδων, ξεκινώντας από το δίκτυο εκτίμησης των Hadfield et al. [2]. Στους Πίνακες 5.1 και 5.15 παρουσιάζουμε τα αποτελέσματα που προκύπτουν από την εκπαίδευση διάφορων παραλλαγών αρχιτεκτονικών του δικτύου. Παραθέτουμε αναλυτικά παρακάτω το τελικό δίκτυο, το οποίο πετύχαινε τα καλύτερα αποτελέσματα. Παρουσιάζουμε τη δομή του δικτύου και στο Σχήμα 5.3.

```
LSTM_Net(
    (FC1): Sequential(
        (0): Linear(in_features=58, out_features=400, bias=True)
        (1): Dropout(p=0.5, inplace=False)
        (2): ReLU()
    )
    (FC2): Sequential(
        (0): Linear(in_features=400, out_features=400, bias=True)
```

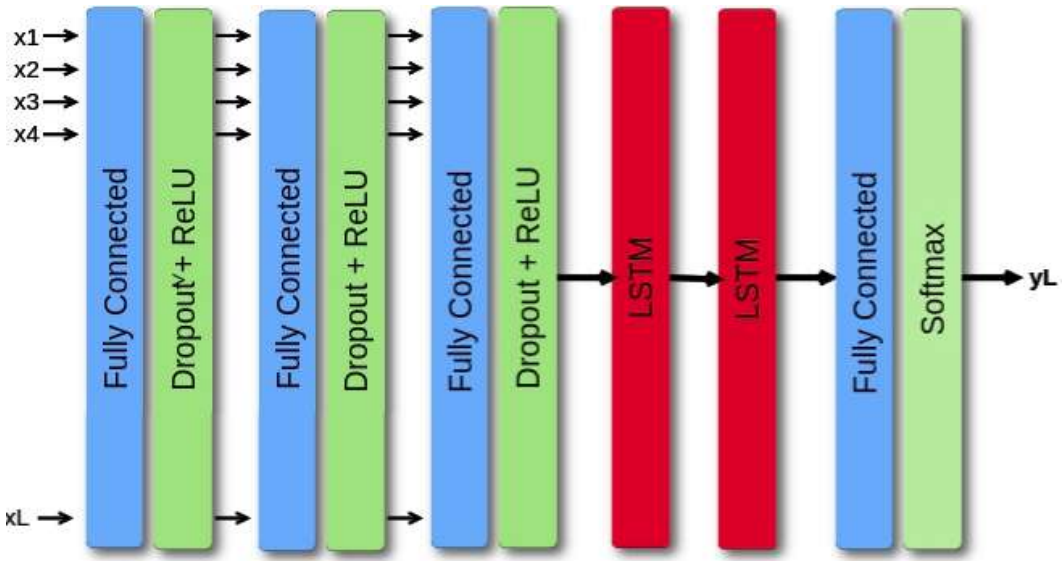


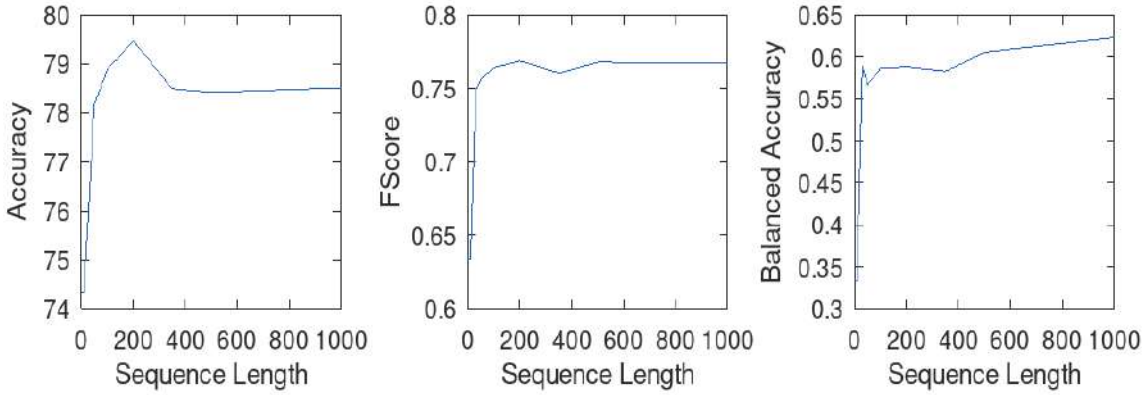
Figure 5.3: Το LSTM δίκτυο που χρησιμοποιούμε για την εκτίμηση του engagement.

```
        (1): Dropout(p=0.5, inplace=False)
        (2): ReLU()
    )
    (FC3): Sequential(
        (0): Linear(in_features=400, out_features=400, bias=True)
        (1): Dropout(p=0.5, inplace=False)
        (2): ReLU()
    )
    (LSTM1): Sequential(
        (0): LSTM(400, 400)
    )
    (LSTM2): Sequential(
        (0): LSTM(400, 200)
    )
    (FC4): Sequential(
        (0): Linear(in_features=200, out_features=3, bias=True)
    )
)
```

Net	Accuracy	W. Fscore	Balanced Accuracy
common class	74.32	63.38	33.33
3 FC, 1 LSTM, FC	76.23	74.15	53.78
3FC, 2 LSTM (2C, C), (C, C), FC	78.08	75.33	56.24
3FC, 2 LSTM (2C, 2C), (2C, C), FC	**78.78**	**76.21**	**56.61**
2 FC, 1 LSTM, FC	77.78	75.36	56.65
2 FC, 2 LSTM (2C, 2C), (2C,C), FC	78.10	75.33	54.97
2 FC, 3 LSTM (2C, 2C), (2C,C),(C,C), FC	77.56	72.80	52.14

Πίνακας 5.1: Αναζήτηση βέλτιστης αρχιτεκτονικής αναδρομικού δικτύου. Αποτελέσματα εκτίμησης στο σύνολο δεδομένων TD-Joint Attention.

Θα ήταν ενδιαφέρον να βγάλουμε συμπεράσματα για το με ποιο τρόπο επηρεάζει την εκτίμηση του engagement το χρονικό διάστημα που βλέπουν οι LSTM στιβάδες κάθε φορά, δηλαδή το μέγεθος της σειράς στιγμιοτύπων που εισάγονται κάθε φορά στο δίκτυο κατά την εκπαίδευση. Για να εξετάσουμε αυτή την εξάρτηση κρατώντας σταθερές όλες τις υπόλοιπες παραμέτρους του δικτύου, δοκιμάζουμε να το εκπαιδεύσουμε και να πάρουμε αντίστοιχες εκτιμήσεις για διάφορες τιμές του sequence length. Τα αποτελέσματα της διαδικασίας αυτής φαίνονται στα τρία διαγράμματα του Σχήματος 5.4.



Σχήμα 5.4: Μεταβολή των μετρικών αξιολόγησης ανάλογα με το χρονικό διάστημα που εισάγεται στο δίκτυο για το σύνολο δεδομένων TD-Joint Attention και εκτίμηση με LSTM δίκτυο.

Τα καρέ των video είναι δειγματοληπτημένα με ρυθμό 30 frame per second (fps). Ξεκινάμε με sequence length = 10, δηλαδή $\frac{1}{3}$ sec και φτάνουμε έως sequence length = 1000, δηλαδή περίπου 30secs. Δεν είναι πρακτικό να αυξήσουμε παραπάνω το μέγεθος του χρονικού διαστήματος καθώς πολλαπλασιάζεται ο χρόνος εκπαίδευσης. Παρατηρούμε ότι από τα 3 δευτερόλεπτα και για μεγαλύτερα sequence length επιτυγχάνουμε αρκετά καλό accuracy, το οποίο βελτιώνεται ακόμη περισσότερο στην περιοχή των 6-7 δευτερολέπτων. Όσον αφορά στις μετρικές που λαμβάνουν υπόψιν τους την ανισοκατανομή των κλάσεων βλέπουμε ότι το fscore επίσης από τα 3-4 δευτερόλεπτα και έπειτα παραμένει σχετικά σταθερό ενώ το balanced accuracy αυξάνεται συνεχώς με μία μικρή κλίση. Σίγουρα, υπάρχει συσχέτιση μεταξύ του χρονικού διαστήματος που βλέπουν τα LSTM και των τελικών αποτελεσμάτων της εκτίμησης. Φαίνεται πως 3 δευτερόλεπτα του video περιέχουν σημαντική πληροφορία για το engagement, αναγκαία και ικανή ώστε να είναι δυνατή η εκτίμησή του. Το συμπέρασμα αυτό ταυτίζεται με τα συμπεράσματα στα οποία έχουν καταλήξει ψυχολόγοι, ότι τα 3 δευτερόλεπτα είναι επαρκής χρονική περίοδος για να γίνει διακριτό το επίπεδο engagement ενός ανθρώπου. Στο διπλό χρόνο μπορούμε συνεπώς να παρατηρήσουμε και αλλαγή του επιπέδου του engagement. Συνεπώς, για την υπόλοιπη εργασία χρησιμοποιήσαμε στα δίκτυά μας μέγεθος χρονικής ακολουθίας 200 timestamps, δηλαδή 6-7 secs. Το συμπέρασμα αυτό επιβεβαιώθηκε αργότερα και για τα συνελικτικά δίκτυα.

Στη συνέχεια πραγματοποιήσαμε ορισμένες δοκιμές για να επιλέξουμε καλύτερες παραμέτρους για το δίκτυο. Τα αποτελέσματα παρουσιάζονται συνοπτικά στους πίνακες 5.2, 5.3 και 5.4. Ύστερα από μία βασική μελέτη για τις τιμές των παραμέτρων καταλήγουμε σε ορισμένες τιμές που παρουσιάζουν σχετικά καλύτερα αποτελέσματα εκτίμησης. Έτσι, το δίκτυο λαμβάνει τα δεδομένα σε ομάδες των 128 (batch_size = 128), ενώ το learning

rate του δικτύου ορίζεται ίσο με 10^{-4}. Επίσης, θέτουμε το μέγεθος των στιβάδων του δικτύου $C = 200$ για τις δύο τελευταίες και $2C = 400$ για τις υπόλοιπες.

batch_size	learning_rate	Accuracy	W. Fscore	Balanced Acc
common class		74.32	63.38	33.33
128	10^{-3}	74.52	0.6956	0.3349
128	10^{-4}	**78.78**	**76.21**	**56.61**
128	10^{-5}	76.22	70.06	51.09

Πίνακας 5.2: Ενδεικτικά αποτελέσματα αναζήτησης βέλτιστου learning rate του LSTM δικτύου για την εκτίμηση στο σύνολο δεδομένων TD-Joint Attention.

batch_size	learning_rate	Accuracy	Fscore	Balanced Acc
common class		74.32	63.38	33.33
32	10^{-4}	76.34	72.70	50.98
64	10^{-4}	77.45	75.43	54.62
128	10^{-4}	**78.78**	**76.21**	**56.61**
256	10^{-4}	**79.32**	76.01	50.66

Πίνακας 5.3: Ενδεικτικά αποτελέσματα αναζήτησης βέλτιστου batch size του LSTM δικτύου για την εκτίμηση στο σύνολο δεδομένων TD-Joint Attention.

C	Accuracy	W. Fscore	Balanced Acc
common class	74.32	63.38	33.33
25	75.92	75.45	55.41
50	78.00	75.98	58.03
100	78.78	76.21	56.61
200	**79.47**	**76.88**	**58.82**
500	78.71	76.44	56.50

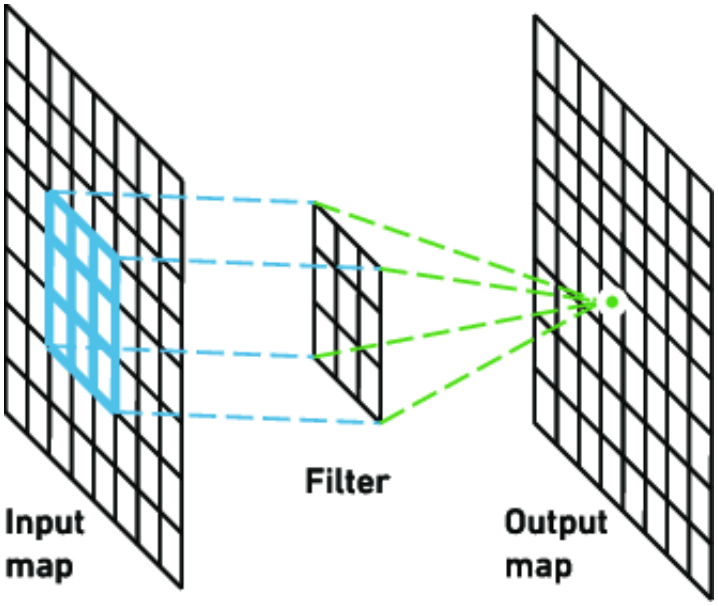
Πίνακας 5.4: Ενδεικτικά αποτελέσματα αναζήτησης βέλτιστου μεγέθους στιβάδων C του LSTM δικτύου για την εκτίμηση στο σύνολο δεδομένων TD-Joint Attention.

5.3 Εκτίμηση με συνελικτικό δίκτυο

Στη συνέχεια, κάναμε προσπάθειες να εκτιμήσουμε το engagement αξιοποιώντας συνελικτικά νευρωνικά δίκτυα. Στην αρχή δοκιμάσαμε με μονοδιάστατες συνελίξεις κατασκευάζοντας ένα βαθύ συνελικτικό δίκτυο πολλών καναλιών κατά τα πρότυπα του [7] των Javed et al.

Έπειτα, εμπνευσμένοι από το "2D/3D Pose Estimation and Action Recognition using Multitask Deep Learning" [6], το οποίο αναφέρουμε στο Κεφάλαιο 3, σχεδιάσαμε δίκτυα με στόχο να αξιοποιούνται ταυτόχρονα η υψηλού επιπέδου πληροφορία που εμπεριέχεται στις θέσεις των μελών του σώματος, οι μεγάλες δυνατότητες εξαγωγής συμπερασμάτων των CNN και η πληροφορία από τη χρονική αλληλουχία των στιγμιοτύπων.

5.3.1 CNN



Σχήμα 5.5: Απεικόνιση συνελικτικής στιβάδας [12]

Τα δίκτυα CNN αποτελούν μία κατηγορία βαθιών νευρωνικών δικτύων που χρησιμοποιούνται ευρέως στην Όραση Υπολογιστών, στην αναγνώριση από οπτικά δεδομένα, στην επεξεργασία εικόνων. Τα δίκτυα αυτά έχουν την ιδιότητα να μειώνουν αποτελεσματικά το μέγεθος πολύ μεγάλων συνόλων δεδομένων, διατηρώντας ταυτόχρονα τα χαρακτηριστικά των εικόνων.

Η λειτουργία ενός τέτοιου δικτύου βασίζεται στο συνδυασμό συνελικτικών στιβάδων (convolutional layers), στιβάδων συγκέντρωσης - υποδειγματοληψίας (pooling layers) καθώς και πλήρων στιβάδων (fully connected layers).

Οι στιβάδες δειγματοληψίας χρησιμοποιούν πολλαπλά φίλτρα μικρών διαστάσεων με τα οποία πραγματοποιείται συνέλιξη των εικόνων εισόδου κατά πλάτος και ύψος της εικόνας. Κάθε φίλτρο χρησιμοποιείται για να εντοπίζει την ύπαρξη συγκεκριμένων χαρακτηριστικών ή προτύπων στις εικόνες εισόδου. Έτσι, μέσω της συνελικτικής στιβάδας παράγεται ένα σύνολο από χάρτες ενεργοποίησης. Ένα παράδειγμα λειτουργίας της συνελικτικής στιβάδας παρουσιάζεται στο Σχήμα 5.5 [12].

Οι στιβάδες συγκέντρωσης τοποθετούνται ανάμεσα στις συνελικτικές στιβάδες. Οι στιβάδες αυτές χρησιμοποιώντας συγκεκριμένα φίλτρα πραγματοποιούν υποδειγματοληψία, δηλαδή μειώνουν το πλήθος των παραμέτρων στο δίκτυο. Με αυτό τον τρόπο, μειώνοντας δηλαδή το μέγεθος του δικτύου στο χώρο, αντιμετωπίζουν φαινόμενα overfitting. Για παράδειγμα, οι στιβάδες max pooling που χρησιμοποιούμε εδώ, κρατούν από ένα μικρό σύνολο τιμών μόνο τη μέγιστη τιμή και απορρίπτουν όλες τις υπόλοιπες, όπως φαίνεται στο Σχήμα 5.6 [13].

Στο τέλος των συνελικτικών δικτύων τοποθετούνται ορισμένες πλήρεις στιβάδες, στις οποίες τα δεδομένα εισόδου φτάνουν αφού έχουν συμπτυχθεί από τις προηγούμενες στιβάδες του δικτύου, αλλά με αναλλοίωτα τα τα βασικά τους χαρακτηριστικά, ώστε να μπορούν να εξαχθούν χρήσιμα συμπεράσματα. Όπως πάντα τις πλήρεις στιβάδες ακολουθεί τέλος και μία στιβάδα ενεργοποίησης. Στο Σχήμα 5.9 [60] παραθέτουμε το δίκτυο AlexNet, ένα ιδιαίτερα διαδεδομένο και σχετικά απλό συνελικτικό δίκτυο, τη δομή του οποίου χρησιμοποιούμε και στην παρούσα διπλωματική.

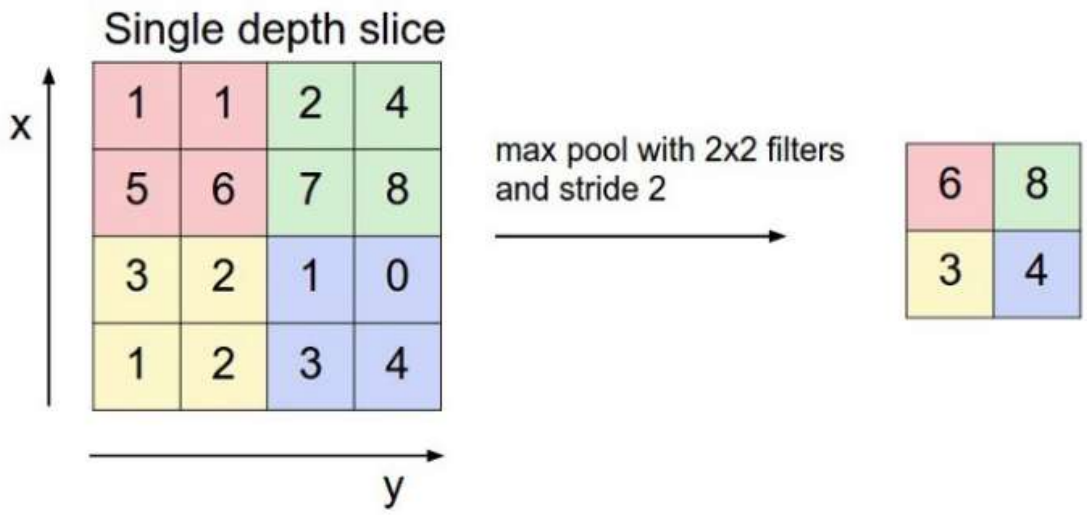
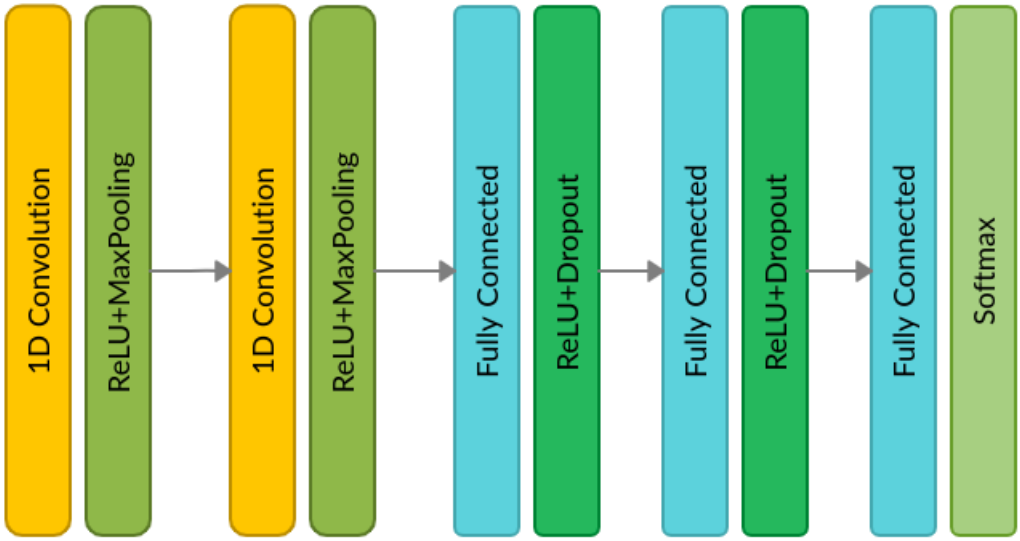


Figure 5.6: Απεικόνιση στιβάδας max pooling [12]



Σχήμα 5.7: Το 1DCNN δίκτυο που χρησιμοποιούμε για την εκτίμηση του engagement.

5.3.2 Υλοποίηση της μεθόδου με 1D CNNs

Και πάλι με τη χρήση της βιβλιοθήκης PyTorch προχωρήσαμε στην υλοποίηση του δικτύου. Πρόκειται για ένα δίκτυο στα πρότυπα του [7], με βασική μετατροπή την προσθήκη στιβάδων MaxPooling. Παρουσιάζουμε τη δομή του δικτύου αυτού στο Σχήμα 5.7.

```
CNN_Net(
    (features): Sequential(
        (0): Conv1d(sequence_length, 64, kernel_size=(3,), stride=(1,),
        padding=(1,))
        (1): ReLU(inplace=True)
        (2): MaxPool1d(kernel_size=3, stride=1, padding=0, dilation=1,
        ceil_mode=False)
        (3): Dropout(p=0.2, inplace=False)
```

```
        (4): Conv1d(64, 128, kernel_size=(3,), stride=(1,), padding=(1,))
        (5): ReLU(inplace=True)
        (6): MaxPool1d(kernel_size=3, stride=1, padding=0, dilation=1,
        ceil_mode=False)
        (7): Dropout(p=0.2, inplace=False)
    )
    (avgpool): AdaptiveAvgPool2d(output_size=(6))
    (classifier): Sequential(
        (1): Linear(in_features=36, out_features=256, bias=True)
        (2): ReLU(inplace=True)
        (3): Dropout(p=0.2, inplace=False)
        (4): Linear(in_features=256, out_features=256, bias=True)
        (5): ReLU(inplace=True)
        (3): Dropout(p=0.2, inplace=False)
        (6): Linear(in_features=256, out_features=3, bias=True)
    )
)
```

Πρόκειται για ένα αρκετά απλό δίκτυο, το οποίο δε λαμβάνει υπόψιν τη χρονική συνέχεια των στιγμιοτύπων και άρα δε μπορεί να αντλήσει πληροφορίες από αυτή. Όπως δείχνουμε παρακάτω, καταφέρνουμε με το δίκτυο αυτό να πάρουμε ορισμένα αποτελέσματα, τα οποία όμως δεν ξεπερνούν το LSTM ή τα 2D CNN δίκτυα.

5.3.3 Υλοποίηση της μεθόδου με AlexNet

Αρχικά, πραγματοποιήσαμε κατάλληλη αναδιάταξη της μορφής των δεδομένων εισόδου, μετατρέποντας ουσιαστικά τα σύνολα των μελών που αποτελούν τις πόζες σε αναπαραστάσεις που μοιάζουν με εικόνες. Ο κάθετος άξονας αντιπροσωπεύει το χρόνο, ενώ ο οριζόντιος αντιπροσωπεύει τα σημεία του σκελετού. Τέλος, οι συντεταγμένες κάθε σημείου είτε είναι 2D - (x,y) όπως στα δεδομένα BabyAffect είτε είναι 3D - (x, y, z), όπως στα δεδομένα Joint Attention αποτελούν τα κανάλια του δικτύου. Η αναδιάταξη αυτή φαίνεται στο Σχήμα 5.8.

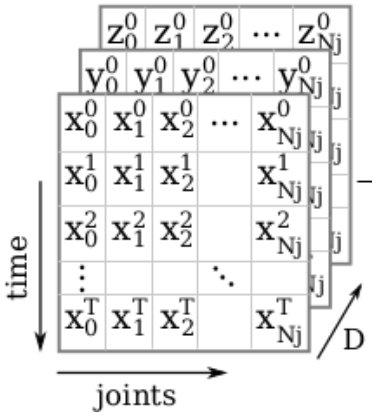


Figure 5.8: Αναδιάταξη των δεδομένων του σκελετού σε τρισδιάστατη αναπαράσταση [6].

Με αυτό τον τρόπο είναι δυνατό να χρησιμοποιήσουμε κλασικές δισδιάστατες συνελίξεις ώστε εξάγουμε μοτίβα απευθείας από την χρονική αλληλουχία των σημείων του

σκελετού. Έτσι, χρησιμοποιούμε ένα βαθύ συνελικτικό νευρωνικό δίκτυο για να εξάγουμε χαρακτηριστικά από τις πόζες εισόδου.

Επιλέγουμε να εκπαιδεύσουμε ένα δίκτυο της δομής του AlexNet [60]. Το AlexNet διαγωνίστηκε στο ImageNet Large Scale Visual Recognition Challenge το 2012. Πέτυχε top-5 error 15.3%, περισσότερο από 10.8% χαμηλότερο από το προηγούμενο state of the art δίκτυο. Είναι ένα ισχυρό συνελικτικό δίκτυο, που έχει χρησιμοποιηθεί ευρέως από τότε και έχει σχετικά απλή δομή γεγονός που εξυπηρετεί τα πειράματά μας. Μετατρέπουμε κατάλληλα τα μεγέθη των στιβάδων και των παραμέτρων τους για να ταιριάζουν στα δεδομένα μας. Έτσι,υλοποιούμε και εκπαιδεύουμε με τη χρήση της βιβλιοθήκης PyTorch το παρακάτω δίκτυο:

```
AlexNet(
    (features): Sequential(
        (0): Conv2d(3, 64, kernel_size=(5, 5), stride=(1, 1), padding=(1, 1))
        (1): ReLU(inplace=True)
        (2): MaxPool2d(kernel_size=3, stride=1, padding=0, dilation=1,
        ceil_mode=False)
        (3): Conv2d(64, 192, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
        (4): ReLU(inplace=True)
        (5): MaxPool2d(kernel_size=3, stride=1, padding=0, dilation=1,
        ceil_mode=False)
        (6): Conv2d(192, 384, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
        (7): ReLU(inplace=True)
        (8): Conv2d(384, 256, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
        (9): ReLU(inplace=True)
        (10): Conv2d(256, 256, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
        (11): ReLU(inplace=True)
        (12): MaxPool2d(kernel_size=3, stride=1, padding=0, dilation=1,
        ceil_mode=False)
    )
    (avgpool): AdaptiveAvgPool2d(output_size=(6, 6))
    (classifier): Sequential(
        (0): Dropout(p=0.5, inplace=False)
        (1): Linear(in_features=9216, out_features=4096, bias=True)
        (2): ReLU(inplace=True)
        (3): Dropout(p=0.5, inplace=False)
        (4): Linear(in_features=4096, out_features=4096, bias=True)
        (5): ReLU(inplace=True)
        (6): Linear(in_features=4096, out_features=3, bias=True)
    )
)
```

Για να επιβεβαιώσουμε το συμπέρασμα μας για το χρονικό διάστημα που χρειάζεται να λαμβάνουν κάθε φορά τα δίκτυά μας για να επιτυγχάνουν καλή εκτίμηση του engagement πραγματοποιήσαμε και με το συνελικτικό δίκτυο αντίστοιχα πειράματα με εκείνα για το αναδρομικό δίκτυο. Έτσι, εκπαιδεύσαμε συνελικτικά δίκτυα και πήραμε αντίστοιχες εκτιμήσεις για διάφορες τιμές του sequence length. Τα αποτελέσματα της διαδικασίας αυτής φαίνονται στο διάγραμμα του Σχήματος 5.10. Και από αυτά επιβεβαιώνουμε ότι τα 3 δευτερόλεπτα περιέχουν σημαντική πληροφορία για το engagement, αναγκαία και ικανή

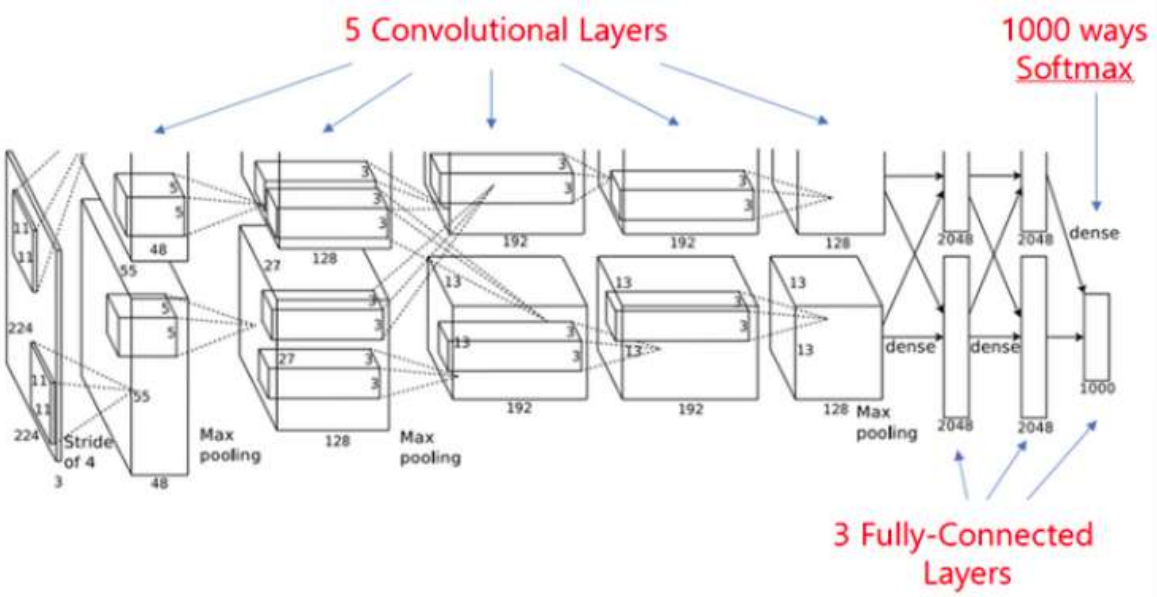


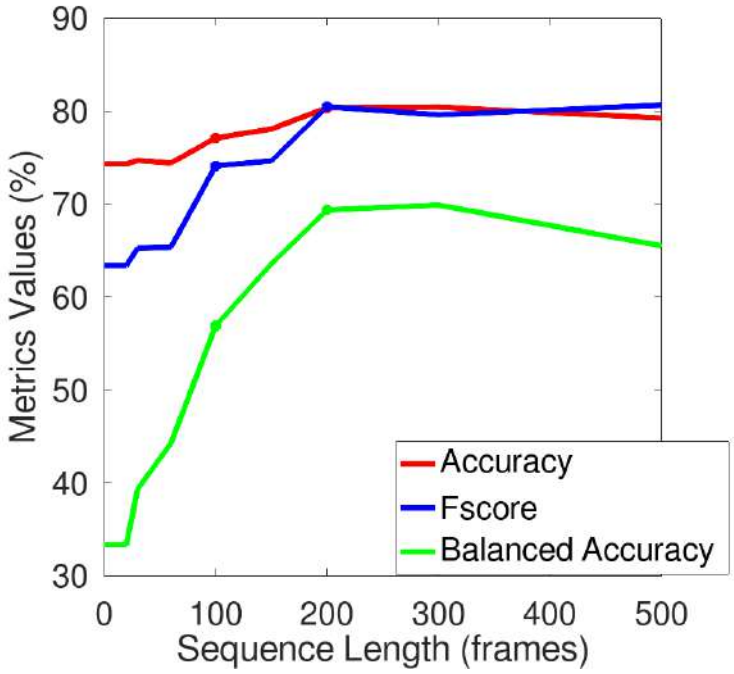
Figure 5.9: Το δίκτυο AlexNet [13].

για την εκτίμησή του.

5.3.4 Υλοποίηση της μεθόδου με απλούστερο 2D CNN

Το δίκτυο αυτό απαιτούσε σημαντικό χρόνο και πόρους για την εκπαίδευσή του. Για το λόγο αυτό δοκιμάσαμε να εκπαιδεύσουμε και απλούστερα δισδιάστατα συνελικτικά δίκτυο για την εκτίμηση. Ένα από αυτά το οποίο οδηγούσε σε αποτελέσματα αρκετά κοντά στο AlexNet δίκτυο είναι το ακόλουθο. Παρουσιάζουμε τη δομή του και στο Σχήμα 5.11

```
CNN_Net(
    (features): Sequential(
        (0): Conv2d(3, 26, kernel_size=(5,5), stride=(1,1), padding=(1,1))
        (1): ReLU(inplace=True)
        (2): MaxPool1d(kernel_size=3, stride=1, padding=0, dilation=1,
        ceil_mode=False)
        (3): Conv2d(26, 80, kernel_size=(3,3), stride=(1,1), padding=(1,1))
        (4): ReLU(inplace=True)
        (5): Conv2d(80, 160, kernel_size=(3,3), stride=(1,1), padding=(1,1))
        (6): ReLU(inplace=True)
        (7): MaxPool2d(kernel_size=3, stride=1, padding=0, dilation=1,
        ceil_mode=False)
    )
    (avgpool): AdaptiveAvgPool2d(output_size=(6,6))
    (classifier): Sequential(
        (0): Dropout(p=0.5, inplace=False)
        (1): Linear(in_features=5760, out_features=2048, bias=True)
        (2): ReLU(inplace=True)
        (4): Linear(in_features=2048, out_features=3, bias=True)
    )
)
```



Σχήμα 5.10: Μεταβολή των μετρικών αξιολόγησης ανάλογα με το χρονικό διάστημα που εισάγεται στο δίκτυο για το σύνολο δεδομένων TD-Joint Attention και εκτίμηση με συνελικτικό δίκτυο AlexNet.

Παρακάτω, στον Πίνακα 5.5 παρουσιάζουμε ορισμένα αποτελέσματα από όλα τα παραπάνω δίκτυα για τα δεδομένα TD-Joint Attention. Το δίκτυο AlexNet μας δίνει τα καλύτερα αποτελέσματα με πολύ υψηλό accuracy και ταυτόχρονα πολύ καλές προβλέψεις για όλες τις κλάσεις των δεδομένων. Τα υπόλοιπα δίκτυα δίνουν επίσης καλά αποτελέσματα, ωστόσο με χαμηλότερα balanced accuracy, γεγονός που μας δείχνει πως δεν κατατάσσουν εξίσου καλά τα στιγμιότυπα της σπανιότερης κλάσης.

Net	Accuracy	W. Fscore	Balanced Acc	W.Precision
common class	74.32	63.38	33.33	55.25
1D CNN	77.44	75.15	58.28	76.30
2D CNN	78.93	76.46	60.13	77.43
LSTM	79.47	76.88	58.82	78.04
AlexNet	**80.36**	**80.48**	**69.37**	**80.71**

Πίνακας 5.5: Αποτελέσματα εκτίμησης για τις διάφορες αρχιτεκτονικές των δικτύων στο σύνολο δεδομένων TD-Joint Attention.

Στο Σχήμα 5.12 παρουσιάζουμε τέσσερα στιγμιότυπα από μία αλληλεπίδραση του συνόλου δεδομένων TD-Joint Attention. Στα στιγμιότυπα αυτά έχει εξαχθεί ο σκελετός του παιδιού ενώ ταυτόχρονα αναγράφονται και το επισημειωμένο επίπεδο engagement (Ground Truth), αλλά και η εκτίμηση του επιπέδου engagement από το δίκτυο AlexNet (Our Method). Και στα τέσσερα στιγμιότυπα το δίκτυό μας εκτιμά σωστά το επίπεδο του engagement. Στο πρώτο στιγμιότυπο, το οποίο βρίσκεται χρονικά στην αρχή της αλληλεπίδρασης το παιδί ήδη φαίνεται πως παρακολουθεί το ρομπότ, καθώς έχει στραμ-

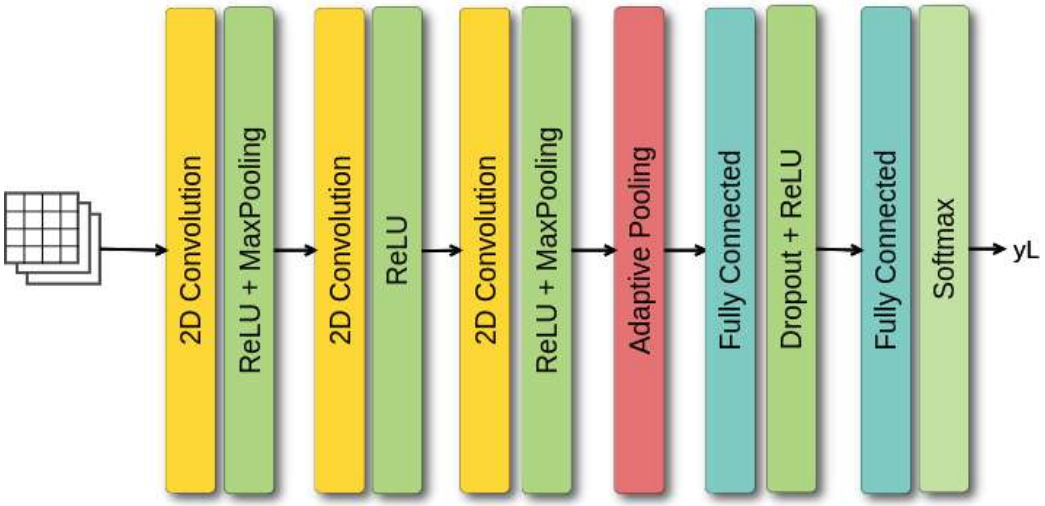


Figure 5.11: Το απλούστερο 2DCNN δίκτυο που χρησιμοποιούμε για την εκτίμηση του engagement.

μένο το σώμα του και το κεφάλι του προς αυτό. Το επίπεδο του engagement εδώ είναι 1. Στο δεύτερο στιγμιότυπο το ρομπότ έχει ξεκινήσει να κινείται ώστε να δείξει στο παιδάκι ότι προσπαθεί να πιάσει το τουβλάκι. Και εδώ το επίπεδο του engagement είναι 1. Στα επόμενα δύο στιγμιότυπα το παιδί πιάνει το τουβλάκι και στη συνέχεια το δίνει στο ρομπότ. Εδώ το παιδί είναι fully engaged καθώς δρα για να βοηθήσει το ρομπότ (επίπεδο engagement 2). Σε αυτή την αλληλεπίδραση το δίκτυό μας εκτιμά σωστά τόσο το μεσαίο όσο και υψηλότερο επίπεδο του engagement.

5.4 Εξατομίκευση του δικτύου

Στη συνέχεια, δοκιμάζουμε να εξατομικεύσουμε το δίκτυο μας σε κάθε παιδί ξεχωριστά, ώστε να εξετάσουμε πως επηρεάζονται τα αποτελέσματα της εκτίμησης. Στην Ενότητα 2.4 περιγράφονται ορισμένες επιτυχείς προσεγγίσεις που εκτιμούν το engagement χρησιμοποιώντας εξατομικευμένα δίκτυα.

Χρησιμοποιούμε τα TD δεδομένα Joint Attention της BabyRobot για τα οποία έχουμε παρουσιάσει ορισμένα αποτελέσματα παραπάνω. Το test set με το οποίο δουλεύουμε περιλαμβάνει τα δεδομένα από τις καταγραφές πέντε παιδιών. Χωρίζουμε εκ νέου τα δεδομένα κάθε παιδιού σε train και test set, με το train set να αποτελείται από το 15% των στιγμιοτύπων και το test set από το 85% των στιγμιοτύπων, καθώς επιθυμούμε απλά το δίκτυο να εκπαιδευτεί περεταίρω για λίγες εποχές σε λίγα μόνο στιγμιότυπα από κάθε παιδί του test set. Στη συνέχεια εκπαιδεύουμε ξεχωριστά για κάθε παιδί το δίκτυό μας με τα νέα train δεδομένα. Παρουσιάζουμε τα αποτελέσματα ενδεικτικά για ένα παιδί στον Πίνακα 5.6. Τα αποτελέσματα και για τα υπόλοιπα παιδιά ήταν αντίστοιχα.

Βλέπουμε πως η βελτίωση καθώς εξατομικεύουμε την εκπαίδευση σε κάθε παιδί είναι ελάχιστη. Μάλιστα, σε δύο από τα πέντε παιδιά δεν καταφέραμε να βελτιώσουμε καθόλου την εκτίμηση με αυτόν τον τρόπο. Συνεπώς, το σημαντικό υπολογιστικό κόστος που συνεπάγεται η εξειδίκευση του δικτύου σε κάθε παιδί ξεχωριστά δεν αντισταθμίζεται από την ελάχιστη βελτίωση στην ικανότητα εκτίμησης του engagement κάθε παιδιού. Για το λόγο αυτό, στη συνέχεια της παρούσας εργασίας εστιάσαμε στην εκπαίδευση μοντέλων που θα μπορούν να χρησιμοποιηθούν για την εκτίμηση του engagement όσο το δυνατόν

Σχήμα 5.12: Εκτίμηση του επιπέδου engagement από το AlexNet δίκτυό μας για τέσσερα στιγμιότυπα TD-Joint Attention. Σε κάθε στιγμιότυπο αναγράφονται το πραγματικό επίπεδο του engagement (ground truth) και η τιμή που εκτιμά το δίκτυό μας (our method). Στα δύο πάνω στιγμιότυπα ground truth:1, our method:1, στα δύο κάτω στιγμιότυπα ground truth:2 our method:2.

Child1	Accuracy	W. Fscore	Balanced Acc
common class	89.57	78.78	33.33
generic net	95.91	95.47	86.95
personalized net	95.96	95.52	87.09

Πίνακας 5.6: Συγκριτικά αποτελέσματα εκτίμησης σε ένα από τα παιδιά του TD-Joint Attention για το γενικό και το εξατομικευμένο μοντέλο.

περισσότερων παιδιών.

5.5 Επέκταση σε παιδιά με διαταραχές αυτιστικού φάσματος - ASD Joint Attention

Εφόσον καταλήξαμε σε ορισμένα δίκτυα που προβλέπουν με σημαντική επιτυχία το επίπεδο engagement στα πειράματα Joint Attention με τυπικώς αναπτυσσόμενα παιδιά, προχωρήσαμε ώστε να δοκιμάσουμε τα δίκτυά μας και στα υπόλοιπα σύνολα δεδομένων μας, σε παιδιά με διαταραχές αυτιστικού φάσματος και σε διαφορετικές συνθήκες, διαφορετικά περιβάλλοντα ή διαφορετικούς τύπους παιχνιδιών.

Αρχικά, δοκιμάσαμε να εκτιμήσουμε το engagement των δεδομένων ASD-Joint Attention με τα προεκπαιδευμένα μοντέλα στις αλληλεπιδράσεις TD-Joint Attention. Παρουσιάζουμε τα πρώτα αποτελέσματα στον παρακάτω στον Πίνακα 5.7. Παρατηρούμε ότι τα προεκπαιδευμένα μοντέλα σε παιδιά τυπικώς ανεπτυγμένα αποτυγχάνουν να εκτιμήσουν το engagement παιδιών με διαταραχές αυτιστικού φάσματος. Στις περισσότερες περιπτώσεις

τα accuracy δεν ξεπερνούν το accuracy που θα πετυχαίναμε αν κατατάσσαμε όλα τα στιγμιότυπα στην συνηθέστερη κλάση, ενώ ταυτόχρονα και τα μεγέθη των άλλων μετρικών είναι σχετικά χαμηλά.

Net	Accuracy	W. Fscore	Balanced Acc
common class	57.44	41.15	33.33
1D CNN	50.77	39.51	47.28
2D CNN	**59.58**	**60.39**	**54.32**
LSTM	44.01	37.43	48.71
AlexNet	33.40	32.75	49.40

Πίνακας 5.7: Αποτελέσματα για τις διάφορες αρχιτεκτονικές των δικτύων στα δεδομένα ASD-Joint Attention κάνοντας απευθείας εκτίμηση του engagement με τα προεκπαιδευμένα δίκτυα στα TD παιδιά.

Τα αποτελέσματα αυτά ήταν αναμενόμενα σε ένα βαθμό για μία σειρά λόγων, που αναφέρονται και νωρίτερα στην παρούσα εργασία. Τα παιδιά που εμφανίζουν διαταραχές του αυτιστικού φάσματος δεν παρουσιάζουν κατά τις αλληλεπιδράσεις τους ορισμένα τυπικά σημάδια του επιπέδου του engagement. Χαρακτηριστικό τέτοιο παράδειγμα αποτελεί η έλλειψη σταθερότητας στο βλέμμα, η συνεχής μετακίνησή του. Επιπρόσθετα, συχνά τα παιδιά με αυτισμό εμφανίζουν ορισμένες στερεοτυπικές κινήσεις και συμπεριφορές κατά τη διάρκεια των αλληλεπιδράσεων, οι οποίες δεν εμφανίζονται στις αλληλεπιδράσεις τυπικώς αναπτυσσόμενων παιδιών. Συνεπώς, τα μοτίβα, τα οποία έχουν ξεχωρίσει τα δίκτυά μας και τα οποία τους επιτρέπουν να αναγνωρίζουν με επιτυχία το engagement τυπικώς αναπτυσσόμενων παιδιών, δεν μπορούν να εφαρμοστούν αυτούσια για να πραγματοποιηθεί αντίστοιχα επιτυχής αναγνώριση του επιπέδου engagement των παιδιών με αυτισμό.

Τα δεδομένα ASD-Joint Attention δεν είναι αρκετά ώστε να μπορούμε να εκπαιδεύσουμε τα δίκτυα εξαρχής αποκλειστικά με αυτά. Συνεπώς, επιχειρούμε ξεκινώντας από τα προεκπαιδευμένα δίκτυα για τα τυπικώς αναπτυσσόμενα παιδιά να συνεχίσουμε την εκπαίδευση για μερικές ακόμη εποχές ώστε να τα εξειδικεύσουμε σε αλληλεπιδράσεις με παιδιά με διαταραχές αυτιστικού φάσματος. Παρουσιάζουμε τα αποτελέσματα στον Πίνακα 5.8. Τα αποτελέσματα αυτά είναι πολύ ενθαρρυντικά καθώς μέσα σε πολύ λίγες μόνο εποχές (λιγότερες από δέκα) τα δίκτυα μας πετυχαίνουν πολύ καλά αποτελέσματα, σχεδόν παρόμοια με τα αποτελέσματα για τα τυπικώς αναπτυσσόμενα παιδιά. Πιο συγκεκριμένα παρατηρούμε ότι το accuracy του δικτύου AlexNet είναι 20 ποσοστιαίες μονάδες μεγαλύτερο από εκείνο της συνηθέστερης κλάσης και ταυτόχρονα έχει διπλασιαστεί σε σχέση με την εκτίμηση του δικτύου που είχε εκπαιδευτεί μόνο στα TD δεδομένα. Αντίστοιχα, αυξάνονται σημαντικά όλες οι μετρικές γεγονός που δείχνει ότι η εκτίμηση βελτιώνεται ραγδαία όχι μόνο για τη συνηθέστερη αλλά και για τις άλλες κλάσεις. Οι τιμές των μετρικών παραμένουν λίγο χαμηλότερες από ότι στα αντίστοιχα δίκτυα για τα τυπικώς αναπτυσσόμενα παιδιά, όμως κάτι τέτοιο είναι λογικό δεδομένης της μεγαλύτερης δυσκολίας στην περίπτωση των παιδιών με αυτισμό.

Στη συνέχεια, πραγματοποιήσαμε ορισμένα πειράματα στα οποία εκπαιδεύσαμε τα δίκτυα από την αρχή χρησιμοποιώντας δεδομένα εκπαίδευσης τόσο από TD όσο και από ASD παιδιά συγχρόνως. Είναι σημαντικό να παρατηρήσουμε ότι στα ASD δεδομένα έχουμε περίπου μισά στιγμιότυπα σε σχέση με τα σημεία των TD δεδομένων. Επομένως, παρ΄ όλο που τα ASD δεδομένα δεν επαρκούν για να εκπαιδευτεί από την αρχή ένα δίκτυο αποτελεσματικά έχουν συγκρίσιμο όγκο με εκείνο των TD δεδομένων. Συγκρίνουμε τα αποτελέσματα των νέων δικτύων με τα προηγούμενα αποτελέσματα στα αντίστοιχα test

Net	Accuracy	W. Fscore	Balanced Acc	W.Precision
common class	57.44	41.15	33.33	32.26
1D CNN	67.91	67.98	58.73	68.10
2D CNN	70.88	71.32	60.64	73.63
LSTM	71.76	71.55	**62.11**	71.49
AlexNet	**78.73**	**73.39**	59.10	**82.38**

Πίνακας 5.8: Αποτελέσματα εκτίμησης για τις διάφορες αρχιτεκτονικές των δικτύων στα δεδομένα ASD-Joint Attention εκπαιδεύοντας για λίγες εποχές τα προεκπαιδευμένα στα TD παιδιά δίκτυα.

set των TD και των ASD παιδιών. Παρουσιάζουμε τα αποτελέσματα στους Πίνακες 5.9 και 5.10.

Σε γενικές γραμμές, παρατηρήσαμε πως η από κοινού εκπαίδευση TD και ASD παιδιών δε βελτίωσε στις περισσότερες περιπτώσεις τα αποτελέσματα (με μοναδική εξαίρεση το 2D CNN δίκτυο αλλά μόνο για τα ASD παιδιά). Για παράδειγμα το accuracy για όλα τα δίκτυα στα TD δεδομένα είναι δύο ποσοστιαίες μονάδες χαμηλότερο όταν τα δίκτυα εκπαιδεύτηκαν ταυτόχρονα με TD και με ASD δεδομένα. Κάτι τέτοιο, δεν είναι μη αναμενόμενο καθώς όπως έχει αναφερθεί υπάρχουν σημαντικές διαφορές στον τρόπο με τον οποίο εξωτερικεύεται το engagement σε τυπικώς αναπτυσσόμενα και σε παιδιά με διαταραχές αυτιστικού φάσματος. Επομένως, για τη συνέχεια κρατήσαμε τα αποτελεσματικότερα δίκτυα ξεχωριστά εκπαιδευμένα σε TD και ASD δεδομένα.

Net	Accuracy	W. Fscore	Balanced Acc
common class	74.32	63.38	33.33
LSTM(together)	77.05	73.92	56.55
LSTM(best)	**79.47**	**76.88**	**58.82**
2D CNN(together)	76.79	73.54	56.78
2D CNN(best)	**78.93**	**76.46**	**60.13**
AlexNet(together)	78.02	74.23	60.22
AlexNet(best)	**80.36**	**80.48**	**69.37**

Πίνακας 5.9: Αποτελέσματα εκτίμησης για τις διάφορες αρχιτεκτονικές των δικτύων στο σύνολο δεδομένων TD-Joint Attention έπειτα από κοινού εκπαίδευση ASD και TD δεδομένων.

5.6 Επέκταση στο σύνολο δεδομένων BabyAffect

Στη συνέχεια πραγματοποιήσαμε αντίστοιχα πειράματα με τα δεδομένα BabyAffect. Στα δεδομένα αυτά δεν υπάρχει πληροφορία για το βάθος καθώς όπως περιγράφουμε και στην Ενότητα 3.2. η καταγραφή έχει γίνει από μία απλή συμβατική κάμερα. Το γεγονός αυτό θα μπορούσε να αντιμετωπιστεί ως εμπόδιο για την αξιοποίηση των δεδομένων στο πλαίσιο αυτής της εργασίας. Αντίθετα θεωρήσαμε ιδιαίτερη πρόκληση την εκτίμηση του engagement με οποιεσδήποτε συνθήκες καταγραφής του εκάστοτε πειράματος. Για το λόγο αυτό εστιάσαμε στην εξαγωγή των δισδιάστατων σημείων της πόζας και στην εκτίμηση του engagement με τη βοήθεια αυτών.

Net	Accuracy	W. Fscore	Balanced Acc
common class	57.44	41.15	33.33
LSTM(together)	61.78	64.58	54.22
LSTM(best)	**71.76**	**71.55**	**62.11**
2D CNN(together)	**73.70**	**73.96**	**71.72**
2D CNN(best)	70.88	71.32	60.64
AlexNet(together)	73.98	67.05	50.02
AlexNet(best)	**78.73**	**73.39**	**59.10**

Πίνακας 5.10: Αποτελέσματα εκτίμησης για τις διάφορες αρχιτεκτονικές των δικτύων στο σύνολο δεδομένων ASD-Joint Attention έπειτα από κοινού εκπαίδευση ASD και TD δεδομένων.

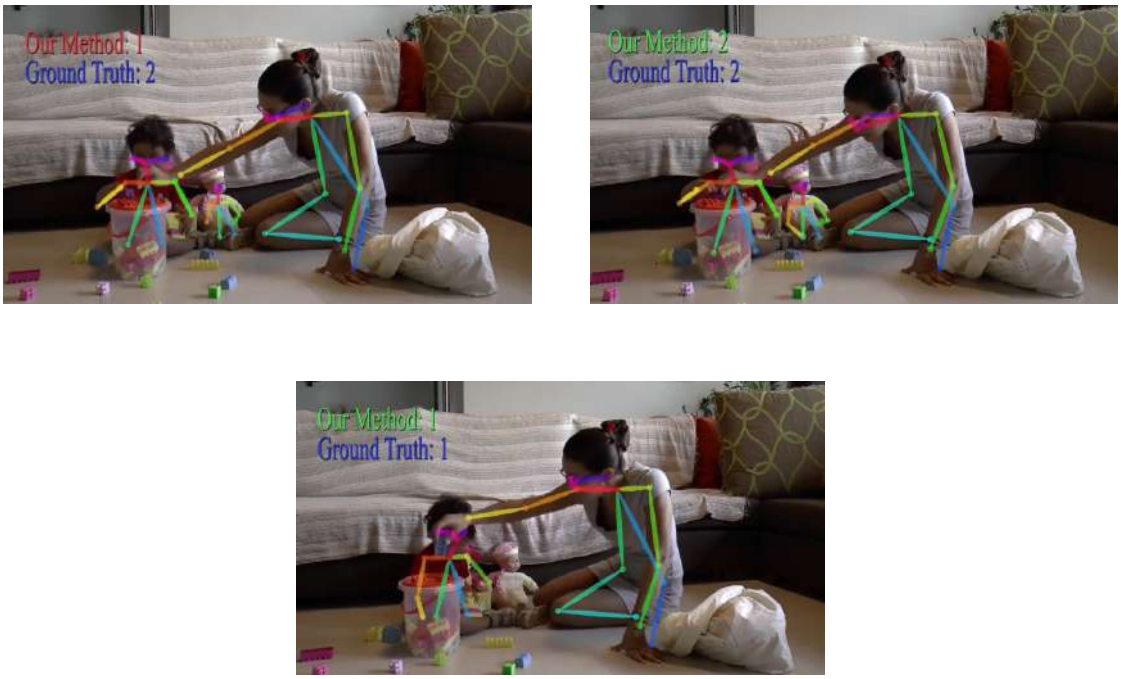
Συνεπώς, εκπαιδεύουμε εκ νέου τα μοντέλα στα TD και ASD δεδομένα Joint Attention, αγνοώντας εντελώς τον άξονα του βάθους ώστε να ελέγξουμε την δυνατότητα εκτίμησης των δικτύων μας και σε τέτοια δεδομένα. Στους πίνακες 5.11 και 5.12 παρουσιάζουμε τις επιδόσεις των δικτύων που βλέπουν μόνο τα 2D δεδομένα σε σύγκριση με τα προηγούμενα δίκτυα που λαμβάνουν ως είσοδο τα 3D δεδομένα. Τα αποτελέσματα είναι σαφώς καλύτερα όταν παρέχουμε στα δίκτυα την 3D πληροφορία. Μάλιστα, το μονοδιάστατο CNN δίκτυο δεν κατάφερνε να εκτιμήσει το engagement βλέποντας μόνο τα 2D δεδομένα.

Παρά τις διαφορές, για τα υπόλοιπα δίκτυα, κρίνουμε τα αποτελέσματα ικανοποιητικά και μπορούμε να συνεχίσουμε με τα 2D δίκτυα για τα BabyAffect δεδομένα. Για παράδειγμα, το δίκτυο AlexNet εξακολουθεί να εκτίμα το engagement στα δεδομένα TD-Joint Attention με accuracy μεγαλύτερο από 80% και w. Fscore που πλησιάζει το 80%. Μάλιστα, το 2D CNN δίκτυο έδωσε καλύτερα αποτελέσματα όταν του δίνονταν τα δεδομένα μόνο με τις διαστάσεις ύψους και πλάτους, χωρίς το βάθος. Για παράδειγμα, στα δεδομένα ASD-Joint Attention το accuracy έφτασε το 75% με πολύ υψηλά w. Fscore και balanced accuracy, τα οποία έχουν τιμές 75.39% και 73.13% αντίστοιχα. Για αυτό μπορεί να οφείλεται το γεγονός ότι το μέγεθος του διανύσματος εισόδου ταίριαξε απόλυτα με τις διαστάσεις των στιβάδων του δικτύου μας, δίνοντας μεγαλύτερες δυνατότητες εξαγωγής συμπερασμάτων στο δίκτυο.

Net	Accuracy	W. Fscore	Balanced Acc
common class	74.32	63.38	33.33
2D CNN(2D)	**80.55**	**77.94**	**60.39**
2D CNN(3D)	78.93	76.46	60.13
LSTM(2D)	75.01	72.25	48.83
LSTM(3D)	**79.47**	**76.88**	**58.82**
AlexNet(2D)	80.42	78.10	57.96
AlexNet(3D)	**80.36**	**80.48**	**69.37**

Πίνακας 5.11: Αποτελέσματα εκτίμησης για τις διάφορες αρχιτεκτονικές των δικτύων στο σύνολο δεδομένων TD-Joint Attention με (3D) και χωρίς (2D) συντεταγμένη βάθους.

Εκπαιδεύοντας, λοιπόν, τα δίκτυα που δίνουν καλές εκτιμήσεις για τα ASD παιδιά του Joint Attention, στα δεδομένα BabyAffect λαμβάνουμε τα παρακάτω αποτελέσματα, τα



Σχήμα 5.13: Εκτίμηση του επιπέδου engagement από το AlexNet δίκτυό μας για τρία στιγμιότυπα του συνόλου δεδομένων BabyAffect. Σε κάθε στιγμιότυπο αναγράφονται το πραγματικό επίπεδο του engagement (ground truth) και η τιμή που εκτιμά το δίκτυό μας (our method). Πάνω αριστερά ground truth:2, our method:1, πάνω δεξιά ground truth:2, our method:2 και κάτω ground truth:1, our method:1.

οποία παρουσιάζουμε στον Πίνακα 5.13. Μπορούμε να παρατηρήσουμε ότι οι τιμές των μετρικών αξιολόγησης εδώ είναι χαμηλότερες τόσο από τις αντίστοιχες των Joint Attention TD όσο και από εκείνες των Joint Attention ASD. Ωστόσο, εξακολουθούμε να λαμβάνουμε για το accuracy τιμή αισθητά καλύτερη από την κατάταξη όλων των στιγμιοτύπων στην συχνότερη κλάση (προσεγγίζει το 80%), ενώ ταυτόχρονα τα W. Fscore και Balanced Accuracy είναι αρκετά υψηλά, γεγονός που μας δείχνει ότι τα δίκτυα αναγνωρίζουν σημαντικό ποσοστό στιγμιότυπων όλων των επιπέδων engagement ακόμη και του λιγότερο συχνού. Επομένως, με δεδομένο ότι τα παιδιά σε αυτό το σύνολο δεδομένων είναι στο περιβάλλον του σπιτιού τους, παίζοντας εντελώς ελεύθερα με τις μητέρες τους και τα τρία από τα τέσσερα είναι παιδιά με αυτιστική διαταραχή, θεωρούμε τα αποτελέσματα αυτά ενθαρρυντικά.

Στο Σχήμα 5.13 παρουσιάζουμε τρία στιγμιότυπα από τα δεδομένα BabyAffect. Στα στιγμιότυπα αναγράφονται και τα επισημειωμένα επίπεδα του engagement αλλά και οι εκτιμήσεις του δικτύου μας. Στα πρώτα δύο στιγμιότυπα το παιδί είναι πλήρως engaged στην αλληλεπίδραση με τη μητέρα του, καθώς προσπαθούν μαζί να τοποθετήσουν σωστά ένα τουβλάκι. Στο ένα από τα δύο το δίκτυο μας εκτιμά σωστά το επίπεδο του engagement, ενώ στο άλλο όχι. Στο τρίτο στιγμιότυπο έχουμε μέσο επίπεδο engagement καθώς παρ΄ ότι το παιδί κρατά και επεξεργάζεται ένα από τα τουβλάκια που του έδωσε η μητέρα του, δεν φαίνεται να την προσέχει ούτε προσπαθεί να το τοποθετήσει στο κουτί όπως εκείνη.

Συγκρίνοντας τα στιγμιότυπα μεταξύ τους βλέπουμε τη δυσκολία της εκτίμησης του engagement του παιδιού σε αυτές τις αλληλεπιδράσεις. Κατ΄ αρχάς, η πόζα εδώ είναι ιδιαίτερα σημαντική καθώς το πρόσωπο του παιδιού δε φαίνεται καλά στα περισσότερα στιγμιότυπα. Δεύτερον, μητέρα και παιδί βρίσκονται πολύ κοντά ο ένας στον άλλο δυσκολεύοντας την εκτίμηση της πόζας. Για παράδειγμα, και στα τρία στιγμιότυπα η μητέρα

Net	Accuracy	W. Fscore	Balanced Acc
common class	57.44	41.15	33.33
2D CNN(2D)	**75.82**	**75.39**	**73.13**
2D CNN(3D)	70.88	71.32	60.64
LSTM(2D)	61.53	60.48	47.70
LSTM(3D)	**71.76**	**71.55**	**62.11**
AlexNet(2D)	67.18	64.26	53.58
AlexNet(3D)	**78.73**	**73.39**	**59.10**

Πίνακας 5.12: Αποτελέσματα εκτίμησης για τις διάφορες αρχιτεκτονικές των δικτύων στο σύνολο δεδομένων ASD-Joint Attention με (3D) και χωρίς (2D) συντεταγμένη βάθους.

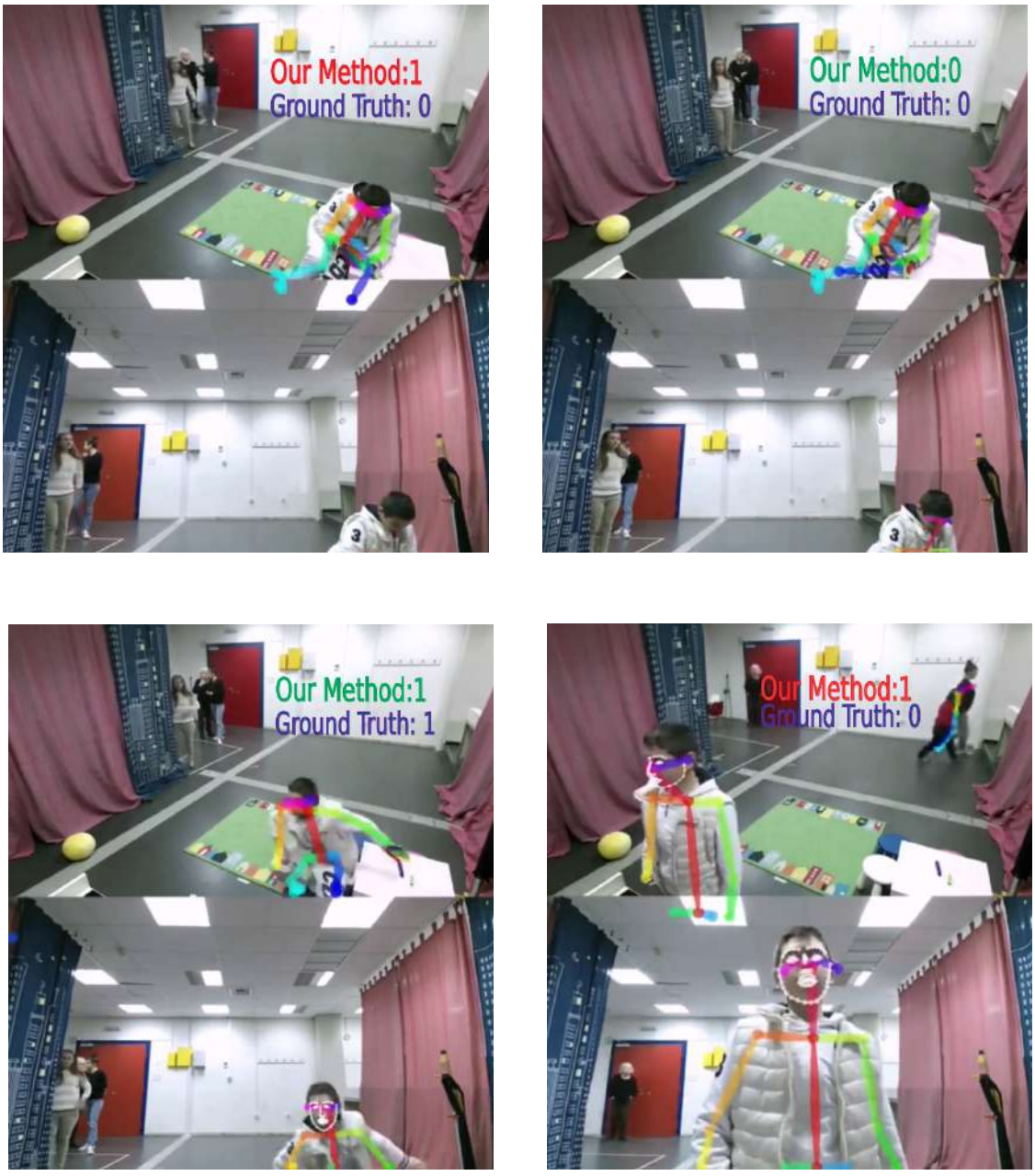
Net	Accuracy	W. Fscore	Balanced Acc	W.Precision
common class	71.48	59.79	33.33	51.10
LSTM	70.55	62.68	37.13	56.20
2D CNN	**77.59**	**71.73**	**53.67**	**74.67**
AlexNet	74.24	69.51	47.27	68.03

Πίνακας 5.13: Αποτελέσματα εκτίμησης για τις διάφορες αρχιτεκτονικές των δικτύων στο σύνολο δεδομένων BabyAffect εκπαιδεύοντας για λίγες εποχές τα προεκπαιδευμένα στα δεδομένα Joint Attention δίκτυα.

κρύβει την εικόνα του παιδιού με το χέρι της, με αποτέλεσμα να είναι δύσκολος ο διαχωρισμός ανάμεσα στα στιγμιότυπα στα οποία το παιδί προσπαθεί πράγματι να τοποθετήσει το τουβλάκι στη σωστή θέση και σε αυτά που απλά παίζει με το τουβλάκι. Η ίδια η αλληλεπίδραση δεν είναι δομημένη όπως στα υπόλοιπα σύνολα δεδομένων, αλλά το παιχνίδι είναι ελεύθερο, γεγονός που καθιστά ακόμη δυσκολότερη την εκτίμηση για το αν το παιδί συμβάλλει κάθε στιγμή σε έναν κοινό στόχο του παιχνιδιού.

5.7 Επέκταση στα δεδομένα ASD-Games

Στη συνέχεια προχωρήσαμε τα πειράματα με τα δεδομένα ASD-Games, δηλαδή τα δεδομένα από τα υπόλοιπα παιχνίδια των 7 ASD παιδιών του ASD-Joint Attention. Για τα δεδομένα αυτά εμπλουτίσαμε το διάνυσμα χαρακτηριστικών, όπως αναφέρουμε και στο Κεφάλαιο 3, με τα σημεία του προσώπου που παράγει το OpenPose. Παρακάτω, στον Πίνακα 5.14 παρουσιάζουμε τα αποτελέσματα της εκπαίδευσής μας. Παρ΄ όλο που η εκτίμηση και εδώ είναι λιγότερο ακριβής από τα Joint Attention δεδομένα τα αποτελέσματα εξακολουθούν να είναι αρκετά καλά. Η αιτία για τη διαφοροποίηση αυτή βρίσκεται στο γεγονός ότι εδώ η εκτίμηση γίνεται για πολύ πιο πολύπλοκες συνθήκες σε σχέση με την απλούστερη συνθήκη στην οποία το ρομπότ ζητά από το παιδί να του δώσει ένα παιχνίδι (Joint Attention). Σε αυτή την περίπτωση καλούμαστε να εκτιμήσουμε το engagement σε ένα σύνολο δεδομένων που περιλαμβάνει τέσσερα διαφορετικά παιχνίδια με διαφορετικές κινήσεις στο χώρο και εντελώς διαφορετικούς τρόπους με τους οποίους πρέπει να αντιδράσουν τα παιδιά στο καθένα. Για παράδειγμα, στο παιχνίδι Pantomime τα παιδιά πρέπει να κινούνται μιμούμενα την κίνηση ενώ κοιτούν προς τα κάτω που βρίσκεται το ρομπότ Nao και λοξά από τις κάμερες. Αντίθετα, στο παιχνίδι Express the Feeling πρέπει να δείξουν το συναίσθημα με το πρόσωπό τους και με το σώμα τους συγχρόνως κοι-



Σχήμα 5.14: Εκτίμηση του επιπέδου engagement από το AlexNet δίκτυό μας για τέσσερα στιγμιότυπα του παιχνιδιού Pantomime.Σε κάθε στιγμιότυπο αναγράφονται το πραγματικό επίπεδο του engagement (ground truth) και η τιμή που εκτιμά το δίκτυό μας (our method). Πάνω αριστερά ground truth:0, our method:1, πάνω δεξιά ground truth:0, our method:0, κάτω αριστερά ground truth:1, our method:1 και κάτω δεξιά ground truth:0, our method:1.

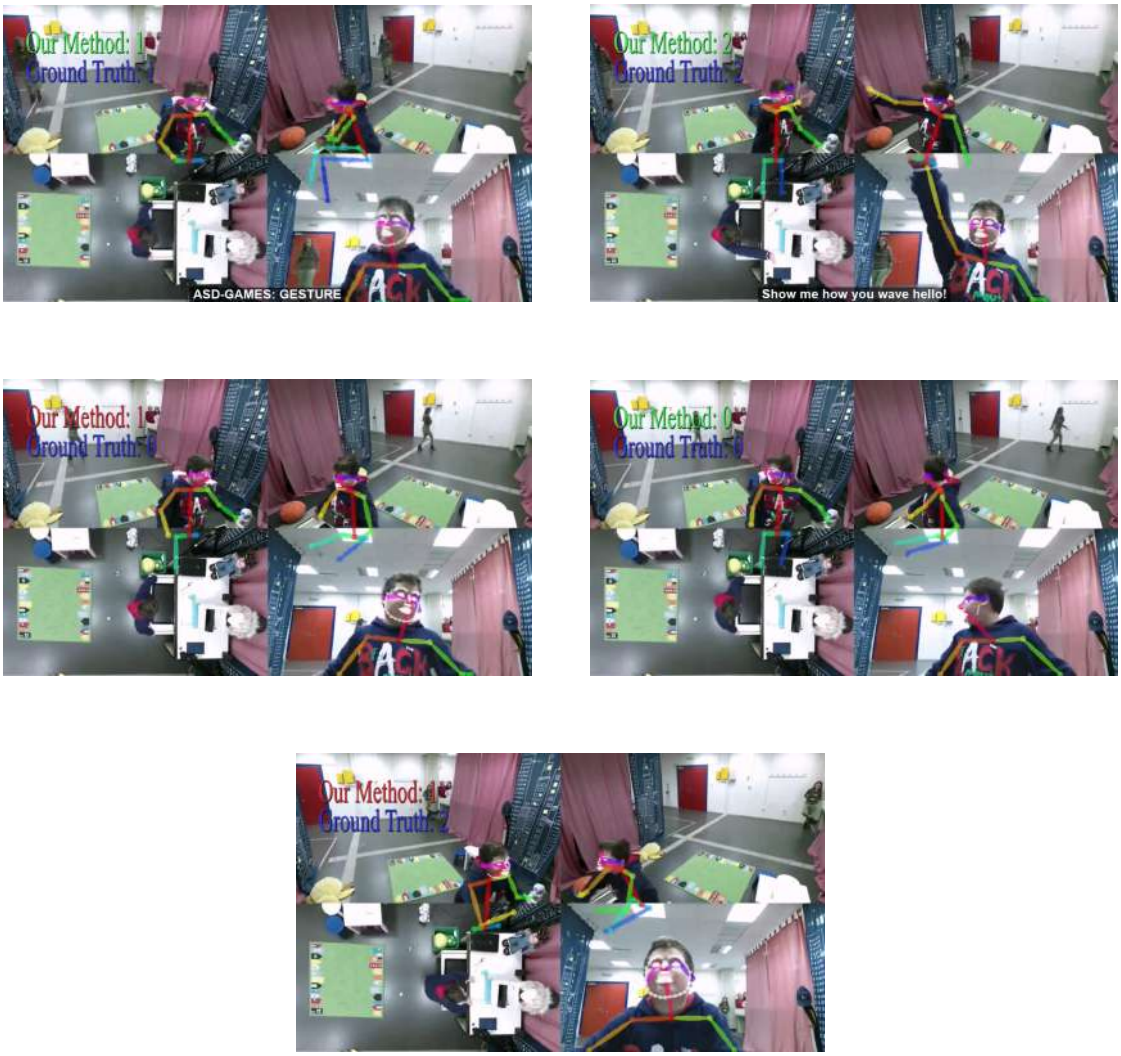
τώντας απευθείας τις κάμερες προς το ρομπότ Furhat. Στο παιχνίδι Guess the Object τα παιδιά πρέπει να κοιτάξουν και να ψάξουν όλο το χώρο ώστε να βρουν το αντικείμενο που τους περιγράφει το ρομπότ. Στην εκτίμηση του engagement σε αυτά τα δεδομένα το συνελικτικό δίκτυο AlexNet πετυχαίνει accuracy και w. Fscore που προσεγγίζουν το 70% (68.34% και 67.57% αντίστοιχα). Τα αποτελέσματα αυτά δείχνουν ότι η μέθοδός μας μπορεί με επιτυχία να χρησιμοποιηθεί για την εκτίμηση του engagement ASD παιδιών για ποικίλες διαφορετικές αλληλεπιδράσεις κατά τη διάρκεια των οποίων τα παιδιά καλούνται να συζητήσουν, να γνέψουν, να κινηθούν στο δωμάτιο ή να αλληλεπιδράσουν με την οθόνη.

Net	Accuracy	W. Fscore	Balanced Acc	W. Precision
common class	62.28	47.81	33.33	38.80
LSTM	64.26	55.55	39.60	62.80
2D-CNN	67.28	56.09	40.92	51.65
AlexNet	**68.34**	**67.57**	**52.38**	**65.07**

Πίνακας 5.14: Αποτελέσματα εκτίμησης για τις διάφορες αρχιτεκτονικές των δικτύων στο σύνολο δεδομένων ASD-Games.

Στο Σχήμα 5.14 παρουσιάζουμε τέσσερα στιγμιότυπα από μία αλληλεπίδραση του συνόλου ASD-Games καθώς και τα επισημειωμένα αλλά και τα εκτιμώμενα επίπεδα του engagement για τα στιγμιότυπα αυτά. Πρόκειται για το παιχνίδι Pantomime στο οποίο συμμετέχει το ρομπότ Nao. Το παιχνίδι λαμβάνει χώρα και με τη βοήθεια της οθόνης αφής από την οποία το παιδί διαλέγει κάρτες. Στα δύο πρώτα από τα στιγμιότυπα του Σχήματος το παιδί δεν ανταποκρίνεται στο κάλεσμα του Nao να διαλέξει μία κάρτα από την οθόνη, αλλά ούτε κοιτά ή φαίνεται να προσέχει το ρομπότ με κάποιο άλλο τρόπο. Συνεπώς, το παιδί είναι πλήρως disengaged (επίπεδο engagement 0). Σε ένα από τα δύο στιγμιότυπα το δίκτυό μας αποτυγχάνει να εκτιμήσει σωστά το engagement του παιδιού καθώς κατατάσσει το στιγμιότυπο στο επίπεδο 1. Αξίζει να παρατηρήσουμε για τη δυσκολία της εκτίμησης ότι στην κάτω όψη του στιγμιότυπου δεν έχουμε εκτίμηση για το σκελετό του παιδιού. Στο τρίτο στιγμιότυπο το παιδί σηκώνεται για να πλησιάσει την οθόνη και επομένως ξεκινά να συμμετέχει στη διαδικασία. Εδώ το επίπεδο του engagement είναι 1 και το δίκτυό μας το εκτιμά σωστά. Συμπεριλαμβάνουμε και το τελευταίο στιγμιότυπο καθώς επίσης παρουσιάζει δυσκολία. Εδώ το engagement έχει επισημειωθεί ως 0, καθώς παρ΄ ότι το παιδί βρίσκεται μπροστά από την οθόνη, εκείνη τη στιγμή το Nao του έχει ζητήσει να το κοιτάξει ώστε να του δείξει πως χορεύει αλλά το παιδί δεν ανταποκρίνεται. Το δίκτυό μας έχει εκτιμήσει το engagement ως μέσου επιπέδου.

Στο Σχήμα 5.15 παρουσιάζουμε άλλα πέντε στιγμιότυπα με τα αντίστοιχα επίπεδα engagement από το παιχνίδι Show me the Gesture στο οποίο το ρομπότ Furhat ζητά από το παιδί να του δείχνει ορισμένες χειρονομίες. Στο πρώτο στιγμιότυπο η εκτίμηση είναι απλή καθώς το παιδί κοιτάζει και παρακολουθεί το ρομπότ που του μιλά και έτσι τόσο το επισημειωμένο όσο και το εκτιμώμενο επίπεδο engagement είναι 1. Στο δεύτερο στιγμιότυπο το ρομπότ έχει ζητήσει από το παιδί να του δείξει πως χαιρετάει και το παιδί πράγματι κοιτώντας το ρομπότ χαιρετά. Τη στιγμή αυτή είναι πλήρως engaged και το δίκτυό μας εκτιμά σωστά επίπεδο engagement 2. Στο τρίτο στιγμιότυπο το ρομπότ έχει ζητήσει από το παιδί να το καλέσει κοντά του και περιμένει. Το παιδί κοιτάζει αμήχανα δεξιά και αριστερά και επομένως είναι πλήρως disengaged. Εδώ το επίπεδο του engagement ανιχνεύεται λανθασμένα ως μέσο. Στο επόμενο στιγμιότυπο, στο οποίο το παιδί εξακολουθεί να είναι disengaged και μάλιστα γυρνά το κεφάλι του και κοιτάζει προς



Σχήμα 5.15: Εκτίμηση του επιπέδου engagement από το AlexNet δίκτυό μας για πέντε στιγμιότυπα του παιχνιδιού Show me the Gesture. Σε κάθε στιγμιότυπο αναγράφονται το πραγματικό επίπεδο του engagement (ground truth) και η τιμή που εκτιμά το δίκτυό μας (our method). Πάνω αριστερά ground truth:1, our method:1, πάνω δεξιά ground truth:2, our method:2, στη μέση αριστερά ground truth:0, our method:1, στη μέση δεξιά ground truth:0, our method:0 και κάτω ground truth:2, our method:1.



Σχήμα 5.16: Τρία στιγμιότυπα από το σύνολο δεδομένων PInSoRo [14]. Τα παιδιά βρίσκονται καθισμένα καθόλη τη διάρκεια των αλληλεπιδράσεων γεγονός που καθιστά δυσκολότερη την εκτίμηση του επιπέδου του engagement με τη χρήση της πόζας.

τα πίσω, πιθανώς αναζητώντας κάποιον ενήλικα, το δίκτυο εκτιμά σωστά το επίπεδο του engagement ως χαμηλό (1). Στο τελευταίο στιγμιότυπο υπάρχει μία ακόμη δυσκολία. Το ρομπότ έχει ζητήσει από το παιδί να του δείξει το τραπέζι και το παιδί ανταποκρίνεται δείχνοντας το τραπέζι μπροστά του και κοιτάζοντας ταυτόχρονα το ρομπότ. Είναι δηλαδή πλήρως engaged. Προφανώς, η κίνηση αυτή (το άγγιγμα δηλαδή του τραπεζιού) δεν εκτιμάται σωστά από το δίκτυό μας και γι΄ αυτό εκτιμά το engagement ως επιπέδου 1 αντί ως επιπέδου 2.

5.8 Επέκταση στο σύνολο δεδομένων PInSoRo

Παράλληλα, πραγματοποιήσαμε πειράματα και για τα δεδομένα της PInSoRo. Στον παρακάτω Πίνακα 5.15 παρουσιάζουμε και για τα δεδομένα αυτά τις επιδόσεις διαφορετικών LSTM δικτύων. Η αρχιτεκτονική του δικτύου που επελέγη και παραπάνω φαίνεται να εξυπηρετεί καλύτερα σε σχέση με άλλες αρχιτεκτονικές την εκτίμηση του engagement και στην περίπτωση των δεδομένων της PInSoRo, παρ΄ όλο που εδώ τα αποτελέσματα δεν είναι τόσο ικανοποιητικά.

Στη συνέχεια αναζητούμε καλύτερες τιμές των παραμέτρων του δικτύου και για τη βάση δεδομένων της PInSoRo, καθώς τα αποτελέσματα από το δίκτυο στο οποίο έχουμε καταλήξει για τα BabyRobot και BabyAffect δεδομένα δεν είναι αρκετά ικανοποιητικά. Έτσι, καταλήγουμε το δίκτυο να λαμβάνει τα δεδομένα σε ομάδες των 128 (batch_size = 128), ενώ το learning rate του δικτύου ορίζεται ίσο με 10^{-5}. Επίσης, θέτουμε το μέγεθος των στιβάδων του δικτύου $C = 100$ για τις δύο τελευταίες και $2C = 200$ για τις υπόλοιπες. Παρακάτω στον Πίνακα 5.16 παρουσιάζουμε ορισμένα ενδεικτικά αποτελέσματα.

Net	Accuracy	W. Fscore	Balanced Acc
common class	53.55	37.35	25.00
3 FC, 1 LSTM, FC	**54.92**	**53.42**	**38.96**
3FC, 2 LSTM (2C, C), (C, C), FC	51.82	51.79	37.61
3FC, 2 LSTM (2C, 2C), (2C, C), FC	**54.18**	**53.27**	**38.25**
2 FC, 1 LSTM, FC	54.01	51.63	38.10
2 FC, 2 LSTM (2C, 2C), (2C,C), FC	52.73	51.86	36.33
3 FC, 3 LSTM (2C, 2C), (2C,C),(C,C), FC	53.00	52.23	36.96

Πίνακας 5.15: Αναζήτηση βέλτιστης αρχιτεκτονικής αναδρομικού δικτύου. Αποτελέσματα εκτίμησης στο σύνολο δεδομένων PInSoRo.

C	batch_size	learning_rate	Accuracy	W. Fscore	Balanced Acc
common class			53.55	37.35	25.00
100	128	10^{-4}	50.48	47.25	34.30
50	128	10^{-5}	56.28	45.99	37.49
200	128	10^{-5}	55.20	53.64	37.43
100	64	10^{-5}	52.54	48.23	31.27
100	128	10^{-5}	**57.07**	**55.82**	**40.01**

Πίνακας 5.16: Ενδεικτικά αποτελέσματα αναζήτησης βέλτιστων παραμέτρων για το LSTM δίκτυο εκτίμησης στο σύνολο δεδομένων PInSoRo.

Οι τιμές των μετρικών αξιολόγησης των δικτύων μας στην PInSoRo είναι αρκετά χαμηλές. Βρισκόμαστε εδώ αντιμέτωποι με αλληλεπιδράσεις που πραγματοποιούνται κατά κύριο λόγο γύρω από ένα τραπέζι και συνεπώς η κίνηση των παιδιών είναι πολύ περιορισμένη σε σχέση με τα προηγούμενα σύνολα δεδομένων. Αυτό φαίνεται και στο Σχήμα 5.16 στο οποίο παρουσιάζουμε ορισμένα ενδεικτικά στιγμιότυπα των δεδομένων PInSoRo από τον ιστότοπο της βάσης [14]. Η ποικιλία της πόζας στις αλληλεπιδράσεις αυτές είναι συντριπτικά μικρότερη. Αυτοί μπορεί να είναι κάποιοι από τους λόγους για τους οποίους οι μέθοδοί μας δεν εκτιμούν σε αυτό το σύνολο δεδομένων με την ίδια επιτυχία όπως προηγουμένως το engagement των παιδιών που συμμετέχουν. Το ίδιο πρόβλημα αντιμετωπίζουμε και στο σύνολο δεδομένων ASD-School οπότε θα επανέλθουμε σε αυτό.

5.9 Επέκταση στα δεδομένα ASD-School

Τέλος, πραγματοποιούμε αντίστοιχα πειράματα και για τα δεδομένα από το σχολείο του Πειραιά (ASD-School). Οι καταγραφές εδώ μοιάζουν πολύ με τα δεδομένα της PInSoRo, καθώς και εδώ η δράση λαμβάνει χώρα μπροστά από μία οθόνη και τα παιδιά βρίσκονται καθισμένα μπροστά σε ένα τραπέζι. Εδώ, η πόζα από μόνη της δε θα μας δώσει αρκετή πληροφορία για το engagement των παιδιών και συνεπώς εξάγουμε και χρησιμοποιούμε και τα σημεία του προσώπου και των χεριών. Και εδώ, όπως φαίνεται στον Πίνακα 5.17, προς το παρόν τα αποτελέσματα που πετυχαίνουμε δεν είναι ικανοποιητικά. Το αναδρομικό δίκτυο και το συνελικτικό δίκτυο AlexNet δεν καταφέρνουν να γενικεύσουν καθώς εκπαιδεύονται με τα δεδομένα ASD-School. Το συνελικτικό δίκτυο 2DCNN παρ΄ όλο που έχει accuracy μικρότερο από το accuracy που πετυχαίνουμε όταν κατατάσσουμε όλα τα στιγμιότυπα στη συνηθέστερη κλάση (60.50% έναντι 65.19%) γενικεύει αποτελεσματικότερα

καθώς καταφέρνει να εκτιμήσει σωστά το engagement και σε ορισμένα στιγμιότυπα που ανήκουν στις άλλες κλάσεις, πετυχαίνοντας W. Fscore 58.87%, balanced accuracy 41.11% και w. Precision 57.82%.

Το γεγονός ότι τα παιδιά βρίσκονται καθισμένα καθόλη τη διάρκεια της αλληλεπίδρασης δυσχεραίνει τη δυνατότητα των δικτύων μας να γενικεύσουν σωστά αξιοποιώντας τα δεδομένα της πόζας. Πιθανώς, να χρειάζονται σε αυτά τα δεδομένα επιπλέον χαρακτηριστικά για την εκτίμηση του engagement όπως η φωνή παιδιού και ρομπότ ή ο χρόνος αντίδρασης του παιδιού.

Net	Accuracy	W. Fscore	Balanced Acc	W. Precision
common class	65.19	51.45	33.33	42.50
LSTM	63.27	51.58	33.33	43.28
2D-CNN	**60.59**	**58.87**	**41.11**	**57.82**
AlexNet	64.84	52.86	33.78	44.27

Πίνακας 5.17: Αποτελέσματα εκτίμησης για τις διάφορες αρχιτεκτονικές των δικτύων στο σύνολο δεδομένων ASD-School.

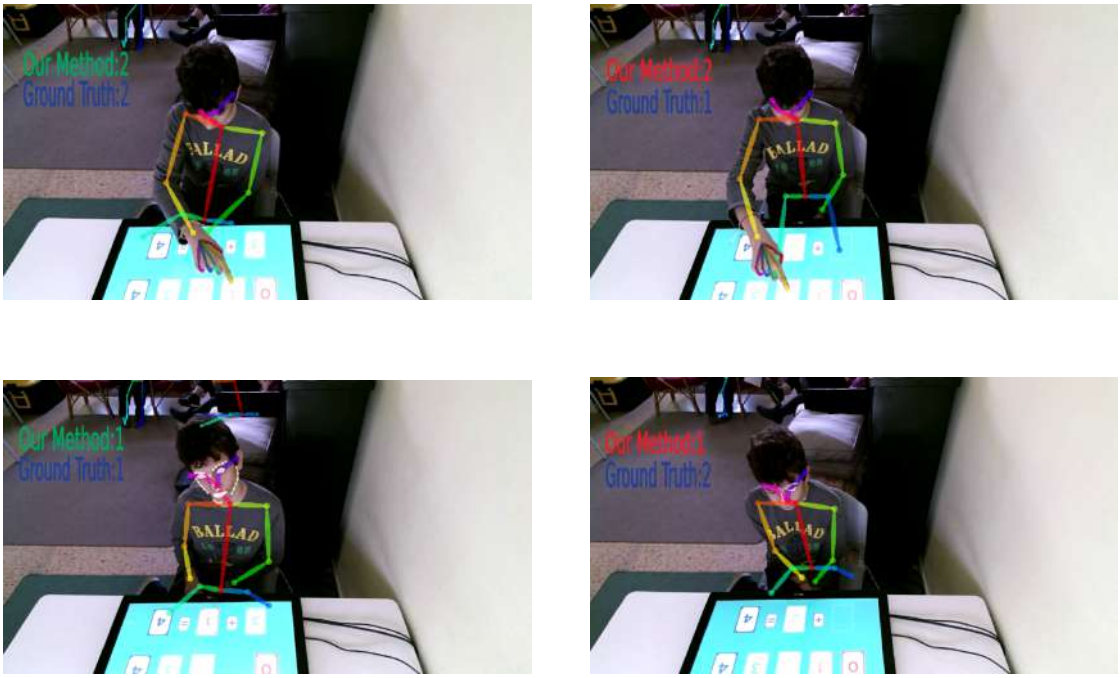
Στο Σχήμα 5.17 περιλαμβάνουμε τέσσερα στιγμιότυπα από τα δεδομένα ASD-School. Τα πρώτα δύο στιγμιότυπα φαίνονται πανομοιότυπα, ωστόσο έχουμε κατατάξει το πρώτο στην κλάση 2, ενώ το δεύτερο στην κλάση 1. Αυτό συμβαίνει γιατί στο πρώτο στιγμιότυπο το ρομπότ έχει ζητήσει από το παιδί να το βοηθήσει με την πράξη, ενώ στο δεύτερο στιγμιότυπο έχει πει στο παιδί ότι τώρα θα δοκιμάσει μόνο του να βρει το σωστό αριθμό. Τα δύο αυτά στιγμιότυπα αναδεικνύουν τη δυσκολία του προβλήματος αλλά και τη σημασία της χρονικής συνέχειας για το engagement. Το 2DCNN δίκτυο έχει κατατάξει και τα δύο αυτά στιγμιότυπα στην κλάση 2.

Τα επόμενα δύο στιγμιότυπα επίσης μοιάζουν πολύ μεταξύ τους, με τις πόζες του παιδιού να είναι ίδιες. Στο πρώτο το επίπεδο του engagement έχει επισημειωθεί ως 1, καθώς το ρομπότ έχει ζητήσει από το παιδί να το παρηγορήσει για το λάθος του αλλά το παιδί δεν ανταποκρίνεται. Στο δεύτερο στιγμιότυπο το επίπεδο του engagement έχει επισημειωθεί ως 2, καθώς εδώ το παιδί πράγματι παρηγορεί το ρομπότ. Το 2DCNN δίκτυο έχει κατατάξει και τα δύο αυτά στιγμιότυπα στην κλάση 1. Από τα τέσσερα στιγμιότυπα του Σχήματος αναδεικνύεται η επιπλέον δυσκολία που παρουσιάζουν τα δεδομένα ASD-School για το σωστό εντοπισμό του engagement.

Από τις προσπάθειες εκτίμησης του επιπέδου του engagement στα δύο τελευταία σύνολα δεδομένων PInSoRo και ASD-School συμπεραίνουμε πως σε καταστάσεις και συνθήκες που το παιδί είναι στατικό και κάθεται συνεχώς μπροστά στο ρομπότ η εκτίμηση του engagement δε μπορεί να βασιστεί αποκλειστικά στην πόζα. Συνεπώς, σε μια περαιτέρω μελέτη θα πρέπει να δοθεί έμφαση σε άλλα χαρακτηριστικά όπως για παράδειγμα τα RGB δεδομένα του προσώπου των παιδιών ή όπως αναφέραμε παραπάνω η ομιλία και ο χρόνος αντίδρασης των παιδιών.

5.10 Εκτίμηση με τη χρήση RGB δεδομένων

Τέλος, έχει ιδιαίτερη χρησιμότητα να συγκρίνουμε τα παραπάνω αποτελέσματα με τη χρήση της πόζας με τα αντίστοιχα αποτελέσματα που μπορούμε να λάβουμε αξιοποιώντας απευθείας τα RGB δεδομένα. Πραγματοποιούμε τη σύγκριση για τα TD-Joint Attention δεδομένα του BabyRobot.



Σχήμα 5.17: Εκτίμηση του επιπέδου engagement από το 2DCNN δίκτυό μας για τέσσερα στιγμιότυπα ASD-School. Σε κάθε στιγμιότυπο αναγράφονται το πραγματικό επίπεδο του engagement (ground truth) και η τιμή που εκτιμά το δίκτυό μας (our method). Πάνω αριστερά ground truth:2, our method:2, πάνω δεξιά ground truth:1, our method:2, κάτω αριστερά ground truth:1, our method:1 και κάτω δεξιά ground truth:2, our method:1 .

Χωρίζουμε σε frames τα TD Joint Attention videos με ρυθμό 30fps. Στη συνέχεια δειγματοληπτούμε τα δεδομένα κρατώντας ένα ανά πέντε frames.

Πραγματοποιήσαμε πειράματα με ορισμένα προεκπαιδευμένα δίκτυα και τα βάρη τους, που είναι αρκετά διαδεδομένα και έχουν καλά αποτελέσματα στο σύνολο δεδομένων ImageNet. Τα δίκτυα αυτά είναι το AlexNet, παρόμοια δομή με του οποίου έχουμε χρησιμοποιήσει και προηγουμένως για τα δεδομένα πόζας, το ResNet-18 [15] καθώς και το ResNet-50 [15]. Χρησιμοποιούμε όλα τα μοντέλα προεκπαιδευμένα στο ImageNet, όπως είναι υλοποιημένα στη βιβλιοθήκη torchvision. Τα μοντέλα ResNet προτάθηκαν στο Deep Residual Learning for Image Recognition. Ο αριθμός που τα προσδιορίζει αντιστοιχεί στο πλήθος των στιβάδων του δικτύου. Τα ResNet-18 και ResNet-50 που αξιοποιούμε εδώ έχουν top-5-error στο σύνολο δεδομένων ImageNet 10.92 και 7.13 αντίστοιχα. Τα δίκτυα αυτά παρουσιάστηκαν το 2015 και η δομή τους φαίνεται αναλυτικά στο Σχήμα 5.18. Πρόκειται για βαθιά συνελικτικά δίκτυα με μεγάλο αριθμό συνεχόμενων συνελικτικών στιβάδων, που σήμερα χρησιμοποιούνται ευρέως στην αναγνώριση εικόνας.

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Σχήμα 5.18: Η δομή των βαθιών νευρωνικών δικτύων ResNet[15].

Net	Accuracy	W. Fscore	Balanced Acc	W. Precision
common class	74.32	63.38	33.33	55.25
LSTM(pose)	79.47	76.88	58.82	78.04
AlexNet(pose)	**80.36**	**80.48**	**69.37**	**80.71**
AlexNet(RGB)	73.20	65.85	37.18	55.32
ResNet-18(RGB)	74.83	65.99	36.32	57.21
ResNet-50(RGB)	**74.52**	**64.52**	**45.49**	**63.28**

Πίνακας 5.18: Αποτελέσματα εκτίμησης με τη χρήση πόζας ή απευθείας RGB δεδομένων για τις διάφορες αρχιτεκτονικές των δικτύων στο σύνολο δεδομένων TD-Joint Attention.

Στον Πίνακα 5.18 παρουσιάζουμε τα αποτελέσματα της εκτίμησης του engagement αποκλειστικά από τα RGB δεδομένα, συγκριτικά και με τα αποτελέσματα που έχουμε πετύχει χρησιμοποιώντας τα δεδομένα της πόζας. Τα RGB δίκτυα απαιτούν περισσότερες από 250 εποχές για να συγκλίνουν ή και δεν καταφέρνουν να συγκλίνουν. Ακόμη, παρατηρούμε ότι η ακρίβεια (accuracy) που πετυχαίνουμε αξιοποιώντας τα RGB δεδομένα

αυτούσια είναι περίπου ίση (ελάχιστα ξεπερνάει) την ακρίβεια της επιλογής της συνηθέστερης κλάσης. Μόνο στο ResNet-50 η ακρίβεια αυτή συνοδεύεται και από βελτιωμένο balanced accuracy, δηλαδή μόνο σε αυτή την περίπτωση το δίκτυό μας πραγματικά εκπαιδεύεται ώστε να αναγνωρίζει τις λιγότερο συνήθεις κλάσεις. Τα αποτελέσματα αυτά αναδεικνύουν τα σημαντικά πλεονεκτήματα που μας προσφέρει η πόζα για την εκτίμηση του engagement.

Συνεπώς, για να αποκτήσουμε ικανοποιητικότερα αποτελέσματα εκτίμησης απευθείας από τα RGB δεδομένα χρειάζεται μελέτη περισσότερων αρχιτεκτονικών, περισσότερα δεδομένα, καθώς και περισσότερος χρόνος εκπαίδευσης. Για παράδειγμα, θα μπορούσαμε να αξιοποιήσουμε και 3D αρχιτεκτονικές που έχουν επιτύχει πολύ σημαντική πρόοδο στην αναγνώριση ανθρώπινης δράσης όπως οι αρχιτεκτονικές ResNet 3D και (2+1)D [61] ή η αρχιτεκτονική I3D [62]. Ακόμη, θα μπορούσαμε να εστιάσουμε μόνο στα RGB στοιχεία της περιοχής του προσώπου των παιδιών και να συνδυάσουμε τα χαρακτηριστικά αυτά με την πόζα.

Κεφάλαιο 6

Συμπεράσματα και Μελλοντικές Επεκτάσεις

6.1 Ορισμένα Συμπεράσματα

Στην ενότητα αυτή παρουσιάζουμε ορισμένα συμπεράσματα που προκύπτουν από τα παραπάνω πειράματα σε σχέση με την εκτίμηση του engagement. Τα σημαντικότερα αποτελέσματα μαζί με τα συμπεράσματα αυτά έχουμε υποβάλλει προς δημοσίευση στα πλαίσια του ICRA 2020 [63]. Στην εξαγωγή των συμπερασμάτων με βάση τα αποτελέσματά μας μας βοήθησε και η καθηγήτρια ψυχολόγος κα. Χριστίνα Παπαηλιού.

Η μέθοδος που σχεδιάσαμε και υλοποιήσαμε για την εκτίμηση του engagement με τη βοήθεια της πόζας και δισδιάστατων συνελικτικών δικτύων είναι επιτυχημένη και οδηγεί σε ακριβέστερα αποτελέσματα σε σχέση με προηγούμενες μεθόδους που συναντάμε στη βιβλιογραφία. Τελικός στόχος πρέπει να είναι η δυνατότητα εκτίμησης του engagement από ένα ενοποιημένο μοντέλο με μικρές και σύντομες προσαρμογές σε τελείως διαφορετικά περιβάλλοντα (με διαφορετικά παιδιά -τυπικώς αναπτυσσόμενα ή με διαταραχές αυτιστικού φάσματος, μεγαλύτερα ή μικρότερα σε ηλικία-, αλληλεπιδράσεις με ρομπότ, ενήλικες ή άλλα παιδιά, διαφορετικές συνθήκες αλληλεπίδρασης -ελεύθερο παιχνίδι ή πιο καθορισμένες αλληλεπιδράσεις που ρυθμίζονται από κανόνες-, σε εργαστηριακό, σχολικό ή οικιακό περιβάλλον κ.α.).

Οι διαφοροποιήσεις στα ποσοστά επιτυχίας εκτίμησης του επιπέδου του engagement μετά από τα πειράματά μας για τα διαφορετικά σύνολα δεδομένων είναι οι αναμενόμενες. Τα δίκτυά μας επιτυγχάνουν τα μεγαλύτερα ποσοστά επιτυχίας όταν εκτιμούν το engagement στα δεδομένα TD-Joint Attention του BabyRobot, δηλαδή σε τυπικώς αναπτυσσόμενα παιδιά που συμμετέχουν σε μία από κοινού με το Nao δραστηριότητα σε περιβάλλον εργαστηρίου και κινούνται ελεύθερα με αποτέλεσμα να υπάρχει δυνατότητα έκφρασης της συναισθηματικής τους κατάστασης και άρα και του engagement μέσα από τη στάση του σώματός τους. Για την εκτίμηση σε αυτά τα δεδομένα τόσο το accuracy όσο και τα w. Fscore και w. precision ξεπερνούν το 80%, ενώ και το balanced accuracy φτάνει το 70%. Η εκτίμησή μας σε αυτές τις αλληλεπιδράσεις είναι ιδιαίτερα ακριβής.

Ακολουθούν τα δίκτυα που εκπαιδεύσαμε στα δεδομένα ASD-Joint Attention του BabyRobot, με accuracy που προσεγγίζει το 80% και w. precision και w. Fscore που ξεπερνούν το 80% και 70% αντίστοιχα (82.38% και 73.39%). Εδώ έχουμε τις ίδιες συνθήκες με προηγουμένως καθώς τα παιδιά συμμετέχουν σε μία καθορισμένη από κοινού δραστηριότητα με το ρομπότ κινούμενα ελεύθερα σε περιβάλλον εργαστηρίου. Ωστόσο, τα παιδιά που συμμετέχουν είναι παιδιά με διαταραχές αυτιστικού φάσματος. Το γεγονός

αυτό, για πολλούς λόγους που περιγράψαμε παραπάνω καθιστά δυσκολότερο τον ορθό εντοπισμό του επιπέδου engagement. Για το λόγο αυτό είναι ιδιαίτερα ικανοποιητικό το γεγονός ότι τα αποτελέσματα της εκτίμησης στα δεδομένα αυτά προσεγγίζουν σε ακρίβεια τα αποτελέσματα της εκτίμησης στα τυπικώς αναπτυσσόμενα παιδιά.

Στη συνέχεια, ακολουθούν τα δίκτυα για την εκτίμηση του engagement στα δεδομένα ASD-Games του BabyRobot και στα δεδομένα BabyAffect για τα οποία έχουμε λίγο μικρότερα ποσοστά επιτυχίας από εκείνα που έχουμε στις δύο προηγούμενες ομάδες δεδομένων. Όσον αφορά στα δεδομένα ASD-Games εδώ η δυσκολία βρίσκεται στο γεγονός ότι οι στόχοι και τα παιχνίδια παρουσιάζουν σημαντική ποικιλομορφία σε σχέση με την απλή συνθήκη Joint Attention. Έτσι, accuracy και w. Fscore προσεγγίζουν το 70% (68.34% και 67.57% αντίστοιχα). Όσον αφορά στις αλληλεπιδράσεις στα δεδομένα BabyAffect σε αυτές συμμετέχουν παιδιά με διαταραχές αυτιστικού φάσματος μικρότερων ηλικιών, που δε βρίσκονται πλέον στο καθορισμένο περιβάλλον του εργαστηρίου αλλά στο οικιακό τους περιβάλλον και αλληλεπιδρούν απευθείας με τις μητέρες τους. Οι παράγοντες αυτοί καθιστούν δυσκολότερη την εκτίμηση του επιπέδου του engagement. Εδώ, το accuracy είναι 77.59% με w. Fscore 71.73%. Επιπλέον, και στις δύο αυτές κατηγορίες αλληλεπιδράσεων δεν ήταν διαθέσιμα δεδομένα βάθους στις καταγραφές με αποτέλεσμα να δουλεύουμε με δισδιάστατα και όχι τρισδιάστατα σημεία σκελετού. Με δεδομένες όλες τις αναφερθείσες επιπρόσθετες δυσκολίες κρίνουμε τα αποτελέσματα της εκτίμησης και στα δύο αυτά σύνολα δεδομένων ως επιτυχημένα.

Τέλος, τα ποσοστά επιτυχίας στο σύνολο δεδομένων Pinsoro και στα δεδομένα του σχολείου στον Πειραιά δεν είναι όσο ικανοποιητικά θα θέλαμε. Τα δεδομένα αυτά παρουσιάζουν μία σημαντική διαφοροποίηση σε σχέση με όλες τις προηγούμενες ομάδες δεδομένων. Τα παιδιά που συμμετέχουν στα παιχνίδια-αλληλεπιδράσεις, βρίσκονται καθισμένα μπροστά από ένα τραπέζι με μία οθόνη αφής. Το γεγονός αυτό περιορίζει πολύ σημαντικά τις κινήσεις τους και κατά συνέπεια και την ποικιλία στα δεδομένα πόζας που επεξεργαζόμαστε. Οι πολύ πιο περιορισμένες πόζες των παιδιών αυτών είναι δυνατό να δημιουργούν αυτή τη διαφοροποίηση στα αποτελέσματα που λαμβάνουμε σε σχέση με τις άλλες ομάδες δεδομένων.

Από το σύνολο των πειραμάτων μας, εξάγουμε τα παρακάτω συμπεράσματα:

- Τα δεδομένα της πόζας επιτρέπουν και μπορούν να βοηθήσουν σημαντικά την εκτίμηση χαρακτηριστικών, όπως το engagement, που αφορούν την εσωτερική ψυχική-νοητική κατάσταση του ατόμου. Αυτό φαίνεται εντονότερα και όταν συγκρίνουμε τα αποτελέσματα εκτίμησης με τη χρήση πόζας σε σχέση με την εκτίμηση με τη χρήση απευθείας RGB δεδομένων.

- Τα δεδομένα της πόζας μπορούν να εμπλουτιστούν και με επιπλέον χαρακτηριστικά, που εξάγονται από αυτά όπως η απόσταση από τον παρτενέρ, οι κατευθύνσεις κεφαλιού, σώματος κλπ ώστε να βελτιωθεί η εκτίμηση.

- Η αξιοποίηση συνελικτικών δικτύων, τα οποία τροφοδοτούμε με δεδομένα πόζας, αναδιαταγμένα ώστε να έχουν τη μορφή εικόνας, μπορεί να βοηθήσει σημαντικά στην εκτίμηση του engagement. Με αυτό τον τρόπο αξιοποιούμε την ικανότητα των συνελικτικών δικτύων να γενικεύουν πολύ επιτυχημένα όταν εκπαιδεύονται με σύνολα εικόνων και ταυτόχρονα την απαραίτητη πληροφορία που περιέχει η χρονική ακολουθία των στιγμιοτύπων για το επίπεδο του engagement. Χρησιμοποιώντας τέτοια δίκτυα πετυχαίνουμε σε όλα τα σύνολα δεδομένων με τα οποία δουλέψαμε μεγαλύτερα ποσοστά επιτυχίας από ότι όταν χρησιμοποιούμε δίκτυα που λαμβάνουν την πόζα ως μονοδιάστατο διάνυσμα.

- Η πόζα μπορεί να οδηγήσει σε ιδιαίτερα καλή εκτίμηση του engagement σε αλληλεπιδράσεις κατά τις οποίες τα παιδιά έχουν τη δυνατότητα να κινούνται στο χώρο, να εκφράζονται μέσω ποικιλίας κινήσεων. Το γεγονός αυτό είναι πολύ σημαντικό καθώς εφόσον η πόζα είναι αρκετή και δεν απαιτείται η εξαγωγή άλλων πολύπλοκων χαρακτηριστικών η εκτίμηση του engagement μπορεί ενσωματωθεί σε ρομποτικές εφαρμογές καθώς ο χρόνος ανάλυσης για τον υπολογισμό της πόζας και συνεπώς ο χρόνος απόκρισης του ρομπότ δεν θα είναι μεγάλος. Η δυνατότητα αυτή της πόζας είναι πιο περιορισμένη σε περιπτώσεις που τα παιδιά είναι στατικά και επομένως πρέπει εδώ τα δεδομένα να εμπλουτιστούν και από άλλες πηγές (παραπάνω δευτερογενή χαρακτηριστικά ή και RGB δεδομένα για το πρόσωπο κ.α.).

- Ξεκινώντας από μοντέλα που είναι εκπαιδευμένα σε τυπικώς αναπτυσσόμενα παιδιά σε συγκεκριμένες συνθήκες αλληλεπίδρασης, είναι δυνατόν να γενικεύσουμε πετυχαίνοντας καλά αποτελέσματα και σε παιδιά με διαταραχές αυτιστικού φάσματος και σε διαφορετικές συνθήκες και τύπους αλληλεπιδράσεων, εκπαιδεύοντας τα μοντέλα μας με λιγότερα δεδομένα. Αυτό είναι ένα από τα σημαντικότερα αποτελέσματα της παρούσας διπλωματικής εργασίας, καθώς τα δεδομένα με παιδιά με διαταραχές αυτιστικού φάσματος είναι σημαντικά δυσκολότερο να συγκεντρωθούν σε μεγάλη κλίμακα.

- Το χρονικό διάστημα των 3 δευτερολέπτων περιλαμβάνει αναγκαία και επαρκή πληροφορία ώστε να μπορεί ένα δίκτυο να εκτιμήσει το επίπεδο του engagement. Το συμπέρασμα αυτό προέκυψε μέσα από τα πειράματά μας για διαφορετικές αρχιτεκτονικές δικτύων και επιβεβαιώνεται από αντίστοιχες μελέτες ψυχολόγων ερευνητών.

Με βάση τα συμπεράσματα αυτά, στην επόμενη Ενότητα προτείνουμε ορισμένους άξονες για μελλοντικές προεκτάσεις της παρούσας εργασίας.

6.2 Μελλοντικές Επεκτάσεις

Το πρόβλημα εκτίμησης του engagement παρουσιάζει ιδιαίτερο ενδιαφέρον, τόσο λόγω των προκλήσεων που παρουσιάζει, όσο και λόγω των σημαντικών εφαρμογών που μπορεί να έχει. Η δυνατότητα ακριβούς εκτίμησης του engagement μπορεί να βελτιώσει θεαματικά την ποιότητα της διαδικασίας αλληλεπίδρασης μεταξύ παιδιών και ρομπότ, επιτρέποντας την διάδοση της χρήσης βοηθών ρομπότ, για παράδειγμα στην καλύτερη αντιμετώπιση των συνεπειών των διαταραχών του αυτιστικού φάσματος. Η προσπάθεια που κάναμε σε αυτή τη διπλωματική εργασία για την επίλυση του προβλήματος εκτίμησης του engagement θα μπορούσε ενδεικτικά να επεκταθεί προς τις ακόλουθες κατευθύνσεις:

- Ενσωμάτωση της μεθόδου μας σε ρομποτική εφαρμογή ώστε να πραγματοποιηθούν πειράματα με αλληλεπιδράσεις παιδιών και ρομπότ στις οποίες το ρομπότ θα εκτιμά το επίπεδο του engagement των παιδιών και θα προσαρμόζει τη συμπεριφορά του ανάλογα με την εκτίμηση αυτή. Αξιολόγηση με τον τρόπο αυτό της αποδοτικότητας των μοντέλων μας σε πραγματικές συνθήκες.

- Εξέταση παραγόντων που θα μπορούσαν να βελτιώσουν την εκτίμηση και στην περίπτωση στατικών αλληλεπιδράσεων, για παράδειγμα εκπαίδευση με συνδυασμό πόζας και RGB δεδομένων του προσώπου των παιδιών.

- Αξιοποίηση και άλλων χαρακτηριστικών για να εμπλουτιστούν τα χαρακτηριστικά που προκύπτουν από την πόζα. Ένα από αυτά τα χαρακτηριστικά θα μπορούσε να είναι το αν μιλάει ή όχι το παιδί ή το αν μιλάει ή όχι το ρομπότ.
- Πειράματα και με Graph Convolutional Networks για την εκτίμηση του engagement. Συνδυασμός δικτύων για ακριβέστερη εκτίμηση.

Σχήμα 6.1: Στιγμιότυπο από το παιχνίδι "Show me the Gesture", "Δείξε μου πως θα με καλέσεις κοντά σου!"

Bibliography

[1] F. Del Duchetto, P. Baxter, and M. Hanheide, "Are you still with me? continuous engagement assessment from a robot's point of view," *arXiv preprint arXiv:2001.03515*, 2020.

[2] 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.

[3] S. Lemaignan, C. E. R. Edmunds, E. Senft, and T. Belpaeme, "The pinsoro dataset: Supporting the data-driven study of child-child and child-robot social dynamics," *PLOS ONE*, vol. 13, pp. 1–19, 2018.

[4] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh, "Openpose: realtime multi-person 2d pose estimation using part affinity fields," *arXiv preprint arXiv:1812.08008*, 2018.

[5] P. P. Filntisis, N. Efthymiou, P. Koutras, G. Potamianos, and P. Maragos, "Fusing body posture with facial expressions for joint recognition of affect in child–robot interaction," *IEEE Robotics and Automation Letters*, vol. 4, pp. 4011–4018, 2019.

[6] D. C. Luvizon, D. Picard, and H. Tabia, "2d/3d pose estimation and action recognition using multitask deep learning," in *Proc. CVPR*, 2018.

[7] H. Javed, W. Lee, and C. H. Park, "Toward an automated measure of social engagement for children with autism spectrum disorder–a personalized computational modeling approach," *Frontiers in Robotics and AI*, vol. 7, p. 43, 2020.

[8] T. Mittal, P. Guhan, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "Emoticon: Context-aware multimodal emotion recognition using frege's principle," in *Proc CVPR*, 2020.

[9] U. Bhattacharya, T. Mittal, R. Chandra, T. Randhavane, A. Bera, and D. Manocha, "Step: Spatial temporal graph convolutional networks for emotion perception from gaits," *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[10] C. Wang, Y. Wang, and A. L. Yuille, "An approach to pose-based action recognition," in *Proc. CVPR*, 2013.

[11] D. Thakur, "Lstm and its equations," https://medium.com/@divyanshu132/lstm-and-its-equations-5ee9246d04af.

[12] H. Yakura, S. Shinozaki, R. Nishimura, Y. Oyama, and J. Sakuma, "Malware analysis of imaged binary samples by convolutional neural network with attention mechanism," in *Proceedings of ACM Conference on Data and Application Security and Privacy*, 2018.

[13] U. Udofia, "Basic overview of convolutional neural network (cnn)," https://medium.com/dataseries/basic-overview-of-convolutional-neural-network-cnn-4fcc7dbb4f17.

[14] "The pinsoro dataset," https://freeplay-sandbox.github.io/coding-scheme.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016.

[16] N. Efthymiou, P. P. Filntisis, P. Koutras, A. Tsiami, J. Hadfield, G. Potamianos, and P. Maragos, "Childbot: Multi-robot perception and interaction with children," *arXiv preprint arXiv:2008.12818*, 2020.

[17] T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, and F. Tanaka, "Social robots for education: A review," *Science robotics*, vol. 3, 2018.

[18] Y. Feng, Q. Jia, M. Chu, and W. Wei, "Engagement evaluation for autism intervention by robots based on dynamic bayesian network and expert elicitation," *IEEE Access*, vol. 5, pp. 19 494–19 504, 2017.

[19] J. Wainer, B. Robins, F. Amirabdollahian, and K. Dautenhahn, "Using the humanoid robot kaspar to autonomously play triadic games and facilitate collaborative play among children with autism," *IEEE Transactions on Autonomous Mental Development*, vol. 6, pp. 183–199, 2014.

[20] E. S. Kim, L. D. Berkovits, E. P. Bernier, D. Leyzberg, F. Shic, R. Paul, and B. Scassellati, "Social robots as embedded reinforcers of social behavior in children with autism," *Journal of autism and developmental disorders*, vol. 43, pp. 1038–1049, 2013.

[21] S. Ali, F. Mehmood, D. Dancey, Y. Ayaz, M. J. Khan, N. Naseer, R. D. C. Amadeu, H. Sadia, and R. Nawaz, "An adaptive multi-robot therapy for improving joint attention and imitation of asd children," *IEEE Access*, vol. 7, pp. 81 808–81 825, 2019.

[22] A. Taheri, M. Alemi, A. Meghdari, H. Pouretemad, and S. Holderread, "Clinical application of humanoid robots in playing imitation games for autistic children in iran," *Procedia-Social and Behavioral Sciences*, vol. 176, pp. 898–906, 2015.

[23] S. Tariq, S. Baber, A. Ashfaq, Y. Ayaz, M. Naveed, and S. Mohsin, "Interactive therapy approach through collaborative physical play between a socially assistive humanoid robot and children with autism spectrum disorder," in *International Conference on Social Robotics*, 2016, pp. 561–570.

[24] Y. Zhang, W. Song, Z. Tan, H. Zhu, Y. Wang, C. M. Lam, Y. Weng, S. P. Hoi, H. Lu, B. S. M. Chan *et al.*, "Could social robots facilitate children with autism spectrum disorders in learning distrust and deception?" *Computers in Human Behavior*, vol. 98, pp. 140–149, 2019.

[25] I. Giannopulu, K. Terada, and T. Watanabe, "Communication using robots: a perception-action scenario in moderate asd," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 30, pp. 603–613, 2018.

[26] S. Lemaignan, F. Garcia, A. Jacq, and P. Dillenbourg, "From real-time attention assessment to "with-me-ness" in human-robot interaction," in *ACM International Conference on Human-Robot Interaction (HRI)*. IEEE, 2016.

[27] A. Chorianopoulou, E. Tzinis, E. Iosif, A. Papoulidi, C. Papailiou, and A. Potamianos, "Engagement detection for children with autism spectrum disorder," in *Proc. ICASSP*, 2017.

[28] O. Rudovic, Y. Utsumi, J. Lee, J. Hernandez, E. C. Ferrer, B. Schuller, and R. W. Picard, "Culturenet: A deep learning approach for engagement intensity estimation from face images of children with autism," in *Proc. IROS*, 2018.

[29] C. L. Sidner, C. Lee, and N. Lesh, "The role of dialog in human robot interaction," *International workshop on language understanding and agents for real world interaction*, 2003.

[30] M. Khamassi, G. Chalvatzaki, T. Tsitsimis, G. Velentzas, and C. Tzafestas, "A framework for robot learning during child-robot interaction with human engagement as reward signal," in *Proc. BAILAR Workshop, in International Conference on Robot and Human Interactive Communication*, 2018.

[31] I. Poggi, *Mind, hands, face and body: a goal and belief view of multimodal communication*. Weidler, 2007.

[32] Z. Zheng, H. Zhao, A. R. Swanson, A. S. Weitlauf, Z. E. Warren, and N. Sarkar, "Design, development, and evaluation of a noninvasive autonomous robot-mediated joint attention intervention system for young children with asd," *IEEE Transactions on Human-Machine Systems*, vol. 48, pp. 125–135, 2018.

[33] O. Rudovic, J. Lee, M. Dai, B. Schuller, and R. Picard, "Personalized machine learning for robot perception of affect and engagement in autism therapy," *Science*, vol. 3, 2018.

[34] O. Rudovic, H. W. Park, J. Busche, B. Schuller, C. Breazeal, and R. W. Picard, "Personalized estimation of engagement from videos using active learning with deep reinforcement learning," in *Proc. CVPR Workshop*, 2019.

[35] M. Khamassi, G. Velentzas, T. Tsitsimis, and C. Tzafestas, "Robot fast adaptation to changes in human engagement during simulated dynamic social interaction with active exploration in parameterized reinforcement learning," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, pp. 881–893, 2018.

[36] C. Rich, B. Ponsler, A. Holroyd, and C. L. Sidner, "Recognizing engagement in human-robot interaction," in *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2010.

[37] G. Castellano, A. Pereira, I. Leite, A. Paiva, and P. Mcowan, "Detecting user engagement with a robot companion using task and social interaction-based features," in *Proc. international conference on Multimodal interfaces*, 2009.

[38] C. L. Sidner, C. Lee, C. D. Kidd, N. Lesh, and C. Rich, "Explorations in engagement for humans and robots," in *Artificial Intelligence*, 2005.

[39] O. Rudovic, J. Lee, M. M., B. Schuller, and R. Picard, "Measuring engagement in robot-assisted autism therapy: A cross-cultural study," *Frontiers in Robotics and AI*, vol. 4, p. 36, 2017.

[40] E. S. Kim, R. Paul, F. Shic, and B. Scassellati, "Bridging the research gap: Making hri useful to individuals with autism," *Journal of Human-Robot Interaction Steering Committee*, vol. 1, p. 26–54, 2012.

[41] D. Gouaillier, V. Hugel, P. Blazevic, C. Kilner, J. Monceaux, P. Lafourcade, B. Marnier, J. Serre, and B. Maisonnier, "Mechatronic design of nao humanoid," in *Proc. ICRA*, 2009.

[42] "Furhat Robotics," http://www.furhatrobotics.com/.

[43] P. J. Besl and N. D. McKay, ""a method for registration of 3-d shapes"," *IEEE Transanctions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, p. 239–256, 1992.

[44] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *IEEE International Conference on Automatic Face Gesture Recognition*, 2018.

[45] L. Sigal, "Human pose estimation," in *Encyclopedia of Computer Vision*. Springer, 2011.

[46] D. Ramanan, D. A. Forsyth, and A. Zisserman, "Strike a pose: Tracking people by finding stylized poses," in *CVPR*, 2005.

[47] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in *British Machine Vision Conference*, 2010.

[48] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Poselet conditioned pictorial structures," in *Proc. CVPR*, 2013.

[49] A. Bulat and G. Tzimiropoulos, "Human pose estimation via convolutional part heatmap regression," in *ECCV*, 2016.

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

[51] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in *Proc. CVPR*, 2017.

[52] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. CVPR*, 2016.

[53] E. Marinou, M. Zanfir, V. Olaru, and C. Sminchisescu, "3d human sensing, action and emotion recognition in robot assisted therapy of children with autism," in *Proc. CVPR*, 2018.

[54] E. Groff, "Laban movement analysis: Charting the ineffable domain of human movement," *Journal of Physical Education, Recreation & Dance*, vol. 66, pp. 27–30, 1995.

[55] J. Yu, Y. Yoon, and M. Jeon, "Predictively encoded graph convolutional network for noise-robust skeleton-based action recognition," *arXiv preprint arXiv:2003.07514*, 2020.

[56] J. Gao, T. He, X. Zhou, and S. Ge, "Focusing and diffusion: Bidirectional attentive graph convolutional networks for skeleton-based action recognition," *arXiv preprint arXiv:1912.11521*, 2019.

[57] A. Gaschler, S. Jentzsch, M. Giuliani, K. Huth, J. de Ruiter, and A. Knoll, "Social behavior recognition using body posture and head pose for human-robot interaction," in *Proc. IROS*, 2012.

[58] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, 2019, pp. 8024–8035.

[59] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–80, 1997.

[60] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," *Neural Information Processing Systems*, vol. 25, 2012.

[61] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc CVPR*, 2018.

[62] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proc CVPR*, 2017.

[63] D. Anagnostopoulou, N.Efthymiou, C.Papailiou, and P. Maragos, "Engagement estimation during child robot interaction using deep convolutional networks focusing on asd children," *Submitted to ICRA*, 2021.