Αναγνώριση Ανθρώπινης Δραστηριότητας σε Πραγματικό Χρόνο με CNN-LSTM

Κασιωτάκης Ηλίας
Τμήμα Ψηφιακών Συστημάτων
Πανεπιστήμιο Πειραιώς
Αθήνα, Ελλάδα
eliaskasiotakis@gmail.com

***ΠΕΡΙΛΗΨΗ*— Λόγω της εξέλιξης της τεχνολογίας και της Τεχνητής Νοημοσύνης (ΤΝ), τα τελευταία χρόνια γίνεται έρευνα στην ανάπτυξη συνεργατικών ρομπότ, cobot, προκειμένου να ενταχτούν στη βιομηχανία. Για αυτό τον σκοπό, είναι απαραίτητη η ακριβής και αποδοτική αναγνώριση ανθρώπινης δραστηριότητας (Human Activity Recognition HAR) σε πραγματικό χρόνο (real time) ώστε η αλληλεπίδραση των ρομπότ με τον άνθρωπο να είναι η επιθυμητή. Αρκετοί ερευνητές αναπτύσσουν μεθόδους αναγνώρισης ανθρώπινης δραστηριότητας χρησιμοποιώντας ως δεδομένα τη θέση των αρθρώσεων (joints) του ανθρώπινου σκελετού. Παρόλα αυτά, η αποκλειστική χρήση των σκελετικών δεδομένων (skeletal data) αγνοεί την πληροφορία των αλληλεπιδράσεων του ατόμου με αντικείμενα στο χώρο που σχετίζονται με τη δράση, χάνοντας σημαντική οπτική πληροφορία. Στη συγκεκριμένη εργασία, χρησιμοποιούμε ένα CNN-LSTM δίκτυο το οποίο δέχεται ως είσοδο τα σκελετικά δεδομένα εξαγόμενα από το εργαλείο Mediapipe και χωρική πληροφορία σχετικά με τα αντικείμενα (object information) για την αναγνώριση ανθρώπινης δραστηριότητα σε πραγματικό χρόνο.**

Λέξεις Κλειδιά—CNN, LSTM, pose , landmarks, Mediapipe, πραγματικός χρόνος (real time), αναγνώριση ανθρώπινης δραστηριότητας (human activity recognition), σκελετικά δεδομένα (skeletal data), object information, αρθρώσεις (joints)

I. ΕΙΣΑΓΩΓΗ

Στα πλαίσια της βιομηχανίας (industry 4.0), η συνεργασία μεταξύ ανθρώπων και συνεργατικών ρομπότ (cobots) γίνεται ιδιαίτερα σημαντική στην παραγωγική διαδικασία. Προκειμένου το ρομπότ να συνεισφέρει σημαντικά κατά τη διάρκεια της αλληλεπίδρασης είναι σημαντικό να μπορούν να αντιλαμβάνονται οπτικά τον άνθρωπο συνεργάτη τους και να αναγνωρίσουν την δραστηριότητά του. Προκειμένου να επιτευχθεί ο συγκεκριμένος στόχος, ενσωματώνονται στο ρομπότ μέθοδοι αναγνώρισης ανθρώπινης δραστηριότητας [1].

Για την αναγνώριση της ανθρώπινης δραστηριότητας χρησιμοποιείται κυρίως οπτική πληροφορία και σύνηθες εργαλείο καταγραφής αποτελεί κοινή κάμερα που παράγει RGB εικόνες [1][2]. Ωστόσο, αυτές οι προσεγγίσεις αντιμετωπίζουν αρκετές προκλήσεις καθώς τα σύνολα δεδομένων τείνουν να είναι μικρότερα σε σύγκριση με άλλους τομείς, όπως η ανίχνευση αντικειμένων, με αποτέλεσμα να μην υπάρχουν αρκετά περίπλοκα περιβάλλοντα. Ως αποτέλεσμα, οι RGB μέθοδοι τείνουν σε υπερκεπαίδευση (overfit) και έχουν μικρές ικανότητες γενίκευσης (generalization). Για αυτό το λόγο, υιοθετούνται μέθοδοι αναγνώρισης της θέσης των αρθρώσεων του ανθρώπινου σκελετού καθώς οδηγούν σε ανεξαρτησία από το υπόλοιπο περιβάλλον και συνεπώς σε καλύτερα αποτελέσματα γενίκευσης.

Παρόλα αυτά, η εξαγωγή σκελετικών δεδομένων αντιμετωπίζει αρκετές προκλήσεις όπως μεγάλη ποικιλία ανθρώπινων στάσεων (poses), των πολλών βαθμών ελευθερίας των ανθρώπινων κινήσεων και των περιπτώσεων απόκρυψης (occlusion) μέρους της εικόνας [3]. Επίσης, τα σκελετικά δεδομένα αγνοούν την υπόλοιπη χωρική πληροφορία [1].

Λόγω της μεγάλης ομοιότητας μεταξύ κινήσεων και της περιπλοκότητας ορισμένων δράσεων, η ενσωμάτωση αναγνώρισης αντικειμένων που συμμετέχουν στη δράση οδηγεί σε αύξηση της αποδοτικότητας και της ακρίβειας του μοντέλου [4].

Η συγκεκριμένη εργασία αξιοποιεί μεθόδους τεχνητής νοημοσύνης για την αναγνώριση ανθρώπινης δραστηριότητας σε πραγματικό χρόνο χρησιμοποιώντας ένα σύνολο δεδομένων βίντεο. Για τους σκοπούς της εργασίας χρησιμοποιήθηκε το σύνολο δεδομένων (dataset) IKEA ASM dataset [5] και εξετάστηκαν 24 κλάσεις.

Τα βίντεο υποβλήθηκαν σε προεπεξεργασία για την δημιουργία μικρότερης διάρκειας βίντεο των 24 frames και για την εξαγωγή των θέσεων των αρθρώσεων και των επιθυμητών αντικειμένων πάνω σε αυτά.

Η μέθοδος που αναπτύχθηκε χρησιμοποιεί ένα CNN-LSTM μοντέλο βασισμένο στο [6] όπου το CNN βρίσκει τα χωρικά χαρακτηριστικά (spatial features) μεταξύ των frames, των joints και των αντικειμένων, τα οποία εισάγονται ως είσοδος στο LSTM για να εξάγει τα χρονικά χαρακτηριστικά (temporal feautures).

Παρουσιάζεται η απόδοση του CNN LSTM δικτύου και αν επιτυγχάνεται η πρόβλεψη σε πραγματικό χρόνο.

II. Μεθοδολογια

A. Επιλογή Συνόλου Δεδομένων (Dataset)

Η αλληλεπίδραση μεταξύ ανθρώπων και ρομπότ πραγματοποιείται συνήθως σε σύνολα δεδομένων συναρμολόγησης (assemblies). Επιπλέον, η προοπτική παρακολούθησης της δραστηριότητας πρέπει να είναι παρόμοια με την προοπτική ενός cobot. Επίσης, πρέπει να υπάρχουν διαθέσιμα δεδομένα τόσο για την θέση των αρθρώσεων του ανθρώπου στο χώρο όσο και για για τη θέση των αντικειμένων που χρησμοποιούνται..

Για αυτούς τους λόγους αποφασίστηκε να χρησιμοποιηθεί το σύνολο δεδομένων συναρμολόγησης επίπλων IKEA ASM, το οποίο είναι ένα πολύ εκτεταμένο σύνολο δεδομένων, καταγεγραμμένο από πολλαπλές προοπτικές. Αποτελείται από σκηνές όπου διαφορετικοί τύποι επίπλων συναρμολογούνται σε διάφορα περιβάλλοντα, καθιστώντας το σύνολο δεδομένων αρκετά απαιτητικό. Επιπλέον, σε περιέχει επίσης επισημασμένες μάσκες αντικειμένων.

Το IKEA ASM αποτελείται από 371 διακριτές διαδικασίες συναρμολόγησης, στις οποίες 48 συμμετέχοντες κλήθηκαν να συναρμολογήσουν έναν από τους τέσσερις διαφορετικούς τύπους επίπλων. Κάθε διαδικασία συναρμολόγησης καταγράφεται από τρεις διαφορετικές κάμερες Kinect V2, με αποτέλεσμα να υπάρχουν 1113 βίντεο και 35 ώρες υλικού με 24 frames ανά δευτερόλεπτο (fps). Για την αναγνώριση ανθρώπινων δράσεων, υπάρχουν 17.000 επισημασμένα στιγμιότυπα δράσεων, κατανεμημένα σε 33 ατομικές κατηγορίες.

Τα κομμάτια των επίπλων περιλαμβάνουν επτά διαφορετικές κατηγορίες αντικειμένων. Για τις εμφανίσεις των επισημάνσεων των αντικειμένων, το 1% των δεδομένων επισημάνθηκε χειροκίνητα και οι υπόλοιπες αποκτήθηκαν μέσω της υπερεκπαίδευσης (overfitting) αρκετών μοντέλων στα επισημασμένα δεδομένα, με αποτέλεσμα να προκύψουν δεδομένα ψευδο-αληθινής βάσης (pseudo ground truth data). Ωστόσο, οι μάσκες αντικειμένων είναι διαθέσιμες μόνο για μία όψη, όπως φαίνεται στο Σχήμα 1. Οι επισημάνσεις περιλαμβάνουν segmentation masks και τις συντεταγμένες και το μέγεθος των boxes.

Για τα δεδομένα σκελετού, το IKEA ASM παρέχει προβλέψεις σκελετού 2D, που λήφθηκαν μέσω του OpenPose και του Keypoint R-CNN, από τα οποία Το σύνολο δεδομένων παρέχει επίσημους διαχωρισμούς για εκπαίδευση (training) και δοκιμή (test) [5].

Σχήμα 1: Η τμηματοποίηση των αντικειμένων (instance segmentation) του Mask R-CNN με το Swin-Tiny backbone, που εκπαιδεύτηκε το IKEA ASM.

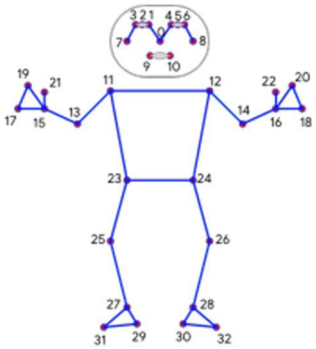
Λόγω του μεγάλου μεγέθους του dataset αποφασίστηκε να χρησιμοποιηθούν τα βίντεο των δράσεων από 7 διαφορετικά άτομα για training set και 3 διαφορετικών ατόμων για το test set από τον επίσημο διαχωρισμό αντίστοιχα για την κατασκευή των επίπλων Kallax_Shelf_Drawer και Lack_Side_Table. Έτσι, επιτεύχθηκε διαχωρισμός test set 25.64% και training set 74.36%.

B. Ανίχνευση Αρθρώσεων μέσω του Mediapipe

Το framework της MediaPipe χρησιμοποιείται για τη δημιουργία μηχανικής μάθησης η οποία μπορεί να επεξεργαστεί χρονοσειριακά δεδομένα όπως βίντεο, ήχο κ.λπ. Αντίθετα με τα περισσότερα frameworks μηχανικής μάθησης που απαιτούν υψηλή υπολογιστική ισχύ, το MediaPipe μπορεί να λειτουργεί αποτελεσματικά σε συσκευές με χαμηλή υπολογιστική ισχύ, όπως Androids και ενσωματωμένες συσκευές Internet of Things (IoT).
Το εργαλείο της MediaPipe αποτελείται από το framework MediaPipe και τις λύσεις MediaPipe. Το framework MediaPipe έχει αναπτυχθεί μέσω των γλωσσών προγραμματισμού C++, Java και Objective C και αποτελείται από τα εξής βασικά application programming interfaces (API):

- Calculator API
- Graph Construction API
- Graph Execution API)

Οι λύσεις MediaPipe περιλαμβάνουν 16 προ-εκπαιδευμένα μοντέλα TensorFlow και TensorFlow Lite, κατασκευασμένα για συγκεκριμένες χρήσεις. Σε αυτήν την εργασία, χρησιμοποιήθηκε η λύση MediaPipe για να εξάγει τις (x,y) θέσεις των αρθρώσεων του σκελετού του ανθρώπου που συμμετέχει στην δραστηριότητα. Αυτό το μοντέλο εξάγει 33 τρισδιάστατα σημεία αναγνώρισης (όπως φαίνεται στο Σχήμα 3) από κάθε frame.



Σχήμα 2: Οι 33 αρθρώσεις του Mediapipe

Για να το πετύχει αυτό, χρησιμοποιούνται ταυτόχρονα δύο εξαρτημένα μοντέλα. Πρώτα, ένα μοντέλο ανίχνευσης ανθρώπου (a Person Detection Model) που εστιάζει στην ανίχνευση του bounding box του ανθρώπινου προσώπου κάνοντας την ισχυρή, αλλά έγκυρη, υπόθεση ότι το κεφάλι του ατόμου θα είναι πάντα ορατό στην περίπτωση χρήσης [3].

Στη συνέχεια το μοντέλο ενός Νευρωνικού Δικτύου εξάγει τα 33 joints. Το μοντέλο είναι καλά εκπαιδευμένο και εύρωστο και, συνεπώς, μπορεί να ανιχνεύει και να αναγνωρίζει ακριβώς τις αρθρώσεις, ακόμα και σε μερικώς ορατές στάσεις σώματος στις περισσότερες περιπτώσεις [3].

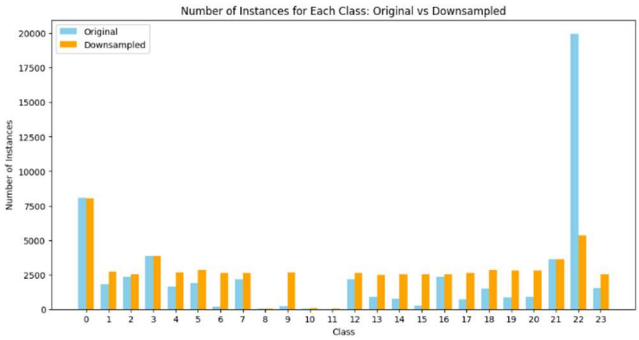
Το μοντέλο ανίχνευσης πόζας από το MediaPipe εφαρμόστηκε σε όλα τα συλλεγμένα frames από τα video. Το μοντέλο εξήγαγε τις συντεταγμένες x, y και z των joints αλλά χρησιμοποιήθηκαν μόνο οι x και y. Τα x και y κανονικοποιούνται σε [0.0, 1.0] από το πλάτος και το ύψος της εικόνας αντίστοιχα [3].

C. Προεπεξεργασία Δεδομένων

Αρκετά σημαντική διαδικασία ήταν η προεπεξεργασία των δεδομένων για την εκτέλεση της συγκεκριμένης διαδικασίας. Αρχικά, επειδή το μοντέλο πρέπει να λειτουργεί σε πραγματικό χρόνο δεν μπορεί να έχει διαθέσιμη ολόκληρη τη χρονοσειρά της δράσης αλλά ένα συγκεκριμένο μέρος της του οποίου το μέγεθος κρίνεται από τον ρυθμό με τον οποίο παράγει frames η κάμερα. Επειδή οι κάμερες που χρησιμοποιήθηκαν για τη συλλογή του dataset παρέχουν 24 fps, κάθε video χωρίστηκε σε μικρότερα video των 24 frames. Σε κάθε νέο video αντιστοιχήθηκαν οι κλάσεις στις οποίες ανήκει κάθε frame.

Στη συνέχεια, εξάχθηκαν κανονικοποιημένες με το πλάτος και το ύψος του frame οι συντεταγμένες x και y κάθε joint αντίστοιχα. Προκειμένου να εντάξουμε χωρική πληροφορία που σχετίζεται με τα αντικείμενα, εισάγουμε ως επιπλέον joints τις κανονικοποιημένες συντεταγμένες x και y των boxes των κλάσεων των αντικειμένων και σημειώνονται οι συντεταγμένες των αντικειμένων που βρίσκονται πιο κοντά στο δεξί και στο αριστερό χέρι του ατόμου κάνοντας την υπόθεση ότι ένας άνθρωπος μπορεί να αλληλοεπιδρά μόνο με ένα αντικείμενο σε κάθε χέρι όπως γίνεται στο [1]. Θεωρούμε ότι το task του object detection είναι ήδη λυμένο και χρησιμοποιούμαι τις πληροφορίες από το pseudo ground truth. Στη συνέχεια, μετατοπίστηκε το κέντρο αρχής του συστήματος συντεταγμένων στη θέση ανάμεσα στα joints 11 και 12 και τα δεδομένα υπέστησαν MinMax κανονικοποίηση για να βρεθούν ξανά εντός του εύρους 0-1.

Επειδή το dataset που δημιουργήθηκε ήταν αρκετά imbalanced, πραγματοποιήθηκε συνδυασμός των τεχνικών downsampling και oversampling.



Σχήμα 3: Το πλήθος των frames ανά κλάση πριν (μπλε) και μετά τις τεχνικές των downsampling και oversampling (πορτοκαλί).

D. Η Αρχιτεκτονική του Μοντέλου

Ακολουθεί η περιγραφή της αρχιτεκτονικής του CNN-LSTM μοντέλου:

Επίπεδο Εισόδου: Το μοντέλο ξεκινά με ένα επίπεδο εισόδου σε σχήμα 3D τανυστή με διαστάσεις (Nfr=24, Njoint=45, Nattr=2), που αναπαριστά 24 frames, 45 αρθρώσεις (33 αρθρώσεις σκελετού συν 12 αρθρώσεις που αναπαριστούν αντικείμενα) και 2 χαρακτηριστικά ανά άρθρωση (x,y συντεταγμένες).

Πρώτο Συνελικτικό Μπλοκ:

Επίπεδο Conv2D (128 φίλτρα, μέγεθος 5x5, ενεργοποίηση 'relu'): Αυτό το επίπεδο εφαρμόζει 128 φίλτρα με πυρήνα 5x5 στα δεδομένα εισόδου. Η μη γραμμική συνάρτηση ενεργοποίησης ReLU βοηθά στην επιτάχυνση της διαδικασίας εκπαίδευσης εισάγοντας μη γραμμικότητα.

Επίπεδο MaxPooling2D (παράθυρο 1x2): Εφαρμόζεται μέγιστη συμπίεση σε κάθε περιοχή 1x2, μειώνοντας τη διάσταση των δεδομένων και διατηρώντας τις πιο σημαντικές πληροφορίες.

Επίπεδο BatchNormalization: αποφυγή overfit.

Δεύτερο Συνελικτικό Μπλοκ:

Επίπεδο Conv2D (256 φίλτρα, μέγεθος 5x5, ενεργοποίηση 'relu'): Εφαρμόζει 256 φίλτρα με 5x5 πυρήνα, ακολουθούμενο από MaxPooling2D και BatchNormalization.

MaxPooling2D (παράθυρο 1x2): Όπως και πριν, μειώνει τη διάσταση των δεδομένων.

BatchNormalization

Χρονική Διανομή και Flatten:

Επίπεδο TimeDistributed (Flatten): Μετατρέπει την έξοδο

από τα προηγούμενα επίπεδα σε επίπεδο διανύσματος, προετοιμάζοντας την είσοδο για τα επόμενα επίπεδα.

TimeDistributed Dense Layer (256 νευρώνες, ενεργοποίηση 'relu'): Περιλαμβάνει 256 νευρώνες με ενεργοποίηση ReLU και εφαρμόζεται BatchNormalization.

LSTM Layer:

Επίπεδο LSTM (256 νευρώνες, ενεργοποίηση 'tanh', recurrent dropout 0.3): Επεξεργάζεται τις χρονοσειρές των δεδομένων, βοηθώντας στην εκμάθηση χρονικών σχέσεων μεταξύ των εισόδων.

Τελικά Επίπεδα Dense:

TimeDistributed Dense Layer (256 νευρώνες, ενεργοποίηση 'relu'): Ένα άλλο επίπεδο με 256 νευρώνες, ακολουθούμενο από Dropout (0.2) για την πρόληψη του overfitting.

TimeDistributed Dense Layer (64 νευρώνες, ενεργοποίηση 'relu'): Επίπεδο με 64 νευρώνες και Dropout (0.3) για επιπλέον προστασία από το overfitting.

Τελικό Επίπεδο Εξόδου:

TimeDistributed Dense Layer (classes, ενεργοποίηση 'softmax'): Το τελικό επίπεδο παρέχει τις πιθανότητες για τις 24 κλάσεις, επιτρέποντας την κατηγοριοποίηση της εξόδου.

Αυτή η αρχιτεκτονική συνδυάζει συνελικτικά, χρονικά και πλήρως συνδεδεμένα επίπεδα για να διαχειριστεί τις χρονοσειρές και τις πολυδιάστατες εισόδους.

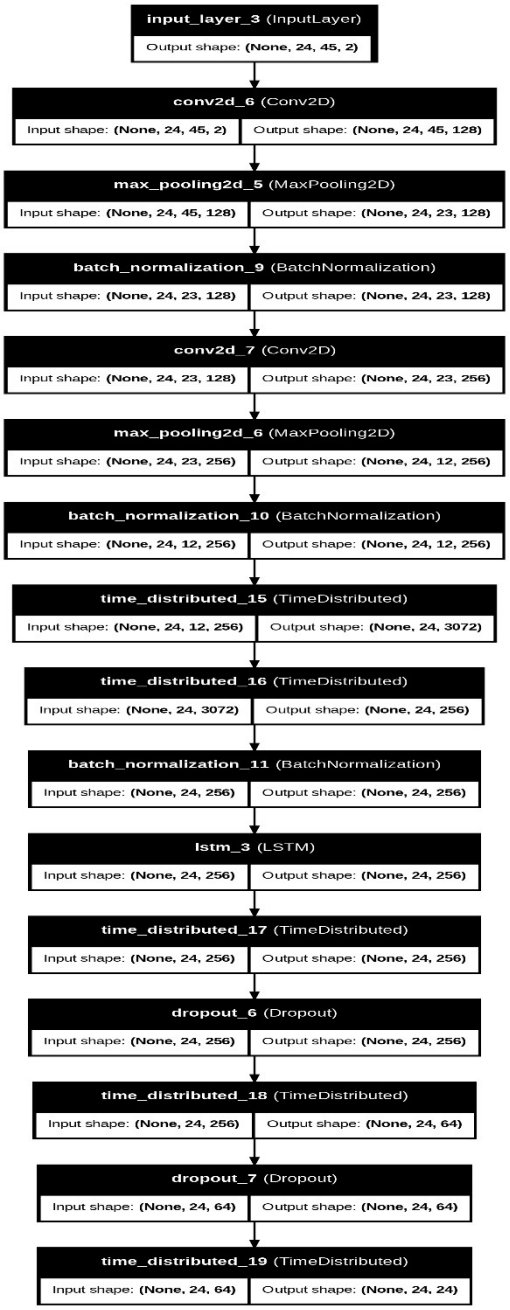
Το μοντέλο εκπαιδεύτηκε με τον βελτιστοποιητή Adam και τη συνάρτηση απώλειας κατηγοριοποίησης (categorical cross entropy), ενώ η απόδοση αξιολογήθηκε με τη μετρική ακρίβειας (accuracy). Επιπλέον ορισμένα από τα βάρη αρχικοποιήθηκαν βάσει της κατανομής He.

Model: "functional"

Layer (type)	Output Shape	Param #
input_layer (InputLayer)	(None, 24, 45, 2)	0
conv2d (Conv2D)	(None, 24, 45, 128)	2,432
conv2d_1 (Conv2D)	(None, 24, 45, 256)	295,168
max_pooling2d (MaxPooling2D)	(None, 24, 23, 256)	0
conv2d_2 (Conv2D)	(None, 24, 23, 256)	590,080
dropout (Dropout)	(None, 24, 23, 256)	0
time_distributed (TimeDistributed)	(None, 24, 5888)	0
time_distributed_1 (TimeDistributed)	(None, 24, 256)	1,507,584
lstm (LSTM)	(None, 24, 256)	525,312
time_distributed_2 (TimeDistributed)	(None, 24, 256)	65,792
dropout_1 (Dropout)	(None, 24, 256)	0
time_distributed_3 (TimeDistributed)	(None, 24, 128)	32,896
dropout_2 (Dropout)	(None, 24, 128)	0
time_distributed_4 (TimeDistributed)	(None, 24, 64)	8,256
dropout_3 (Dropout)	(None, 24, 64)	0
time_distributed_5 (TimeDistributed)	(None, 24, 24)	1,560

Total params: 3,029,080 (11.56 MB)
Trainable params: 3,029,080 (11.56 MB)
Non-trainable params: 0 (0.00 B)

Σχήμα 4: Συνοπτική αναπαράσταση της αρχιτεκτονικής σχηματικά και σε μορφή πίνακα

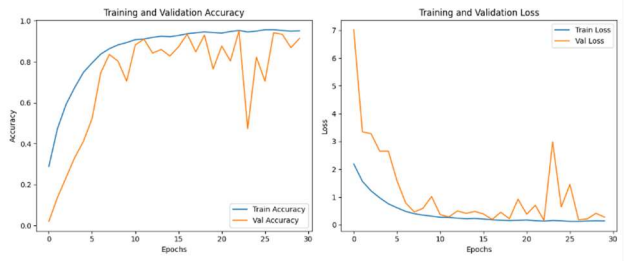


III. ΑΠΟΤΕΛΕΣΜΑΤΑ

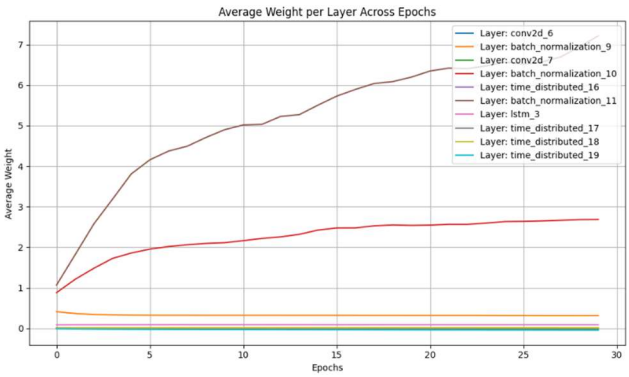
Από τις 100 επαναλήψεις που ήταν προγραμματισμένες για την εκπαίδευση του μοντέλου, πραγματοποιήθηκαν μόνο 36 λόγω της διαδικασίας 'early stopping' που ενεργοποιήθηκε. Αυτό συνέβη διότι η απώλεια στα δεδομένα επικύρωσης δεν μειώθηκε για συνεχόμενες 7 επαναλήψεις, επομένως το μοντέλο σταμάτησε νωρίτερα για να αποφευχθεί το overfitting και να διατηρηθούν τα βέλτιστα βάρη. Προκειμένου να γίνονται παρακολουθείτε η απόδοση του μοντέλου από το training set μέσω random split 80-20 δημιουργήθηκε ένα validation set

Παρακάτω φαίνονται τα αποτελέσματα της εκπαίδευσης.

Το μοντέλο φαίνεται να αποδίδει πολύ καλά τόσο στα δεδομένα εκπαίδευσης όσο και στα δεδομένα επαλήθευσης, με την ακρίβεια να φτάνει σχεδόν στο 100% μετά από μερικές εποχές και η απώλεια είναι πολύ χαμηλή, υποδεικνύοντας ότι το μοντέλο μαθαίνει αποτελεσματικά.



Σχήμα 5: Η ακρίβεια (accuracy) (αριστερά) και η απώλεια (loss) (δεξιά) του training set (μπλε) και του validation set (πορτοκαλί) ανά εποχή.



Σχήμα 6: Το μέσο βάρος κάθε layer ανά εποχή.

Παρόλα αυτά, η ακρίβεια στο test set είναι πολύ χαμηλή. Επειδή τα δεδομένα στο test set ήταν επίσης imbalanced χρησιμοποιήθηκε η μετρική F1 weighted η οποία δεν ξεπερνάει το 5%. Το μοντέλο όχι μόνο επηρεάστηκε από overfitting (λογικά λόγω του oversampling) αλλά αν παρατηρήσουμε το μέσω βάρος κάθε layer ανά εποχή στο Σχήμα 6 παρατηρούμε μία στασιμότητα, ενδεικτικό ότι το μοντέλο υπέστη το φαινόμενο του vanishing gradient. Ο χρόνος ανταπόκρισής του είναι 0.043788 seconds ανά 24 frames με αποτέελσμα να είναι real time.

Σχήμα 7: Ο confusion Matrix.

IV. ΜΕΛΛΟΝΤΙΚΕΣ ΚΑΤΕΥΘΥΝΣΕΙΣ

Προκειμένου να εξερευνηθούν τα πλεονεκτήματα του μοντέλου θα γίνει προσπάθεια πειραματισμού με περισσότερες αρχιτεκτονικές. Επίσης, θα χρησιμοποιηθεί validation set που δε θα προκύπτει από random split καθώς μπορεί και το ίδιο μπορεί να είχε κοινά δεδομένα με το training set λόγω του oversampling με αποτέλεσμα να μην είναι το κατάλληλο κριτήριο παρακολούθησης της απόδοσης του μοντέλου. Ίσως αντικατασταθεί το oversampling με data augmentation. Επιπρόσθετα, αντί να χρησιμοποιηθούν οι θέσεις των boxes των αντικειμένων έτοιμα από το dataset θα εξεταστεί η ανάπτυξη μοντέλου object detection σε πραγματικό χρόνο.

V. ΣΥΜΠΕΡΑΣΜΑΤΑ

Ο σκοπός αυτής της μελέτης ήταν να αναπτύξει ένα μοντέλο CNN-LSTM που να πραγματοποιεί αναγνώριση ανθρώπινης δραστηριότητας σε πραγματικό χρόνο χρησιμοποιώντας skeletal και object info. Ενώ το μοντέλο είναι πραγματικό χρόνου η απόδοσή του είναι πολύ χαμηλή για πραγματική χρήση. Περαιτέρω έρευνα πρέπει να πραγματοποιηθεί για τη βελτίωση του μοντέλου.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Aganian, D. *et al.* (2023) *How object information improves skeleton-based human action recognition in assembly tasks*, *arXiv.org*. Available at: https://arxiv.org/abs/2306.05844 (Accessed: 30 September 2024).

[2] (No date a) *A systematic literature review on vision based gesture recognition techniques | request PDF*. Available at: https://www.researchgate.net/publication/324805847_A_systematic_literature_review_on_vision_based_gesture_recognition_techniques (Accessed: 30 September 2024).

[3] Bazarevsky, V., Grishchenko, I., Raveendran, K., Zhu, T., Zhang, F., & Grundmann, M. (2020, June 17). *BlazePose: On-device Real-time Body Pose tracking*. arXiv.org. https://arxiv.org/abs/2006.10204

[4] Author links open overlay panelRoshan Singh a *et al.* (2022) *Recent trends in Human activity recognition – a comparative study*, *Cognitive Systems Research*. Available at: https://www.sciencedirect.com/science/article/pii/S138904172200047X (Accessed: 30 September 2024).

[5] Ben-Shabat, Y. *et al.* (2023) *The IKEA ASM dataset: Understanding people assembling furniture through actions, objects and pose*, *arXiv.org*. Available at: https://arxiv.org/abs/2007.00394 (Accessed: 30 September 2024).

[6] (No date) *(PDF) real-time human action recognition using Deep Learning*. Available at: https://www.researchgate.net/publication/366523828_Real-Time_Human_Action_Recognition_Using_Deep_Learning (Accessed: 30 September 2024).