ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Σύστημα Υπολογιστικής Όρασης για την Αυτόματη Αξιολόγηση της Ικανότητας Δαχτυλοσυλλαβισμού στην Ελληνική Νοηματική Γλώσσα

Διπλωματική Εργασία

Άγγελος Παντόπουλος

Επιβλέπων: Γεράσιμος Ποταμιάνος

Σεπτέμβριος 2022

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Σύστημα Υπολογιστικής Όρασης για την Αυτόματη Αξιολόγηση της Ικανότητας Δαχτυλοσυλλαβισμού στην Ελληνική Νοηματική Γλώσσα

Διπλωματική Εργασία

Άγγελος Παντόπουλος

Επιβλέπων: Γεράσιμος Ποταμιάνος

Σεπτέμβριος 2022

UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

A Computer Vision System for Automatic Evaluation of Greek Sign Language Fingerspelling Proficiency

Diploma Thesis

Angelos Pantopoulos

Supervisor: Gerasimos Potamianos

September 2022

Εγκρίνεται από την Επιτροπή Εξέτασης:

Επιβλέπων **Γεράσιμος Ποταμιάνος**
Αναπληρωτής Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Μέλος **Νικόλαος Μπέλλας**
Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Μέλος **Αντώνιος Αργυρίου**
Αναπληρωτής Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Ευχαριστίες

Αρχικά θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Γεράσιμο Ποταμιάνο για την ευκαιρία να εκπονήσω τη παρούσα διπλωματική εργασία και την πολύτιμη καθοδήγηση καθ' όλη τη διάρκεια της συγγραφής και της ανάπτυξης της εφαρμογής. Ακόμα να ευχαριστήσω τη διδακτορική φοιτήτρια Κατερίνα Παπαδημητρίου και τους νηματιστές του προγράμματος SL-ReDu για τη συλλογή των ελληνικών δεδομένων που χρησιμοποιήθηκαν. Κλείνοντας θα ήθελα να εκφράσω την ευγνωμοσύνη μου στην οικογένεια μου και στους φίλους μου για την έμπρακτη υποστήριξη όλα αυτά τα χρόνια.

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ ΠΕΡΙ ΑΚΑΔΗΜΑΪΚΗΣ ΔΕΟΝΤΟΛΟΓΙΑΣ ΚΑΙ ΠΝΕΥΜΑΤΙΚΩΝ ΔΙΚΑΙΩΜΑΤΩΝ

«Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ρητά ότι η παρούσα διπλωματική εργασία, καθώς και τα ηλεκτρονικά αρχεία και πηγαίοι κώδικες που αναπτύχθηκαν ή τροποποιήθηκαν στα πλαίσια αυτής της εργασίας, αποτελεί αποκλειστικά προϊόν προσωπικής μου εργασίας, δεν προσβάλλει κάθε μορφής δικαιώματα διανοητικής ιδιοκτησίας, προσωπικότητας και προσωπικών δεδομένων τρίτων, δεν περιέχει έργα/εισφορές τρίτων για τα οποία απαιτείται άδεια των δημιουργών/δικαιούχων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον και πληρούν τους κανόνες της επιστημονικής παράθεσης. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο, αρχεία ή/και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή. Δηλώνω επίσης ότι τα αποτελέσματα της εργασίας δεν έχουν χρησιμοποιηθεί για την απόκτηση άλλου πτυχίου. Αναλαμβάνω πλήρως, ατομικά και προσωπικά, όλες τις νομικές και διοικητικές συνέπειες που δύναται να προκύψουν στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής».

Ο Δηλών

Άγγελος Παντόπουλος

Περίληψη

Η νοηματική γλώσσα αποτελεί το κύριο μέσο επικοινωνίας των κωφών και των ατόμων με προβλήματα στην ακοή. Το πρόβλημα της αναγνώρισης της νοηματικής γλώσσας και παράλληλα του δαχτυλοσυλλαβισμού αποτελεί συνδυασμό της επιστήμης της υπολογιστικής όρασης και της επιστήμης της βαθιάς μάθησης. Οι δυσκολίες κατά την αναγνώριση της νοηματικής γλώσσας από τον υπολογιστή συναντιούνται κατά την εξακρίβωση της ανθρώπινης κίνησης, την αναγνώριση των γραμματικών συμβολισμών και τον υπολογισμό της σωστής νοηματικής ερμηνείας με τη χρήση βαθιάς μάθησης. Στην παρούσα διπλωματική εργασία εστιάζουμε στην ανάπτυξη ενός μοντέλου αναγνώρισης δαχτυλοσυλλαβισμού σύμφωνα με το αλφάβητο της ελληνικής νοηματικής γλώσσας μέσα από την τεχνολογία της υπολογιστικής όρασης και την εφαρμογή του σε μια εκπαιδευτική πλατφόρμα αυτοαξιολόγησης αξιοποιώντας βιβλιοθήκες βαθιάς μάθησης των τελευταίων ετών.

Στο πρώτο στάδιο της διπλωματικής εξετάζεται η ανίχνευση χεριών και η εξαγωγή χρήσιμης πληροφορίας από τα δεδομένα που έχουμε στην διάθεση μας. Στην συνέχεια εξετάζεται η χρήση αρχιτεκτονικών συνελικτικών και επαναληπτικών νευρωνικών δικτύων για την πρόβλεψη γραμμάτων σύμφωνα με την πληροφορία που δίνεται από την κάμερα του υπολογιστή. Τέλος παρουσιάζεται η τεχνολογία ανάπτυξης του εργαλείου αυτοαξιολόγησης και η χρήση του ως εκπαιδευτικό εργαλείο εκμάθησης της ελληνικής νοηματικής γλώσσας.

Abstract

Sign language constitutes the primary means of communication for the deaf and the hearing impaired people. The problem of sign language recognition and fingerspelling combines the science of computer vision and the science of deep learning. The difficulties found in sign language recognition consist of verifying human motion, the recognition of letter symbols, and determing the signing meaning with the use of deep learning technology. In this thesis, we focus on developing a model to recognize fingerspelling based on the Greek signing language alphabet with computer vision tools and its use in order to develop an educational self-testing platform by employing deep learning libraries developed in recent years.

In the first part of the thesis, we study the hand detection system and the extraction of useful information from our available data. Next we examine the implementation of convolutional and recurrent neural networks for the letter prediction procedure according to the information captured by the computer's webcam. Lastly, we present the technology used in developing the self-evaluation application, as a Greek sign language educational tool.

Πίνακας Περιεχομένων

Ευχαριστίες **ix**

Περίληψη **xii**

Abstract **xiii**

Πίνακας Περιεχομένων **xv**

Κατάλογος σχημάτων **xvii**

Συντομογραφίες **xix**

1 Εισαγωγή **1**

1.1 Συνεισφορά 1

1.2 Δομή διπλωματικής 2

2 Επισκόπηση Σχετικής Βιβλιογραφίας **3**

2.1 Μέθοδοι αναγνώρισης χειρονομιών 3

2.1.1 Μέθοδοι βαθιάς μάθησης 3

2.1.2 Παραδοσιακές μέθοδοι μηχανικής μάθησης 5

2.2 Εργαλεία εκμάθησης νοηματικής γλώσσας 7

3 Μοντέλα Βαθιάς Μάθησης **11**

3.1 Συνελικτικά Νευρωνικά Δίκτυα 11

3.1.1 Εισαγωγή στα Συνελικτικά Νευρωνικά Δίκτυα 11

3.1.2 Επίπεδο συνέλιξης 12

3.1.3 Επίπεδο συσσώρευσης 12

3.1.4 Πλήρως συνδεδεμένα επίπεδα 13

3.2 MobileNets 13
3.3 Επαναληπτικά Νευρωνικά Δίκτυα 15
3.3.1 Δίκτυα Μακράς-Βραχείας Μνήμης 16
3.4 Μοντέλα στη διπλωματική 17

4 Σύνολο Δεδομένων Διπλωματικής 19

5 Ανάπτυξη του Συστήματος Αναγνώρισης 21
5.1 Μεθοδολογία 21
5.2 Εξαγωγή πληροφορίας 21
5.3 Εκπαίδευση μοντέλου CNN 23
5.4 Εκπαίδευση μοντέλου RNN 25
5.5 Αποτελέσματα 25

6 Περιγραφή Γραφικού Περιβάλλοντος 29
6.1 Φιλοσοφία εφαρμογής 29
6.2 Δυνατότητες εφαρμογής 29
6.2.1 Τεχνολογίες και εργαλεία που χρησιμοποιήθηκαν 30
6.3 Παρουσίαση εφαρμογής 30
6.3.1 Αρχική διεπαφή 30
6.3.2 Εμφάνιση αποτελέσματος σε εισαγωγή λάθος απάντησης 31
6.3.3 Εμφάνιση αποτελέσματος σε εισαγωγή σωστής απάντησης 31
6.3.4 Εμφάνιση τελικού αποτελέσματος 32

7 Συμπεράσματα 35

Βιβλιογραφία 37

Κατάλογος Σχημάτων

2.1 Αρχιτεκτονική CNN που αναπτύχθηκε στο [1] 4

3.1 Θεμελιώδεις επίπεδα CNN. Σχήμα από [2] 11
3.2 Γραφική περιγραφή του επιπέδου συνέλιξης. Σχήμα από [3] 12
3.3 Περιγραφική διαδικασία λειτουργίας του επιπέδου συσσώρευσης 13
3.4 Γραφική αναπαράσταση πλήρως συνδεδεμένου επιπέδου. Σχήμα από [4] . . 14
3.5 Συνέλιξη βάθους στην αρχιτεκτονική MobileNet. Σχήμα από [5] 15
3.6 Σχηματική απεικόνιση 'ξεδιπλώματος' RNN. Σχήμα από [6] 16
3.7 Αναπαράσταση LSTM. Σχήμα από [7] 17

4.1 Στιγμιότυπα του ελληνικού συνόλου δεδομένων 19
4.2 Λέξεις με τις περισσότερες εμφανίσεις στο σύνολο δεδομένων 20

5.1 Διάγραμμα γενικής μεθοδολογίας που χρησιμοποιήθηκε 22
5.2 Σημεία αναγνώρισης χεριού από τη βιβλιοθήκη MediaPipe. Εικόνα από [8] 23
5.3 Εξαγωγή πληροφορίας για το γράμμα Η 23
5.4 Εξαγωγή πληροφορίας για το γράμμα Κ 24
5.5 Αναπαράσταση λειτουργίας του CNN στο σύνολο δεδομένων 24
5.6 Αναπαράσταση λειτουργίας του συνελικτικού νευρωνικού δικτύου. Διάγραμμα τροποποιημένο από [9] 25
5.7 Αναπαράσταση λειτουργίας του επαναληπτικού νευρωνικού δικτύου. Διάγραμμα από [9] 25
5.8 Διάγραμμα ακρίβειας ανά epoch για τα δεδομένα εκπαίδευσης 26
5.9 Διάγραμμα απώλειας ανά epoch για τα δεδομένα εκπαίδευσης 26
5.10 Διάγραμμα ακρίβειας ανά epoch για τα δεδομένα δοκιμής 27
5.11 Διάγραμμα απώλειας ανά epoch για τα δεδομένα δοκιμής 27

6.1 Αρχικό περιβάλλον της εφαρμογής 30
6.2 Στιγμιότυπο δαχτυλοσυλλαβισμού 31
6.3 Στιγμιότυπο της εφαρμογής κατά την εισαγωγή λανθασμένης απάντησης . 32
6.4 Στιγμιότυπο της εφαρμογής κατά την εισαγωγή σωστής απάντησης 33
6.5 Εμφάνιση τελικού αποτελέσματος 33

Συντομογραφίες

3D	Three Dimensional (Τρισδιάστατο)
ASL	American Sign Language (Αμερικανική Νοηματική Γλώσσα)
BiLSTM	Bidirectional LSTM (Μνήμη Μακράς-Βραχείας Διάρκειας Δύο Κατευθύνσεων)
CNN	Convolutional Neural Network (Συνελικτικό Νευρωνικό Δίκτυο)
CTC	Connectionist Temporal Classification
DNA	Deoxyribonucleic Acid (Δεσοξυριβονουκλεϊκό Οξύ)
DNN	Deep Neural Network (Βαθύ Νευρωνικό Δίκτυο)
FPS	Frames Per Second (Καρέ Ανά Δευτερόλεπτο)
GRU	Gated Recurrent Units
GSL	Greek Sign Language (Ελληνική Νοηματική Γλώσσα)
HMM	Hidden Markov Model (Κρυφό Μαρκοβιανό Μοντέλο)
LSTM	Long Short-Term Memory (Μνήμη Μακράς-Βραχείας Διάρκειας)
PCA	Principal Component Analysis (Ανάλυση Κυρίων Συνιστωσών)
ReLU	Rectified Linear Unit (Ανορθωμένη Γραμμική Συνάρτηση Ράμπας)
RGB	Red-Green-Blue (Κόκκινο-Πράσινο-Μπλε)
RNN	Recurrent Neural Network (Επαναληπτικό Νευρωνικό Δίκτυο)
SCRF	Segmental Conditional Random Field (Τμηματικό υπό Συνθήκη Τυχαίο Πεδίο)
sEMG	Surface Electromyograph (Ηλεκτρομυογράφημα Επιφάνειας)
SIFT	Scale Invariant Feature Transform
SVM	Support-Vector Machine (Μηχανή Διανυσμάτων Υποστήριξης)
VQ	Vector Quantization (Διανυσματική Κβαντοποίηση)

Κεφάλαιο 1

Εισαγωγή

Η σημαντική ανάπτυξη του υπολογιστικού υλικού των τελευταίων δεκαετιών και η συνεχής προσπάθεια των ερευνητών να αναπτύξουν συστήματα τεχνητής νοημοσύνης έχουν οδηγήσει στην άνθηση του κλάδου της υπολογιστικής όρασης και μαζί με αυτήν της αναγνώρισης της ανθρώπινης κίνησης. Τα συστήματα αναγνώρισης δαχτυλοσυλλαβισμού που αναπτύσσονται τα τελευταία έτη έχουν ως στόχο την βελτίωση της επικοινωνίας των ατόμων με ειδικές ανάγκες και τη διαδραστική εκμάθηση των νοηματικών γλωσσών. Τα πρώτα συστήματα αναγνώρισης χειρομορφών απαιτούσαν τη χρήση εξωτερικών αισθητήρων ανίχνευσης κίνησης ή τη χρήση ειδικά χρωματισμένων γαντιών προκειμένου να γίνει η ανίχνευση των χεριών. Σήμερα η τεχνολογία της βαθιάς μάθησης και η επιστήμη των δεδομένων επιτρέπουν στους ερευνητές τη δημιουργία μοντέλων στιγμιαίας αναγνώρισης χειρονομιών και κίνησης μέσα από κάμερες χαμηλού κόστους. Στόχος της παρούσας διπλωματικής αποτελεί η ανάπτυξη ενός μοντέλου αναγνώρισης δαχτυλοσυλλαβισμού της ελληνικής νοηματικής γλώσσας και η εφαρμογή του σε μια εκπαιδευτική πλατφόρμα αυτοαξιολόγησης.

1.1 Συνεισφορά

Η συνεισφορά της διπλωματικής συνοψίζεται ως εξής:

1. Μελετήθηκαν σε βάθος οι αρχιτεκτονικές των συνελικτικών και επαναληπτικών νευρωνικών δικτύων και οι εφαρμογές τους σε σύγχρονα συστήματα βαθιάς μάθησης.

2. Υλοποιήθηκαν συστήματα εξαγωγής πληροφορίας για την κίνηση και θέση των ανθρώπινων χεριών με στόχο την τροφοδότηση τους σε μοντέλα τεχνητής νοημοσύνης.

3. Ενσωματώθηκαν σε συνδυασμό οι αρχιτεκτονικές των MobileNet και Keras RNN για την ανάλυση των δεδομένων της ελληνικής νοηματικής βάσης δεδομένων και τη δημιουργία προβλέψεων σε ζωντανό χρόνο.

4. Αναπτύχθηκε μια εφαρμογή αυτοαξιολόγησης δαχτυλοσυλλαβισμού με βάση το ελληνικό αλφάβητο σε ζωντανό χρόνο, με στόχο την εκπαίδευση και διάδοση της ελληνικής νοηματικής γλώσσας.

1.2 Δομή διπλωματικής

Η δομή της διπλωματικής εργασίας ορίζεται ως εξής:

1. Το κεφαλαίο 2 περιγράφει ερευνητικές εργασίες και μεθοδολογίες που έχουν χρησιμοποιηθεί σε σχετικές εργασίες των τελευταίων ετών.

2. Το κεφάλαιο 3 αναφέρει τις βασικές δομές των νευρωνικών δικτύων στην επιστήμη της βαθιάς μάθησης και τα μοντέλα που χρησιμοποιήθηκαν στη διπλωματική.

3. Το κεφάλαιο 4 αναλύει το σύνολο δεδομένων της ελληνικής νοηματικής γλώσσας που χρησιμοποιήθηκε για την εκπαίδευση του μοντέλου.

4. Το κεφάλαιο 5 περιγράφει τη μεθοδολογία ανάλυσης των δεδομένων, τις βιβλιοθήκες αναγνώρισης χεριών και τους τρόπους χρήσης των νευρωνικών δικτύων κατά την εκπαίδευση του μοντέλου.

5. Το κεφάλαιο 6 παρουσιάζει τη τελική εφαρμογή αυτοαξιολόγησης που αναπτύχθηκε στα πλαίσια της διπλωματικής, τις δυνατότητες της και το γραφικό περιβάλλον της.

6. Στο 7ο κεφάλαιο αναφέρονται τα συμπεράσματα της διπλωματικής, τα σχόλια του αποτελέσματος και πιθανές μελλοντικές επεκτάσεις της εφαρμογής.

Κεφάλαιο 2

Επισκόπηση Σχετικής Βιβλιογραφίας

Η εξέλιξη της επιστήμης της μηχανικής μάθησης τα τελευταία χρόνια και η συνεχής πρόοδος στο πεδίο της αναγνώρισης ανθρώπινων κινήσεων έχει οδηγήσει πολλούς ερευνητές να ασχοληθούν με την αναγνώριση δαχτυλοσυλλάβισης. Σε αυτή την ενότητα θα παρουσιαστούν μερικά από τα εργαλεία και τις μεθόδους που χρησιμοποιούνται σήμερα στο τομέα αυτό.

2.1 Μέθοδοι αναγνώρισης χειρονομιών

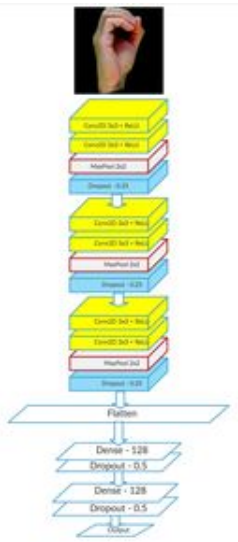
2.1.1 Μέθοδοι βαθιάς μάθησης

Το 2017 στο [1] προτάθηκε η ανάπτυξη ενός συνελικτικού νευρωνικού δικτύου (CNN) με στόχο την αναγνώριση των γραμμάτων της αμερικανικής νοηματικής γλώσσας (ASL) και των αριθμητικών ενδείξεων μέχρι τον αριθμό 10. Η ακρίβεια που προκύπτει σύμφωνα με τους ερευνητές είναι 97% για τις αριθμητικές ενδείξεις και 82.5% για τις αλφαβητικές.

Στο [10] το 2015 παρουσιάστηκε μια εκδοχή συστήματος αναγνώρισης αγγλικής δαχτυλοσυλλάβισης βασισμένο στην ASL με τη μέθοδο της βαθιάς μάθησης. Η ερευνητική τους δημοσίευση χρησιμοποιεί εικόνες βάθους για να εκπαιδεύσει ένα CNN. Η προσέγγιση που χρησιμοποιήθηκε αναγνωρίζει το αγγλικό αλφάβητο και τις αριθμητικές ενδείξεις πέραν των γραμμάτων J και Z τα οποία δεν υποστηρίζονταν από τη μέθοδο που επιλέχθηκε. Η ακρίβεια του συστήματος προσεγγίζει το 85% σε δοκιμές που πραγματοποιήθηκαν.

Παρόμοια την ίδια χρονιά μια ακόμα εκδοχή ενός CNN παρουσιάστηκε στο [11]. Για την εκπαίδευση του συστήματος μετάφρασης δαχτυλοσυλλάβισης χρησιμοποιήθηκε η αρχιτεκτονική GoogLeNet με χρήση των δημόσιων δεδομένων ILSVRC2012 και των δεδομένων

που προσέφεραν τα πανεπιστήμια των Surrey [12] και Massey [13]. Στα προαναφερθέντα δεδομένα εφαρμόστηκε η τεχνική του transfer learning, ωστόσο το μοντέλο αδυνατεί να αναγνωρίσει το γράμμα J, ενώ έχει εκπαιδευτεί στα γράμματα A-Y όπου και παρουσιάζει ακρίβεια της τάξης του 91.63%.



Σχήμα 2.1: Αρχιτεκτονική CNN που αναπτύχθηκε στο [1]

Το 2019 στα πλαίσια του [14] προτάθηκε η ανάπτυξη ενός μοντέλου αναγνώρισης δακτυλοσυλλαβισμού στην ASL με τη χρήση ενός επαναληπτικού νευρωνικού δικτύου (RNN). Στο σύστημα χρησιμοποιήθηκε αρχιτεκτονική με συνελικτικούς κωδικοποιητές/αποκωδικοποιητές με μηχανισμούς προσοχής και συνάρτηση τετραγωνικής ευθυγράμμισης. Για την εκπαίδευση του μοντέλου έγινε χρήση των δεδομένων TTIC/UChicago [15]. Το μοντέλο παρουσίασε αυξημένη ακρίβεια σε σύγκριση με προηγούμενες αρχιτεκτονικές που είχαν προταθεί πάνω στην ίδια συλλογή δεδομένων.

Το 2018 στο [16] προτάθηκε ένα σύστημα αναγνώρισης δακτυλοσυλλάβισης στην ASL βασισμένο σε κωδικοποιητές/αποκωδικοποιητές με μηχανισμούς προσοχής. Για την εκπαίδευση του μοντέλου έγινε χρήση της μεγαλύτερης έως τότε συλλογής δακτυλοσυλλάβισης από διάφορες πηγές του διαδικτύου με πολύπλοκα περιβάλλοντα και μη σταθερή ποιότητα φωτισμού. Η συνάρτηση CTC που επιλέχθηκε για την εκπαίδευση του RNN με τα τελικά δεδομένα του ChicagoFSVid εμφάνισε ακρίβεια της τάξης του 42%.

Το 2017 στα πλαίσια του [15] αναπτύχθηκε ένα σύστημα αναγνώρισης συμβολισμών ASL με τη χρήση ενός βαθιού νευρωνικού δικτύου (DNN) και τη τεχνική του τμηματικού υπό συνθήκη τυχαίου πεδίου (SCRF) χωρίς τη χρήση λεξικού. Στα δεδομένα προστέθηκαν χειροκίνητα τα όρια συμβολισμών ως πληροφορία για τη γρηγορότερη αναγνώριση από το

σύστημα. Η συγκεκριμένη πρόταση παρουσίασε ακρίβεια 92% στις περιπτώσεις όπου διερμηνείς σημείωναν τα όρια των συμβολισμών στα δεδομένα και 83% σε περιπτώσεις όπου εμφανίζονταν δύο ή περισσότεροι διερμηνείς στα δεδομένα.

Στο [17] το 2017 προτάθηκε ο σχεδιασμός ενός συστήματος αναγνώρισης ASL με τη χρήση ενός μηχανισμού αυτο-κωδικοποιητή για την εξαγωγή των κύριων χαρακτηριστικών των δαχτυλοσυλλαβισμών. Η αρχιτεκτονική RNN κωδικοποιητών/αποκωδικοποιητών με μηχανισμούς προσοχής αύξησε σε μεγάλο βαθμό την ακρίβεια του συστήματος σε σχέση με διαφορετικές υλοποιήσεις που είχαν προταθεί τα προηγούμενα χρόνια, ειδικότερα σε περιπτώσεις δεδομένων χωρίς χειροκίνητη οριοθέτηση των συμβολισμών.

2.1.2 Παραδοσιακές μέθοδοι μηχανικής μάθησης

Σε αυτές τις δημοσιεύσεις οι συγγραφείς κάνουν χρήση κλασσικής μηχανικής μάθησης όπως ταξινομητές μηχανών διανυσμάτων υποστήριξης (SVM), κρυφά Μαρκοβιανά μοντέλα (HMM) και ρηχά νευρωνικά δίκτυα.

Το 2007 στο [18] προτάθηκε ένα σύστημα αναγνώρισης χειρονομιών χρησιμοποιώντας φίλτρα Gabor και ομαδοποίηση Fuzzy-c-mean βασισμένο στο αλφάβητο της ASL. Στο σύστημα που παρουσιάστηκε έγινε χειροκίνητη αποκοπή χεριών από εικόνες με ομαλό και καλά φωτισμένο φόντο προκειμένου να επιβεβαιωθεί το καλύτερο δυνατό αποτέλεσμα. Στις εικόνες που προέκυψαν έγινε εφαρμογή φίλτρων Gabor ακολουθούμενη από ανάλυση κυρίων συνιστωσών (PCA) με στόχο τη μείωση δισδιάστατου θορύβου και την εξαγωγή των κύριων χαρακτηριστικών που χρησιμοποιήθηκαν κατά την εκπαίδευση του Fuzzy-c-mean αλγορίθμου. Η αποτελεσματικότητα του μοντέλου σύμφωνα με τις δοκιμές που πραγματοποιήθηκαν προσέγγισε το 93.2%.

Στο σύστημα που δημοσιεύθηκε το 2015 στο [19] έγινε χρήση ηλεκτρομυογραφημάτων επιφάνειας (sEMG) με ταξινομητές SVM με στόχο την αναγνώριση των γραμμάτων της ASL. Τα σήματα sEMG που χρησιμοποιήθηκαν προέκυψαν από την τοποθέτηση εξωτερικών αισθητήρων στα χέρια των διερμηνέων. Από τα δεδομένα των αισθητήρων έγινε εξαγωγή μεγεθών όπως η μέση απόλυτη τιμή, η μέση αλλαγή εύρους και το απλό ολοκλήρωμα τετραγώνου με στόχο την εκπαίδευση του ταξινομητή SVM. Το σύστημα στις σχετικές δοκιμές που διεξήχθησαν κατάφερε να προσεγγίσει ακρίβεια 91.73%.

Το 2013 στο [20] σχεδιάστηκε ένα σύστημα αναγνώρισης δαχτυλοσυλλαβισμού βασισμένο στο αλφάβητο της ASL κάνοντας χρήσης της τεχνολογίας Kinect. Στο συγκεκριμένο

σύστημα η κάμερα βάθους Kinect χρησιμοποιήθηκε για την αναγνώριση θέσης του χεριού. Για την ανάλυση των δεδομένων RGB που καταγράφηκαν έγινε χρήση του αλγορίθμου SIFT (Scale Invariant Feature Transform). Τα τελικά δεδομένα που προέκυψαν από την καταγραφή 120,000 εικόνων τροφοδοτούνται σε ένα ταξινομητή SVM για εκπαίδευση. Η αποτελεσματικότητα του συστήματος όταν εξετάστηκε σε πραγματικές εικόνες με συμβολισμούς ASL έφτασε την ακρίβεια του 91.26%.

Το 2011 στα πλαίσια του [21] αναπτύχθηκε εφαρμογή αναγνώρισης δαχτυλοσυλλαβισμού βασισμένη στο αμερικανικό αλφάβητο με στόχο τη χρήση του σε εφαρμογές και βιντεοπαιχνίδια. Το σύστημα μέσω της τεχνικής Bag-of-Features έχει τη δυνατότητα αναγνώρισης δέρματος και της θέση του χεριού, για την εξαγωγή των σημαντικών χαρακτηριστικών γίνεται εφαρμογή του αλγορίθμου SIFT, και στη συνέχεια πραγματοποιείται διανυσματικός κβαντισμός (VQ) όπου σημειώνονται τα κύρια σημεία της χειρομορφής χρησιμοποιώντας ομαδοποίηση k-means. Τα τελικά διανύσματα που προκύπτουν τροφοδοτούνται σε έναν ταξινομητή SVM, ο οποίος εκπαιδεύεται στη διανυσματική θέση των δαχτύλων. Η ακρίβεια της συγκεκριμένης μεθόδου προσεγγίζει το 90%.

Στο [22] ερευνητές ανέπτυξαν το δικό τους σύστημα αναγνώρισης δαχτυλοσυλλαβισμού βασισμένο στην ASL. Η προσέγγισή τους αποτελείται από την εφαρμογή Γκαουσιανού μέσου για την εξαγωγή της χειρονομίας από τα δεδομένα των καταγεγραμμένων εικόνων. Οι σχηματικές περιγραφές και οι κινήσεις των χεριών εξάγονται ως διανύσματα και τροφοδοτούνται σε ένα νευρωνικό δίκτυο 3 επιπέδων για την εκπαίδευσή του. Η ακρίβεια του μοντέλου που περιγράφηκε σύμφωνα με δοκιμές αγγίζει το 100%.

Το 1999 στο [23] προτάθηκε η δημιουργία μοντέλου αναγνώρισης δαχτυλοσυλλαβισμού βασισμένο και αυτό στην ASL κάνοντας χρήση τεχνικών νευρωνικών δικτύων. Για την εισαγωγή των δεδομένων έγινε χρήση ειδικού γαντιού με ειδικούς χρωματικούς κωδικούς για τη συλλογή πρόσθετων πληροφοριών. Στα δεδομένα που συλλέχθηκαν, μεταξύ αυτών και δεδομένα της ιαπωνικής νοηματικής γλώσσας, εφαρμόστηκε μορφολογική ανάλυση κύριων συνιστωσών με στόχο την εξαγωγή των κύριων χαρακτηριστικών. Το νευρωνικό δίκτυο που χρησιμοποιήθηκε αποτελούνταν από 3 επίπεδα πρόσθιας τροφοδότησης με σιγμοειδή συνάρτηση ενεργοποίησης. Η ακρίβεια του μοντέλου σύμφωνα με τις δοκιμές ήταν 89.06%.

Το 1996 στο [24] παρουσιάστηκε η ανάπτυξη συστήματος αναγνώρισης του αλφαβήτου της ASL σε εικόνες με σύνθετο και περίπλοκο φόντο. Η μέθοδός τους εφαρμόζει φίλτρα Gabor με στόχο την εξαγωγή των κύριων χαρακτηριστικών της αλφαβητικής ένδειξης. Στην

συνέχεια πραγματοποίησαν την τεχνική ελέγχου ομοιότητας Elastic Bunch- Graph Matching για την αναγνώριση της ένδειξης. Ως δεδομένα χρησιμοποιήθηκαν 10 αλφαβητικές ενδείξεις της ASL καταγεγραμμένες από 24 άτομα σε 3 διαφορετικά υπόβαθρα με διαφορετικά επίπεδα φωτισμού. Η ακρίβεια του μοντέλου ύστερα από διεξαγωγή ελέγχου σε 239 εικόνες προσέγγισε το 86.2%.

Τη δική τους πρόταση για ένα σύστημα ανεπτυγμένο με τη χρήση ενός Bayesian δικτύου παρουσίασαν ερευνητές στο [25] βασισμένοι στο αλφάβητο της βρετανικής νοηματικής γλώσσας. Για την ανάλυση των δεδομένων που προήλθαν από κάμερες υπολογιστή (webcam) χαμηλής ποιότητας χρησιμοποιήθηκαν η τεχνική του Histogram of Gradients και το μοντέλο HMM. Η μέθοδος στοχεύει στην αναγνώριση της κίνησης της αλφαβητικής ένδειξης και όχι του σχήματος της σε αντίθεση με προηγούμενες δημοσιεύσεις. Σύμφωνα με τις δοκιμές των ερευνητών το σύστημα προσεγγίζει ακρίβεια 98.9%.

Στα πλαίσια του [26] αναπτύχθηκε ένα σύστημα αναγνώρισης δαχτυλοσυλλάβισης ASL σε πραγματικό χρόνο με τη χρήση ενός HMM. Για την εισαγωγή των δεδομένων της αλφαβητικής ένδειξης στο σύστημα ο χρήστης καλείται να φορέσει ένα γάντι συγκεκριμένων χρωματικών ενδείξεων προκειμένου η μετάφραση σε πραγματικό χρόνο να καταστεί δυνατή. Η ακρίβεια της συγκεκριμένης αρχιτεκτονικής για δοκιμές που έγιναν με το προαναφερθέν γάντι αγγίζει το 99.2%.

Στο [27] το 2013 αναπτύχθηκε ένα σύστημα αναγνώρισης δαχτυλοσυλλαβισμού βασισμένου στην ASL με τη χρήση ημιμαρκοβιανής συνθήκης τυχαίων πεδίων. Για την επεξεργασία των δεδομένων έγινε χειροκίνητη οριοθέτηση των συμβολισμών από διερμηνείς. Για τον υπολογισμό του κατάλληλου γράμματος χρησιμοποιήθηκε η μέθοδος SCRF. Το ποσοστό λανθασμένης πρόβλεψης που εμφάνισε η συγκεκριμένη πρόταση ήταν 11.6%.

2.2 Εργαλεία εκμάθησης νοηματικής γλώσσας

Σε αυτήν την ενότητα περιγράφονται δημοσιεύσεις των τελευταίων ετών γύρω από την ανάπτυξη εφαρμογών και συστημάτων με στόχο την εκπαίδευση νοηματικής γλώσσας ανά τον κόσμο.

Στο [28] αναπτύχθηκε το εκπαιδευτικό παιχνίδι MatLIBRAS Racing με στόχο τη διαδραστική εκμάθηση της βραζιλιάνικης νοηματικής γλώσσας από τους μαθητές. Το παιχνίδι είναι σχεδιασμένο με γνώμονα την καλύτερη εμπειρία των χρηστών, την πρόσβαση του παιχνι-

διού από μεγάλο αριθμό πλατφορμών και τη χρήση του λογισμικού του ως εργαλείο. Όσον αφορά τον τρόπο παιχνιδιού ο παίκτης καλείται να νοηματίσει στην κάμερα της συσκευής με τα χέρια του τους αριθμούς 0-9 προκειμένου να διαχειριστεί την κίνηση ενός αμαξιού κατά τη διάρκεια αγώνα. Σύμφωνα με έρευνες που διεξήγαγε η ομάδα του πανεπιστήμιου το παιχνίδι παρουσίασε θετικά στοιχεία ως εργαλείο μάθησης.

Το 2019 στο [29] παρουσιάστηκε το εκπαιδευτικό εργαλείο εκμάθησης νοηματικής γλώσσας SiLearn. Το SiLearn αποτελεί μια εφαρμογή σχεδιασμένη για κινητά τηλέφωνα και λειτουργεί ως ένα ψηφιακό αποθετήριο μεταφρασμένων λέξεων. Με τη χρήση τεχνητής νοημοσύνης η εφαρμογή διαθέτει την ικανότητα αναγνώρισης κειμένου και φυσικών αντικειμένων με στόχο την άμεση παρουσίαση τους ως νοήματα. Το τελικό αποτέλεσμα είναι σχεδιασμένο ως εργαλείο χρήσης για παιδιά, γονείς και καθηγητές, ενώ αρχικές έρευνες που πραγματοποιήθηκαν με βάση την ινδική νοηματική γλώσσα αποδεικνύουν τη θετική επίδραση στην εκμάθηση νοηματικής γλώσσας και βελτίωση των δεξιοτήτων των μαθητών.

Στα πλαίσια ανάπτυξης εκπαιδευτικής εφαρμογής νοηματικής γλώσσας [30] παρουσιάστηκε το σύστημα Virtual Sign. Το Virtual Sign αποτελεί ένα αμφίδρομο μεταφραστή μεταξύ της πορτογαλικής νοηματικής γλώσσας και πορτογαλικού κειμένου. Ο μεταφραστής για την είσοδο νοηματικών δεδομένων βασίζεται σε δύο συσκευές, την κάμερα Microsoft Kinect και ειδικά γάντια 5DT sensor προκειμένου να λάβει τα πλήρη δεδομένα που χρειάζεται. Το σύστημα του μεταφραστή χρησιμοποιεί νευρωνικά δίκτυα ως τεχνολογικό υπόβαθρο ενώ η ακρίβεια του μοντέλου αγγίζει το 90%. Στόχος της ομάδας για την τεχνολογία του Virtual Sign συνιστά η εγκατάστασή του σε σχολικές μονάδες βοηθώντας την επικοινωνία και βελτιώνοντας την εκπαίδευση των μαθητών με προβλήματα ακοής.

Την ίδια χρονιά στα πλαίσια ακαδημαϊκής έρευνας [31] αναπτύχθηκε μια εκπαιδευτική πλατφόρμα με στόχο την εκμάθηση της νοηματικής γλώσσας στους μαθητές της Μαλαισίας. Η εφαρμογή υιοθετεί ένα εύχρηστο και απλό γραφικό περιβάλλον και με τη χρήση ενός χαρακτήρα κινουμένων σχεδίων διδάσκει στους μαθητές δαχτυλοσυλλάβιση και νοηματικούς συμβολισμούς διαφόρων λέξεων. Στόχος των ερευνητών αποτελεί η δωρεάν και διαδραστική διδασκαλία νεαρών παιδιών γύρω από την τοπική νοηματική γλώσσα, ενώ θεωρούν πως η τεχνολογία τους μπορεί να συμπληρώσει τους υπάρχοντες δια ζώσης τρόπους εκπαίδευσης.

Στο [32] παρουσιάστηκε από ερευνητές μια εκδοχή ενός ψηφιακού εργαλείου εκμάθησης νοηματικής γλώσσας βασισμένο στο Σλοβενικό αλφάβητο. Το σύστημα που αναπτύχθηκε αποτελεί ένα διαδικτυακό αποθετήριο οπτικού υλικού όπου διερμηνείς παρουσιάζουν τις νο-

ηματικές κινήσεις της Σλοβενικής νοηματικής γλώσσας και παράλληλα συνιστά ένα μέσο αξιολόγησης όπου προσφέρονται διάφοροι τρόποι εξέτασης πάνω στα βίντεο που παρακολουθήθηκαν. Η πλατφόρμα χρησιμοποιεί πληθώρα τεχνικών λύσεων προκειμένου η πρόσβαση στο μεταφρασμένο υλικό και στα διαγωνίσματα να πραγματοποιείται από μεγάλο εύρος συσκευών.

Το 2012 στο [33] προτάθηκε ένα εκπαιδευτικό εργαλείο εκμάθησης νοηματικής γλώσσας με στόχο την εκμετάλλευση της δημοτικότητας των συσκευών με οθόνες αφής. Το σύστημα που παρουσιάστηκε ως εφαρμογή κινητού τηλεφώνου αποτελεί έναν ψηφιακό μεταφραστή νοηματικής γλώσσας με βάση την τεχνολογία X3D [34]. Συνοπτικά η εφαρμογή με την υιοθέτηση ενός ψηφιακού avatar δίνει τη δυνατότητα στο χρήστη να μεταφράσει κείμενο σε νοήματα σε πραγματικό χρόνο στοχεύοντας στη βελτίωση της επικοινωνίας των ατόμων με ειδικές ανάγκες. Ταυτόχρονα σύμφωνα με τους ερευνητές το τρισδιάστατο μοντέλο παρουσίασης των νοηματικών κινήσεων λειτουργεί και ως εκπαιδευτικό εργαλείο για τους χρήστες που επιθυμούν να μάθουν νέες χειρομορφές.

Στο [35] αναπτύσσεται ένα διαδραστικό εργαλείο εκμάθησης της ASL με το όνομα Funsigns. Το Funsigns μέσα από μια πληθώρα εκπαιδευτικών μεθόδων όπως τα εκπαιδευτικά βίντεο, οι αξιολογήσεις, η μάθηση από χαρακτήρες κινουμένων σχεδίων κ.α. στοχεύει στην εκμάθηση της νοηματικής γλώσσας σε παιδιά ηλικίας 2-5 που αντιμετωπίζουν προβλήματα ακοής. Η εφαρμογή χρησιμοποιεί τη θεωρία χρωμάτων και την υιοθέτηση παιδικών χαρακτήρων προκειμένου να κάνει τη μάθηση διαδραστική και ευχάριστη. Στόχος της πλατφόρμας είναι η προσθήκη τεχνολογίας επαυξημένης πραγματικότητας για αναγνώριση αντικειμένων από το άμεσο περιβάλλον του μαθητή και μετάφρασή τους στην ASL.

Το 2016 στο [36] παρουσιάστηκε η εφαρμογή εκμάθησης ASL με το όνομα SmartSignPlay. Η εφαρμογή κινητών τηλεφώνων SmartSignPlay δημιουργήθηκε με στόχο την εκμάθηση της ASL σε παιδιά με προβλήματα ακοής βασιζόμενη στη δημοτικότητα των συσκευών με οθόνες αφής. Μέσα από ειδικά διαμορφωμένα μαθήματα λεξικού και παρακολούθηση των κινήσεων του ψηφιακού avatar οι μαθητές καλούνται να εξασκήσουν τις γνώσεις τους μέσα από παιχνίδια και διαγωνίσματα που παρουσιάζονται από χαρακτήρες κινουμένων σχεδίων. Οι ερευνητές έχουν ως στόχο τη διαδραστική και ευχάριστη εκπαίδευση των παιδιών αλλά και των γονιών στην πρώτη τους επαφή με την ASL.

Στα πλαίσια του [37] παρουσιάστηκε το εκπαιδευτικό εργαλείο δαχτυλοσυλλαβισμού της ινδικής νοηματικής γλώσσας με το όνομα SignQuiz. Το SignQuiz αποτελεί μια διαδι-

κτυακή εφαρμογή αυτοαξιολόγησης των γνώσεων του χρήστη βασιζόμενο στην τεχνολογία της βαθιάς μάθησης για την αναγνώριση των χειρομορφών. Ο χρήστης κατά την διάρκεια της αξιολόγησης καλείται να σχηματίσει το γράμμα που φαίνεται στην οθόνη και βαθμολογείται ανάλογα με την απόδοσή του. Σύμφωνα με τη δημοσίευση τα αποτελέσματα των ερευνών κρίνουν πως το SignQuiz συνιστά αποδοτικότερη μέθοδο μάθησης από τη γραπτή πληροφορία.

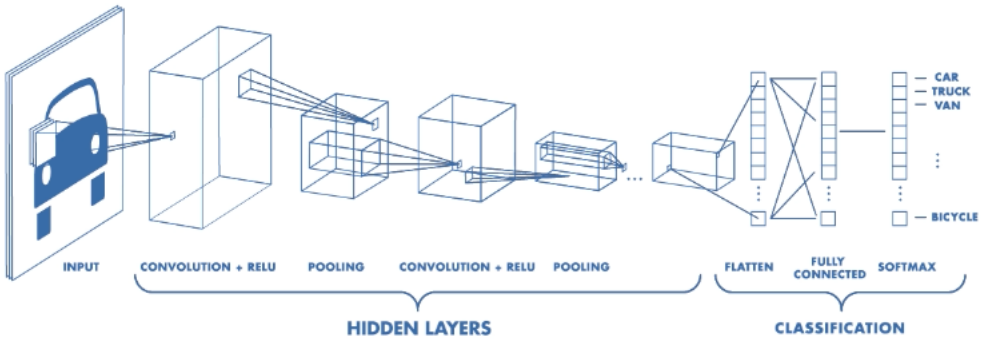
Κεφάλαιο 3

Μοντέλα Βαθιάς Μάθησης

3.1 Συνελικτικά Νευρωνικά Δίκτυα

3.1.1 Εισαγωγή στα Συνελικτικά Νευρωνικά Δίκτυα

Τα CNN αποτελούν τη δημοφιλέστερη αρχιτεκτονική νευρώνων στην επιστήμη της βαθιάς μάθησης. Χρησιμοποιούνται σημαντικά σε τομείς τεχνητής νοημοσύνης όπως η αναγνώριση προσώπου, η μελέτη κλιματικών συμπεριφορών, η ανάπτυξη μοντέλων γλώσσας κ.α. Στόχος των CNN είναι η εκμετάλλευση της τοπικής χωρικότητας και η μείωση των διαστάσεων των δεδομένων σύμφωνα με την ιδιότητα πως γειτονικά δεδομένα έχουν μεγαλύτερη σχέση από ότι τα μακρινά.



Σχήμα 3.1: Θεμελιώδεις επίπεδα CNN. Σχήμα από [2]

Η γενική δομή των CNN συντελείται από ένα η περισσότερα επίπεδα συνέλιξης σε πολλές περιπτώσεις ακολουθούμενα από επίπεδα υπερδειγματοληψίας (pooling layers) και πλήρως

συνδεδεμένα επίπεδα (fully connected layers) στην έξοδο του δικτύου (βλέπε Σχήμα 3.1).

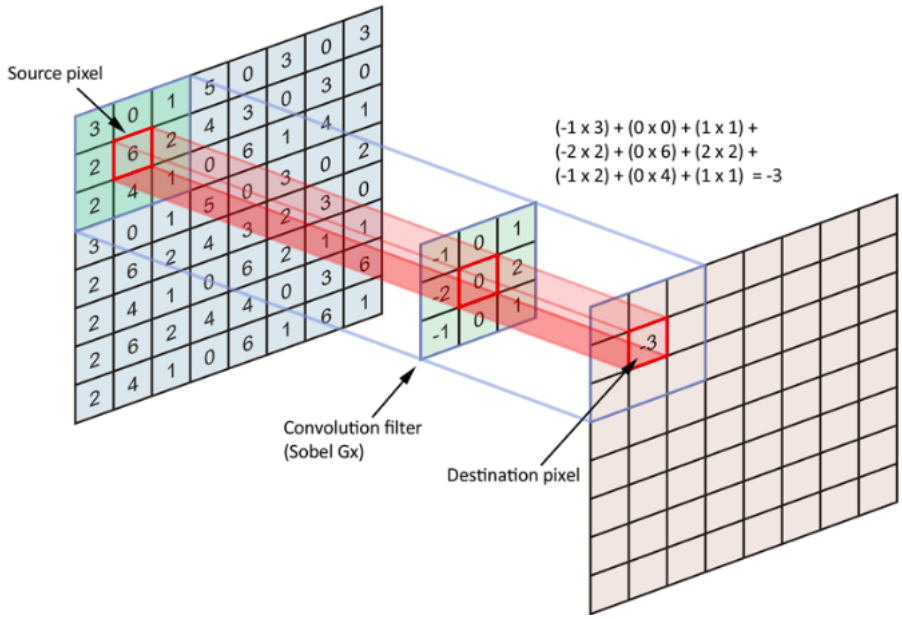
3.1.2 Επίπεδο συνέλιξης

Ο όρος συνέλιξη προκύπτει από τη μαθηματική πράξη συνδυασμού δύο συναρτήσεων για το σχηματισμό μιας τρίτης. Στο πλαίσιο των CNN η συνέλιξη χρησιμοποιείται στα δεδομένα εισόδου για την παραγωγή του χάρτη ενεργοποίησης μέσω μιας σειράς φίλτρων ή πυρήνων (kernel).

Η μαθηματική περιγραφή της τρισδιάστατης συνέλιξης στην περίπτωση εισόδου βίντεο γίνεται από τον τύπο:

$$S_l[i,j,k] = ReLU(b_l + \sum_{m=0}^{M_{l-1}} \sum_{n=0}^{N_{l-1}} \sum_{\tau=0}^{T_{l-1}} W_l[m,n,\tau,:] \circledast I[i-m, j-n, k-\tau, :]) \qquad (3.1)$$

όπου ReLU είναι η μη γραμμική συνάρτηση ενεργοποίησης που συνήθως χρησιμοποιείται στις αρχιτεκτονικές των CNN. Σχηματικά η μαθηματική διαδικασία που συμβαίνει στο επίπεδο συνέλιξης φαίνεται στο Σχήμα 3.2.



Σχήμα 3.2: Γραφική περιγραφή του επιπέδου συνέλιξης. Σχήμα από [3]

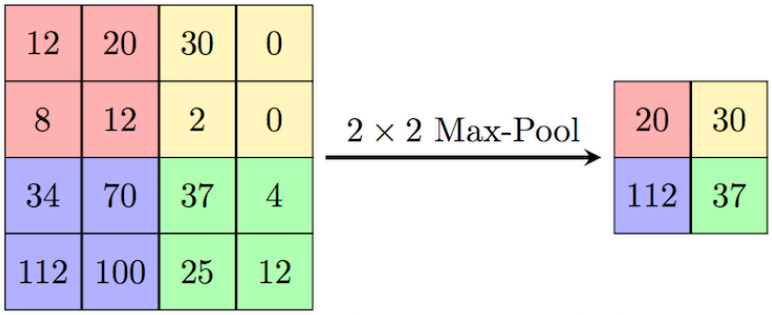
3.1.3 Επίπεδο συσσώρευσης

Στη σημαντική πλειονότητα των αρχιτεκτονικών CNN, προκειμένου να επιτευχθεί μείωση των διαστάσεων του μοντέλου και βελτιστοποίηση της απόδοσης του, υιοθετείται η

προσθήκη στρωμάτων συγκέντρωσης μετά τα επίπεδα συνέλιξης. Στα επίπεδα συσσώρευσης η έξοδος που προκύπτει από τη διαδικασία της συνέλιξης μετασχηματίζεται με τη χρήση ενός τελεστή συγκέντρωσης. Συνήθως ως τελεστής χρησιμοποιείται η συνάρτηση τοπικού μεγίστου, καθώς διατηρεί τη μέγιστη τιμή απόκρισης του φίλτρου. Η μαθηματική περιγραφή του επιπέδου συσσώρευσης περιγράφεται από την εξίσωση:

$$S_l^{Region} = max\{S_l[x]\} \quad \forall x \in Region \tag{3.2}$$

όπου S_l είναι η έξοδος του επιπέδου συνέλιξης. Η διαδικασία του επιπέδου συγκέντρωσης φαίνεται στο Σχήμα 3.3.



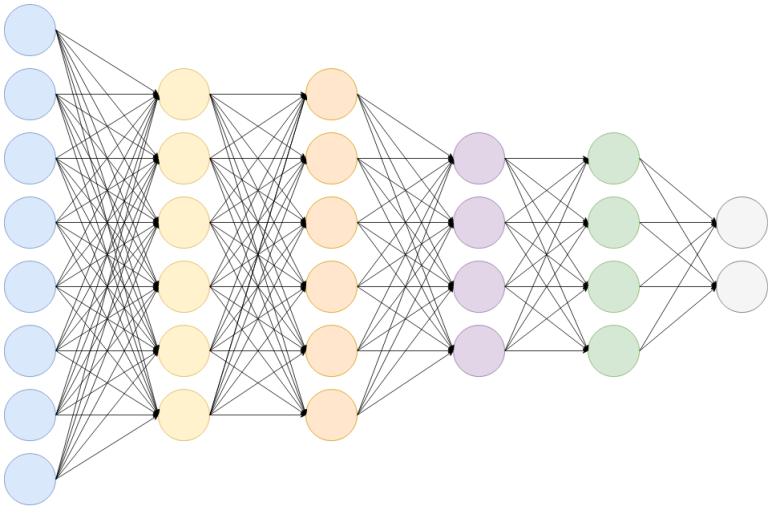
Σχήμα 3.3: Περιγραφική διαδικασία λειτουργίας του επιπέδου συσσώρευσης

3.1.4 Πλήρως συνδεδεμένα επίπεδα

Τα πλήρως συνδεδεμένα επίπεδα στις αρχιτεκτονικές των CNN αποσκοπούν στην σύνδεση όλων των εισόδων ενός επιπέδου με όλες τις εξόδους του προηγούμενου επιπέδου. Μέσω αυτής της ιδιότητας το μοντέλο ολοκληρώνει τη διαδικασία της ταξινόμησης. Η διαδικασία του μετασχηματισμού των δεδομένων εξόδου σε δεδομένα εισόδου και παράλληλα μείωση των διαστάσεων τους από το πλήρως συνδεδεμένο επίπεδο καλείται ισοπέδωση (flatten). Η συμπεριφορά του επιπέδου γραφικά παρουσιάζεται στο Σχήμα 3.4.

3.2 MobileNets

Από το 2012 όπου η αρχιτεκτονική AlexNet βραβεύτηκε στον διαγωνισμό του ImageNet ILSVRC 2012 τα CNN αποτελούν αναπόσπαστο κομμάτι της επιστήμης της βαθιάς μάθησης με πολλές αρχιτεκτονικές να εμφανίζουν σημαντική δημοτικότητα λόγω της υψηλής

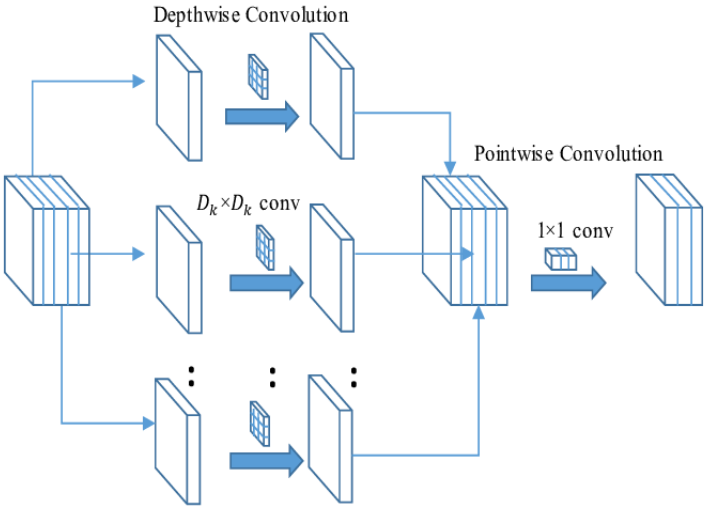


Σχήμα 3.4: Γραφική αναπαράσταση πλήρως συνδεδεμένου επιπέδου. Σχήμα από [4]

ακρίβειας που επιτυγχάνουν, όπως τα GoogleNet, τα VGGNet και τα ResNet. Προκειμένου τα προαναφερθέντα δίκτυα να προσεγγίσουν υψηλότερες ακρίβειες και καλύτερα αποτελέσματα έχουν υιοθετηθεί βαθύτερα επίπεδα συνέλιξης κάνοντας τις αρχιτεκτονικές πιο πολυπλοκες και υπολογιστικά ακριβότερες. Στην πραγματικότητα για εφαρμογές βαθιάς μάθησης σε πραγματικό χρόνο όπως η αναγνώριση νοηματικής γλώσσας,η ρομποτική και τα αυτοοδηγούμενα οχήματα απαιτούνται νευρωνικά δίκτυα που τρέχουν σε περιορισμένους πόρους.

Τα MobileNet [38] αποτελούν βαθιά συνελικτικά νευρωνικά δίκτυα που αναπτύχθηκαν από ερευνητές της Google με στόχο τη χρήση τους σε κινητές και ενσωματωμένες συσκευές για εφαρμογές υπολογιστικής όρασης. Η αρχιτεκτονική των MobileNets βασίζεται στην έννοια της διαχωρίσιμης συνέλιξης βάθους (depth-wise separable convolution). Η συμβατική λειτουργία συνέλιξης έχει ως αποτέλεσμα το φιλτράρισμα χαρακτηριστικών που βασίζονται στους συνελικτικούς πυρήνες και το συνδυασμό χαρακτηριστικών προκειμένου να παραχθεί μια νέα αναπαράσταση. Αυτά τα βήματα φιλτραρίσματος και συνδυασμού μπορούν να χωριστούν σε δύο βήματα χρησιμοποιώντας παραγοντικές συνελίξεις που ονομάζεται διαχωρίσιμη συνέλιξη από άποψη βάθους (βλέπε Σχήμα 3.5).

Αυτή η τροποποίηση του συνελικτικού νευρωνικού δικτύου επέτρεψε στα MobileNets να επιτύχουν ισοδύναμη ακρίβεια με οποιοδήποτε CNN υψηλής απόδοσης με καλύτερη αποτελεσματικότητα όσον αφορά την ταχύτητα και την κατανάλωση υπολογιστικών πόρων.



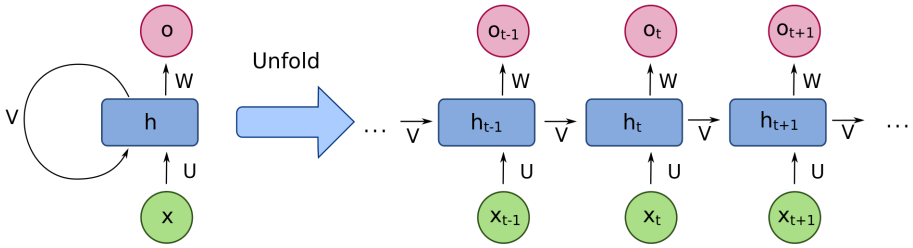
Σχήμα 3.5: Συνέλιξη βάθους στην αρχιτεκτονική MobileNet. Σχήμα από [5]

3.3 Επαναληπτικά Νευρωνικά Δίκτυα

Τα RNN συντελούν μια ειδική κατηγορία νευρωνικών δικτύων σχεδιασμένα να επεξεργάζονται αποτελεσματικά ακολουθιακά δεδομένα όπως βίντεο, ήχο, κείμενο, ακολουθίες DNA κ.α. Πιο συγκεκριμένα τα RNN διαθέτουν εσωτερική μνήμη που επιτρέπει σε προηγούμενες εξόδους να χρησιμοποιηθούν εκ νέου ως είσοδος ενώ παράλληλα μετασχηματίζει την είσοδο σύμφωνα με τις προηγούμενες εισόδους που δόθηκαν. Η επαναλαμβανόμενη αρχιτεκτονική των RNN οδηγεί στην εξάρτηση όλων των δεδομένων εισόδου που περιγράφεται από τη σχέση:

$$h_t = f(x_t, h_{t-1}) = f(x_t, f(x_{t-1}, h_{t-2})) = ... = f(x_t, f(x_{t-1}, ..., f(x_1, h_0)...) \quad (3.3)$$

όπου x_t είναι η είσοδος τη χρονική στιγμή t, $f(x)$ η συνάρτηση που περιγράφει την μετατροπή της εισόδου από το επαναληπτικό δίκτυο και h_t η έξοδος. Λόγω αυτής της ιδιότητας τα RNN έχουν τη δυνατότητα να περιγράψουν χρονικές εξαρτήσεις μεταξύ των εξόδων ακόμα και σε μη συνεχόμενα δεδομένα αφού λόγω της επαναληπτικής αρχιτεκτονικής τους δημιουργείται ένας μηχανισμός μνήμης (βλέπε Σχήμα 3.6). Στη γενική μορφή τους τα επαναληπτικά νευρωνικά δίκτυα εμφανίζουν προβλήματα και περιορισμούς στις σχέσεις μακράς διάρκειας. Ο λόγος φαίνεται και στη σχέση 3.3 όπου ο υπολογισμός της κλίσης περιλαμβάνει την πράξη της παραγώγισης μιας σύνθεσης συναρτήσεων που αυξάνει σημαντικά τον απαραίτητο υπολογιστικό φόρτο σε περιπτώσεις μεγάλων ακολουθιών. Ο περιορισμός αυτός στην πράξη αντιμετωπίζεται με την εισαγωγή των νευρώνων Μακράς-Βραχείας Μνήμης (LSTM) και των Gated Recurrent Units (GRU).



Σχήμα 3.6: Σχηματική απεικόνιση 'ξεδιπλώματος' RNN. Σχήμα από [6]

3.3.1 Δίκτυα Μακράς-Βραχείας Μνήμης

Τα LSTM αποτελούν νευρωνικά δίκτυα που παρουσιάζουν κοινά στοιχεία με τα προαναφερθέντα RNN, με κάποιες στοιχειώδεις διαφορές.

Για να περιγραφούν οι διαφορές με τα RNN και οι κύριες ιδιότητες των LSTM αρχικά ορίζουμε τα εξής διανύσματα:

- i_t: Η θύρα εισόδου
- f_t: Η θύρα λησμόνησης
- c_t: Το κύτταρο μνήμης
- o_t: Η θύρα εξόδου
- h_t: Η κρυφή κατάσταση

Τα διανύσματα αυτά προκύπτουν από τις εξισώσεις:

$$i_t = (W^i x_t + U^i h_{t-1} + b^i) \tag{3.4}$$

$$f_t = (W^f x_t + U^f h_{t-1} + b^f) \tag{3.5}$$

$$o_t = (W^f x_t + U^f h_{t-1} + b^o) \tag{3.6}$$

$$u_t = tanh(W^u x_t + U^u h_{t-1} + b^u) \tag{3.7}$$

$$c_t = i_t \odot u_t + f_t \odot c_{t-1} \tag{3.8}$$

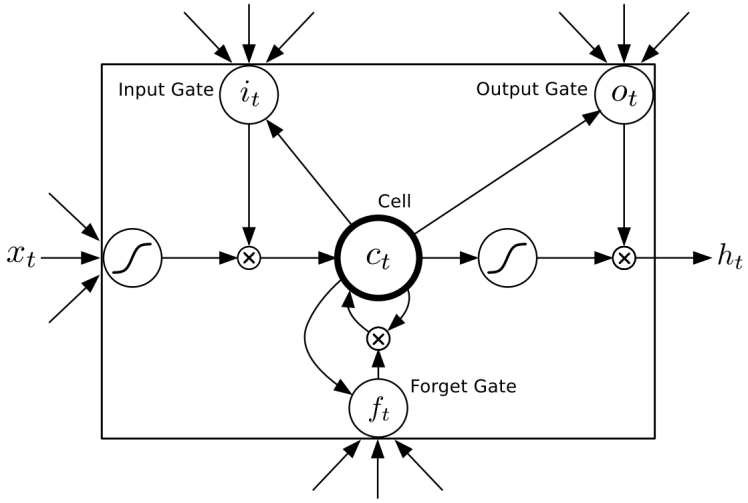
$$h_t = o_t \odot tanh(c_{t-1}) \tag{3.9}$$

όπου $\sigma()$ είναι η σιγμοειδής συνάρτηση και $tanh()$ η υπερβολική εφαπτομένη.

Διαισθητικά η διαδικασία λειτουργίας των LSTM μέσα από τις παραπάνω εξισώσεις φαίνεται πως είναι:

- Το i_t ελέγχει ποια διανύσματα θα μεταβληθούν και σε ποιο βαθμό.
- Το f_t ελέγχει αν το περιεχόμενο της μνήμης c_t πρέπει να διατηρηθεί στο επόμενο βήμα λειτουργίας.
- Το o_t ρυθμίζει την έκθεση της μνήμης στο νευρωνικό δίκτυο.
- Το h_t αντικατοπτρίζει την αποθηκευμένη πληροφορία του κυττάρου μνήμης η οποία επηρεάζεται από την καινούργια είσοδο x_t και τις προηγούμενες επαναληπτικές ενέργειες.

Σχηματικά η λειτουργία του LSTM φαίνεται στο Σχήμα 3.7.



Σχήμα 3.7: Αναπαράσταση LSTM. Σχήμα από [7]

3.4 Μοντέλα στη διπλωματική

Για την αναγνώριση των κινήσεων των νοηματιστών κατά το δαχτυλοσυλλαβισμό και τη δημιουργία προβλέψεων γραμμάτων του ελληνικού αλφαβήτου, έγινε χρήση δύο αρχιτεκτονικών νευρωνικών δικτύων σε συνδυασμό για την εκπαίδευση του μοντέλου.

Πιο συγκεκριμένα χρησιμοποιήθηκαν τα δίκτυα:

- MobileNetV2 με προεκπαιδευμένα βάρη στο ImageNet
- RNN με LSTM με χρήση του module Sequential

Τα δύο μοντέλα προέρχονται από την βιβλιοθήκη Keras [39] του TensorFlow, και τροποποιήθηκαν σχηματικά στις εισόδους ώστε να τροφοδοτηθούν με τα δεδομένα του MediaPipe που θα επεξηγηθούν σε επόμενο κεφάλαιο. Η επιλογή του προεκπαιδευμένου μοντέλου στο ImageNet έγινε με στόχο την επίτευξη μεγαλύτερης ακριβείας.

Κεφάλαιο 4

Σύνολο Δεδομένων Διπλωματικής

Για τα πλαίσια εκπαίδευσης του μοντέλου της εφαρμογής αυτοαξιολόγησης χρησιμοποιήθηκε το σύνολο δεδομένων ελληνικού δαχτυλοσυλλαβισμού του προγράμματος SL-ReDu του Πανεπιστημίου Θεσσαλίας [40]. Τα δεδομένα έχουν καταγραφεί από 12 διερμηνείς και σχηματίζουν 685 λέξεις της ελληνικής γλώσσας σε διαφορετικά περιβάλλοντα και καταστάσεις φωτισμού. Το συγκεκριμένο σύνολο δεδομένων δεν είναι ακόμη δημοσιευμένο για ελεύθερη χρήση. Παραδείγματα φαίνονται στο Σχήμα 4.1.

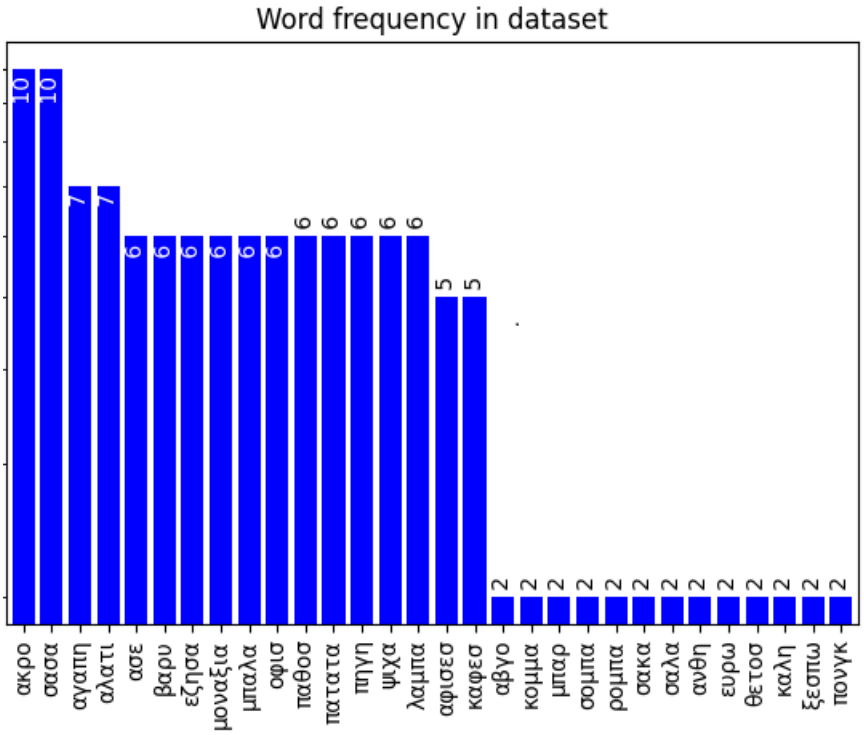
Σχήμα 4.1: Στιγμιότυπα του ελληνικού συνόλου δεδομένων

Πιο αναλυτικά η δομή του συνόλου δεδομένων είναι:

- 12 κατηγορίες καταγεγραμμένου υλικού χωρισμένο για κάθε νοηματιστή.
- 978 συνολικά καταγεγραμμένες ακολουθίες.

- Βίντεο δαχτυλοσσυλαβισμού με ετικέτα τη λέξη που συμβολίζεται από τον νοηματιστή.
- Ανάλυση εικόνας 640x480.
- Η διάρκεια των βίντεο κυμαίνεται στα 8-12 δευτερόλεπτα.
- Η καταγραφή των δεδομένων έγινε με στατική κάμερα και με ρυθμό 30 καρέ ανά δευτερόλεπτο (fps).
- Χρήση δεξιού χεριού από τους νοηματιστές.
- Το υλικό είναι αποθηκευμένο σε έγχρωμο βίντεο RGB μορφής webm.

Η κατανομή των 30 λέξεων με τις περισσότερες εμφανίσεις στο σύνολο δεδομένων φαίνονται στο Σχήμα 4.2



Σχήμα 4.2: Λέξεις με τις περισσότερες εμφανίσεις στο σύνολο δεδομένων

Κεφάλαιο 5

Ανάπτυξη του Συστήματος Αναγνώρισης

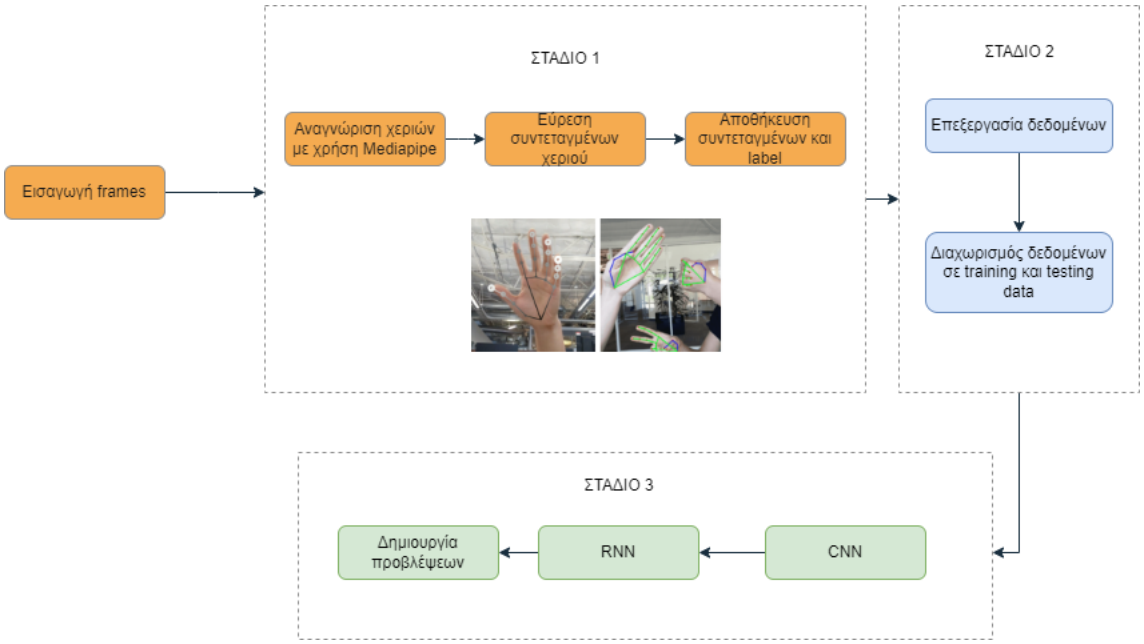
5.1 Μεθοδολογία

Η γενική μεθοδολογία μας αποτελείται από τα εξής βήματα (βλέπε επίσης Σχήμα 5.1):

- Αρχικά γίνεται χρήση του MediaPipe για την εξαγωγή και αποθήκευση συντεταγμένων των χεριών στο σύνολο δεδομένων και αποθήκευσή τους με το όνομα της λέξης.
- Επεξεργασία και κανονικοποίηση δεδομένων, διαχωρισμός τους σε δεδομένα μάθησης και δεδομένα δοκιμής.
- Προώθηση των συντεταγμένων και των ονομάτων σε ένα CNN για εκπαίδευση του μοντέλου.
- Προώθηση των αποτελεσμάτων του προηγούμενου βήματος σε ένα RNN για την εκπαίδευση του τελικού μοντέλου.
- Επαλήθευση μοντέλου και ακρίβειας μέσω των δεδομένων δοκιμής.
- Δημιουργία προβλέψεων δαχτυλοσυλλαβισμού για τη λειτουργία της εφαρμογής αυτοαξιολόγησης

5.2 Εξαγωγή πληροφορίας

Για την εξαγωγή της πληροφορίας από το σύνολο δεδομένων ελληνικής νοηματικής γλώσσας του προγράμματος SL-ReDu χρησιμοποιήθηκε η προεκπαιδευμένη βιβλιοθήκη αναγνώρισης ανθρώπινου σώματος MediaPipe [41].



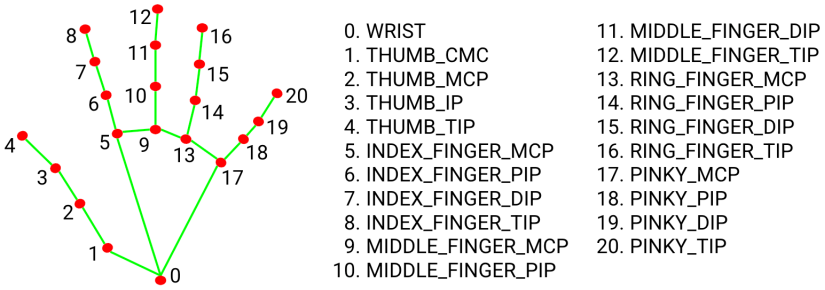
Σχήμα 5.1: Διάγραμμα γενικής μεθοδολογίας που χρησιμοποιήθηκε

Το MediaPipe είναι ένα framework σχεδιασμένο για την ανάπτυξη εφαρμογών μηχανικής μάθησης για την ανάλυση δεδομένων πολλαπλών μέσων όπως τα βίντεο και ο ήχος. Η βιβλιοθήκη συναρτήσεων του MediaPipe παρέχει μεγάλη συλλογή μοντέλων αναγνώρισης του ανθρώπινου σώματος και της κίνησης του, μερικές εκ των οποίων είναι οι Face Mesh, Pose, Face Detection, Holistic, Iris tracking και Hand tracking. Τα προαναφερθέντα μοντέλα έχουν εκπαιδευτεί με τη χρήση γιγαντιαίων όγκων δεδομένων σε μεγάλη ποικιλία διαφορετικών περιβαλλόντων από τους ερευνητές της Google εκμεταλλευόμενοι την τεχνολογία του TensorFlow Lite, και είναι σχεδιασμένα ώστε να επιστρέφουν 3D συντεταγμένες για διάφορα μέρη του ανθρώπινου σώματος.

Το Hand Tracking module της MediaPipe που χρησιμοποιείται για την εξαγωγή πληροφορίας στην παρούσα διπλωματική συντελείται από δύο διαφορετικά μοντέλα που δουλεύουν σε συνδυασμό μεταξύ τους, το μοντέλο αναγνώρισης παλάμης και το μοντέλο αναγνώρισης αρθρώσεων. Το πρώτο παρέχει ακριβείς περικοπές εικόνων παλάμης που τροφοδοτούνται στο μοντέλο αναγνώρισης αρθρώσεων. Με τη χρήση της συγκεκριμένης μεθοδολογίας μειώνεται ο υπολογιστικός φόρτος εργασίας που απαιτείται για την αναγνώριση των ανθρώπινων χεριών σε σχέση με άλλες αρχιτεκτονικές.

Πιο συγκεκριμένα με τη χρήση του MediaPipe Hands [42] γίνεται εξαγωγή 21 3D σημείων ανά χέρι που συμβολίζουν το ύψος, το πλάτος και το βάθος για διάφορα σημεία του χεριού (βλέπε Σχήμα 5.3) που θα χρησιμοποιηθούν για την πρόβλεψη των γραμμάτων. Κατά

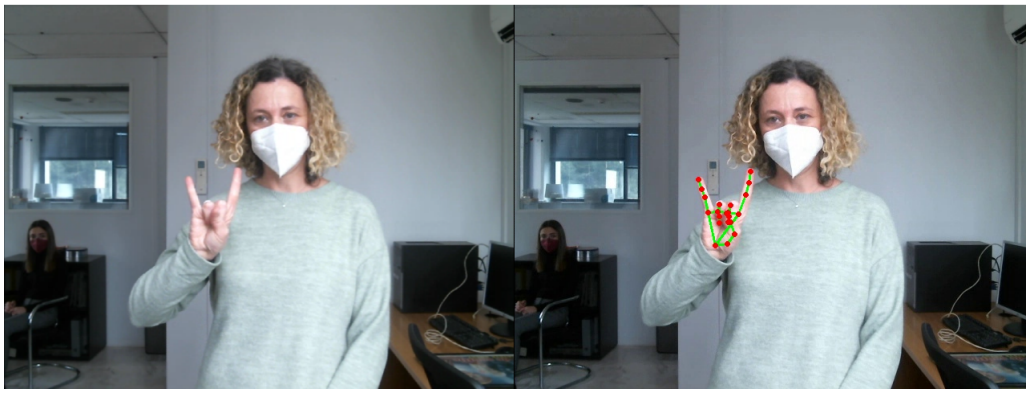
την ανάλυση των δεδομένων και επεξεργασία των μεμονωμένων πλαισίων που προκύπτουν από τα βίντεο δαχτυλοσυλλάβισης αποθηκεύονται οι συντεταγμένες των διάφορων κινήσεων των χεριών κατά τον δαχτυλοσυλλαβισμό των λέξεων. Στα πλαίσια της διπλωματικής εργασίας χρησιμοποιήθηκε ελάχιστη πιθανότητα αναγνώρισης παλάμης ίση με 0.7.



Σχήμα 5.2: Σημεία αναγνώρισης χεριού από τη βιβλιοθήκη MediaPipe. Εικόνα από [8]

Οι συντεταγμένες που προκύπτουν για κάθε βίντεο δαχτυλοσυλλαβισμού αποθηκεύονται μαζί με το όνομα της λέξης που συμβολίζεται από τον νοηματιστή ως πίνακας διαστάσεων 21 $\times$ 2 προκειμένου να τροφοδοτηθεί αργότερα στο συνελικτικό νευρωνικό δίκτυο MobileNet που έχει επιλεχθεί στα πλαίσια της παρούσας διπλωματικής.

Στιγμιότυπα από την εξαγωγή συντεταγμένων κατά το συμβολισμό των λέξεων φαίνονται στα Σχήματα 5.3 και 5.4.



Σχήμα 5.3: Εξαγωγή πληροφορίας για το γράμμα Η

5.3 Εκπαίδευση μοντέλου CNN

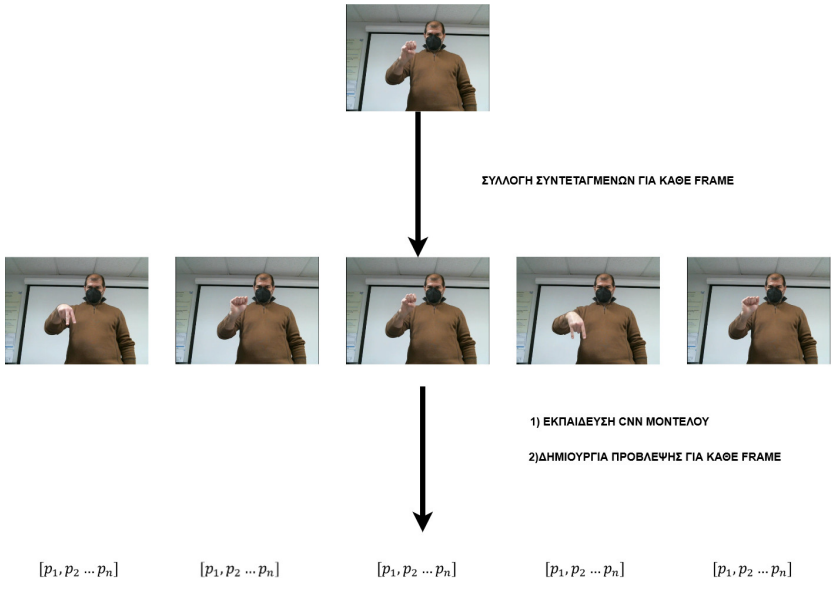
Αρχικά τα δεδομένα που έχουν επεξεργαστεί και εξαχθεί σε αρχεία csv σε συνδυασμό με τις ετικέτες των λέξεων τροφοδοτούνται στο μοντέλο MobileNet με στόχο την αρχική εκπαίδευση του μοντέλου. Το δίκτυο CNN κατά την εκπαίδευση του δημιουργεί προβλέψεις

Σχήμα 5.4: Εξαγωγή πληροφορίας για το γράμμα K

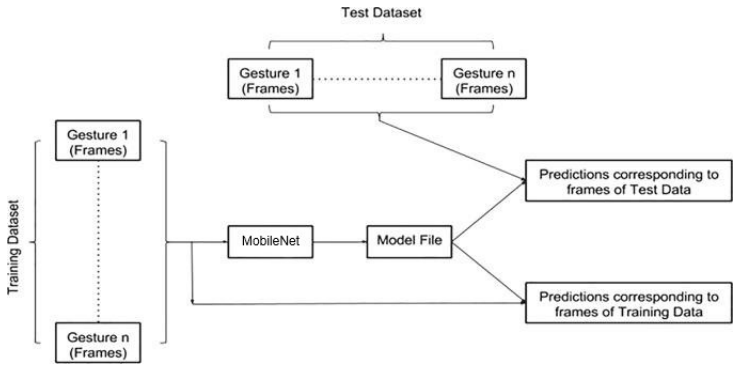
για τις συντεταγμένες κάθε καρέ που έχουν προκύψει κατά το δαχτυλοσυλλαβισμό των λέξεων και αποθηκεύει την πρόοδο των προβλέψεων του σε ένα αρχείο pickle που μετέπειτα θα χρησιμοποιηθεί από το RNN.

Στην γραφική αναπαράσταση του Σχήματος 5.5 φαίνεται η μεθοδολογία που χρησιμοποιείται από το νευρωνικό δίκτυο για τη δημιουργία των προβλέψεων και την εκπαίδευση του μοντέλου, με βάση τη πληροφορία που δέχεται σε κάθε καρέ.

Η γραφική απεικόνιση του Σχήματος 5.6 απεικονίζει τη συνολική μεθοδολογία που ακολουθήθηκε κατά την εκπαίδευση του μοντέλου. Το διάγραμμα παρουσιάζει τα βήματα εξαγωγής πληροφορίας από τους καταγεγραμμένους συμβολισμούς των λέξεων, την προώθηση της στο στο δίκτυο MobileNet και το διαχωρισμό της σε σύνολα εκπαίδευσης και δοκιμής.



Σχήμα 5.5: Αναπαράσταση λειτουργίας του CNN στο σύνολο δεδομένων

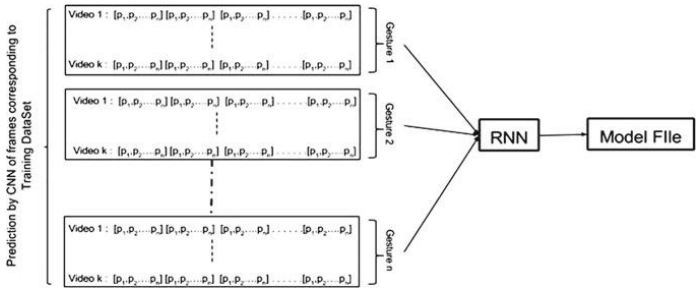


Σχήμα 5.6: Αναπαράσταση λειτουργίας του συνελικτικού νευρωνικού δικτύου. Διάγραμμα τροποποιημένο από [9]

5.4 Εκπαίδευση μοντέλου RNN

Με τη σειρά του καθώς το MobileNet έχει τελειώσει με τη διαδικασία της εκπαίδευσης, τα ακολουθιακά δεδομένα του προηγούμενου βήματος τροφοδοτούνται αυτή τη φορά στο μοντέλο RNN με στόχο να αναγνωριστούν οι ακολουθιακές κινήσεις των νοηματιστών σε συνδυασμό με τις ετικέτες των βίντεο. Το RNN επεξεργάζεται τις προβλέψεις που δημιουργήθηκαν στο προηγούμενο στάδιο της εκπαίδευσης και αποθηκεύτηκαν στο αρχείο pickle με στόχο τη δημιουργία του τελικού μοντέλου.

Η διαδικασία περιγράφεται γραφικά στο Σχήμα 5.7.

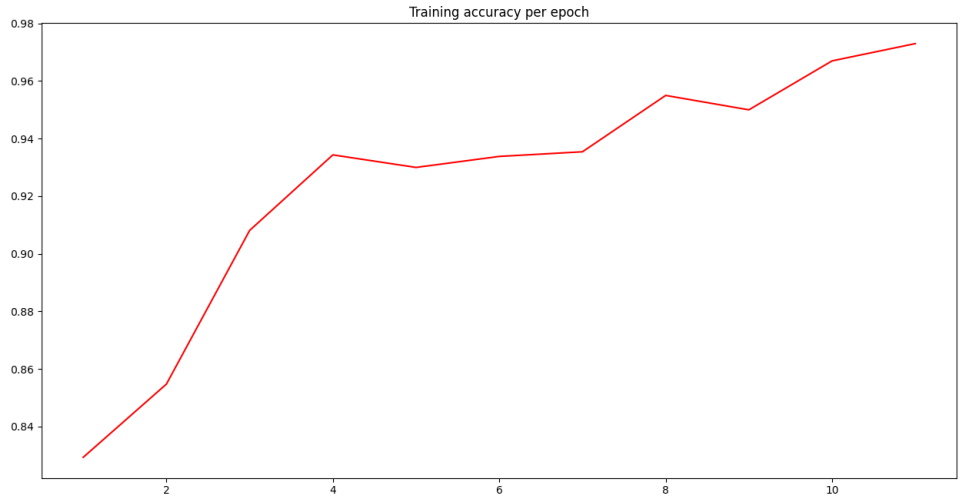


Σχήμα 5.7: Αναπαράσταση λειτουργίας του επαναληπτικού νευρωνικού δικτύου. Διάγραμμα από [9]

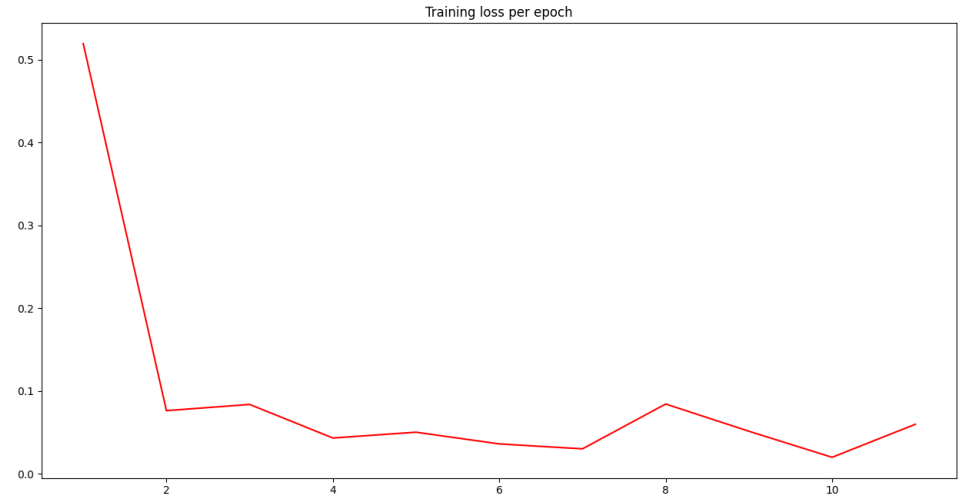
5.5 Αποτελέσματα

Ο συνδυασμός των προαναφερθέντων τεχνικών οδήγησε στην επιτυχή εκπαίδευση του μοντέλου με ακρίβεια της τάξης 97.3% και loss 6% (βλέπε Σχήματα 5.8 και 5.9) ενώ κατά την διάρκεια της δοκιμής επιτεύχθηκε ποσοστό ακρίβειας 97.8% και loss 2% (βλέπε Σχή-

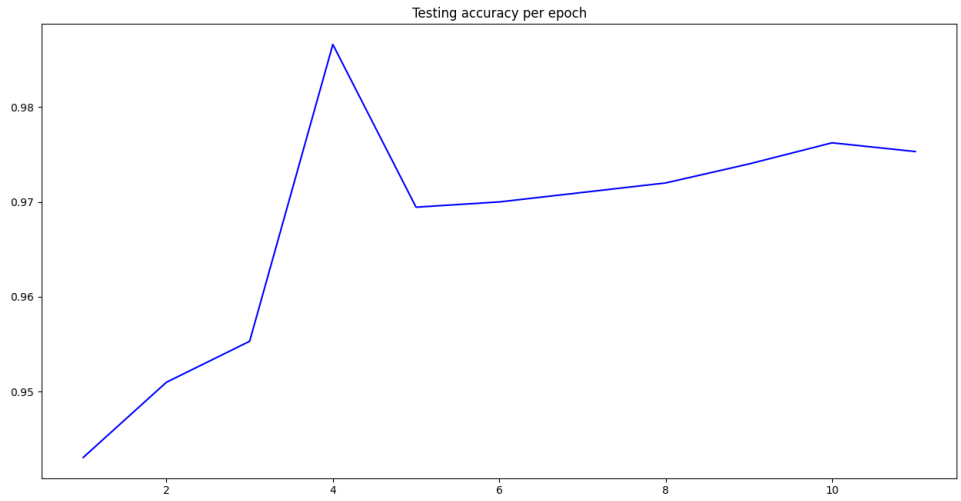
ματα 5.10 και 5.11) φέρνοντας ικανοποιητικά αποτελέσματα στην αναγνώριση των ελληνικών συμβολισμών.



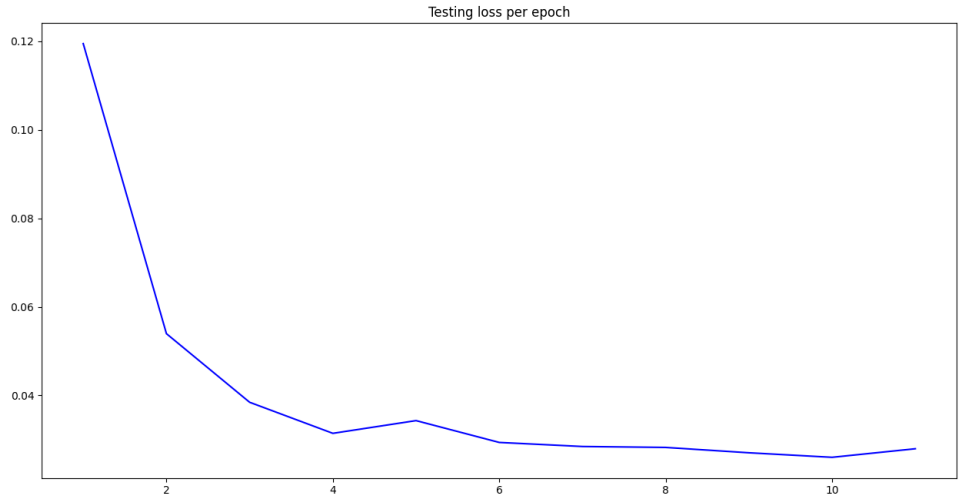
Σχήμα 5.8: Διάγραμμα ακρίβειας ανά epoch για τα δεδομένα εκπαίδευσης



Σχήμα 5.9: Διάγραμμα απώλειας ανά epoch για τα δεδομένα εκπαίδευσης



Σχήμα 5.10: Διάγραμμα ακρίβειας ανά epoch για τα δεδομένα δοκιμής



Σχήμα 5.11: Διάγραμμα απώλειας ανά epoch για τα δεδομένα δοκιμής

Κεφάλαιο 6

Περιγραφή Γραφικού Περιβάλλοντος

6.1 Φιλοσοφία εφαρμογής

Όπως έχει προαναφερθεί στόχος της παρούσας διπλωματικής εργασίας αποτελεί η ανάπτυξη μίας εφαρμογής αυτόματης αξιολόγησης δακτυλοσυλλαβισμού, βασισμένης στο αλφάβητο της ελληνικής νοηματικής γλώσσας (GSL) με τη χρήση τεχνικών μηχανικής μάθησης. Σκοπός της εφαρμογής αποτελεί η επιμόρφωση του χρήστη γύρω από την GSL, η συνεχής εξάσκηση των δεξιοτήτων του σε αυτή και η διάδοση της GSL σε μεγαλύτερο μέρος του πληθυσμού. Η εφαρμογή αναπτύχθηκε με γνώμονα το εύχρηστο και κατανοητό γραφικό περιβάλλον προκείμενου η εκπαίδευση της νοηματικής γλώσσας να είναι προσιτή προς όλους.

6.2 Δυνατότητες εφαρμογής

Η εφαρμογή αυτόματης αξιολόγησης της ικανότητας δακτυλοσυλλαβισμού προσφέρει τη δυνατότητα στο χρήστη να ελέγξει τις γνώσεις του μέσα από μια εικονική αξιολόγηση σε πάνω από 3000 λέξεις της ελληνικής γλώσσας σε πραγματικό χρόνο, κάνοντας χρήση της κάμερας που διαθέτει ο χρήστης. Στην παρούσα φάση η εφαρμογή είναι προγραμματισμένη να δημιουργεί εξετάσεις των 10 λέξεων κάθε φορά, προσφέροντας στο χρήστη το τελικό του αποτέλεσμα αλλά και την ακρίβεια γραμμάτων (letter accuracy) προκειμένου να γνωρίζει ο εξεταζόμενος πόσο μακριά βρισκόταν από την σωστή λύση.

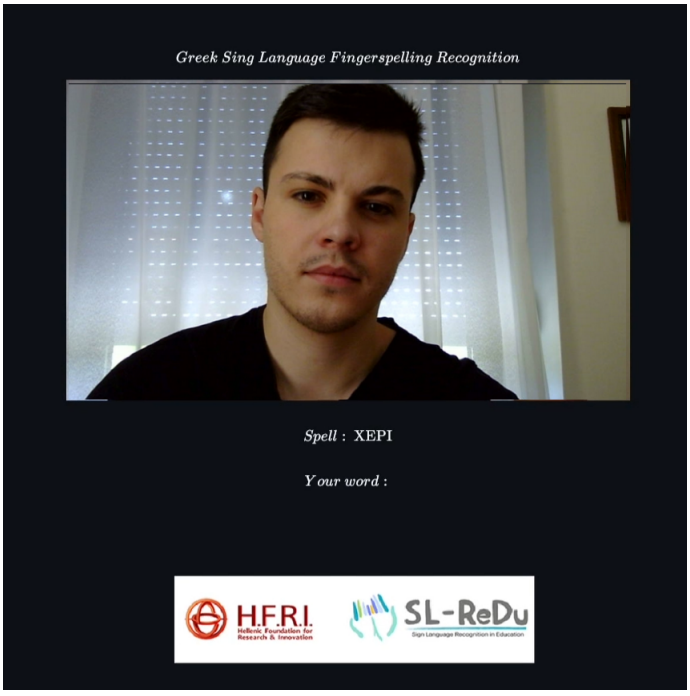
6.2.1 Τεχνολογίες και εργαλεία που χρησιμοποιήθηκαν

Για την ανάπτυξη της εφαρμογής έγινε χρήση της προγραμματιστικής γλώσσας Python 3.10 και πολυάριθμων βιβλιοθηκών της μεταξύ άλλων της MediaPipe [43] για την εξαγωγή δεδομένων και ανίχνευση των χεριών, των TensorFlow [44] και Cuda [45] για την εκπαίδευση του νευρωνικού δικτύου, της OpenCV [46] για την λήψη της ζωντανής ροής από την κάμερα του χρήστη και του Streamlit framework για τη δημιουργία του γραφικού περιβάλλοντος που θα παρουσιαστεί αναλυτικότερα παρακάτω.

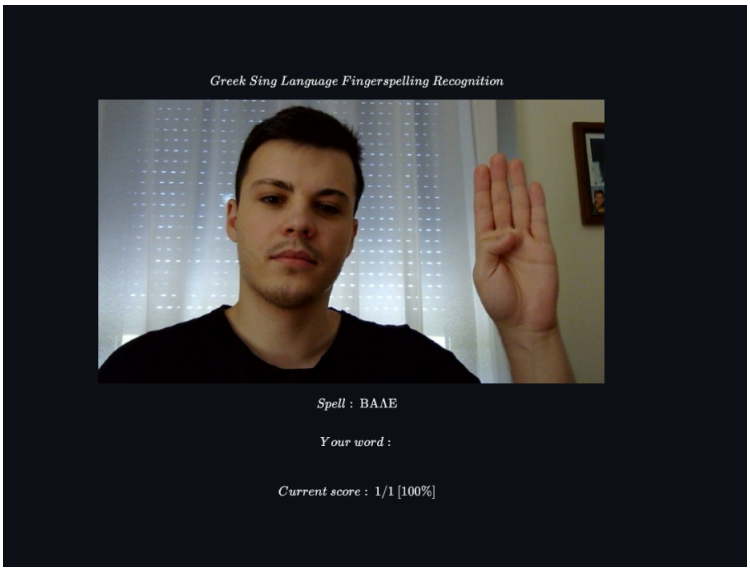
6.3 Παρουσίαση εφαρμογής

6.3.1 Αρχική διεπαφή

Προκειμένου η εφαρμογή αξιολόγησης να διατηρήσει απλότητα και εύκολη πλοήγηση, η αρχική επαφή του χρήστη με την εφαρμογή αποτελείται από την πρώτη λέξη που καλείται να δακτυλοσυλλαβίσει στην κάμερα. Ο χρήστης λαμβάνει την εξεταστέα λέξη στο πεδίο “Spell: ” ενώ η πρόοδος κατά τη δαχτυλοσυλλάβιση φαίνεται στο πεδίο “Your word: ”. Στο Σχήμα 6.1 φαίνεται η αρχική επαφή του χρήστη με το γραφικό περιβάλλον καθώς καλείται να δαχτυλοσυλλαβίσει τη λέξη “ΧΕΡΙ”.



Σχήμα 6.1: Αρχικό περιβάλλον της εφαρμογής



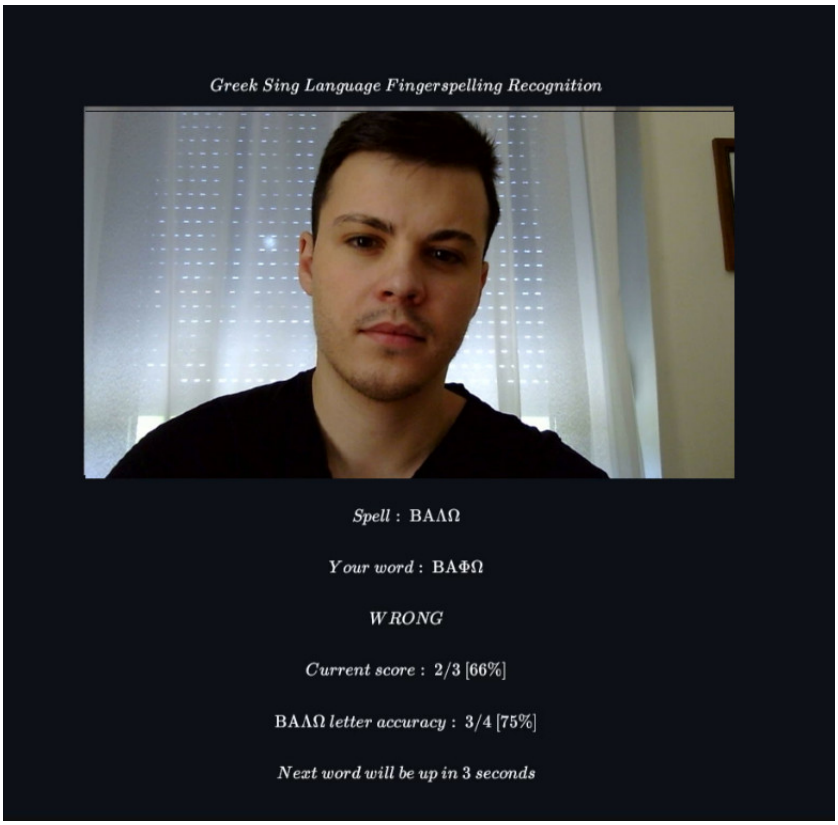
Σχήμα 6.2: Στιγμιότυπο δαχτυλοσυλλαβισμού

6.3.2 Εμφάνιση αποτελέσματος σε εισαγωγή λάθος απάντησης

Σε περίπτωση σχηματισμού λανθασμένου νοήματος από τον εξεταζόμενο η εφαρμογή εμφανίζει το πεδίο "WRONG", ακολουθούμενο από το πεδίο "Current Score" που δείχνει τον αριθμό των σωστών απαντήσεων του χρήστη μέχρι εκείνη τη στιγμή. Παράλληλα ο χρήστης μέσω του πεδίου "Word's letter accuracy" έχει τη δυνατότητα να δει πόσο απείχε από τη σωστή απάντηση (βλέπε Σχήμα 6.3). Η επόμενη λέξη της αξιολόγησης εμφανίζεται τρία δευτερόλεπτα αφού εμφανιστεί το μήνυμα του αποτελέσματος.

6.3.3 Εμφάνιση αποτελέσματος σε εισαγωγή σωστής απάντησης

Εφόσον ο χρήστης ολοκληρώσει το σχηματισμό της λέξης, επιβεβαιώνεται από το σύστημα της εφαρμογής η σωστή ορθογραφία της λέξης και σε περίπτωση σωστού αποτελέσματος προστίθεται ένας πόντος στο τελικό αποτέλεσμα της εξέτασης. Ταυτόχρονα στο γραφικό περιβάλλον της εφαρμογής εμφανίζεται το πεδίο "CORRECT" συνοδευόμενο από το πεδίο "Current: " που προβάλει τον αριθμό των μέχρι τώρα σωστών απαντήσεων του χρήστη (βλέπε Σχήμα 6.2). Όπως και στη περίπτωση του λάθους εμφανίζεται το πλαίσιο "Word's letter accuracy: " που υπολογίζει ποσοτικά τα γράμματα της λέξης που δαχτυλοσυλλάβισε ο χρήστης σωστά (βλέπε Σχήμα 6.4). Στην περίπτωση επιτυχίας ο χρήστης λαμβάνει τη μεγαλύτερη δυνατή ακρίβεια. Σε περίπτωση που ο χρήστης δεν έχει φτάσει τον αριθμό των 10 ερωτήσεων, η επόμενη λέξη θα εμφανιστεί μετά το πέρας των τριών δευτερολέπτων.

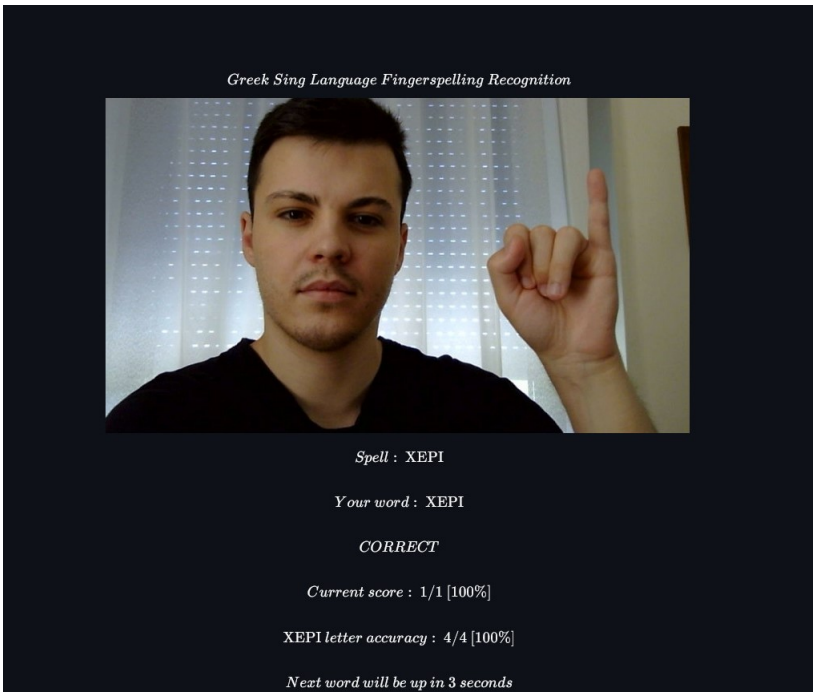


Σχήμα 6.3: Στιγμιότυπο της εφαρμογής κατά την εισαγωγή λανθασμένης απάντησης

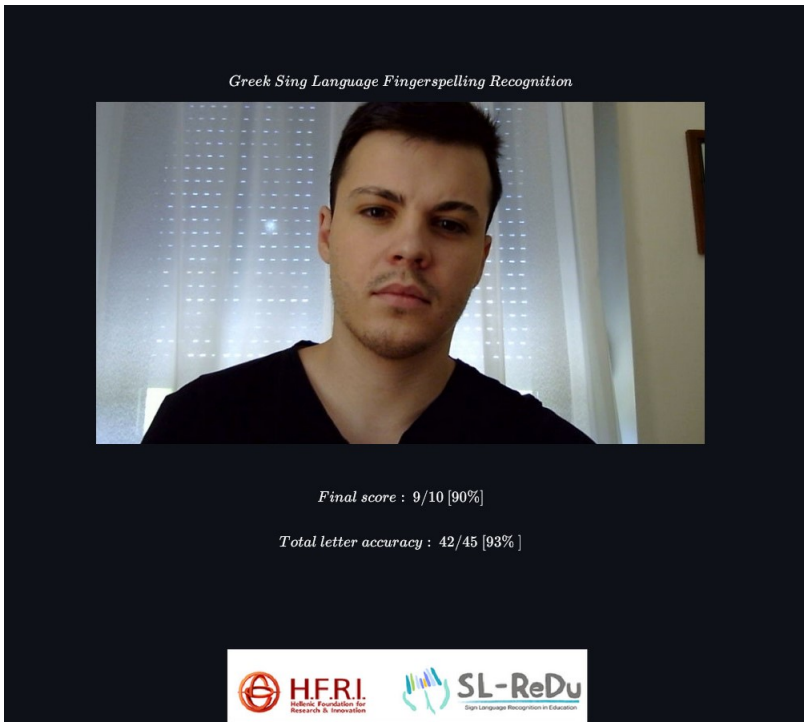
6.3.4 Εμφάνιση τελικού αποτελέσματος

Ως τελικό μέρος της εξέτασης, αφού ο χρήστης έχει ολοκληρώσει το συλλαβισμό των δέκα λέξεων του ελληνικού αλφαβήτου εμφανίζεται το τελικό αποτέλεσμα στο πεδίο “Final score: ”. Το αποτέλεσμα του χρήστη εμφανίζεται ως βαθμολογία με άριστα το 10 και ελάχιστη το μηδέν και παράλληλα ως ποσοστό με άριστη βαθμολογία το 100% και ελάχιστη το 0%.

Παράλληλα στο πεδίο “Total letter accuracy: ” εμφανίζεται η συνολική ακρίβεια σε επίπεδο των συλλαβισμένων λέξεων σε σχέση με τις αντίστοιχες σωστές απαντήσεις. Όπως και το τελικό αποτέλεσμα, η συνολική ακρίβεια δαχτυλοσυλλάβισης παρουσιάζεται και σε μορφή ποσοστού (βλέπε Σχήμα 6.5).



Σχήμα 6.4: Στιγμιότυπο της εφαρμογής κατά την εισαγωγή σωστής απάντησης



Σχήμα 6.5: Εμφάνιση τελικού αποτελέσματος

Κεφάλαιο 7

Συμπεράσματα

Κατά τη διάρκεια της παρούσας διπλωματικής εργασίας ερευνήθηκε η αναγνώριση των ανθρώπινων χεριών και των συμβολισμών τους με στόχο την ανάπτυξη ενός μοντέλου αναγνώρισης ελληνικής νοηματικής γλώσσας και τη χρήση του ως εκπαιδευτικό εργαλείο αυτοαξιολόγησης. Για την εκπαίδευση του μοντέλου χρησιμοποιήθηκαν τεχνικές εξαγωγής 3D συντεταγμένων των ανθρώπινων χεριών για τη μελέτη των συμβολισμών και της κίνησης που εμφανίζονται κατά τη διάρκεια του δαχτυλοσυλλαβισμού. Τα προαναφερθέντα δεδομένα μετά την επεξεργασία τους τροφοδοτήθηκαν στα δίκτυα MobileNet και Keras RNN που χρησιμοποιήσαμε για την εκπαίδευση του μοντέλου. Η χρήση της προεκπαιδευμένης βιβλιοθήκης MediaPipe για την τροφοδότηση της πληροφορίας και τη δημιουργία προβλέψεων και η εκμετάλλευση του TensorFlow Lite έχουν παρουσιαστεί ως λύσεις σε προβλήματα που εμφανίζονται σε πολλά παρόμοια συστήματα αναγνώρισης νοηματικής γλώσσας όπως οι διάφορες συνθήκες φωτισμού, η χαμηλή ποιότητα κάμερας και η ελάχιστη αναγκαία υπολογιστική δύναμη που απαιτεί η πρόβλεψη σε ζωντανό χρόνο. Τέλος ως μέρος της διπλωματικής αναπτύχθηκε ένα εκπαιδευτικό εργαλείο αυτοαξιολόγησης δαχτυλοσυλλαβισμού της ελληνικής νοηματικής γλώσσας με γνώμονα τη χρήση του από μαθητές όλων των επιπέδων.

Όσον αφορά τις μελλοντικές επεκτάσεις και ιδέες βελτίωσης της εφαρμογής αυτοαξιολόγησης πιθανές δράσεις είναι η διάθεση του ως δωρεάν διδακτικό εργαλείο στο διαδίκτυο και η δημιουργία νέων λειτουργιών για εκπαιδευτικούς σκοπούς όπως η συμπλήρωση φράσεων με δαχτυλοσυλλαβισμό και η ανάπτυξη παιχνιδιών δαχτυλοσυλλάβισης για χρήστες μικρότερης ηλικίας. Ενώ τέλος στα μελλοντικά σχέδια αναβάθμισης της εφαρμογής είναι η δημιουργία ενός μοντέρνου και κατανοητού γραφικού περιβάλλοντος για την ευκολότερη

χρήση του εργαλείου.

Πιθανά μελλοντικά σχέδια αποτελούν:

- Η χρήση διαφορετικών αρχιτεκτονικών CNN και RNN (όπως για παράδειγμα οι αρχιτεκτονικές VGG, EfficientNet, BiLSTM) για σύγκριση ακρίβειας και υπολογιστικών απαιτήσεων.
- Εφαρμογή διαφορετικών frameworks για την εξαγωγή συντεταγμένων ανθρώπινων χεριών όπως το Yoha και το OpenPose με στόχο τη βελτιστοποίηση του μοντέλου σε διάφορες συνθήκες φωτισμού.
- Η δημιουργία ενός μεγαλύτερου σύνολου δεδομένων σε διάφορες αποστάσεις και γωνίες βιντεοσκόπησης με στόχο το ευρύ φάσμα δεδομένων.
- Η αξιολόγηση του συστήματος από επαρκή αριθμό χρηστών.

Βιβλιογραφία

[1] Vivek Bheda and Dianna Radpour. Using deep convolutional networks for gesture recognition in American sign language. *arXiv preprint arXiv:1710.06836*, 2017.

[2] Sumit Saha. A comprehensive guide to convolutional neural networks-the eli5 way. `https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53`, Dec 2018.

[3] Arden Dertat. Applied deep learning - part 4: Convolutional neural networks. `https://towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2`, Nov 2017.

[4] Arc. Convolutional neural network. `https://towardsdatascience.com/convolutional-neural-network-17fb77e76c05`, Dec 2018.

[5] Delwar Hossain, Masudul Haider Imtiaz, Tonmoy Ghosh, Viprav Bhaskar, and Edward Sazonov. Real-time food intake monitoring using wearable egocnetric camera. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 4191–4195. IEEE, 2020.

[6] Weijiang Feng, Naiyang Guan, Yuan Li, Xiang Zhang, and Zhigang Luo. Audio visual speech recognition with multimodal recurrent neural networks. In *2017 International Joint Conference on neural networks (IJCNN)*, pages 681–688. IEEE, 2017.

[7] Alex Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

[8] Google. MediaPipe Hands. `https://google.github.io/mediapipe/solutions/hands.html`. Accessed on 08.07.2022.

[9] Sarfaraz Masood, Adhyan Srivastava, Harish Chandra Thuwal, and Musheer Ahmad. Real-time sign language gesture (word) recognition from video sequences using CNN and RNN. In *Intelligent Engineering Informatics*, pages 623–632. Springer, 2018.

[10] Byeongkeun Kang, Subarna Tripathi, and Truong Q. Nguyen. Real-time sign language fingerspelling recognition using convolutional neural networks from depth map. *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 136–140, 2015.

[11] Brandon Garcia and Sigberto Alarcon Viesca. Real-time American sign language recognition with convolutional neural networks. *Convolutional Neural Networks for Visual Recognition*, 2:225–232, 2016.

[12] Eng-Jon Ong, Helen Cooper, Nicolas Pugeault, and Richard Bowden. Sign language recognition using sequential pattern trees. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2200–2207. IEEE, 2012.

[13] Andre Barczak, Napoleon Reyes, M. Abastillas, A. Piccio, and Teo Susnjak. A new 2D static hand gesture colour image dataset for ASL gestures. *Research Letters in the Information and Mathematical Sciences*, 15, 2011.

[14] Katerina Papadimitriou and Gerasimos Potamianos. End-to-end convolutional sequence learning for ASL fingerspelling recognition. In *Proc. Interspeech 2019*, pages 2315–2319, 2019.

[15] Taehwan Kim, Jonathan Keane, Weiran Wang, Hao Tang, Jason Riggle, Gregory Shakhnarovich, Diane Brentari, and Karen Livescu. Lexicon-free fingerspelling recognition from video: Data, models, and signer adaptation. *Computer Speech & Language*, 46:209–232, 2017.

[16] Bowen Shi, Aurora Martinez Del Rio, Jonathan Keane, Jonathan Michaux, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. American sign language fingerspelling recognition in the wild. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 145–152. IEEE, 2018.

[17] Bowen Shi and Karen Livescu. Multitask training with unlabeled data for end-to-end sign language fingerspelling recognition. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 389–396. IEEE, 2017.

[18] M. Ashraful Amin and Hong Yan. Sign language finger alphabet recognition from Gabor-PCA representation of hand gestures. In *2007 International Conference on Machine Learning and Cybernetics*, volume 4, pages 2218–2223. IEEE, 2007.

[19] Celal Savur and Ferat Sahin. Real-time American sign language recognition system using surface EMG signal. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pages 497–502. IEEE, 2015.

[20] K. Otiniano Rodríguez and G. Cámara Chávez. Finger spelling recognition from RGB-D information using kernel descriptor. In *2013 XXVI Conference on Graphics, Patterns and Images*, pages 1–7. IEEE, 2013.

[21] Nasser H. Dardas and Nicolas D. Georganas. Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques. *IEEE Transactions on Instrumentation and Measurement*, 60(11):3592–3607, 2011.

[22] Priyanka Mekala, Ying Gao, Jeffrey Fan, and Asad Davari. Real-time sign language recognition based on neural network architecture. In *2011 IEEE 43rd Southeastern symposium on system theory*, pages 195–199. IEEE, 2011.

[23] Markus V. Lamari, Md Shoaib Bhuiyan, and Akira Iwata. Hand alphabet recognition using morphological PCA and neural networks. In *IJCNN'99. International Joint Conference on Neural Networks. Proceedings*, volume 4, pages 2839–2844. IEEE, 1999.

[24] Jochen Triesch and Christoph Von Der Malsburg. Robust classification of hand postures against complex backgrounds. In *Proceedings of the second international conference on automatic face and gesture recognition*, pages 170–175. IEEE, 1996.

[25] Stephan Liwicki and Mark Everingham. Automatic recognition of fingerspelled words in British sign language. In *2009 IEEE computer society conference on computer vision and pattern recognition workshops*, pages 50–57. IEEE, 2009.

[26] Thad Starner and Alex Pentland. Real-time American sign language recognition from video using hidden Markov models. In *Motion-based recognition*, pages 227–243. Springer, 1997.

[27] Taehwan Kim, Greg Shakhnarovich, and Karen Livescu. Fingerspelling recognition with semi-Markov conditional random fields. In *Proceedings of the IEEE international conference on computer vision*, pages 1521–1528, 2013.

[28] Herleson Paiva Pontes, João Batista Furlan Duarte, and Plácido Rogério Pinheiro. An educational game to teach numbers in Brazilian sign language while having fun. *Computers in Human Behavior*, 107:105825, 2020.

[29] Jestin Joy, Kannan Balakrishnan, and M. Sreeraj. SiLearn: An intelligent sign vocabulary learning tool. *Journal of Enabling Technologies*, 2019.

[30] Paula Escudeiro, Nuno Escudeiro, Rosa Reis, Jorge Lopes, Marcelo Norberto, Ana Bela Baltasar, Maciel Barbosa, and José Bidarra. Virtual sign–A real time bidirectional translator of Portuguese sign language. *Procedia Computer Science*, 67:252–262, 2015.

[31] Shamsul Anuar Mokhtar, Siti Sarah Shamsul Anuar, and Siti Mashitah Shamsul Anuar. Web-based application for learning Malaysian sign language. In *Proceedings of the 11th International Conference on Ubiquitous Information Management and Communication*, pages 1–6, 2017.

[32] Luka Cempre, Aleksander Bešir, and Franc Solina. Dictionary of the Slovenian sign language on the WWW. In *International Conference on Human Factors in Computing and Informatics*, pages 240–259. Springer, 2013.

[33] Mehrez Boulares and Mohamed Jemni. Mobile sign language translation system for deaf community. In *Proceedings of the international cross-disciplinary conference on web accessibility*, pages 1–4, 2012.

[34] Christoph Feichtenhofer. X3D: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 203–213, 2020.

[35] Shiqi Fang. Funsigns - An interactive educational tool to learn sign language. Thesis, Rochester Institute of Technology, December. 2019.

[36] Ching-Hua Chuan and Caroline Anne Guardino. Designing SmartSignPlay: An interactive and intelligent American sign language app for children who are deaf or hard of

hearing and their families. In *Companion publication of the 21st international conference on intelligent user interfaces*, pages 45–48, 2016.

[37] Jestin Joy, Kannan Balakrishnan, and M. Sreeraj. SignQuiz: A quiz based tool for learning fingerspelled signs in Indian sign language using ASLR. *IEEE Access*, 7:28363–28371, 2019.

[38] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[39] François Chollet et al. Keras. `https://keras.io`, 2015.

[40] Gerasimos Potamianos, Katerina Papadimitriou, Eleni Efthimiou, Stavroula-Evita Fotinea, Galini Sapountzaki, and Petros Maragos. SL-ReDu: Greek sign language recognition for educational applications. project description and early results. In *Proceedings of the 13th ACM International Conference on Pervasive Technologies Related to Assistive Environments*, pages 1–6, 2020.

[41] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. MediaPipe: A framework for perceiving and processing reality. In *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*, 2019.

[42] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. MediaPipe Hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214*, 2020.

[43] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wei Hua, Manfred Georg, and Matthias Grundmann. MediaPipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

[44] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat,

Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[45] Péter Vingelmann and Frank H.P. Fitzek. Cuda, release: 10.2.89. `https://developer.nvidia.com/cuda-toolkit`, 2020.

[46] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.