ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Αναγνώριση Δακτυλο-συλλαβισμού στην Ελληνική Νοηματική Γλώσσα με Χρήση Χαρακτηριστικών Ανθρώπινου Σκελετού και Πόζας

Διπλωματική Εργασία

Αναστασία Ψαρού

Επιβλέπων: Γεράσιμος Ποταμιάνος

Οκτώβριος 2023

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Αναγνώριση Δακτυλο-συλλαβισμού στην Ελληνική Νοηματική Γλώσσα με Χρήση Χαρακτηριστικών Ανθρώπινου Σκελετού και Πόζας

Διπλωματική Εργασία

Αναστασία Ψαρού

Επιβλέπων: Γεράσιμος Ποταμιάνος

Οκτώβριος 2023

UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Fingerspelling Recognition in the Greek Sign Language Using Human Skeleton and Pose Features

Diploma Thesis

Anastasia Psarou

Supervisor: Gerasimos Potamianos

October 2023

Εγκρίνεται από την Επιτροπή Εξέτασης:

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

Μέλος **Δημήτριος Ραφαηλιδης**
Αναπληρωτής Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών Πανεπιστήμιο Θεσσαλίας

Μέλος **Χρήστος Αντωνόπουλος**
Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών Πανεπιστήμιο Θεσσαλίας

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου, κύριο Γεράσιμο Ποταμιάνο, για την καθοδήγηση του και την Κατερίνα Παπαδημητρίου για τις συμβουλές της και την βοήθεια στην εκπόνηση της διπλωματικής εργασίας. Τέλος, θέλω να ευχαριστήσω τους φίλους και την οικογένειά μου για τη συνεχή στήριξη.

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».

Η Δηλούσα

Αναστασία Ψαρού

Διπλωματική Εργασία

Αναγνώριση Δακτυλο-συλλαβισμού στην Ελληνική Νοηματική Γλώσσα με Χρήση Χαρακτηριστικών Ανθρώπινου Σκελετού και Πόζας

Αναστασία Ψαρού

Περίληψη

Η νοηματική γλώσσα είναι μια οπτικο-κινησιακή γλώσσα που αξιοποιεί την κίνηση των χεριών, τη στάση ή την κίνηση του σώματος, καθώς και τις εκφράσεις του προσώπου για να μεταφέρει ένα νόημα. Η νοηματική γλώσσα δεν είναι παγκοσμίως ενιαία, αλλά οι κωφοί σε κάθε χώρα χρησιμοποιούν τη δική τους ξεχωριστή νοηματική γλώσσα. Τμήμα της νοηματικής γλώσσας αποτελεί ο δακτυλοσυλλαβισμός. Δακτυλοσυλλαβισμός είναι η διαδικασία αναπαράστασης των γραμμάτων που ανήκουν σε ένα σύστημα γραφής καθώς και αριθμών χρησιμοποιώντας αποκλειστικά τα χέρια ως μέσο.

Τα τελευταία χρόνια, έχουν αναπτυχθεί αρκετά μοντέλα για την ανίχνευση της στάσης του ανθρώπινου σώματος. Κάποια από αυτά τα μοντέλα είναι το MediaPipe, το PIXIE και το ExPose, τα οποία μπορούν να χρησιμοποιηθούν για τον προσδιορισμό της στάσης των χεριών και κατά συνέπεια του δακτυλοσυλλαβισμού. Παράλληλα, έχουν αναπτυχθεί μοντέλα για την ταξινόμηση εικόνων. Τέτοια μοντέλα είναι το Συνελικτικό Νευρωνικό Δίκτυο (ΣΝΔ) και ο Vision Transformer (ViT) και μπορούν να χρησιμοποιηθούν για την ταξινόμηση των γραμμάτων του ελληνικού αλφαβήτου από εικόνες.

Αυτή η διπλωματική εργασία επικεντρώνεται στην αναγνώριση του δακτυλοσυλλαβισμού της Ελληνικής Νοηματικής Γλώσσας (ΕΝΓ) από βίντεο. Εξετάζει, πιο συγκεκριμένα, το στατικό σενάριο (frame-to-letter), στο οποία μια εικόνα αντιστοιχεί σε ένα γράμμα της ελληνικής αλφαβήτου. Τα μοντέλα που χρησιμοποιούνται για αυτό το σκοπό υποβάλλονται σε εκπαίδευση χρησιμοποιώντας τις μεθόδους MediaPipe, PIXIE, ExPose, CNN και ViT. Στη συνέχεια, πραγματοποιούνται συγκρίσεις μεταξύ αυτών των μοντέλων, με στόχο να αξιολογηθεί η ικανότητά τους να πραγματοποιούν ακριβείς προβλέψεις και να γενικεύουν τα αποτελέσματά τους.

Λέξεις-κλειδιά:

Ελληνική Νοηματική Γλώσσα, Βαθιά μάθηση, Όραση υπολογιστών, Ταξινόμηση εικόνων

Diploma Thesis

Fingerspelling Recognition in the Greek Sign Language Using Human Skeleton and Pose Features

Anastasia Psarou

Abstract

Sign language is a form of communication that is used by Deaf people and utilizes hand movements, body positioning or movement, as well as facial expressions to convey meaning. Sign language is not globally unified; rather, the deaf in each country use their own distinct sign language. Fingerspelling is a key element of this communication method, including the use of hands to represent the letters of a writing system and numerical systems as well.

In recent years, several models have been developed for body pose detection. These models involve MediaPipe, PIXIE, and ExPose, which can be used for hand landmark detection. Meanwhile, ongoing research has been conducted on models used for image classification, including Convolution Neural Network (CNN) and Vision Transformer (ViT). Such models can be used to classify the letters of the Greek alphabet from images.

This thesis focuses on fingerspelling translation of Greek Sign Language from video recordings. It specifically explores the static scenario, where the translation is performed on a frame-to-letter basis. The models used for this purpose undergo training using MediaPipe, PIXIE, ExPose, CNN and ViT methods. Subsequently, comparisons are carried out among these models to gain insights into their prediction and generalization capabilities.

Keywords:

Greek Sign Language, Deep learning, Computer vision, Image Classification

Πίνακας περιεχομένων

Κατάλογος σχημάτων

Κατάλογος πινάκων

Συντομογραφίες

ΕΝΓ	Ελληνική Νοηματική Γλώσσα
3D	3 Dimensional
2D	2 Dimensional
NN	Neural Network
ΣΝΔ	Συνελικτικά Νευρωνικά Δίκτυα
CNN	Convolutional Neural Network
ViT	Vision Transformer
CER	Character Error Rate
RNN	Recurrent Neural Network
RGB	Red Green Blue

Κεφάλαιο 1

Εισαγωγή

Η νοηματική γλώσσα είναι η φυσική γλώσσα για την επικοινωνία των κωφών και βαρήκοων ανθρώπων. Αποτελεί τη μόνη γλώσσα που βασίζεται στην έκφραση του προσώπου, στο συμβολισμό λέξεων με χειρομορφές και φράσεων με κινήσεις των χεριών διαθέτοντας πλήρη συντακτική δομή. Ένα σημαντικό τμήμα της νοηματικής γλώσσας είναι ο δακτυλοσυλλαβισμός, ο οποίος αποτελεί τη διαδικασία αναπαράστασης γραμμάτων και αριθμών χρησιμοποιώντας αποκλειστικά τα χέρια ως μέσο.

Σημαντικό είναι να σημειωθεί ότι δεν υπάρχει μια παγκόσμια, ενιαία νοηματική γλώσσα. Κάθε χώρα διαθέτει τη δική της νοηματική γλώσσα, η οποία αποτελείται από μοναδικά νοήματα και διαθέτει το δικό της ξεχωριστό αλφάβητο. Σε αυτήν τη διπλωματική εργασία, εξετάζεται ο δακτυλοσυλλαβισμός στην Ελληνική Νοηματική Γλώσσα (ΕΝΓ) [12].

1.1 Αντικείμενο της διπλωματικής

Το αντικείμενο της διπλωματικής εργασίας αφορά την αναγνώριση του δακτυλοσυλλαβισμού της ΕΝΓ από βίντεο. Τα βίντεο διαχωρίζονται σε ακολουθίες εικόνων, όπου κάθε εικόνα αντιστοιχεί σε ένα συγκεκριμένο γράμμα. Συνεπώς, εξετάζεται το στατικό σενάριο αναγνώρισης του δακτυλοσυλλαβισμού. Για την επίτευξη αυτού του στόχου, χρησιμοποιούνται δύο κύριες μέθοδοι. Οι πρώτες μέθοδοι που εξετάζονται σχετίζονται με την αναγνώριση της στάσης του σώματος, όπως οι MediaPipe, PIXIE, και ExPose. Επιπλέον, χρησιμοποιούνται μέθοδοι ταξινόμησης εικόνων, όπως ο ViT και το ΣΝΔ. Τέλος, συγκρίνονται και αξιολογούνται τα αποτελέσματα που παράγουν αυτές οι μέθοδοι όσον αφορά την ακρίβεια της ανίχνευσης του δακτυλοσυλλαβισμού στην ΕΝΓ.

1.1.1 Συνεισφορά

Η συνεισφορά της διπλωματικής συνοψίζεται ως εξής:

1. Μελετήθηκαν διαφορετικές μέθοδοι αναγνώρισης της στάσης του ανθρώπινου σώματος με σκοπό την πρόβλεψη του δακτυλοσυλλαβισμού στην ΕΝΓ.

2. Διερευνήθηκαν μέθοδοι ταξινόμησης εικόνων για την πρόβλεψη του δακτυλοσυλλαβισμού από εικόνες.

3. Πραγματοποιήθηκε εκπαίδευση μοντέλων χρησιμοποιώντας τις προαναφερθείσες μεθόδους.

4. Πραγματοποιήθηκε ανάλυση διαφορετικών στρώσεων στο πλαίσιο της αρχιτεκτονικής Συνελικτικών Νευρωνικών Δικτύων (ΣΝΔ) με σκοπό τη βελτίωση της εκπαίδευσης των μοντέλων.

5. Παρουσιάστηκαν τα αποτελέσματα της εκπαίδευσης των παραπάνω μοντέλων και αξιολογήθηκαν τα αποτελέσματα τους.

1.2 Οργάνωση του τόμου

Η δομή της διπλωματικής εργασίας είναι η εξής:

- Το κεφάλαιο 2 αναλύει τη βάση δεδομένων που χρησιμοποιείται στη συγκεκριμένη εργασία.

- Το κεφάλαιο 3 παρουσιάζει τις μεθόδους ανίχνευσης της στάσης του ανθρώπινου σώματος που εφαρμόζονται για τον προσδιορισμό του δακτυλοσυλλαβισμού στην ΕΝΓ.

- Το κεφάλαιο 4 αναλύει τις μεθόδους ταξινόμησης εικόνων που χρησιμοποιούνται για την αναγνώριση του δακτυλοσυλλαβισμού στην ΕΝΓ.

- Το κεφάλαιο 5 παρουσιάζει τη διαδικασία εκπαίδευσης των μοντέλων με βάση τις μεθόδους που αναφέρονται στα κεφάλαια 3 και 4. Παράλληλα, εκτιμάται η επίδοση κάθε ενός μοντέλου και υλοποιείται σύγκριση της απόδοσης τους.

- Το κεφάλαιο 6 περιγράφει τη σύγκριση των μοντέλων που κατασκευάζονται στο κεφάλαιο 5 με χρήση της μετρικής του ποσοστού σφάλματος χαρακτήρων (CER).

- Το κεφάλαιο 7 παρουσιάζει τα συμπεράσματα της διπλωματικής εργασίας, καθώς και προτάσεις για μελλοντικές βελτιώσεις που θα μπορούσαν να εφαρμοστούν.

Κεφάλαιο 2

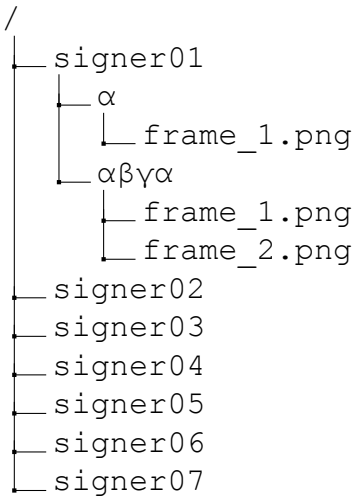
Ανάλυση της βάση δεδομένων

2.1 Περιγραφή της βάσης δεδομένων

Για την υλοποίηση αυτής της εργασίας χρησιμοποιείται ένα σύνολο από βίντεο που καταγράφουν άτομα να εκτελούν δακτυλοσυλλαβισμούς που αντιπροσωπεύουν γράμματα και λέξεις της ελληνικής αλφαβήτας. Αυτά τα δεδομένα παράχθηκαν και χρησιμοποιήθηκαν από το SL-ReDu [13], ένα καινοτόμο πρόγραμμα του Πανεπιστημίου Θεσσαλίας το οποίο αναγνωρίζει αυτόματα την ελληνική νοηματική γλώσσα από βίντεο.

Σε κάθε βίντεο παρουσιάζεται ένα άτομο που εκτελεί μια σειρά από δακτυλοσυλλαβισμούς. Στην αρχή κάθε βίντεο, ο διερμηνέας έχει τα χέρια του στο ύψος της λεκάνης. Όσο το βίντεο προχωράει, ο διερμηνέας μετακινεί τα χέρια του στο ύψος των ώμων και προσεγγίζει όλο και πιο πολύ το δακτυλοσυλλαβισμό. Ο δακτυλοσυλλαβισμός γίνεται όλο και πιο εμφανής καθώς το βίντεο προχωρά. Στο τέλος του βίντεο, τα χέρια του διερμηνέα επιστρέφουν στα πόδια του.

Η βάση δεδομένων περιλαμβάνει 21 διερμηνείς κάθε ένας από τους οποίους νοηματίζει σε 74 βίντεο. Κάθε άτομο νοηματίζει τα 24 γράμματα της ελληνικής αλφαβήτας και 50 ελληνικές λέξεις, που δεν είναι οι ίδιες μεταξύ των διερμηνέων. Τα βίντεο έχουν χωριστεί σε ακολουθίες εικόνων, χρησιμοποιώντας τη βιβλιοθήκη OpenCV της python, η οποία παρουσιάζεται στο [14]. Όσον αφορά τη δομή αποθήκευσης των ακολουθιών των εικόνων, κάθε νοηματίζον αντιπροσωπεύεται από ένα ξεχωριστό φάκελο. Μέσα σε κάθε φάκελο διερμηνέα υπάρχουν υποφάκελοι. Αυτοί οι υποφάκελοι έχουν ονόματα που αντιστοιχούν στο συγκεκριμένο ελληνικό γράμμα ή λέξη που αναπαριστούν οι ακολουθίες των εικόνων που βρίσκονται μέσα σε αυτούς. Αυτή η δομή φαίνεται με σαφήνεια στο Σχήμα 2.1.

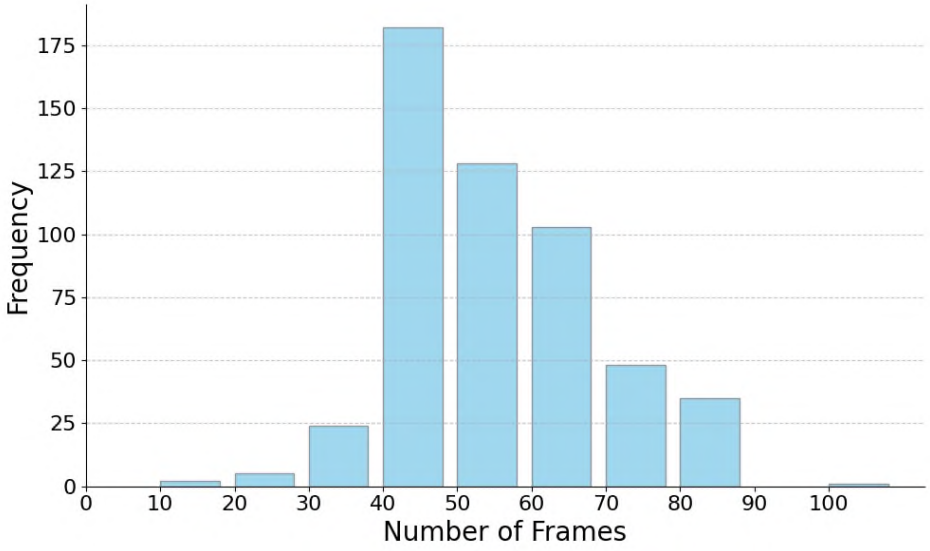


Σχήμα 2.1: Δομή αποθήκευσης των ακολουθιών των εικόνων κάθε διερμηνέα.

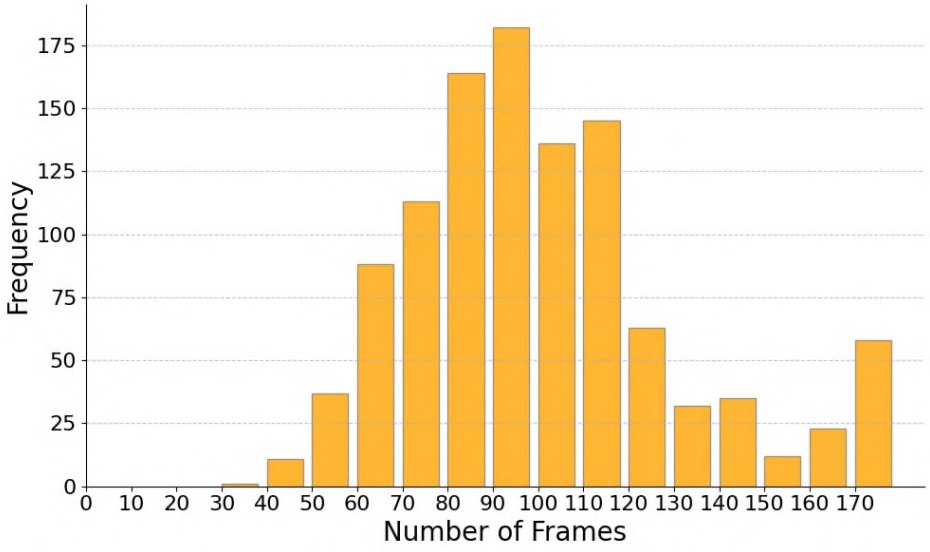
Αξίζει να αναφερθεί ότι υπάρχει ένας διερμηνέας που νοηματίζει με το αριστερό χέρι. Σε αυτή την περίπτωση, οι εικόνες που απεικονίζουν το συγκεκριμένο άτομο έχουν υποστεί αναστροφή κατά μήκος του άξονα y. Μετά την πραγματοποίηση αυτής της αναστροφής, οι εικόνες επεξεργάζονται με τον ίδιο τρόπο όπως και οι υπόλοιπες εικόνες του συνόλου δεδομένων.

2.2 Στατιστικά που αφορούν τη βάση δεδομένων

Αυτή η υποενότητα παρέχει πρόσθετες πληροφορίες για τη βάση δεδομένων. Η ανάλυση όλων των εικόνων είναι 540x960 pixels, και η κατανομή των ακολουθιών των εικόνων ανά βίντεο παρουσιάζεται στα Σχήματα 2.2 και 2.3. Τα βίντεο στα οποία νοηματίζονται γράμματα αποτελούνται επί των πλείστων από 40 έως 50 εικόνες ενώ τα βίντεο στα οποία νοηματίζονται λέξεις αποτελούνται επί των πλείστων από 90 έως 100 εικόνες.



Σχήμα 2.2: Ιστόγραμμα του αριθμού των εικόνων ανά βίντεο στο οποίο νοηματίζεται ένα γράμμα.



Σχήμα 2.3: Ιστόγραμμα του αριθμού των εικόνων ανά βίντεο στο οποίο νοηματίζεται μια λέξη.

Κεφάλαιο 3

Μέθοδοι ανίχνευση της πόζας των χεριών

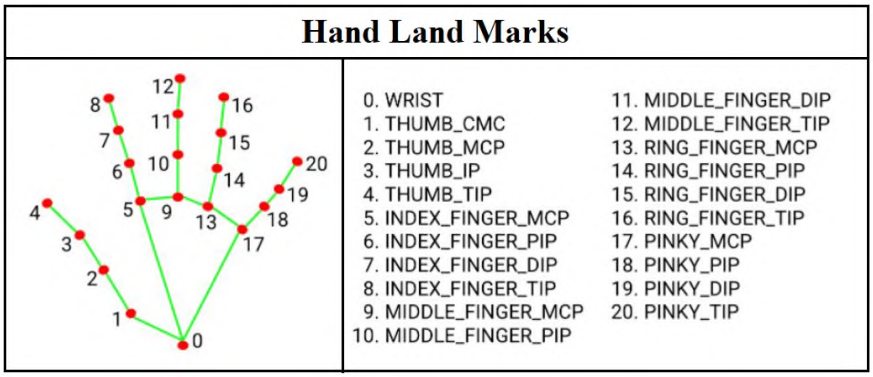
Στο παρόν κεφάλαιο, αναλύονται οι μέθοδοι που χρησιμοποιούνται για την ανίχνευση της στάσης του δεξιού χεριού. Εξηγείται η λειτουργία τους και αναδεικνύονται οι λόγοι που τις ξεχωρίζουν σε σύγκριση με άλλα μοντέλα ανίχνευσης της στάσης του σώματος.

3.1 MediaPipe

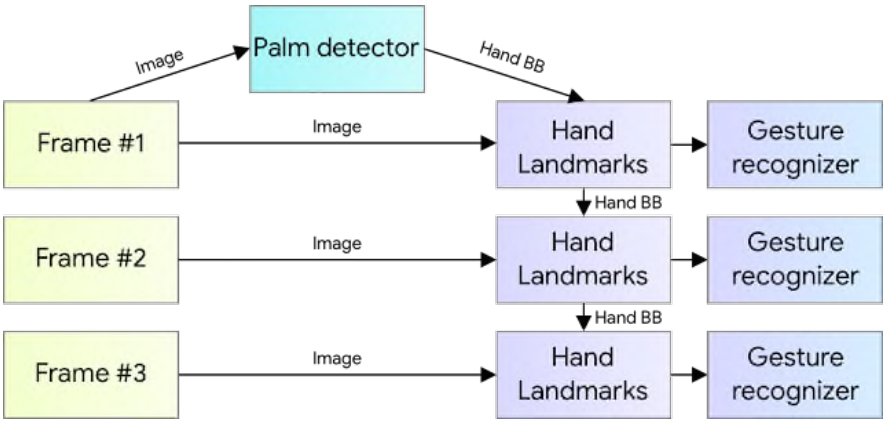
Το MediaPipe [15] είναι μια βιβλιοθήκη ανοιχτού κώδικα που αναπτύχθηκε από την Google και παρέχει ένα μεγάλο αριθμό προ-κατασκευασμένων μοντέλων για την αναγνώριση του ανθρώπινου σώματος και της κίνησής του. Μερικά από αυτά είναι η ανίχνευση του προσώπου και της πόζας του, η αναγνώριση των χεριών και της πόζας τους, ο προσδιορισμός της πόζας ολόκληρου του σώματος και το ολιστικό μοντέλο που συνδυάζει τα παραπάνω, όπως φαίνεται στο Σχήμα 3.1. Σε αυτήν την εργασία, το MediaPipe θα χρησιμοποιηθεί για την ανίχνευση των σημείων της ανθρώπινης παλάμης.

Σχήμα 3.1: Ολιστικό μοντέλο του MediaPipe - Σχήμα από [1].

Το MediaPipe αναγνωρίζει τα ορόσημα του χεριού εντοπίζοντας 21 3D σημεία σε κάθε χέρι, σύμφωνα με το Σχήμα 3.2. Αυτό το μοντέλο αποτελείται από ένα μοντέλο ανίχνευσης παλάμης και ένα μοντέλο ανίχνευσης των ορόσημων του χεριού. Το πρώτο ανιχνεύει το χέρι στην εικόνα, ενώ το δεύτερο αναγνωρίζει τα ορόσημα στο χέρι, όπως φαίνεται στο Σχήμα 3.3.



Σχήμα 3.2: Προσδιορισμός των οροσήμων του χεριού από το MediaPipe - Σχήμα από [2].

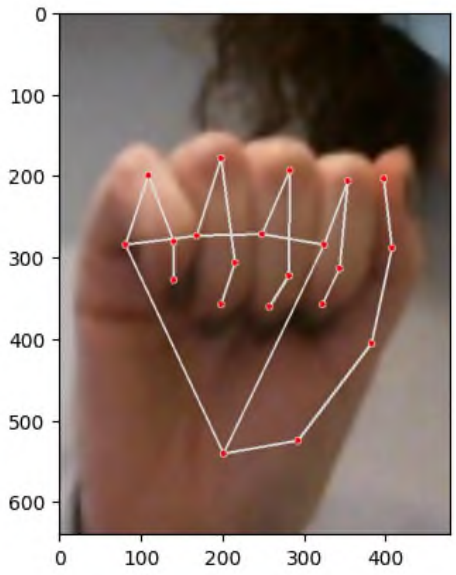


Σχήμα 3.3: Επισκόπηση της διαδικασίας ανίχνευσης των οροσήμων του χεριού από το MediaPipe - Σχήμα από [3].

Οι διαστάσεις x, y και z κάθε ορόσημου αποθηκεύονται σε μια σειρά ενός πίνακα για κάθε εικόνα. Οι x, y διαστάσεις αντιπροσωπεύουν τις οριζόντιες και κάθετες συντεταγμένες κάθε ορόσημου, όπως φαίνεται στο Σχήμα 3.4. Η z διάσταση αντιπροσωπεύει το βάθος κάθε σημείου έχοντας τον καρπό ως σημείο αναφοράς. Όσο μικρότερη η τιμή του z, τόσο πιο κοντά είναι το ορόσημο στην κάμερα. Αυτές οι συντεταγμένες, παράλληλα, υπόκεινται κανονικοποίηση με αποτέλεσμα να παίρνουν τιμές από 0 έως 1, πράγμα που μπορεί να

συμβάλλει στην καλύτερη απόδοση του μοντέλου που θα διοχετευθούν αργότερα.

Επομένως, κάθε δακτυλοσυλλαβισμός σε μια εικόνα αντιπροσωπεύεται από 63 σημεία, τα οποία αποθηκεύονται σε έναν πίνακα μαζί με το γράμμα που αναπαριστούν. Υπάρχουν περιπτώσεις στις οποίες το MediaPipe δεν καταφέρνει να ανιχνεύει δακτυλοσυλλαβισμό, καθώς κάποιες εικόνες είναι θολές. Σε αυτές τις περιπτώσεις, οι συγκεκριμένες εικόνες παραλείπονται και δεν επεξεργάζονται παραπάνω.



Σχήμα 3.4: Αναγνώριση σημείων ανθρώπινης παλάμης από το MediaPipe σε εικόνα που περιλαμβάνεται στη βάση δεδομένων.

3.2 PIXIE

Το PIXIE (Collaborative Regression of Expressive Bodies) [4] είναι ένα μοντέλο που ανακατασκευάζει το τρισδιάστατο σχήμα και τη στάση του σώματος, την άρθρωση των χεριών και την έκφραση του προσώπου του ανθρώπου από μια μόνο εικόνα, όπως παρουσιάζεται στο Σχήμα 3.5. Παράλληλα, είναι ικανό να εκτιμήσει την έκφραση του προσώπου, τον φωτισμό, τη λευκαύγεια[1] και τις τρισδιάστατες αποκλίσεις για την επιφάνεια του προσώπου.

Το PIXIE λειτουργεί προσδιορίζοντας ξεχωριστά το σώμα, το πρόσωπο και τα χέρια και έπειτα συνδυάζει τα επιμέρους αποτελέσματα. Παράλληλα, προσδίδει ένα επίπεδο εμπιστοσύνης σε κάθε επιμέρους συνιστώσα, όπως φαίνεται στο Σχήμα 3.6. Αυτό σημαίνει ότι κάθε

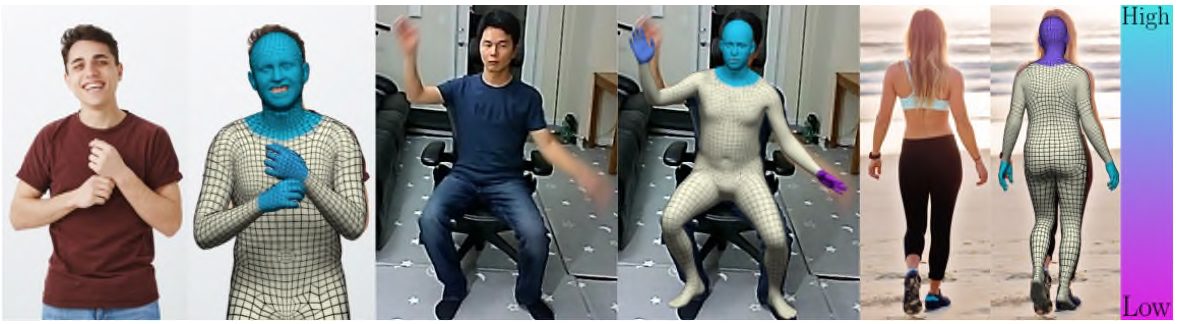
[1] Η λευκαύγεια είναι το μέτρο της ανακλαστικότητας μιας επιφάνειας ή ενός σώματος.

πρόβλεψη που δημιουργείται από το PIXIE συνοδεύεται από μια μετρική, η οποία αντιπροσωπεύει το επίπεδο εμπιστοσύνης του ίδιου του μοντέλου για την εκάστοτε πρόβλεψη. Σε αντίθεση, παρόμοια μοντέλα ενσωματώνουν τα επιμέρους στοιχεία χωρίς να χρησιμοποιούν κάποια τέτοια μετρική.

Σχήμα 3.5: Ανίχνευση της στάσης του σώματος από το PIXIE. Από αριστερά προς τα δεξιά: (1) αρχική εικόνα, (2) πρόβλεψη στάσης χρησιμοποιώντας μόνο το μοντέλο σώματος του PIXIE, (3) πρόβλεψη χρησιμοποιώντας το μοντέλο PIXIE - Σχήμα από [4]

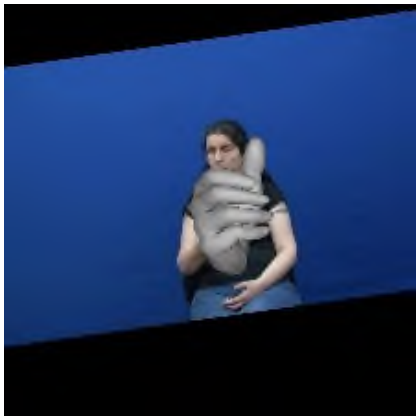
Το PIXIE λαμβάνει υπόψιν ότι για τον προσδιορισμό της στάσης του σώματος απαιτείται η γνώση του φύλου του ατόμου που αναπαριστάται. Για αυτό το λόγο, τα δεδομένα που χρησιμοποιούνται για τη εκπαίδευσή του κατηγοριοποιούνται σε δεδομένα που αναπαριστούν άντρα, γυναίκα ή άτομα που ανήκουν στο μη-δυαδικό φύλο.

Σκοπός αυτής της εργασίας είναι ο προσδιορισμός του δακτυλοσυλλαβισμού στην ΕΝΓ. Για αυτόν το λόγο, αξιοποιείται το τμήμα του PIXIE που σχετίζεται με την αναπαράσταση της στάσης του δεξιού χεριού. Ένα παράδειγμα αυτής της αναπαράστασης εμφανίζεται στο Σχήμα 3.7. Το PIXIE προσδιορίζει την πόζα του δεξιού χεριού επιστρέφοντας 15 πίνακες διαστάσεων 3 * 3. Παράδειγμα λειτουργίας του PIXIE σε εικόνα της βάσης δεδομένων, που χρησιμοποιήθηκε σε αυτήν την εργασία, παρουσιάζεται στο Σχήμα 3.8.



Σχήμα 3.6: Αναγνώριση της στάσης του σώματος από το PIXIE, χρησιμοποιώντας το βαθμό εμπιστοσύνης - Σχήμα από [4]

Σχήμα 3.7: Αναπαράσταση τρισδιάστατης μορφής του χεριού από το PIXIE. Από αριστερά προς τα δεξιά: (1) αρχική εικόνα, (2) προσδιορισμός της τρισδιάστατης μορφής του χεριού από το PIXIE, (3) προσθήκη της υφής του χεριού - Σχήμα από [5].



Σχήμα 3.8: Παράδειγμα λειτουργίας του PIXIE σε εικόνα της βάσης δεδομένων

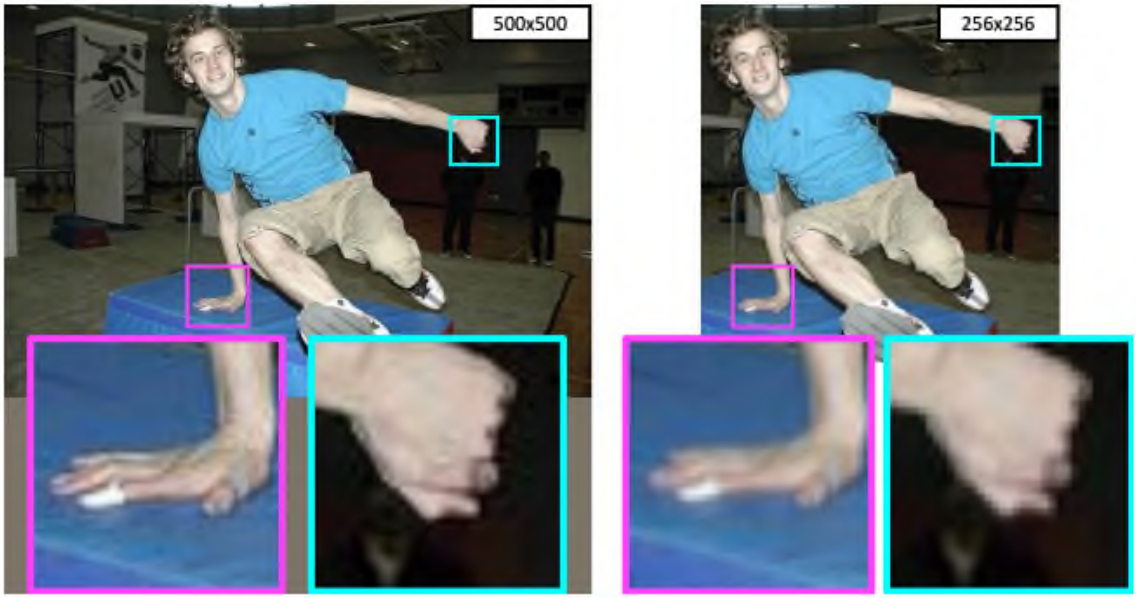
3.3 ExPose

Το ExPose (EXpressive POse and Shape rEgression) [7] είναι ένα μοντέλο που προσδιορίζει την τρισδιάστατη έκφραση του ανθρώπινου σώματος. Ο μηχανισμός του ExPose λειτουργεί παλινδρομώντας το σώμα, το πρόσωπο και τα χέρια, σε μορφή SMPL-X, από μια εικόνα RGB. Το SMPL-X (SMPL eXpressive) [6] είναι ένα μοντέλο που αποτυπώνει από κοινού το σώμα μαζί με το πρόσωπο και τα χέρια. Ουσιαστικά, τα αποτελέσματα του ExPose αποτελούν παραμέτρους του SMPL-X μοντέλου.

Σχήμα 3.9: Παράδειγμα λειτουργίας του μοντέλου SMPL-X - Σχήμα από [6]

Η ανάπτυξη του ExPose απαιτούσε τη διερεύνηση και επίλυση ορισμένων σημαντικών προβλημάτων από τους δημιουργούς του. Αρχικά, τόσο τα χέρια όσο και το πρόσωπο καταλαμβάνουν λίγα pixel σε μια εικόνα σε σχέση με το σώμα. Αυτή η δυσαναλογία μπορεί να δυσχαιράνει την ακρίβεια της εκτίμησης των χεριών και του προσώπου, ιδίως όταν οι εικόνες μειώνονται σε μέγεθος για τις ανάγκες των νευρωνικών δικτύων, όπως φαίνεται στο Σχήμα 3.10.

Για αυτό το λόγο, το ExPose περιλαμβάνει έναν μηχανισμό, την προσοχή. Αυτός ο μηχανισμός χρησιμοποιείται στις περιοχές που εμφανίζεται το πρόσωπο και τα χέρια με σκοπό να εξαχθούν εικόνες με μεγαλύτερη ακρίβεια από την αρχική εικόνα. Ταυτόχρονα, αξιοποιείται η γνώση που αποκομίζεται από τα σύνολα δεδομένων που περιέχουν δεδομένα μόνο προσώπου ή μόνο χεριών. Παράδειγμα της ακρίβειας απεικόνισης των χεριών και του προσώπου παρουσιάζεται στο Σχήμα 3.11. Γενικότερα το ExPose κάνει εκτιμήσεις με μεγαλύτερη ακρίβεια σε σχέση με τα υπάρχοντα μοντέλα με πολύ μικρότερο υπολογιστικό κόστους.



Σχήμα 3.10: Απεικόνιση της μειωμένη ακρίβειας των χεριών μετά την μείωση του μεγέθους της αρχικής εικόνας για τις ανάγκες των νευρωνικών δικτύων - Σχήμα από [7]

Σχήμα 3.11: Απεικόνιση των αποτελεσμάτων του ExPose από διάφορες οπτικές γωνίες, χρησιμοποιώντας το μηχανισμό της προσοχής - Σχήμα από [7]

Σχετικά με την ανάλυση της στάσης του δεξιού χεριού, το ExPose χρησιμοποιεί μια δομή από 15 πίνακες με διαστάσεις 3*3 για την αναπαράστασή της. Αυτοί οι πίνακες λειτουργούν ως μέσο για την πλήρη περιγραφή της τρισδιάστατης θέσης και κατεύθυνσης του χεριού, παρέχοντας στο ExPose τη δυνατότητα να προσδιορίσει το δακτυλοσυλλαβισμό στην ΕΝΓ. Ένα παράδειγμα της λειτουργίας του ExPose σε μια εικόνα, που προέρχεται από τη βάση δεδομένων που χρησιμοποιήθηκε σε αυτήν την εργασία, παρουσιάζεται στο Σχήμα 3.12.

Σχήμα 3.12: Παράδειγμα λειτουργίας του ExPose σε εικόνα της βάσης δεδομένων

3.4 Συνδυασμός μοντέλων PIXIE, MediaPipe και ExPose, MediaPipe

Όπως αναφέρθηκε στην ενότητα 3.1. το τμήμα του MediaPipe που επικεντρώνεται στην αναγνώριση της πόζας των χεριών αποτελείται από δύο μοντέλα. Το πρώτο μοντέλο αφορά την αναγνώριση της παλάμης και το δεύτερο επικεντρώνεται στην αναγνώριση των ορόσημων του χεριού. Σε αυτήν την υποενότητα, περιγράφεται η διαδικασία ενσωμάτωσης του συστήματος αναγνώριση της παλάμης του MediaPipe με το μοντέλο PIXIE και ο συνδυασμός του συστήματος αναγνώρισης της παλάμης του MediaPipe με το μοντέλο ExPose.

Σύμφωνα με το κεφάλαιο 2.1, οι ακολουθίες εικόνων που προέρχονται από ένα βίντεο περιέχουν εικόνες όπου το δεξί χέρι δεν είναι ορατό ή εικόνες που είναι αρκετά θολές. Επομένως, είναι πιθανό το μοντέλο του MediaPipe, το οποίο ανιχνεύει την παρουσία χεριών σε μια εικόνα, να μην αναγνωρίζει κάποιο χέρι σε αυτές τις περιπτώσεις.

Το σενάριο που παρουσιάζεται σε αυτήν την υποενότητα αφορά την αξιοποίηση του μοντέλου του MediaPipe που ανιχνεύει την ύπαρξη ή όχι χεριών σε μια εικόνα. Συγκεκριμένα, το MediaPipe εντοπίζει τις εικόνες στις οποίες παρουσιάζονται χέρια, και στη συνέχεια αυτές οι εικόνες χρησιμοποιούνται ως είσοδος στα μοντέλα PIXIE και ExPose. Σε αυτήν την περίπτωση, τα δεδομένα εισόδου περιέχουν λιγότερο θόρυβο οπότε είναι πιθανό ότι τα μοντέλα που εκπαιδεύονται με βάση αυτά τα δεδομένα να είναι πιο ακριβή στον προσδιορισμό του δακτυλοσυλλαβισμού στην ΕΝΓ.

Κεφάλαιο 4

Μέθοδοι ταξινόμησης εικόνων

Σε αυτήν την ενότητα παρουσιάζονται δύο μέθοδοι ταξινόμησης εικόνων που χρησιμοποιούνται για τον προσδιορισμό του δακτυλοσυλλαβισμού στην ΕΝΓ. Μέσω αυτής της διαδικασίας, εξετάζεται η ικανότητα των μεθόδων ταξινόμησης να αναγνωρίζουν τις διαφορές στη στάση των χεριών του ατόμου.

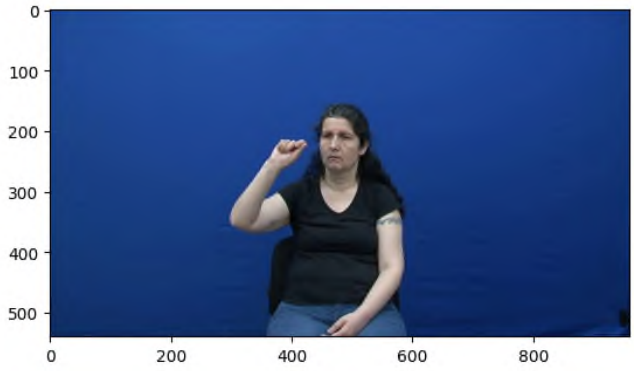
Για την υλοποίηση αυτών των μεθόδων, οι αρχικές εικόνες υποβάλλονται σε μια διαδικασία απομόνωσης των χειρονομιών, προτού περάσουν στη διαδικασία ταξινόμησης. Με αυτόν τον τρόπο, πραγματοποιείται εστίαση στον δακτυλοσυλλαβισμό και μειώνεται ο εξωτερικός θόρυβος.

Για να επιτευχθεί η απομόνωση των χειρονομιών, χρησιμοποιείται η βιβλιοθήκη MediaPipe. Η βιβλιοθήκη, αυτή, ανιχνεύει την παρουσία χεριών σε μια εικόνα. Όταν εντοπίζει το δεξί χέρι, πραγματοποιείται η απομόνωσή του και η αποθήκευσή του ως ξεχωριστή εικόνα, αντίστοιχη με αυτή που παρουσιάζεται στην Εικόνα 4.1. Αυτή η διαδικασία πραγματοποιείται για όλες τις εικόνες του συνόλου δεδομένων πριν την εφαρμογή των μεθόδων ταξινόμησης.

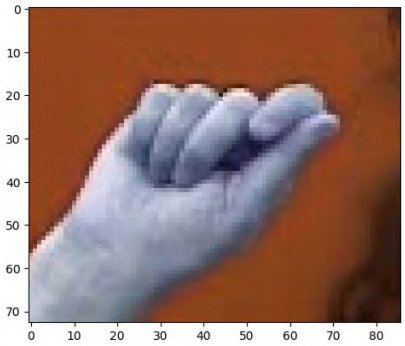
4.1 Vision transformer

Ο Vision Transformer (ViT) αποτελεί μοντέλο ταξινόμησης εικόνων που χρησιμοποιεί αρχιτεκτονική μετασχηματιστή (transformer). Ο μετασχηματιστής αποτελεί μια αρχιτεκτονική βαθιάς μάθησης που επικεντρώνεται στο μηχανισμό πολυκεφαλικής προσοχής (multi-head attention).

Σύμφωνα με το άρθρο [8], η προσοχή μπορεί να περιγραφεί ως την αντιστοίχιση ενός



(α′) Αρχική εικόνα.

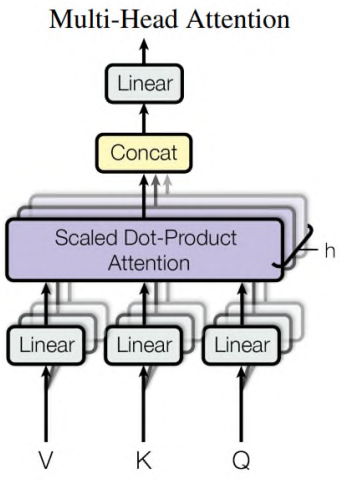


(β′) Εικόνα που περιέχει το απομονωμένο δεξί χέρι.

Σχήμα 4.1: Παράδειγμα της διαδικασίας απομόνωσης του δεξιού χεριού.

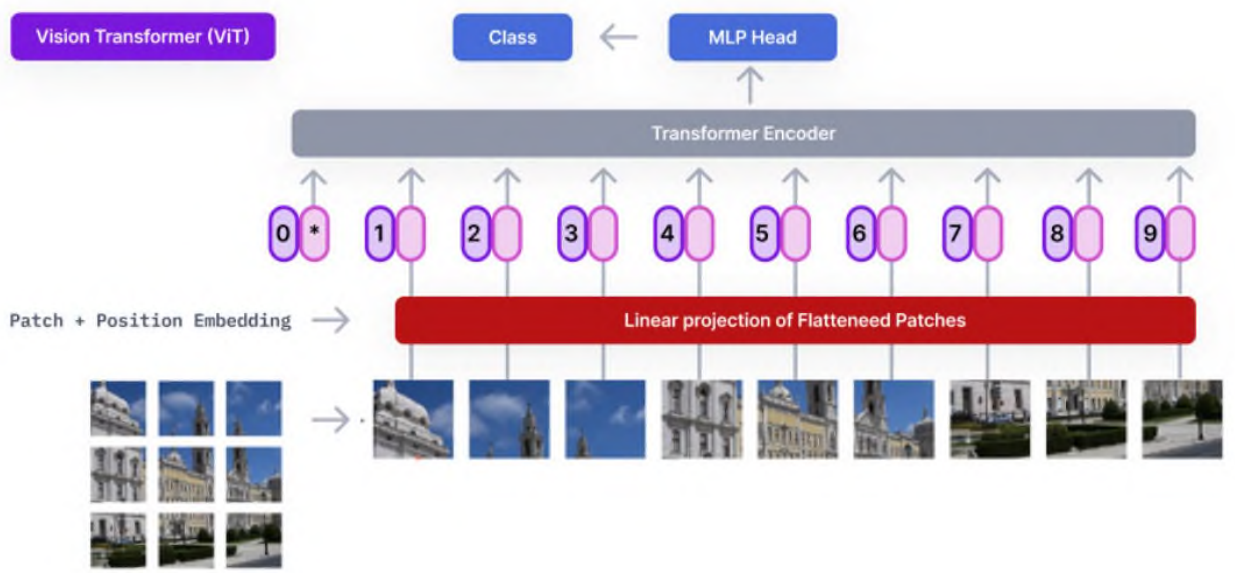
ερωτήματος και ενός συνόλου ζευγών κλειδιών-τιμών σε μια έξοδο, όπου το ερώτημα, τα κλειδιά, οι τιμές και η έξοδος είναι όλα διανύσματα. Η έξοδος υπολογίζεται ως το σταθμισμένο άθροισμα των τιμών, όπου το βάρος που αποδίδεται σε κάθε τιμή υπολογίζεται από μια συνάρτηση συμβατότητας της ερώτησης με το αντίστοιχο κλειδί.

Στην περίπτωση της πολυκεφαλικής προσοχής μειώνονται οι διαστάσεις των διανυσμάτων του ερωτήματος, των κλειδιών και των τιμών. Έπειτα, ο υπολογισμός της συνάρτησης προσοχής υλοποιείται παράλληλα για κάθε συνδυασμό ερωτήματος, κλειδιού και τιμής. Στη συνέχεια, τα αποτελέσματα συνενώνονται και μεταφέρονται στις αρχικές διαστάσεις. Αυτή η διαδικασία περιγράφεται στο Σχήμα 4.2.



Σχήμα 4.2: Αρχιτεκτονική του μηχανισμού πολυκεφαλικής προσοχής (multi-head attention) - Σχήμα από [8]

Σύμφωνα με το άρθρο [16], κατά τη λειτουργία του ViT κάθε εικόνα διαιρείται σε μια ακολουθία μη επικαλυπτόμενων και σταθερού μεγέθους τμημάτων (patches). Αυτά τα τμήματα στη συνέχεια τροφοδοτούνται στον κωδικοποιητή του μετασχηματιστή, όπως φαίνεται στο Σχήμα 4.3. Ο κωδικοποιητής αποτελείται από στρώσεις νευρωνικών δικτύων και λειτουργεί ως βασικό μοντέλο για την εξαγωγή χαρακτηριστικών από τις εικόνες. Τέλος, τα αποτελέσματα που προκύπτουν από τον κωδικοποιητή χρησιμοποιούνται για την πρόβλεψη της τάξης που ανήκει κάθε εικόνα.



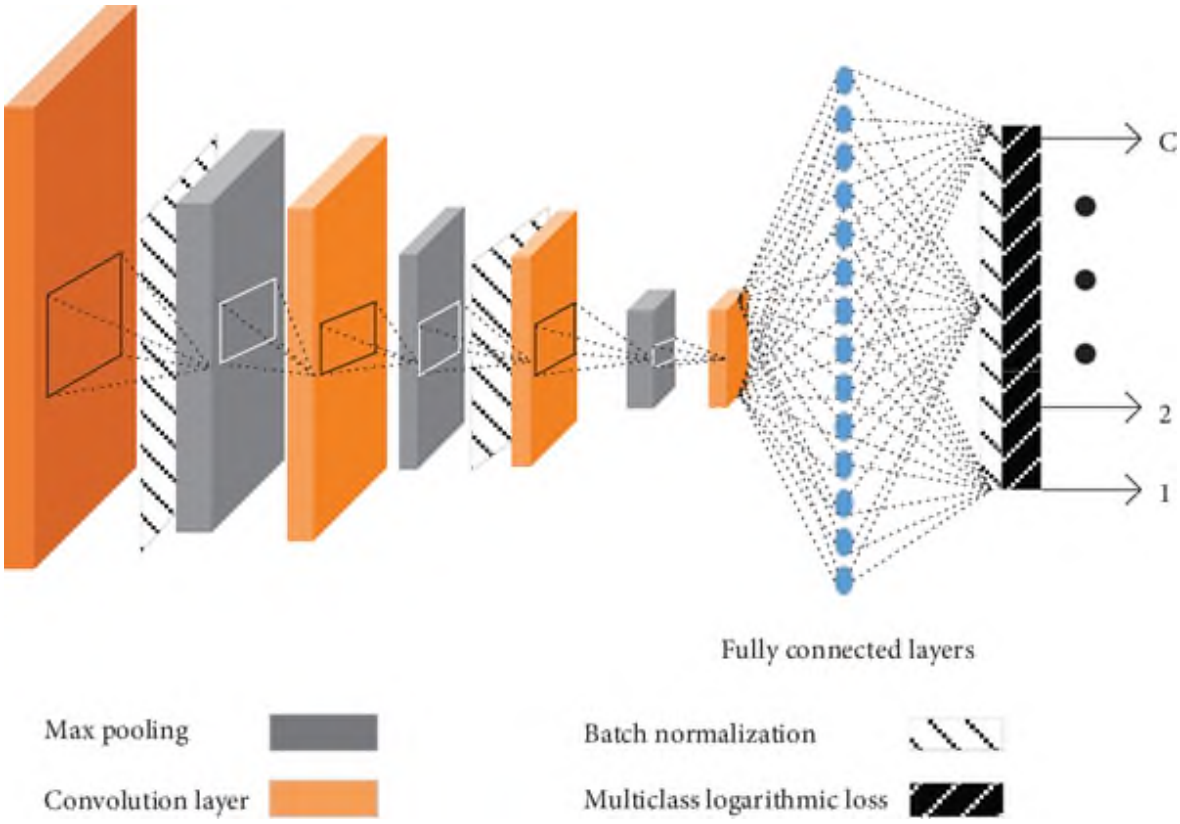
Σχήμα 4.3: Αρχιτεκτονική ViT - Σχήμα από [9]

4.2 Συνελικτικό Νευρωνικό Δίκτυο

Τα Συνελικτικά Νευρωνικά Δίκτυα (ΣΝΔ) συμβάλλουν στη μείωση του μεγέθους των δεδομένων ενώ ταυτόχρονα ενισχύουν την αναγνώριση των πιο σημαντικών χαρακτηριστικών σε μια εικόνα, όπως αναφέρεται στα άρθρα [17], [18]. Ένα ΣΝΔ αποτελείται από ένα σύνολο νευρώνων που εκτελούν συνέλιξη με προκαθορισμένα φίλτρα πάνω στην εικόνα-διάνυσμα που δέχονται ως είσοδο. Αυτή η διαδικασία επιτρέπει την ανακάλυψη σημαντικών χαρακτηριστικών, όπως φωτεινά ή σκοτεινά σημεία και ακμές σε διάφορους προσανατολισμούς, για την ταξινόμηση των εικόνων.

Ένα συνελικτικό νευρωνικό δίκτυο αποτελείται συνήθως από έναν αριθμό διαφορετικών στρώσεων, όπως φαίνεται στο Σχήμα 4.4. Αρχικά, υπάρχει ένας συνδυασμός συνελικτι-

κών και MaxPooling στρώσεων. Οι συνελικτικές στρώσεις περιέχουν φίλτρα που εκτελούν τη διαδικασία της συνέλιξης στα δεδομένα που της δίνονται. Μέσω αυτής της διαδικασίας, αποκαλύπτονται σημαντικά χαρακτηριστικά που περιέχονται στις εικόνες.

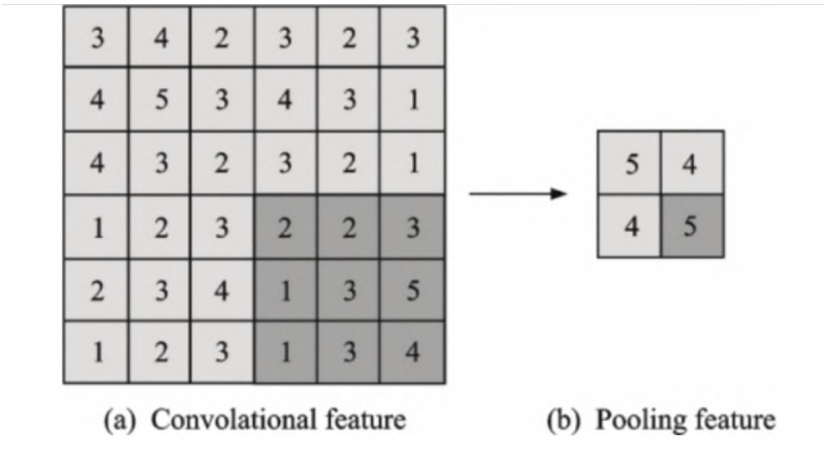


Σχήμα 4.4: Αρχιτεκτονική του ΣΝΔ - Σχήμα από [10]

Η στρώση MaxPooling, που ακολουθεί, συμβάλλει στη μείωση του χωρικού όγκου των χαρακτηριστικών που έχουν εξαχθεί από τη συνελικτική στρώση. Αυτό επιτυγχάνεται με την επιλογή του μεγαλύτερου στοιχείου εντός ενός προκαθορισμένου τμήματος των χαρακτηριστικών, όπως απεικονίζεται στο Σχήμα 4.5.

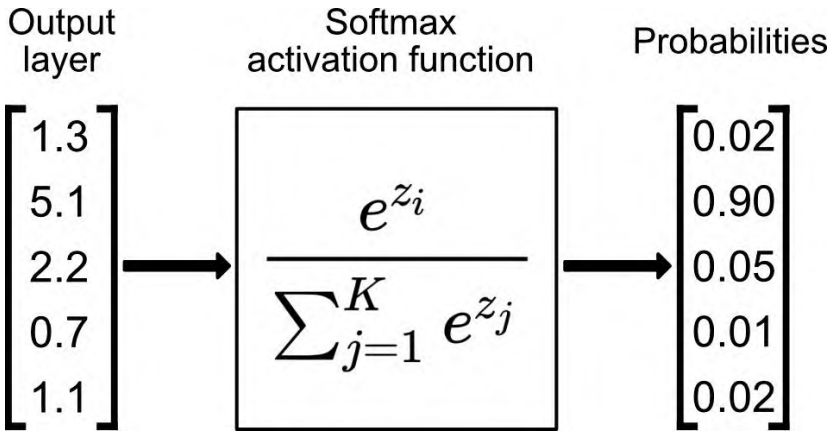
Στη συνέχεια, ακολουθεί μια BatchNormalization στρώση, η οποία κανονικοποιεί τις εισόδους των νευρώνων σε κάθε στρώση, σύμφωνα με το άρθρο [19]. Με αυτόν τον τρόπο συμβάλλει στη βελτίωση της σταθερότητας και στη σύγκλιση της διαδικασίας εκπαίδευσης.

Τέλος, ακολουθεί το πλήρως συνδεδεμένο τμήμα του NN, το οποίο εκτελεί τη διαδικασία ταξινόμησης, όπως παρουσιάζεται στο Σχήμα 4.4. Στη συγκεκριμένη αρχιτεκτονική, το τμήμα αυτό αναλαμβάνει την ταξινόμηση πολλαπλών τάξεων, δηλαδή τα δεδομένα εισόδου ανήκουν σε περισσότερες από μία κατηγορίες. Σε αυτές τις περιπτώσεις, χρησιμοποιείται μια κανονικοποιημένη εκθετική συνάρτηση (softmax), η οποία μετατρέπει το διάνυσμα της



Σχήμα 4.5: Παράδειγμα λειτουργίας της MaxPooling στρώσης - Σχήμα από [11]

τελικής στρώσης του NN σε μια κατανομή πιθανοτήτων, όπως φαίνεται στο σχήμα 4.6. Το μέγεθος αυτής της κατανομής είναι ίσο με τον αριθμό των διαφορετικών τάξεων που περιλαμβάνονται στα δεδομένα. Αυτό σημαίνει ότι το διάνυσμα πιθανοτήτων καθορίζει την πιθανότητα κάθε εισόδου να ανήκει σε μία από τις διαφορετικές τάξεις.



Σχήμα 4.6: Παράδειγμα λειτουργίας της κανονικοποιημένης εκθετικής (softmax) συνάρτησης στα δεδομένα της τελευταίας στρώσης του NN - Σχήμα από [**?**]

Κεφάλαιο 5

Εκπαίδευση και αξιολόγηση των μοντέλων

Στο παρόν κεφάλαιο εξετάζεται η εκπαίδευση μοντέλων για την πρόβλεψη του δακτυλοσυλλαβισμού στην ΕΝΓ από εικόνες. Παράλληλα, πραγματοποιείται αξιολόγηση των αποτελεσμάτων που προκύπτουν. Τα μοντέλα PIXIE και ExPose απαιτούν την αναπαράσταση ολόκληρου του σώματος για την εξαγωγή της πόζα του δεξιού χεριού. Κατά συνέπεια, σε αυτές τις περιπτώσεις, η διαδικασία εκπαίδευσης βασίζεται στη χρήση της πλήρους εικόνας.

Όσον αφορά τις μεθόδους του ViT και του ΣΝΔ, χρησιμοποιούνται εικόνες στις οποίες έχει γίνει απομόνωση του δεξιού χεριού, με σκοπό την εστίαση στον δακτυλοσυλλαβισμό και την αποφυγή περιττού θορύβου που θα μπορούσε να εισαχθεί από την πλήρη εικόνα. Το MediaPipe μπορεί να εκπαιδευτεί χρησιμοποιώντας και τις δύο κατηγορίες εικόνων, καθώς ανιχνεύει τα ορόσημα του χεριού μέσω της ανίχνευση της παλάμης. Στο πλαίσιο αυτής της εργασίας, επιλέγεται να χρησιμοποιηθούν ολόκληρες οι εικόνες για την εξαγωγή των σημείων του χεριού από το MediaPipe.

Για τη διαδικασία εκπαίδευσης και αξιολόγησης των αποτελεσμάτων, χρησιμοποιούνται εικόνες που προέρχονται από βίντεο αναπαράστασης γραμμάτων. Όλες οι εικόνες που αναπαριστούν το ίδιο γράμμα αποθηκεύονται σε έναν κοινό φάκελο. Από αυτές τις εικόνες, το 60% χρησιμοποιείται για την εκπαίδευση των μοντέλων, το 20% χρησιμοποιείται για την επαλήθευση των αποτελεσμάτων και το υπόλοιπο 20% χρησιμοποιείται για τον έλεγχο της απόδοσης των μοντέλων.

5.1 Εκπαίδευση χρησιμοποιώντας τις μεθόδους ανίχνευσης της στάσης του δεξιού χεριού

Σε αυτή την υποενότητα, αναλύεται η αρχιτεκτονική που χρησιμοποιείται για την εκπαίδευση μοντέλων με σκοπό την πρόβλεψη της πόζας του δεξιού χεριού. Αυτά τα μοντέλα υποβάλλονται σε εκπαίδευση χρησιμοποιώντας τις προβλέψεις των οροσήμων των χεριών που παρέχονται από τις βιβλιοθήκες PIXIE, ExPose και MediaPipe, κατά την εφαρμογή τους στις αρχικές εικόνες. Συνεπακόλουθα, πραγματοποιείται αξιολόγηση της απόδοσης των νευρωνικών δικτύων.

5.1.1 Αρχιτεκτονική των μεθόδων ανίχνευσης της στάσης του δεξιού χεριού

Οι βιβλιοθήκες PIXIE και ExPose επιστρέφουν 15 πίνακες διαστάσεων 3*3 για κάθε εικόνα, αντιπροσωπεύοντας την πόζα του δεξιού χεριού. Κατά συνέπεια, το μέγεθος εισόδου που παρέχεται στα νευρωνικά δίκτυα είναι (15, 9, 1). Από την άλλη, η βιβλιοθήκη MediaPipe επιστρέφει 21 σημεία, κάθε ένα από τα οποία περιγράφεται σε τρεις διαστάσεις. Επομένως, το μέγεθος της εισόδου που παρέχεται στα νευρωνικά δίκτυα είναι (21, 3, 1).

Ο κώδικας που περιγράφει την αρχιτεκτονική του νευρωνικού δικτύου (NN) που χρησιμοποιείται για την εκπαίδευση των μοντέλων παρουσιάζεται στο Σχήμα 5.1. Το NN αποτελείται από διάφορες στρώσεις, ξεκινώντας με μια στρώση Reshape, η οποία προσαρμόζει τις διαστάσεις των δεδομένων πριν από την προώθησή τους στο δίκτυο.

Listing 5.1: Αρχιτεκτονική του NN για τα μοντέλα PIXIE και ExPose

```
model = Sequential([
    Reshape((15, 9, 1), input_shape=(135,)),
    Conv2D(16, (3, 3), activation='relu'),
    Conv2D(32, (3, 3), activation='relu'),
    MaxPooling2D((2, 2)),
    Flatten(),
    Dense(1024, activation='relu'),
    Dense(512, activation='relu'),
    Dense(24, activation='softmax')
```

])

Στη συνέχεια, υπάρχουν 2 στρώσεις συνέλιξης, με κάθε μία να χρησιμοποιεί ένα φίλτρο συνέλιξης διαστάσεων (3, 3). Η πρώτη αποτελείται από 16 και η δεύτερη από 32 νευρώνες. Έπειτα, ακολουθεί μια MaxPooling στρώση η οποία συμβάλλει στη μείωση του χωρικού μεγέθους των συνεληγμένων χαρακτηριστικών. Τέλος, εμφανίζεται το τμήμα πλήρως συνδεδεμένων επιπέδων (fully connected) που αποτελείται από Dense στρώσεις και υλοποιεί την ταξινόμηση των γραμμάτων της ελληνικής αλφαβήτας.

5.1.2 Αποτελέσματα εκπαίδευσης

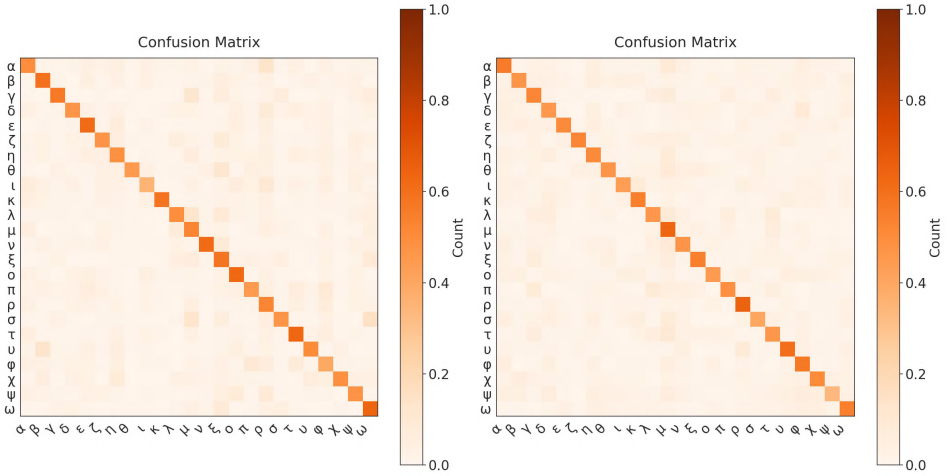
Πληροφορίες μπορούν να συλλεχτούν σχετικά με την απόδοση των προαναφερθέντων μοντέλων μέσω της ανάλυσης των πινάκων σύγχυσης (confusion matrix) που παρουσιάζονται στο Σχήμα 5.1, καθώς και μέσω του υπολογισμού των μετρικών accuracy, precision, recall και f1-score, όπως φαίνεται στον Πίνακα 5.1. Παρατηρείται ότι το μοντέλο MediaPipe παρουσιάζει υψηλότερες τιμές σε αυτές τις μετρικές σε σχέση με τα υπόλοιπα μοντέλα. Αυτή η παρατήρηση επιβεβαιώνεται, επίσης, από τον πίνακα σύγχυσης (Σχήμα 5.1ε′), όπου τα διαγώνια στοιχεία παρουσιάζουν εντονότερο χρώμα στην περίπτωση του MediaPipe, υποδηλώνοντας ότι περισσότερες προβλέψεις είναι σωστές.

Πίνακας 5.1: Σύγκριση accuracy, precision, recall και F1 score για τα μοντέλα που εκπαιδεύτηκαν.

	Accuracy	Precision	Recall	F1 Score
Expose	0.52	0.54	0.52	0.52
Pixie	0.5	0.51	0.5	0.5
Συνδυασμός ExPose και MediaPipe	0.57	0.59	0.57	0.59
Συνδυασμός Pixie και MediaPipe	0.53	0.55	0.53	0.54
MediaPipe	0.74	0.8	0.74	0.76

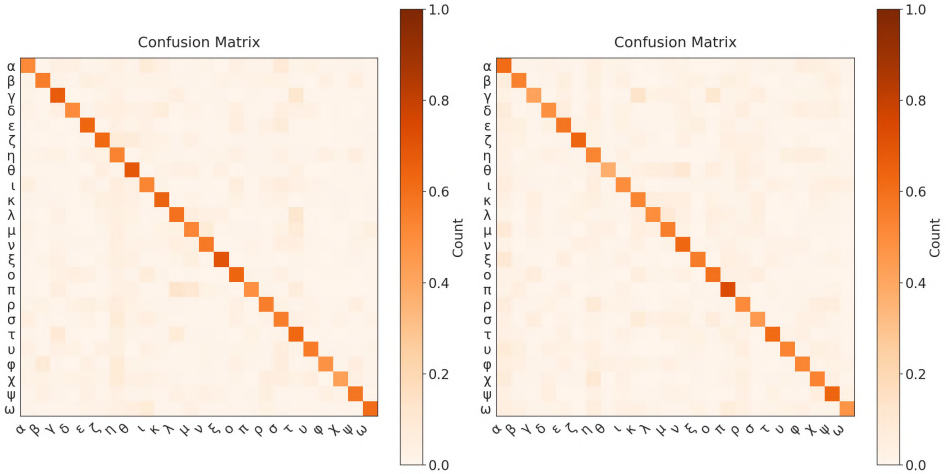
Επιπλέον, αξίζει να σημειωθεί ότι ο συνδυασμός του MediaPipe με το ExPose, καθώς και ο συνδυασμός του MediaPipe με το PIXIE, παρουσιάζουν βελτιωμένη απόδοση σε σχέση με την ανεξάρτητη χρήση του ExPose και του PIXIE. Αυτή η βελτίωση γίνεται εμφανής μέσω των διαγώνιων στοιχείων των πινάκων σύγχυσης του Σχήματος 5.1, τα οποία εμφανίζουν

Σχήμα 5.1: Πίνακες σύγχυσης μεθόδων ανίχνευσης της πόζας του δεξιού χεριού.



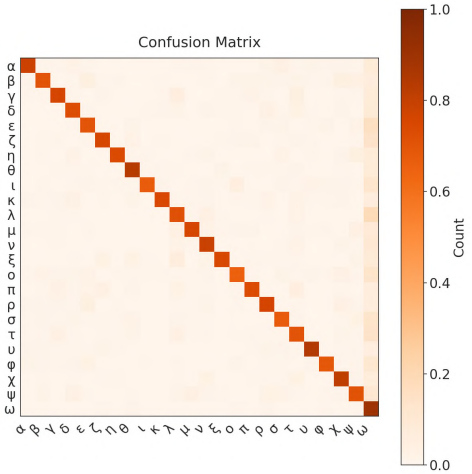
(α′) Χρήση ExPose.

(β′) Χρήση PIXIE.



(γ′) Χρήση ExPose και MediaPipe.

(δ′) Χρήση PIXIE και MediaPipe.



(ε′) Χρήση MediaPipe.

εντονότερο χρώμα στις περιπτώσεις του συνδυασμού MediaPipe με ExPose και MediaPipe με PIXIE. Επιπροσθέτως, αυτή η βελτίωση αντικατοπτρίζεται στη μετρική της ακρίβειας, όπως παρουσιάζεται στο Σχήμα 5.1. Το παραπάνω αποδεικνύει ότι η συνδυασμένη χρήση του MediaPipe με το ExPose και του MediaPipe με το PIXIE επιφέρει βελτίωση στην ακρίβεια και την αξιοπιστία της αναγνώρισης του δακτυλοσυλλαβισμού. Αυτή η βελτίωση υπερβαίνει την απόδοση που επιτυγχάνεται με την ανεξάρτητη χρήση των αντίστοιχων βιβλιοθηκών.

5.2 Εκπαίδευση χρησιμοποιώντας τις μεθόδους ταξινόμησης

Σε αυτήν την υποενότητα εξετάζεται και αξιολογείται η εφαρμογή δύο μεθόδων ταξινόμησης, του ViT και του ΣΝΔ, στο πλαίσιο της ανίχνευσης των γραμμάτων της ελληνικής αλφαβήτας. Πιο αναλυτικά, παρουσιάζεται η αρχιτεκτονική των νευρωνικών δικτύων που χρησιμοποιείται και πραγματοποιείται σύγκριση της απόδοσης των μοντέλων που εκπαιδεύονται.

5.2.1 Αρχιτεκτονική της μεθόδου του ViT

Ο κωδικοποιητής (encoder) αποτελεί βασικό στοιχείο της αρχιτεκτονικής της μεθόδου του ViT, αφού αναλαμβάνει την ανάλυση και την αναπαράσταση της ακολουθίας εισόδου. Ο κωδικοποιητής επεξεργάζεται την ακολουθία αυτή και τη μετατρέπει σε μια μορφή που μπορεί να γίνει κατανοητή από το μετασχηματιστή.

Συγκεκριμένα, η αρχιτεκτονική του κωδικοποιητή (encoder) της μεθόδου ViT παρουσιάζεται στο Σχήμα 5.2. Το NN αποτελείται από μια στρώση LayerNormalization, η οποία κανονικοποιεί τις αθροιστικές εισόδους των νευρώνων σε κάθε στρώση. Έτσι, συμβάλλει στη σταθεροποίηση της διαδικασίας εκπαίδευσης και στη μείωση του χρόνου εκπαίδευσης, σύμφωνα με το άρθρο [20].

Listing 5.2: Αρχιτεκτονική του κωδικοποιητή της μεθόδου ViT

```
x = LayerNormalization()(x)
x = MultiHeadAttention(num_heads=12, key_dim=256)(x, x)
x = LayerNormalization()(x)
```

```
x = Dense(512, activation="gelu")(x)
x = Dropout(0.1)(x)
x = Dense(256)(x)
x = Dropout(0.1)(x)
x = Dense(256)(x)
x = Dropout(0.1)(x)
```

Στη συνέχεια, ακολουθεί μια στρώση MultiHeadAttention. Μέσω αυτής της στρώσης, το μοντέλο έχει τη δυνατότητα να εκτελεί παράλληλα πολλαπλές λειτουργίες προσοχής, οι οποίες αντιστοιχούν στον αριθμό των κεφαλών που έχει. Στη συνέχεια, τα αποτελέσματα συγχωνεύονται για να παράγουν μια συνολική ενιαία έξοδο, όπως παρουσιάζεται και στο άρθρο [8].

Έπειτα, η έξοδος της MultiHeadAttention στρώσης κανονικοποιείται με την προσθήκη μιας στρώσης LayerNormalization. Στη συνέχεια ακολουθεί το πλήρως συνδεδεμένο τμήμα του NN που αποτελείται από τρεις επαναλήψεις του συνδυασμού δύο στρώσεων, μιας Dense και μιας Dropout. Η Dropout στρώση απενεργοποιεί κάποιες μονάδες του NN κατά τη διάρκεια της εκπαίδευσης, σύμφωνα με το άρθρο [21]. Αυτό εμποδίζει τους νευρώνες από το να προσαρμόζονται, και μπορεί να βοηθήσει στη μείωση της υπερ-προσαρμογής (overfitting)..

5.2.2 Αρχιτεκτονική της μεθόδου του ΣΝΔ

Το απόσπασμα του κώδικα που περιγράφει την αρχιτεκτονική του ΣΝΔ παρουσιάζεται στο Σχήμα 5.3. Το NN αποτελείται από τρεις 2D συνελικτικές στρώσεις με 32 νευρώνες η καθεμία. Κάθε συνελικτική στρώση ακολουθείται από μια στρώση MaxPooling και μια στρώση BatchNormalization, που συμβάλλουν στη μείωση της διαστατικότητας των δεδομένων και στην κανονικοποίηση της εξόδου. Τέλος, το πλήρως συνδεδεμένο τμήμα του NN αποτελείται από δύο στρώσεις Dense, όπου η πρώτη έχει 512 και η δεύτερη 24 νευρώνες. Ανάμεσα σε αυτές τις στρώσεις εισάγεται μια Dropout στρώση που συμβάλλει στην αντιμετώπιση της υπερ-προσαρμογής.

Listing 5.3: Αρχιτεκτονική της μεθόδου του ΣΝΔ

```
model = Sequential([
        Conv2D(128, (5,5), 'relu', input_shape=(224, 224, 3)),
        Conv2D(32, (3,3), 'relu'),
```

```
        MaxPooling2D(),
        BatchNormalization(),

        Conv2D(32, (3,3), 'relu'),
        Conv2D(32, (3,3), 'relu'),
        MaxPooling2D(),
        BatchNormalization(),

        Conv2D(32, (3,3), 'relu'),
        Conv2D(32, (3,3), 'relu'),
        MaxPooling2D(),
        BatchNormalization(),

        Conv2D(32, (3,3), 'relu'),
        Conv2D(32, (3,3), 'relu'),
        MaxPooling2D(),

        Flatten(),
        Dense(512, 'relu'),
        Dropout(0.6),
        Dense(24, 'softmax')
])
```

5.2.3 Αποτελέσματα εκπαίδευσης

Στον πίνακα 5.2 παρουσιάζονται τα αποτελέσματα εκπαίδευσης των μοντέλων που περιγράφονται στις υποενότητες 5.2.2 και 5.2.1. Και οι δύο μέθοδοι εμφανίζουν πολύ κοντινά αποτελέσματα όσον αφορά την ακρίβεια και την απώλεια. Οι τιμές αυτών των μετρικών είναι, επίσης, αρκετά ικανοποιητικές, καθώς η ακρίβεια βρίσκεται κοντά στο 1 και η απώλεια προσεγγίζει το 0.

Παραπάνω πληροφορίες σχετικά με την επίδοση των δύο αυτών μεθόδων μπορούν να συλλεχτούν από τους πίνακες σύγχυσης και τις μετρικές accuracy, precision, recall, f1-score,

Πίνακας 5.2: Σύγκριση ακρίβειας και απώλειας των μοντέλων ΣΝΔ και ViT.

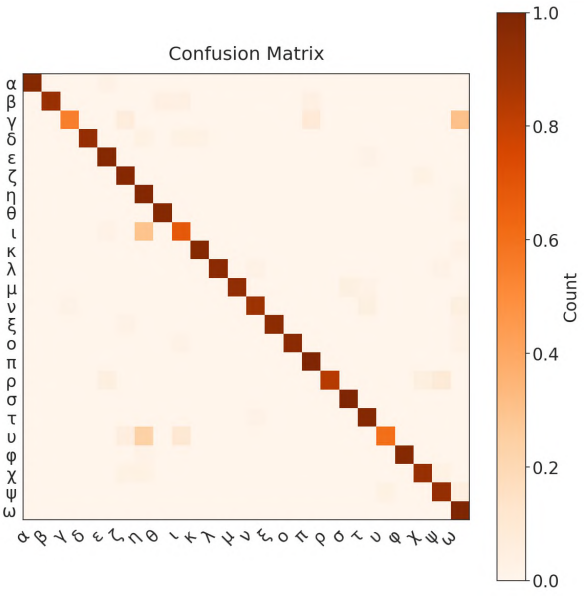
	Ακρίβεια	Απώλεια
ΣΝΔ	0.92	0.32
ViT	0.92	0.29

όπως φαίνονται στα Σχήματα 5.2, 5.3. Παρατηρείται ότι στην περίπτωση του ViT έχουν γίνει πιο ακριβείς προβλέψεις, δεδομένου ότι η διαγώνιος στον πίνακα σύγχυσης παρουσιάζει εντονότερο χρώμα σε σχέση με τον πίνακα σύγχυσης που προέκυψε από το μοντέλο του ΣΝΔ.

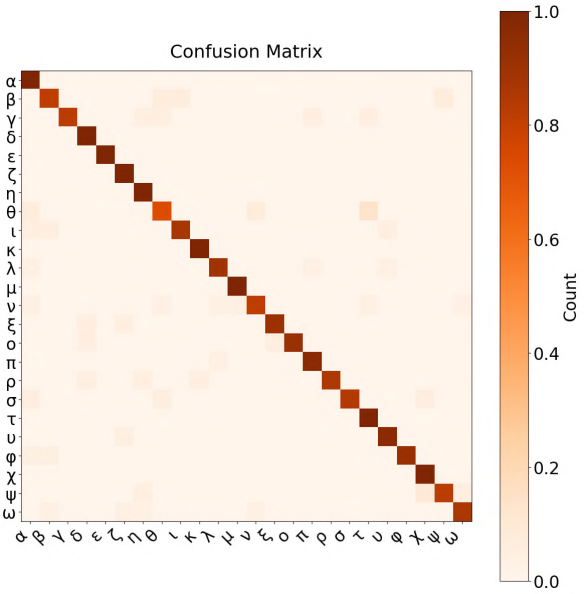
Πίνακας 5.3: Σύγκριση accuracy, precision, recall και F1 score για τα μοντέλα ΣΝΔ και ViT.

	Accuracy	Precision	Recall	F1 Score
ΣΝΔ	0.89	0.91	0.9	0.9
ViT	0.92	0.92	0.92	0.91

Σχήμα 5.2: Πίνακες σύγχυσης μεθόδων ταξινόμησης



(α′) Χρήση μεθόδου ΣΝΔ.



(β′) Χρήση μεθόδου ViT.

Κεφάλαιο 6

Σύγκριση μοντέλων βάσει του ποσοστού σφάλματος χαρακτήρων

Σε αυτήν την ενότητα, υλοποιείται μια επιπλέον σύγκριση μεταξύ των μοντέλων που αναπτύσσονται στο Κεφάλαιο 5, ώστε να επαληθευτεί η απόδοση τους και η ικανότητα γενίκευσης τους σε διαφορετικά σύνολα δεδομένων. Σε αυτήν την περίπτωση, αξιοποιούνται οι ακολουθίες εικόνων που αντιστοιχούν σε μια λέξη. Πιο συγκεκριμένα, χρησιμοποιούνται οι εικόνες που αποτυπώνονται από τον πρώτο διερμηνέα.

Η επιλογή του πρώτου διερμηνέα γίνεται, διότι απαιτούνται χρόνος και υπολογιστικοί πόροι για την υλοποίηση της πρόβλεψης για όλους τους διερμηνείς. Παράλληλα, κάθε διερμηνέας νοηματίζει 50 λέξεις συνολικά. Σύμφωνα με το Σχήμα 2.3, κάθε βίντεο στο οποίο νοηματίζεται λέξη αποτελείται από 40 έως 170 εικόνες, και ο αριθμός των 100 εικόνων είναι η πιο συχνή περίπτωση. Επομένως, το εν λόγω πλήθος εικόνων μπορεί να προσφέρει μια ενδεικτική αναπαράσταση της απόδοσης των μοντέλων.

Η διαδικασία που ακολουθείται είναι η πρόβλεψη του γράμματος που αντιπροσωπεύεται σε κάθε εικόνα. Στη συνέχεια, η ακολουθία γραμμάτων που προκύπτει από αυτήν την πρόβλεψη συγκρίνεται με την ακολουθία των γραμμάτων της λέξης την οποία νοηματίζει ο πρώτος διερμηνέας κάθε φορά.

6.1 Ορισμός του ποσοστού σφάλματος χαρακτήρων

Η μετρική που εφαρμόζεται για τη σύγκριση των ακολουθιών των γραμμάτων είναι το ποσοστό σφάλματος χαρακτήρων (CER) [22]. Για τον υπολογισμού του CER χρησιμοποιεί-

ται η απόσταση Λέβενσταιν (Levenshtein distance). Η μετρική αυτή υπολογίζει τη διαφορά μεταξύ δύο ακολουθιών γραμμάτων, δηλαδή τον ελάχιστο αριθμό εισαγωγών, διαγραφών και αντικαταστάσεων που μπορούν να γίνουν, ώστε μια λέξη να μετατραπεί σε μια άλλη. Έχοντας υπολογίσει την απόσταση Λέβενσταιν μπορεί να υλοποιηθεί το CER για ένα σύνολο δεδομένων με N προβλέψεις και ακολουθίες, σύμφωνα με την Εξίσωση 6.1.

$$CER = \frac{1}{N} \cdot \sum_{n=0}^{N-1} \frac{edit\_distance(n)}{sequence\_length(n)} \tag{6.1}$$

Γενικά, το CER είναι μια μετρική που υποδεικνύει την ομοιότητα μεταξύ δύο ακολουθιών. Όσο μικρότερη είναι η τιμή του CER, τόσο λιγότερες διαφορές υπάρχουν μεταξύ δύο ακολουθιών. Αντίστροφα, όσο μεγαλύτερη είναι η τιμή του CER, τόσο περισσότερες διαφορές παρατηρούνται. Για αυτόν τον λόγο, προτιμάται το CER να έχει όσο το δυνατόν μικρότερη τιμή, καθώς αυτό υποδεικνύει ότι υπάρχουν λιγότερες διαφορές μεταξύ των ακολουθιών.

Ο υπολογισμός του CER υλοποιείται σε ακολουθίες γραμμάτων μεγέθους ίσο με τον αριθμό των εικόνων ανά βίντεο. Συνεπώς, σύμφωνα με το Σχήμα 2.3, κάθε ακολουθία γραμμάτων περιλαμβάνει από 40 έως 170 χαρακτήρες. Λαμβάνοντας υπόψη αυτήν την ποικιλία μεγεθών, χρησιμοποιούνται δύο μέθοδοι κωδικοποίησης των ακολουθιών, για την επιτάχυνση του υπολογισμού του CER και την ελαχιστοποίηση του θορύβου.

Η πρώτη μέθοδος αφορά την παράλειψη χαρακτήρων που δεν επαναλαμβάνονται στην επόμενη και στην προηγούμενη θέση σε σχέση με την τρέχουσα θέση τους στην ακολουθία. Αυτό έχει ως στόχο τη μείωση του θορύβου σε ένα βαθμό. Η δεύτερη μέθοδος αποτελεί τη συγχώνευση συνεχόμενων ίδιων χαρακτήρων. Με αυτόν τον τρόπο, μειώνεται το μήκος των ακολουθιών των γραμμάτων, προσεγγίζοντας περισσότερο τις αρχικές ακολουθίες. Παράλληλα, με την υλοποίηση αυτών των μεθόδων κωδικοποίησης δεν ευνοείται κάποιο μοντέλο παραπάνω σε σχέση με τα υπόλοιπα.

6.2 Σύγκριση του ποσοστού σφάλματος χαρακτήρων των μοντέλων

Οι τιμές του CER υπολογίζονται για τα μοντέλα του ViT, του ΣΝΔ καθώς και για τα μοντέλα που δημιουργούνται με βάση την πόζα των μοντέλων ExPose, PIXIE και MediaPipe. Ο Πίνακας 6.1 αποδεικνύει ότι το μοντέλο του MediaPipe δημιουργεί τις πιο ακριβείς προβλέψεις, καθώς το CER του εμφανίζει τη χαμηλότερη τιμή. Αυτό συνάδει με τα αποτελέσματα των προηγούμενων συγκρίσεων που πραγματοποιούνται στο Κεφάλαιο 5. Τα μοντέλα ExPose και PIXIE, που εκπαιδεύονται με βάση τις φιλτραρισμένες εικόνες από το MediaPipe, επιδεικνύουν βελτιωμένη απόδοση σε σχέση με τα μοντέλα που εκπαιδεύονται στις αρχικές εικόνες, όπως προκύπτει από τους πίνακες σύγχυσης 5.1 στο Κεφάλαιο 5.

Επιπλέον, αξίζει να σημειωθεί ότι η μέθοδος του ViT παρουσιάζει ακριβέστερες προβλέψεις σε σύγκριση με το μοντέλο PIXIE. Τέλος, όσον αφορά τη μέθοδο του ΣΝΔ, εμφανίζει διπλάσιο CER σε σύγκριση με τις άλλες μεθόδους, υποδηλώνοντας σχετικά μειωμένη ακρίβεια στις προβλέψεις της.

Πίνακας 6.1: Σύγκριση του CER των εκπαιδευμένων μοντέλων.

Model	CER
ExPose	244
PIXIE	298
Συνδυασμός ExPose και MediaPipe	225
Συνδυασμός PIXIE και MediaPipe	285
MediaPipe	200
ΣΝΔ	500
ViT	296

Κεφάλαιο 7

Συμπεράσματα

Σε αυτό το κεφάλαιο συνοψίζονται τα αποτελέσματα της σύγκρισης των μεθόδων ταξινόμησης και των μεθόδων ανίχνευσης της πόζας των χεριών για την πρόβλεψη του δακτυλοσυλλαβισμού στην ΕΝΓ. Παράλληλα, προτείνονται μελλοντικές επεκτάσεις που θα μπορούσαν να ενισχύσουν την ανάλυση, παρέχοντας περισσότερες λεπτομέρειες για τα πλεονεκτήματα και τις περιπτώσεις καταλληλότητας κάθε μοντέλου.

7.1 Σύνοψη και συμπεράσματα

Η διαδικασία αναγνώρισης του δακτυλοσυλλαβισμού από βίντεο είναι μια αρκετά πολύπλοκη διαδικασία. Στην συγκεκριμένη εργασία εξετάζεται το στατικό μοντέλο, όπου κάθε εικόνα αντιστοιχεί σε ένα γράμμα της ελληνικής αλφαβήτας. Για το σκοπό αυτό, εφαρμόζονται δύο κατηγορίες μεθόδων, μοντέλα αναγνώρισης της στάσης των χεριών του ανθρώπου και μοντέλα ταξινόμησης εικόνων.

Γενικότερα, παρατηρείται ότι οι μέθοδοι αναγνώρισης της στάσης των χεριών παρουσιάζουν πιο ακριβή αποτελέσματα σε σχέση με τις μεθόδους ταξινόμησης εικόνων, καθώς επικεντρώνονται στην ακριβή αναγνώριση της πόζας των χεριών. Συγκεκριμένα, το MediaPipe υλοποιεί τις πιο ακριβείς προβλέψεις, όπως φαίνεται από τους πίνακες σύγχυσης 5.1, τη μετρική της ακρίβειας και τη μετρική CER 6.1.

7.2 Μελλοντικές επεκτάσεις

Σχετικά με τις μελλοντικές επεκτάσεις μπορεί να εξεταστεί το ακολουθιακό μοντέλο για τη σύγκριση των παραπάνω μεθόδων, όπως γίνεται στα άρθρα [23] και [24]. Σε αυτήν την περίπτωση, κάθε βίντεο, δηλαδή ακολουθία εικόνων αντιστοιχεί σε ένα γράμμα ή λέξη της ελληνικής αλφαβήτας. Παράλληλα, σε αυτήν την περίπτωση απαιτείται η χρήση Επαναλαμβανόμενων Νευρωνικών Δικτύων (RNN) σε αντίθεση με τα ΣΝΔ που χρησιμοποιούνται σε αυτή την εργασία.

Επίσης, μπορούν να χρησιμοποιηθούν δεδομένα στα οποία υπάρχει διαφορετικός φωτισμός και άτομα με διαφορετικούς τόνους δέρματος, ώστε να ελεγχθεί πως λειτουργούν αυτά τα μοντέλα σε διαφορετικές περιβαλλοντικές συνθήκες, ενισχύοντας έτσι την ολοκληρωμένη κατανόηση της απόδοσής τους.

Βιβλιογραφία

[1] Ivan Grishchenko and Valentin Bazarevsky. Mediapipe holistic — simultaneous face, hand and pose prediction, on device. https://ai.googleblog.com/2020/12/mediapipe-holistic-simultaneous-face.html, 2020.

[2] Isaac Osei Agyemang, Isaac Adjei Mensah, Sophyani Banaamwini Yussif, Fiasam Linda Delali, Bernard Cobinnah Mawuli, Bless Lord Y. Agbley, Collins Sey, and Joshua Berkohd. Gesture control of micro-drone: A lightweight-net with domain randomization and trajectory generators, 2023.

[3] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. Mediapipe hands: On-device real-time hand tracking. *CoRR*, abs/2006.10214, 2020.

[4] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Collaborative regression of expressive bodies using moderation. *CoRR*, abs/2105.05301, 2021.

[5] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. https://github.com/yfeng95/PIXIE.

[6] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart1, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black1. Smpl-x: A new joint 3d model of the human body, face and hands together. *CoRR*, abs/1904.05866, 2019.

[7] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart1, Dimitrios Tzionas, and Michael Black. Monocular expressive body regression through body-driven attention. *CoRR*, abs/2008.09062, 2020.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, 2023.

[9] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 12116–12128. Curran Associates, Inc., 2021.

[10] Hongmin Gao, Shuo Lin, Yao Yang, Chenming Li, and Mingxiang Yang. Convolution neural network based on two-dimensional spectrum for hyperspectral image classification. *Journal of Sensors*, 2018:1–13, 08 2018.

[11] Ahmed Alsaffar, Hai Tao, and Mohammed Talab. Review of deep convolution neural network in image classification. pages 26–31, 10 2017.

[12] Galini Sapountzaki. Σαπουντζάκη Γ. (2013). Μαθήματα από την Ελληνική Νοηματική Γλώσσα: Επιφανειακές γλωσσικές δομές και παραμελημένες επικοινωνιακές αξίες. Πρακτικά 1ου Πανελλήνιου Διεπιστημονικού Συμποσίου με θέμα 'Κωφοί και Ακούοντες εν Δράσει: ο ρόλος της δια βίου Μάθησης και της Διάδοσης της Νοηματικής Γλώσσας στην Ποιότητα Ζωής Ανήκοου και Βαρήκοου Πληθυσμού'. Χανιά, 18 Ιουλίου 2013. 2013.

[13] 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. 2020.

[14] Alexander Mordvintsev and Abid K. 2023.

[15] 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. *CoRR*, abs/1906.08172, 2019.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Syl-

vain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020.

[17] Wamidh K. Mutlag, Shaker K. Ali, Zahoor M. Aydam, and Bahaa H. Taher. Feature extraction methods: A review. *Journal of Physics: Conference Series*, 1591(1):012028, jul 2020.

[18] Lihua Luo. Research on image classification algorithm based on convolutional neural network. *Journal of Physics: Conference Series*, 2083(3):032054, nov 2021.

[19] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015.

[20] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. 2016. cite arxiv:1607.06450.

[21] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 06 2014.

[22] Peilu Wang, Ruihua Sun, Hai Zhao, and Kai Yu. A new word language model evaluation metric for character based languages. 8202:315–324, 01 2013.

[23] Agelos Kratimenos, Georgios Pavlakos, and Petros Maragos. Independent sign language recognition with 3d body, hands, and face reconstruction. pages 4270–4274, 2021.

[24] Deep Kothadiya, Chintan Bhatt, Krenil Sapariya, Kevin Patel, Ana Gil, and Juan Corchado Rodríguez. Deepsign: Sign language detection and recognition using deep learning. *Electronics*, 11:1780, 06 2022.