National Technical University of Athens
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

Division of Signals, Control and Robotics
Computer Vision, Speech Communication and Signal Processing Group

3D Body, Face and Hands Reconstruction with application on Sign Language Recognition

Diploma Thesis

Agelos Kratimenos

Supervisor: Prof. Petros Maragos

Agelos Kratimenos

Athens, 30 June 2021

National Technical University of Athens
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

Division of Signals, Control and Robotics
Computer Vision, Speech Communication and Signal Processing Group

3D Body, Face and Hands Reconstruction with application on Sign Language Recognition

Diploma Thesis

Agelos Kratimenos

Supervisor: Prof. Petros Maragos

Approved by the examining committee:

.....	.....	.....
Petros Maragos	Constantinos Tzafestas	Gerasimos Potamianos
Professor	Associate Professor	Associate Professor
NTUA	NTUA	University of Thessaly

Athens, 30 June 2021

.....
Agelos Kratimenos
Electrical and Computer Engineer, NTUA

Ευχαριστίες

Θα ήθελα να ευχαριστήσω πρωτίστως και κυρίως τον Καθηγητή Πέτρο Μαραγκό, για την ανελλιπή βοήθεια του καθόλη τη διάρκεια εκπόνησης της Διπλωματικής. Χωρίς τα συνεχή σχόλια, τις εύστοχες συμβουλές και τις διαρκείς συναντήσεις, το ταξίδι της Διπλωματικής δε θα ήταν διόλου εύκολο και επιτυχημένο.

Στη συνέχεια, θα ήθελα να ευχαριστήσω ιδιαιτέρως τον Γιώργο Παυλάκο, για τη συνεργασία μας στη διάρκεια της Διπλωματικής, χωρίς τον οποίο, η ενασχόληση με τον τομέα της τρισδιάστατης ανακατασκευής ανθρωπίνου σώματος δε θα ήταν καθόλου εύκολη.

Τελος, θα ήθελα να ευχαριστήσω τον συμφοιτητή μου Κλεάνθη Αβραμίδη, για τη συνεργασία μας, παράλληλα με την εκπόνηση της διπλωματικής εργασίας, αλλά και για τη βοήθεια και τις χρήσιμες συμβουλές του για την ίδια τη Διπλωματική, καθόλη αυτή τη διάρκεια.

Abstract

This thesis studies the most prominent 3D methods for the reconstruction of the body, face, and hands from a single image while applying these tools in the problem of Isolated Sign Language Recognition. Sign Language Recognition is a complex visual recognition problem that combines several challenging tasks of Computer Vision due to the necessity to exploit and fuse information from hand gestures, body features, and facial expressions. After analytically studying the state-of-the-art methods for 3D reconstruction, and the technique used for confronting the task of Sign Language Recognition, we employ SMPL-X a contemporary parametric model that enables joint extraction of 3D body shape, face and hands information from a single image. We use this holistic 3D reconstruction for SLR, demonstrating that it leads to higher accuracy than recognition from raw RGB images and their optical flow fed into the state-of-the-art I3D-type network for 3D action recognition and from 2D Openpose skeletons fed into a Recurrent Neural Network. Furthermore, a set of experiments on the body, face, and hand features showed that neglecting any of these, significantly reduces the classification accuracy, proving the importance of jointly modeling body shape, facial expression, and hand pose for Sign Language Recognition. Finally, some experiments with Depth Estimation are conducted, while an analytic comparison between SMPL-X and ExPose is also made.

Keywords: 3D Computer Vision, 3D Body, Face, and Hands Reconstruction, Isolated Sign Language Recognition, SMPL-X, ExPose

Abstract

Αυτή η διπλωματική εργασία μελετά τις πιο σύγχρονες τρισδιάστατες μεθόδους για την ανακατασκευή σώματος, προσώπου και χεριών από μια απλή εικόνα, ενώ παράλληλα εφαρμόζει τα εργαλεία αυτά στο πρόβλημα της αναγνώρισης νοηματικής γλώσσας. Η αναγνώριση νοηματικής γλώσσας είναι ένα σύνθετο οπτικό πρόβλημα αναγνώρισης που συνδυάζει πολλές πτυχές της όρασης υπολογιστών, λόγω της αναγκαιότητας να συνδυαστεί και να εξαχθεί πληροφορία τόσο από τα χέρια και τις εκφράσεις του προσώπου, αλλά και από ολόκληρη τη σωματοδομή. Αφού μελετηθούν αναλυτικά οι state-of-the-art μέθοδοι για την τρισδιάστατη ανακατασκευή καθώς και οι τεχνικές που χρησιμοποιούνται για την αντιμετώπιση του προβλήματος αναγνώρισης νοηματικής γλώσσας, επιστρατεύουμε το SMPL-X, ένα σύγχρονο παραμετρικό μοντέλο που επιτρέπει την εξαγωγή αρθρώσεων για το τρισδιάστατο ανθρώπινο σώμα, τα χέρια και το πρόσωπο από μια εικόνα. Χρησιμοποιούμε αυτό το ολιστικό μοντέλο για την αναγνώριση νοηματικής γλώσσας, δείχνοντας ότι οδηγεί σε υψηλότερες επιδόσεις από απλές εικόνες μαζί με την οπτική τους ροή όταν δίνονται σαν είσοδο σε state-of-the-art I3D-τύπου νευρωνικό δίκτυο, αλλά και από δισδιάστατο Openpose σκελετό όταν δίνεται σαν είσοδο σε ένα Recurrent νευρωνικό δίκτυο. Επιπλέον, ένα σύνολο από πειράματα πάνω στο σώμα, στα χέρια και στο πρόσωπο, δείχνουν ότι η παράλειψη οποιουδήποτε εκ των τριών καναλιών πληροφορίας, μειώνει σημαντικά το ποσοστό αναγνώρισης, αποδεικνύοντας έτσι την σημαντικότητα της συνολικής παραμετροποίησης του ανθρωπίνου σώματος, της έκφρασης και των χεριών στο πρόβλημα αναγνώρισης νοηματικής γλώσσας. Τέλος, μερικά πειράματα με εκτίμηση βάθους πραγματοποιούνται, ενώ γίνεται και αναλυτική σύγκριση μεταξύ των μοντέλων SMPL-X και ExPose.

Λέξεις-Κλειδιά: Τρισδιάστατη όραση υπολογιστών, Τρισδιάστατη ανακατασκευή προσώπου, χεριών και σώματος, Αναγνώριση Νοηματικής Γλώσσας, SMPL-X, ExPose

Εκτεταμένη Περίληψη

0.1 Τρισδιάστατη Ανακατασκευή ανθρωπίνου σώματος για αναγνώριση νοηματικής γλώσσας.

Στην ενότητα αυτή, περιγράφουμε τη διαδικασία με την οποία, το προαναφερθέν εργαλείο SMPL-X [56] θα φανεί εξαιρετικά χρήσιμο στο πρόβλημα της αναγνώρισης νοηματικής γλώσσας. Παρουσιάζουμε την τεχνική μέσω της οποίας μπορούμε να κατασκευάσουμε SMPL-X χαρακτηριστικά, ενώ παράλληλα αναφέρουμε προβληματικές περιπτώσεις του SMPL-ify αλγορίθμου. Στη συνέχεια, ορίζουμε την πειραματική κατασκευή, δηλαδή τις αρχιτεκτονικές των μοντέλων, τις υπόλοιπες μεθόδους παραγωγής χαρακτηριστικών και τις παραμέτρους εκπαίδευσης. Τέλος, παρουσιάζουμε τα αποτελέσματα των πειραμάτων μας και αξιολογούμε τις τεχνικές που χρησιμοποιήσαμε.

0.2 Από το SMPL-X στην Αναγνώριση Νοηματικής Γλώσσας

Το SMPL-X είναι ένα τρισδιάστατο μοντέλο ανακατασκευής ανθρωπίνου σώματος, χεριών και προσώπου, ικανό να διευκολύνει την ανάλυση των ανθρώπινων πράξεων, επικοινωνιών και συναισθημάτων. Βασιζόμενοι σε αυτό, είναι λογικό να εκμεταλλευτούμε αυτό το εργαλείο σε ένα πρόβλημα που απαιτεί αναλυτική και λεπτομερής αναπαράσταση του ανθρωπίνου σώματος, προσώπου και χεριών, δηλαδή το πρόβλημα της αναγνώρισης νοηματικής γλώσσας.

Παρόλο που το SMPL-X μπορεί να ανακατασκευάσει με πολύ μεγάλη ακρίβεια τον άνθρωπο σε μια συγκεκριμένη ακολουθία εικόνων, ο απώτερος στόχος είναι η αναγνώριση των νοημάτων στις ακολουθίες αυτές. Η κύρια προσδοκία είναι ότι η χαμηλών διαστάσεων παραμετρική αναπαράσταση του SMPL-X θα είναι αρκετή για να αιχμαλωτίσει τη πλειοψηφία της πληροφορίας που μεταδίδεται μέσω ενός νοήματος, δηλαδή την πόζα του χεριού, του ανθρωπίνου σώματος καθώς και των εκφράσεων του προσώπου. Η ικανότητα αυτή του SMPL-X πρέπει να το καθιστά μια πολύ αποτελεσματική ενδιάμεση αναπαράσταση για την αναγνώριση νοηματικής γλώσσας. Πιο συγκεκριμένα, το SMPL-X παίρνει κάθε ένα RGB frame ενός βίντεο και ανακατασκευάζει τον άνθρωπο στο frame αυτό επιστρέφοντας 88 παραμέτρους που αντιστοιχούν στην ανακατασκευή αυτή. Επομένως, κάθε βίντεο ή ισοδύναμα κάθε νόημα, μετατρέπεται σε μια ακολουθία από vectors μήκους 88. Η ακολουθία αυτή των SMPL-X παραμέτρων που εκτείνετε σε όλα τα frames ενός νοήματος, μπορεί να χρησιμοποιηθεί σαν είσοδος σε έναν classifier ώστε να κατατάξει το νόημα σε μία από τις διαθέσιμες κατηγορίες. Το σχήμα 2 οπτικοποιεί τη διαδικασία η οποία ακολουθείται για τη παραγωγή της ακολουθίας των vectors, δεδομένου ενός βίντεο.

Το σχήμα 2 1 δίνει ένα σύνολο από εικόνες μαζί με την SMPL-X ανακατασκευή τους,

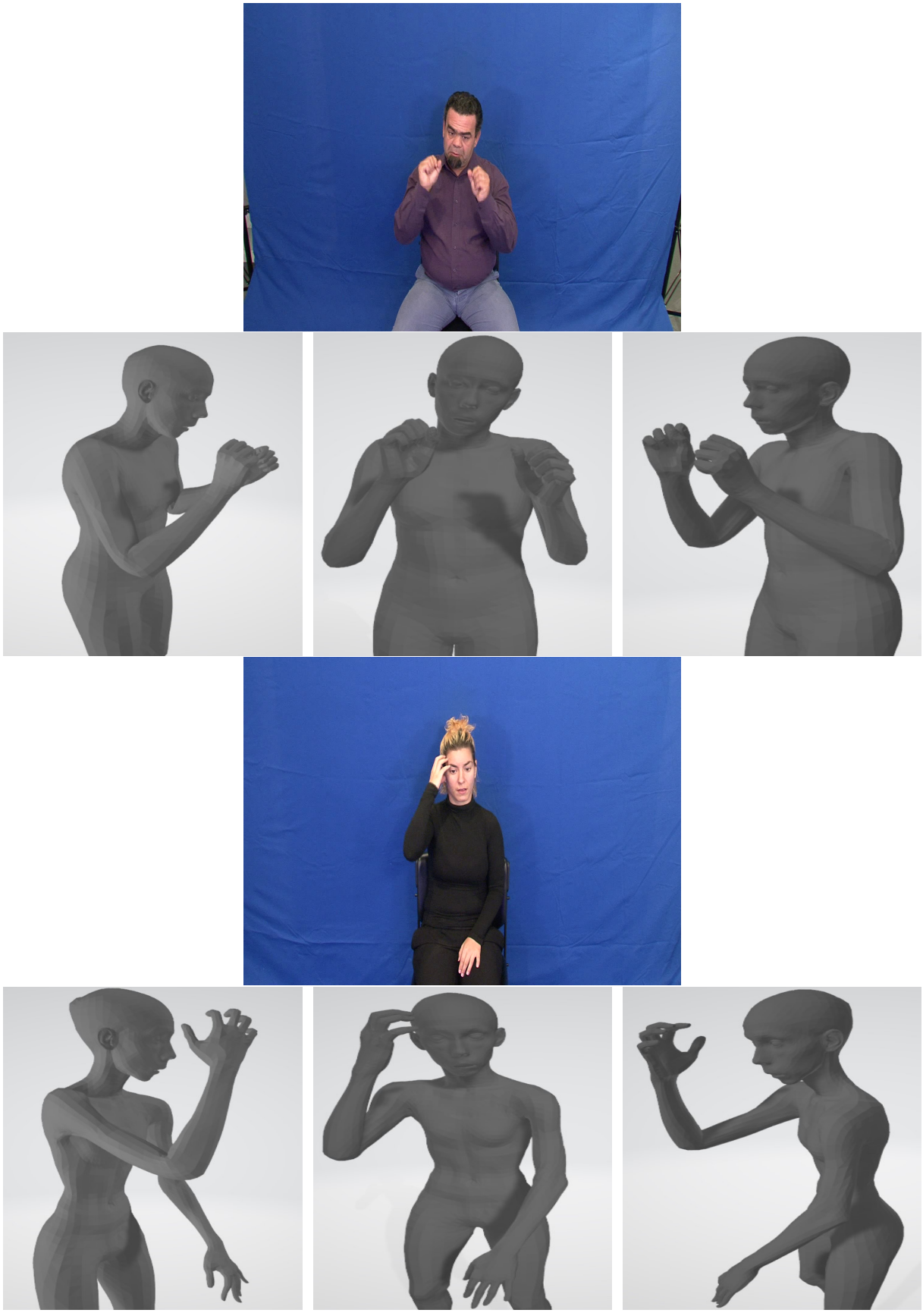

Figure 1: Παράδειγμα από την τρισδιάστατη ανακατασκευή του ανθρωπίνου σώματος, των χεριών και του προσώπου μέσω του SMPL-X μέσω διαφορετικών γωνιών, για ένα RGB frame το οποίο φαίνεται στο πάνω μέρος.

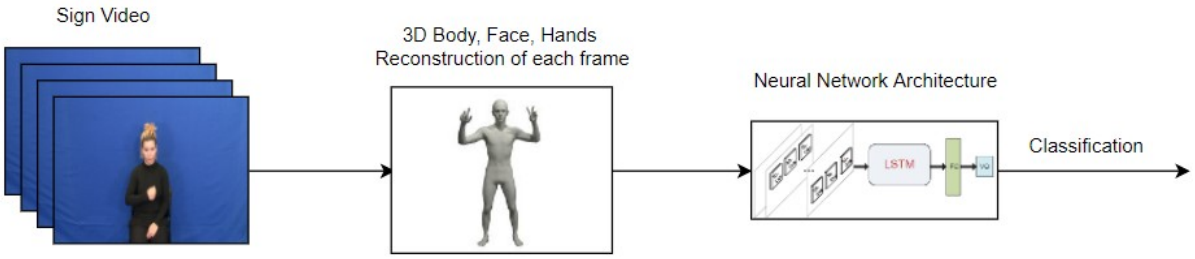

Figure 2: Η αρχιτεκτονική που χρησιμοποιήθηκε για να παραχθεί η ακολουθία από χαρακτηριστικά για ένα βίντεο που περιέχει νοηματική γλώσσα, και ύστερα να χρησιμοποιηθεί για classification.

ώστε να γίνουν κατανοητά τα λεπτομερή αποτελέσματα που παρέχει το μοντέλο του SMPL-X. Πράγματι, το μοντέλο αυτό μπορεί να ανακατασκευάσει με επάρκεια τόσο τη δομή του χεριού όσο και τις λεπτομέρειες του προσώπου, εκτός φυσικά του ίδιου του σώματος, στο περιβάλλον της νοηματικής γλώσσας, γεγονός που θα αποτελέσει κύριο χαρακτηριστικό για μίαν επιτυχή αναγνώριση.

0.2.1 Προβληματικές περιπτώσεις και Χρόνοι εκτέλεσης

Παρόλο που το SMPL-X προσφέρει αρκετά πλεονεκτήματα, χρήσιμα όχι μόνο στο πρόβλημα της αναγνώρισης νοηματικής γλώσσας αλλά γενικότερα στο τομέα της αναγνώρισης ενεργειών, έρχεται και με κάποια μειονεκτήματα. Αρχικά, συγκριτικά με άλλες τρισδιάστατες μεθόδους ανακατασκευής σώματος, χεριών και προσώπου, όπως το HMR [34] ή το ExPose [15], είναι αρκετά πιο αργό. Λόγω της βελτιστοποιητικής διαδικασίας που ακολουθεί, η ανακατασκευή μόνο μιας RGB εικόνας χρειάζεται περίπου ένα λεπτό με τη χρήση μιας κοινής GPU. Το γεγονός αυτό, περιορίζει σημαντικά τη μέθοδο αυτή από το να αναβαθμιστεί σε real-time εφαρμογή. Από την άλλη, το SMPL-X και ο SMPLify-X αλγόριθμος είναι πιθανώς η πιο λεπτομερής και ακριβής μέθοδος ανακατασκευής. Το σχήμα 3 παρέχει μια σύγκριση σε χρόνο και εκφραστικότητα μεταξύ των HMR,ExPose και SMPLify-X.

Από την άλλη, ο SMPLify-X αλγόριθμος χρησιμοποιεί το δισδιάστατο σκελετό του OpenPose [11, 12, 66, 77] για αρχικοποίηση. Παραδόξως, όταν οι γοφοί απουσιάζουν από την εικόνα, το οποίο είναι και το πλέον σύνηθες στο τομέα της αναγνώρισης νοηματικής γλώσσας, ο αλγόριθμος αρχικοποίησης αποτυγχάνει, με αποτέλεσμα ο αλγόριθμος SMPLify-X να μη μπορεί να ελαχιστοποιήσει την συνάρτησης σφάλματος. Αυτό έχει ως αποτέλεσμα, την ανακατασκευή "τεράτων", αντί για τη λεπτομερής τρισδιάστατή αναπαράσταση του ανθρωπίνου σώματος. Το σχήμα 4 δείχνει μερικά παραδείγματα όπου ο SMPLify-X αλγόριθμος αποτυγχάνει να ανακατασκευάσει επιτυχώς το ανθρώπινο σώμα, το πρόσωπο και τις χειρονομίες από μια RGB εικόνα.

0.3 Πειραματικό Μέρος

Μιας και ο στόχος μας είναι να τεστάρουμε την ικανότητα της προτεινόμενης μας μεθόδου να παράξει ικανοποιητική ανακατασκευή τρισδιάστατου προσώπου, χεριών και σώματος, περιορίζουμε την μέθοδο μας σε μη συνεχή αναγνώριση νοηματικής γλώσσας. Η συνεχής νοηματική γλώσσα περιέχει συντακτική και γλωσσολογική δομή που ξεπερνά τα όρια της δουλειάς αυτής. Αυτό σημαίνει ότι απορρίπτουμε από τους πειραματισμούς μας, βάσεις δεδομένων όπως είναι η RWTH-PHOENIX-Weather 2014 [38] και η SIGNUM [76] που

Figure 3: Σύγκριση μεταξύ των τριών state-of-the-art μεθόδων για την τρισδιάστατη ανακατασκευή ανθρωπίνου σώματος, δηλαδή το HMR, το SMPLify-X και το ExPose. Εικόνα από το [15] Supp. Material.

Figure 4: Περιπτώσεις αποτυχίας του SMPLify-X λόγω της αρχικοποιήσης του Openpose, όταν απουσιάζουν οι γοφοί από την RGB εικόνα.

αποτελούνται από ολόκληρες προτάσεις και όχι μεμονωμένες λέξεις. Αντί αυτού, επικεντρωνόμαστε στην Greek Sign Language Lemmas Dataset (GSLL) [1] [43, 72], η οποία αποδείχθηκε ιδανική για τα πειράματα μας. Η MS-ASL [74] βάση δεδομένων αποτελείτε από 222 νοηματιστές σε εξαιρετικά εναλλασσόμενα περιβάλλοντα, κάτι που κάνει τα τη σύγκλιση των Conv3D δικτύων ιδιαιτέρως δύσκολη. Για να προχωρήσουμε σε μια πιο δίκαιη σύγκρ-

[1] Η βάση αυτή μπορεί να βρεθεί στο σύνδεσμο: https://robotics.ntua.gr/gsll-dataset

ιση μεταξύ 3D ανακατασκευής και 3D συνελικτικών δικτύων, επιλέγουμε την GSLL βάση η οποία αποτελείτε μόνο από δύο νοηματιστές και 347 διαφορετικά νοήματα (κλάσεις) σε ένα σύνολο σχεδόν 3500 βίντεο μπροστά από ένα σταθερό μπλε πανό. Ο πίνακας 1 παρουσιάζει περισσότερες λεπτομέρειες για τη βάση και τα επιλεγμένα υποσύνολα.

GSLL Υποσύνολα	Βίντεο	Εικόνες	TrainSet	DevSet	TestSet
50 κλάσεις	538	22808	318	106	114
100 κλάσεις	1038	45437	618	206	214
200 κλάσεις	2038	92599	1218	406	414
300 κλάσεις	3038	140771	1818	606	614
347 κλάσεις	3464	161050	2066	695	703

Table 1: Στατιστικά για την Greek Sign Language Lemmas Dataset και τα αντίστοιχα υποσύνολα της. Ενδεικτικός προτεινόμενος χωρισμός σε train, dev και test σύνολα που χρησιμοποιούνται στα πειράματα του [43].

Η μη συνεχής αναγνώριση νοηματικής γλώσσας μπορεί να θεωρηθεί ως πρόβλημα, συγγενές με την αναγνώρισης πράξεων/κινήσεων (action recognition). Επομένως, αναμένουμε ότι παρόμοιες τεχνικές θα δουλεύουν καλά στην αναγνώρισης νοηματικής γλώσσας επίσης. Επιλέγουμε να μην επέμβουμε στο μήκος των ακολουθιών από χαρακτηριστικά. Έτσι, τα χαρακτηριστικά μας διαφέρουν σε μήκος και εκτείνονται από 10 εικόνες μέχρι και 300. Στη συνέχεια, παρουσιάζουμε τις μεθόδους με τις οποίες επιλέγουμε να αντιμετωπίσουμε το πρόβλημα της αναγνώρισης νοηματικής γλώσσας.

Openpose: Εξάγουμε 411 παραμέτρους για κάθε frame και τα παρέχουμε σαν είσοδο σε ένα Recurrent νευρωνικό δίκτυο που αποτελείται από ένα μόλις Bi-LSTM στρώμα των 256 units και ένα Dense στρώμα για την ταξινόμηση, αφού εφαρμόσουμε standard scaling στα στοιχεία μας. Πιστεύουμε ότι το να παρέχουμε ένα recurrent νευρωνικό δίκτυο σε συνδυασμό με τα Openpose features θα εξαλείψει τις περιττές πληροφορίες όπως το φόντο, τα ρούχα το φωτισμό που περιέχονται σε μια raw εικόνα.

Raw εικόνες και οπτική ροή: Μία τρισδιάστατη state-of-the-art μέθοδος για αναγνώρισης ενεργειών και νοημάτων είναι το I3D network [74, 14]. Μετασχηματίζουμε το μέγεθος κάθε εικόνας σε έναν 175×175 πίνακα και κανονικοποιούμε τα pixels στο διάστημα $[0, 1]$. Δίνουμε τις εικόνες ως είσοδο και σε ένα VGG16-LSTM μοντέλο, το οποίο αρχικοποιείται με τα βάρη του Imagenet [23], για περαιτέρω πειραματισμό. Η εικόνα 5 δείχνει την τρισδιάστατη συνελικτική αρχιτεκτονική που περιεγράφηκε, με πιο λεπτομέρεια.

SMPL-X: Λόγω της ικανότητας του SMPL-X να αναπαραστήσει τη δομή του ανθρωπίνου σώματος με λεπτομέρεια, πιστεύουμε ότι αυτή η μέθοδος θα παρέχει χαρακτηριστικά-κλειδιά για το πρόβλημα της αναγνώρισης της νοηματικής γλώσσας. Επιπλέον, το SMPL-X χρειάζεται τις παραμέτρους του Openpose για να εξάγει τα χαρακτηριστικά του, επομένως υποθέτουμε ότι το πρώτο θα παρέχει πιο ποιοτικά και βαθύτερα χαρακτηριστικά από ότι το δεύτερο. Επιπροσθέτως, το SMPL-X παρέχει τρισδιάστατη πληροφορία, εν συγκρίσει με το Openpose το οποίο δίνει δισδιάστατα keypoints, και επομένως τα εξαγόμενα χαρακτηριστικά θα εμπεριέχουν περισσότερη πληροφορία. Η μέθοδος αυτή εξάγει 88 χαρακτηριστικά για κάθε εικόνα, δημιουργώντας έναν πίνακα (μήκος ακολουθίας) $\times$ 88 για κάθε ακολουθία, η οποία στη συνέχεια υπόκειται standard scaling, όμοια με τα χαρακτηριστικά του Openpose. Παρόμοια με το Openpose επίσης, χρησιμοποιούμε το ίδιο recurrent νευρωνικό δίκτυο, όχι μόνο επειδή τα δύο πειράματα εμπεριέχουν το ίδιος είδος χαρακτηριστικών αλλά και κυρίως επειδή θέλουμε να συγκρίνουμε ευθέως τις φύο μεθόδους, ανεξάρτητα της αρχιτεκτονικής.

Εκπαιδεύομε όλα τα νευρωνικά δίκτυα χρησιμοποιώντας categorical cross-entropy loss.

Stochastic gradient descent χρησιμοποιείται για τη βελτιστοποίηση της συνάρτησης σφάλματος, με αρχικό learning rate στ 0.0001 και 10% decay rate για κάθε εποχή, ενώ το batch size τίθεται 1 λόγω του διαφορετικού μήκους που έχουν οι σειρές. Χρησιμοποιούμε Learning Rate Reduction και Early Stopping μέσω του validation loss με patience 3 και 5 εποχές αντίστοιχα, για να αποφύγουμε το overfitting. Το σχήμα 6 δείχνει ένα παράδειγμα από μια εικόνα, την οπτική της ροή, το δισδιάστατο σκελετό από το Openpose, και την 3D ανακατασκευή του SMPL-X.

0.4 Πειραματική Αποτίμηση

0.4.1 Πειραματικά αποτελέσματα

Με βάση τον πίνακα 2, τα Openpose και SMPL-X μοντέλα, τα οποία αποτελούνται από 1.6 και 0.9 εκατομμύρια παραμέτρους αντίστοιχα, ξεπερνούν σε επίδοση τα Conv3D-LSTM και τα VGG16-LSTM μοντέλα, τα οποία αποτελούνται από 43 και 15 εκατομμύρια παραμέτρους αντίστοιχα. Αυτό μπορεί να ερμηνευτεί με βάση το γεγονός ότι τα δύο πρώτα μπορούν να απορρίψουν την περιττή πληροφορία για κάθε εικόνα, και να κρατήσουν μόνο τη πληροφορία που αφορά στη σωματοδομή του ανθρώπου που νοηματίζει. Συγκεκριμένα, το VGG16 μοντέλο αποτυγχάνει τελείως να συγκλίνει και μηδενίσει το loss του, επιτυγχάνοντας ποσοστό μικρότερο του 10% για όλες τις κλάσεις. Αυτό είναι αναμενόμενο, μιας και οι Joze και Koller in [74], έχουν εκπαιδεύσει ένα VGG16-LSTM μοντέλο στην MS-ASL βάση το οποίο πετυχαίνει 13.33% στο υποσύνολο ASL100 και μόλις 1.47% στο υποσύνολο ASL500. Όπως αναφέρθηκε νωρίτερα, η GSLL βάση χαρακτηρίζεται από ομοιόμορφο περιβάλλον μεταξύ κάθε νοήματος και κάθε νοηματιστή (2 νοηματιστές μπροστά από ένα μπλε χιτώνα). Η MS-ASL βάση από την άλλη, αποτελείται από 222 ξεχωριστούς νοηματιστές όπου κάθε νόημα πραγματοποιείται σε τελείως διαφορετικό περιβάλλον. Πιστεύουμε ότι το Openpose και το SMPL-X θα ξεπεράσουν κατά πολύ σε επίδοση τα συνελικτικά μοντέλα σε αυτές τις βάσεις, που προσομοιώνουν τον πραγματικό κόσμο με μεγαλύτερη ακρίβεια. Τέλος, το SMPL-X φαίνεται να παράγει πιο ποιοτικά χαρακτηριστικά σε σύγκριση με το Openpose, ιδιαίτερα με την ύπαρξη περισσότερων νοημάτων, κάτι που δείχνει ότι μια πιο ποιοτική και λεπτομερής αναπαράσταση του ανθρωπίνου σώματος είναι απαραίτητη για το πρόβλημα της αναγνώρισης νοηματικής γλώσσας. Με την προσθήκη περισσότερων, ποικίλων και δυσκολότερων νοημάτων στο train set, το Openpose αποτυγχάνει να αποτυπώσει τις μικρές λεπτομέρειες

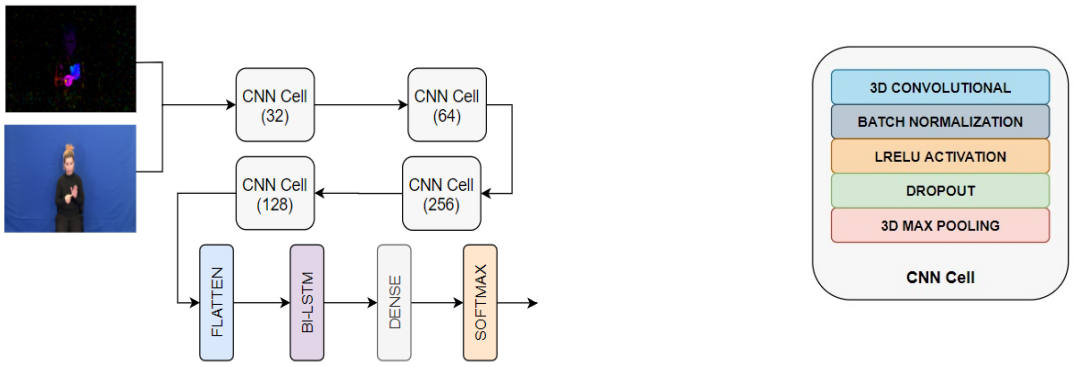

Figure 5: Η αρχιτεκτονική που χρησιμοποιήθηκε για το συνελικτικό I3D-type μοντέλο. Στα αριστερά είναι η προτεινόμενη αρχιτεκτονική των 3D CNN κελιών ακολουθούμενα από ένα Bidirectional LSTM στρώμα. Στα δεξιά φαίνεται το εσωτερικά στρώματα για κάθε ένα από τα 3D cells.

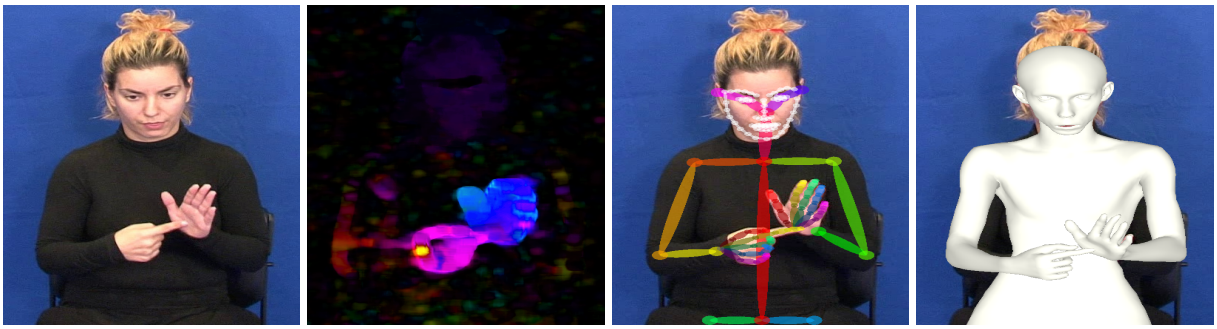

Figure 6: i) Πρώτη εικόνα: RGB εικόνα, ii) Δεύτερη εικόνα: Οπτική ροής της εικόνας, iii) Τρίτη εικόνα: Openpose 2Δ Σκελετό, iv) Τέταρτη εικόνα: 3Δ Ανακατασκευή μέσω SMPL-X.

που διαχωρίζει τα νοήματα αυτά, ενώ το SMPL-X διατηρεί τα ποσοστά ακριβείας του σχεδόν fixed.

0.4.2 Ablation Έρευνα

Για να εξετάσουμε περαιτέρω τα features που παράγονται από το SMPL-X, πειραματιζόμαστε με συνδυασμό ενός υποσυνόλου από αυτά. Συγκεκριμένα, το SMPL-X παράγει ένα σύνολο από 88 παραμέτρους, 10 για το σχήμα, 3 για global orientation, 24 για αριστερό και δεξί χέρι, 3 για το σαγόνι, 6 για αριστερό και δεξί χέρι, 10 για την έκφραση και 32 για τη σωματοδομή. Επιπλέον, είναι ευρέως γνωστό ότι η νοηματική γλώσσα, δεν βασίζεται μόνο στις χειρονομίες αλλά εξίσου και στις κινήσεις ολόκληρου του σώματος και στις εκφράσεις του προσώπου. Για να αναδείξουμε αυτόν τον ισχυρισμό, προχωράμε σε μερικά πειράματα. Αρχικά, αφαιρούμε όλη την πληροφορία που προέρχεται από τις εκφράσεις του προσώπου (σαγόνι, αριστερό και δεξί μάτι και έκφραση) και εκπαιδεύουμε το μοντέλο ξανά, με ένα σύνολο 69 χαρακτηριστικών. Στη συνέχεια, αφαιρούμε μόνο τη πληροφορία της πόζας του σώματος και εκπαιδεύουμε το μοντέλο με ένα σύνολο 50 παραμέτρων. Τέλος, για πληρότητα, αφαιρούμε την πληροφορία από το αριστερό και το δεξί χέρι, και εκπαιδεύουμε ξανά το μοντέλο με 64 παραμέτρους. Εκτελούμε τα ίδια πειράματα και με το Openpose διαχωρίζοντας τα keypoints της πόζας (75 παραμέτρους), τα keypoints του προσώπου (210 παραμέτρους) και τα keypoints αριστερού και δεξιού χεριού (126 παραμέτρους). Ο πίνακας 3 συνοψίζει τα αποτελέσματα για όλα τα προαναφερθέντα πειράματα.

Αρχικά, βλέπουμε ότι η παράλειψη οποιουδήποτε από τα τρια κανάλια πληροφορίας πράγματι μειώνει την ακρίβεια του μοντέλου. Στη πράξη, περιμένουμε ότι η παράλειψη των χαρακτηριστικών του προσώπου να επηρεάσει ακόμα περισσότερο στην αναγνώριση συνεχούς νοηματικής γλώσσας, όπου το πρόσωπο παίζει πολύ σημαντικό ρόλο στο να εκφράζει την ένταση μιας λέξης. Για παράδειγμα, η λέξη "βροχή" και "χιόνι", έχουν ακριβώς την σύνθεση χειρονομίας και σώματος, ενω μόνο το σχήμα του προσώπου αλλάζει. Επιπλέον, παρατηρούμε ότι το να παραλείψουμε την πληροφορία της χειρονομίας στο SMPL-X κοστίζει

Μέθοδος \ GSLL Subset	Subset 50	Subset 100	Subset 200	Subset 300	Πλήρης Βάση	Παράμετροι
3D RGB & Οπτική Ροή	90.41%	86.85%	80.79%	71.36%	65.95%	43.41 εκατ.
2D Openpose Σκελετός	96.49%	94.39%	93.24%	91.86%	88.59%	1.55 εκατ.
3D SMPL-X Ανακατασκευή	**96.52%**	**95.87%**	**95.41%**	**95.28%**	**94.77%**	0.88 εκατ.

Table 2: Σύγκριση των τριών βασικών μεθόδων για εκπαίδευση: i) Απλές RGB εικόνες με την οπτική τους ροή ii) Keypoints του Openpose σκελετού και iii) Τρισδιάστατη ανακατασκευή μέσω SMPL-X.

Παράμετροι	Openpose	SMPL-X
Όλοι	**88.59**%	**94.77**%
Χωρίς Πρόσωπο	88.34%	93.19%
Χωρίς Χέρια	70.20%	89.58%
Χωρίς Σώμα	84.21%	85.02%

Table 3: Πειράματα με υποσύνολα από features που παράχθηκαν με Openpose και SMPL-X.

λιγότερο από το να παραλείψουμε την πόζα του σώματος. Αυτό μπορεί να ερμηνευτεί ως εξής. Όταν υπάρχουν λίγα και απλά διαθέσιμα νοήματα, αυτά μπορούν κυρίως να αποτυπωθούν από την κίνηση του βραχίονα του χεριού, ενώ τα χέρια και τα δάχτυλα παραμένουν κυρίως ευθεία. Παρόλα αυτά, και τα χέρια όπως και η σωματοδομή (κυρίως λόγω των βραχιόνων του χεριού), είναι εξέχουσας σημασίας στην αναγνώριση νοηματικής γλώσσας, ενώ την ίδια στιγμή, η παράλειψη των εκφράσεων του προσώπου επηρεάζει την αποτελεσματικότητα του μοντέλου. Από την άλλη, στο Openpose, λόγω του γεγονότος ότι έχει πολύ λιγότερες παραμέτρους για το σώμα, δηλαδή μόλις 75 από τις 411, είναι πολύ πιο επιβλαβές να αφαιρέσεις τα χέρια, παρά το σώμα.

0.5 Επιπλέον Πειράματα

0.5.1 SMPL-X vs ExPose

Όπως περιεγράφηκε προηγουμένως, το SMPL-X είναι πιθανώς η πιο ποιοτική μέθοδος για την τρισδιάστατη ανακατασκευή της σωματοδομής, των εκφράσεων του προσώπου και των χειρονομιών, από μια RGB εικόνα. Δεν είναι όμως, η πιο γρήγορη μέθοδος. Συγκεκριμένα, η ενός λεπτού διαδικασία βελτιστοποίησης που ακολουθεί ο αλγόριθμος της για κάθε εικόνα, καθιστά τον SMPLify-X αλγόριθμο, μη ικανό για real-time εφαρμογές. Το ExPose, από την άλλη πλευρά, το οποίο δημοσιεύθηκε το 2020, διατηρεί την ποιότητα της ανακατασκευής χρησιμοποιώντας body-driven attention μηχανισμό και τρέχει σε σχεδόν real-time με τη χρήση μιας κοινής GPU. Στη συνέχεια, συγκρίνουμε ποιοτικά τις δύο μεθόδους, ανακατασκευάζοντας τρισδιάστατες αναπαραστάσεις και για το SMPL-X και για το ExPose, σε εικόνες παρμένες από βάσεις νοηματικής γλώσσας.

Επιπλέον, συγκρίνουμε τις δύο αυτές μεθόδους στο πρόβλημα της αναγνώρισης, εφαρμόζοντας την ίδια διαδικασία που περιεγράφηκε στις παραγράφους 0.2 και 0.3. Ο πίνακας 4 δείχνει τα αποτελέσματα για το SMPL-X και το ExPose στα διάφορα υποσύνολα της GSLL βάσης.

Μέθοδος \ GSLL Subset	Subset 50	Subset 100	Subset 200	Subset 300	Παράμετροι
SMPLify-X	96.49%	94.39%	93.24%	91.86%	0.88 εκατομμύρια
ExPose	**99.20%**	**99.46%**	**97.74%**	**96.47%**	1.59 εκατομμύρια

Table 4: Σύγκριση μεταξύ των δύο τρισδιάστατων μεθόδων για ανακατασκευή ανθρωπίνου σώματος, χεριών και προσώπου: i) SMPLify-X ii) ExPose

Παρατηρούμε ότι το ExPose βοηθάει το νευρωνικό δίκτυο να αναγνωρίσει καλύτερα τις διαφορές μεταξύ των νοημάτων σε κάθε βίντεο. Παρά το γεγονός ότι το ExPose, για να καταφέρει να πετύχει σχεδόν real-time ανακατασκευή, μειώνει την εκφραστικότητα και την ποιότητα της, μπορεί να αποκωδικοποιήσει επαρκώς τις λεπτομέρειες του ανθρωπίνου σώματος, των χεριών και του προσώπου, και μάλιστα καλύτερα από το SMPLify-X. Επιπλέον, τα χαρακτηριστικά του ExPose κάνουν το νευρωνικό δίκτυο να συγκλίνει πολύ ταχύτερα σε

σύγκριση με το SMPLify-X. Η σύγκριση μεταξύ των δύο state-of-the-art μεθόδων φαίνεται να παρουσιάζει μεγάλο ερευνητικό ενδιαφέρον και θα πρέπει να διερευνηθεί σε μεγαλύτερες και δυσκολότερες βάσεις επιπλέον.

0.5.2 ExPose στην MS-ASL βάση

Εξετάζουμε περαιτέρω την επίδοση του ExPose, σε μια από τις πιο δύσκολες βάσεις διαθέσιμες για αναγνώριση μη συνεχούς νοηματικής γλώσσας, την MS-ASL. Παρουσιάζουμε τα state-of-the-art αποτελέσματα στη βάση αυτή για τις διάφορες μεθόδους, καθώς παραθέτουμε και τα δικά μας αποτελέσματα χρησιμοποιώντας το ExPose. Εκπαιδεύουμε ένα απλό Recurrent νευρωνικό δίκτυο χρησιμοποιώντας μόνο ένα LSTM επίπεδο και παρουσιάζουμε τα αποτελέσματα μας στον πίνακα 5. Τα πειράματα μας περιορίζονται μόνο σε ένα μικρό υποσύνολο της βάσης. Βλέπουμε ότι το νευρωνικό δίκτυο αρχικοποιημένο με τα ExPose χαρακτηριστικά πετυχαίνει ένα εκπληκτικό 37.39% ακρίβεια. Η μέθοδος αυτή πρέπει να συγκρίνεται με το HCN network, το οποίο είναι ένα σύνθετο recurrent νευρνωικό δίκτυο που χρησιμοποιεί σαν είσοδο τον δισδιάστατο σκελετό του Openpose. Αναμένουμε πως το ExPose συνδυαζόμενο με το ισχυρό HCN δίκτυο θα ξεπεράσει κατά πολύ το παρόν HCN δίκτυο με το Openpose, καθώς και την Re-sign μέθοδο. Τέλος, σε ένα ακόμα μεγαλύτερο υποσύνολο της MS-ASL βάσης, όπου το I3D γρήγορα αποκλίνει, το ExPose δύναται να παραμείνει σταθερό και να το ξεπεράσει, κάνοντας το την καλύτερη διαθέσιμη μέθοδο.

0.5.3 Κανάλι Βάθους

0.5.3.1 MiDaS: Εκτίμηση Βάθους

Το MiDaS είναι ένα ευσταθές μοντέλο εκτίμησης βάθους το οποίο μπορεί να εφαρμοστεί σε ποικίλα περιβάλλοντα, και αναπτύχθηκε από το R. Ranftl και άλλους [63]. Η δουλειά αυτή προτείνει μια πρωτότυπη συνάρτηση σφάλματος που είναι ανεξάρτητη σε μεγάλες πηγές ασυμβατότητας μεταξύ διαφόρων βάσεων δεδομένων, συμπεριλαμβανομένης της αδιευκρίνιστης και διαφοροποιούμενης κλίμακας. Αυτά τα σφάλματα επιτρέπουν την εκπαίδευση δεδομένων που έχουν διαλεχθεί μέσω διαφορετικών καταγραφικών μηχανημάτων όπως είναι οι stereo κάμερες, τα laser scanners, και οι structured light sensors. Ποιοτικά αποτελέσματα του MiDaS φαίνονται στο σχήμα 7.

Ο R. Ranftl και άλλοι, βελτίωσαν το MiDaS μοντέλο προτείνοντας Vision Transformers για Dense πρόβλεψη στο [62]. Οι Dense vision transformers είναι μια αρχιτεκτονική που αναμοχλεύει τους vision transformers με τα συνελικτικά νευρωνικά δίκτυα για προβλήματα που αφορούν σε dense προβλέψεις. Τα tokens συναρμολογούνται από διάφορα στάδια του vision transformer σε αναπαραστάσεις που μοιάζουν με εικόνες διαφόρων αναλύσεων, οι

Μέθοδος \ MS-ASL Subset	Subset 100	Subset 200	Subset 500	Subset 1000
Απλός Classifier	0.99%	0.50%	0.21%	0.11%
VGG+LSTM [21, 19]	13.33%	7.56%	1.47%	-
HCN [46]	46.08%	35.85%	21.45%	15.49%
Re-sign [40]	45.45%	43.22%	27.94%	14.69%
I3D [14]	**81.76%**	**81.97%**	**72.50%**	**57.69%**
ExPose [15]	37.39%	-	-	-

Table 5: Σύγκριση μεταξύ της ExPose μεθόδου με ένα απλό LSTM RNN δίκτυο και των state-of-the-art μεθόδων για αυτή τη βάση.

Figure 7: Ποιοτικά αποτελέσματα του MiDas μοντέλου εκτίμησης βάθους.

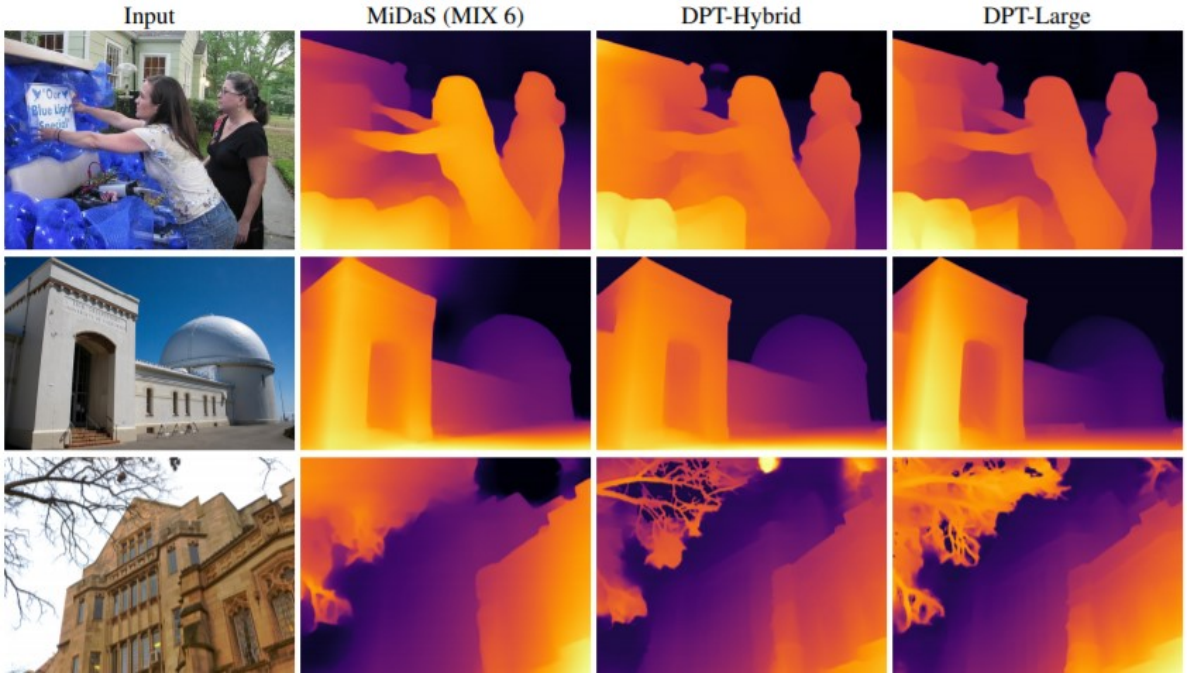

Figure 8: Ποιοτική σύγκριση μεταξύ του πλήρως συνελικτικού δικτύου MiDas και του Depth Vision Transformer στο πρόβλημα της εκτίμησης βάθους.

οποίες σταδιακά συνδυάζονται σε πλήρους ανάλυση προβλέψεις χρησιμοποιώντας συνελικτικούς decoders. Το δίκτυο κορμού του transformer επεξεργάζεται τις αναπαραστάσεις σε σταθερή και σχετικά υψηλή ανάλυση και έχει ένα καθολικό πεδίο αποδοχής σε κάθε στάδιο. Αυτές οι ιδιότητες, επιτρέπουν στον dense vision transformer να παρέχει μια fine-grained και πιο καθολική και ξεκάθαρη πρόβλεψη σε σύγκριση με το πλήρες συνελικτικό νευρωνικό δίκτυο. Συγκεκριμένα, στο πρόβλημα της πρόβλεψης βάθους, οι dense vision transformers πετυχαίνουν βελτίωση της τάξης 28% σε σύγκριση με τους state-of-the-art πλήρως συνελικτικών νευρωνικών μοντέλων MiDaS στο [63]. Μια ποιοτική σύγκριση μεταξύ αυτών των δύο φαίνεται στο σχήμα 8.

Η αρχιτεκτονική του depth vision transformer φαίνεται στο Σχήμα 9. Στα αριστερά φαίνεται η σύνοψη της αρχιτεκτονικής. Η εικόνα της εισόδου μετασχηματίζεται σε tokens (πορτοκαλί) είτε μέσω της εξαγωγής μη επικαλυπτόμενων patches ακολουθούμενα από μια

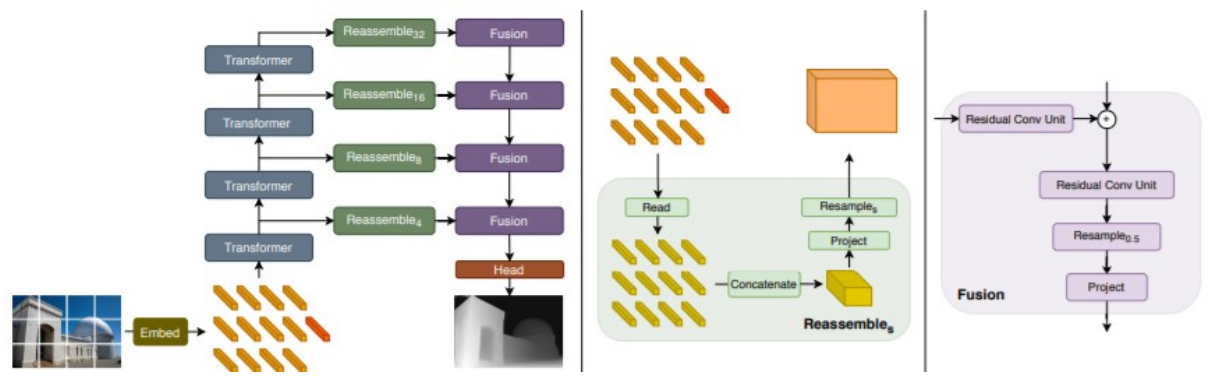

Figure 9: Η αρχιτεκτονική του depth vision transformer για το πρόβλημα της εκτίμησης βάθους.

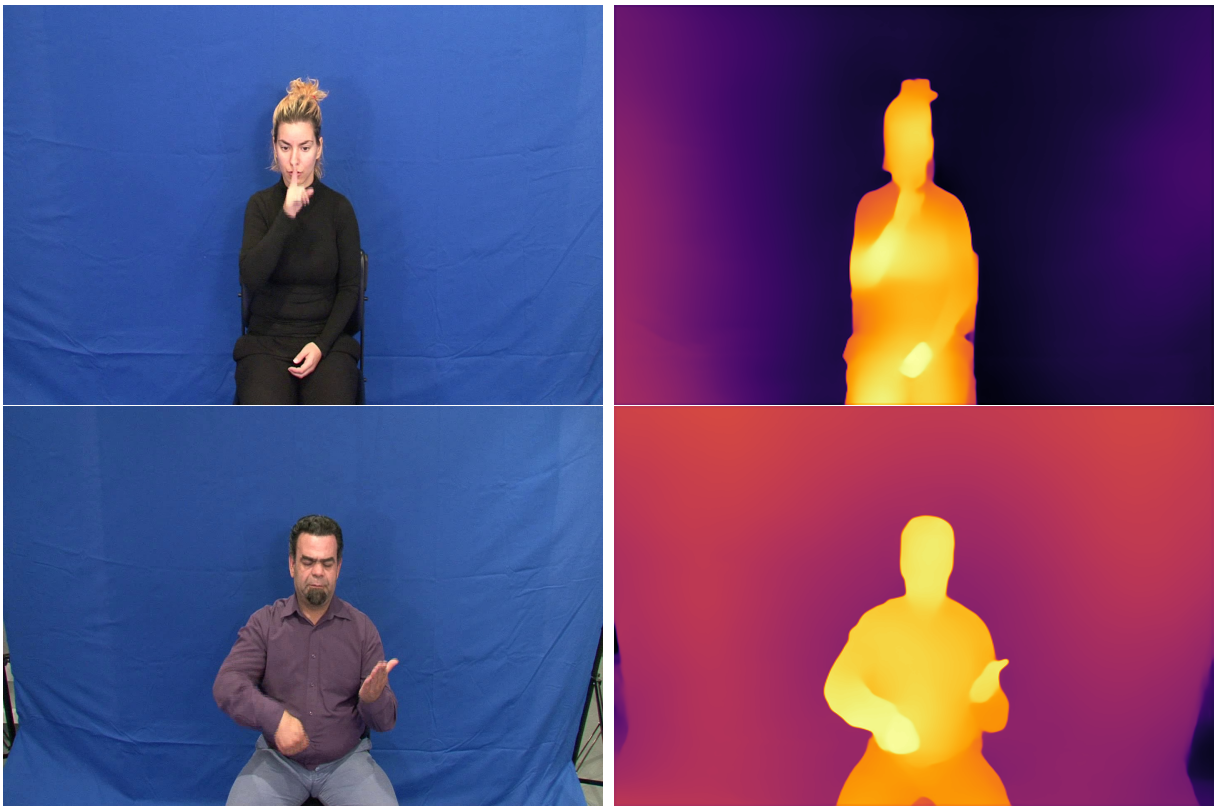

Figure 10: Παραδείγματα από την GSLL βάση μαζί με την Εκτίμηση Βάθους του MiDaS.

γραμμική προβολή στην flattened αναπαράσταση (DPT-Base και DPT-Large) είτε μέσω της εφαρμογής του ResNet-50 εξαγωγέα χαρακτηριστικών (DPT-Hybrid). Η ενσωματωμένη εικόνα προσαυξάνεται με positional embedding και ένα ανεξάρτητο readout token (κόκκινο) προστίθεται. Τα tokens περνάνε μέσω πολλαπλών σταδίων του transformer. Ύστερα, τα tokens επανασυναρμολογούνται από διαφορετικά στάδια ώστε να μοιάζουν με μια αναπαράσταση εικόνας σε πολλαπλές αναλύσεις (πράσινο). Τα fusion μοντέλα (μωβ) σταδιακά ενώνουν και κάνουν upsampling τις αναπαραστάσεις για να παράγουν fine-grained προβλέψεις. Στο κέντρο, βλέπουμε την ανάλυση της $Reassemble_S$ διαδικασίας. Τα tokens ενώνονται σε χάρτες χαρακτηριστικών με χωρική ανάλυση $\frac{1}{s}$ της αρχικής εικόνας. Τέλος, στα δεξιά τα μπλοκ fusion συνδυάζουν τα χαρακτηριστικά χρησιμοποιώντας residual συνελικτικές μονάδες και κάνουν upsampling των χαρτών με χαρακτηριστικά.

Για τα πειράματα μας, χρησιμοποιούμε το πιο σύγχρονο και επιτυχές μοντέλο για εκ-

τίμηση βάθους, δηλαδή το MiDaS μοντέλο που περιέχει τους depth vision transformers.

0.5.3.2 Πειράματα με το κανάλι του βάθους

Πιστεύουμε ότι η πληροφορία βάθους θα αυξήσει την απόδοση του νευρωνικού δικτύου, όταν χρησιμοποιηθεί σαν δεύτερο κανάλι. Μιας και η Greek Sign Language Lemmas Dataset δε περιέχει πληροφορία βάθους, χρησιμοποιούμε το προαναφερθέν εργαλείο MiDaS για εκτίμηση του βάθους. Δύο παραδείγματα από τη βάση αυτή μαζί με την εκτίμηση βάθους φαίνονται στο σχήμα 10.

Για να τεστάρουμε την πληροφορία του βάθους, εκπαιδεύουμε ένα μοντέλο χρησιμοποιώντας ως μόνη πληροφορία το βάθος. Στη συνέχεια, εκπαιδεύουμε ένα δικάναλο CNN-LSTM μοντέλο χρησιμοποιώντας την αρχιτεκτονική του σχήματος 2, ενώ τέλος συνδυάζουμε την SMPL-X πληροφορία με αυτή του βάθους. Τα αποτελέσματα φαίνονται στον πίνακα 6. Παρατηρούμε ότι η χρήση μόνο του βάθος δε βοηθάει το νευρωνικό να εκπαιδευτεί σωστά για το πρόβλημα της αναγνώρισης νοηματικής γλώσσας, σε σύγκριση με άλλα προβλήματα που είναι ιδιαίτερα χρήσιμο. Τα αποτελέσματα με μόνο την RGB πληροφορία έχουν αναφερθεί στο [42]. Ο συνδυασμός RGB πληροφορίας και βάθους δε βοηθάει το νευρωνικό να κρατήσει τη χρήσιμη πληροφορία με αποτέλεσμα να έχει ελαφρώς χειρότερη επίδοση.

Μέθοδος \ GSLL Subset	Subset 50	Subset 100	Subset 200	Subset 300
Μόνο βάθος	11.56%	9.91%	9.1%	6.50%
RGB	**88.59%**	**84.58%**	**71.98%**	**55.37%**
RGB + Βάθος	85.81%	82.59%	70.01%	52.10%

Table 6: Πειραματικά αποτελέσματα για τις τρεις μεθόδους training: i) χρησιμοποιώντας μόνο πληροφορία βάθους, ii) χρησιμοποιώντας μόνο την RGB πληροφορία και iii) συνδυάζοντας RGB εικόνες και την πληροφορία του βάθους.

0.6 Συνεισφορές

Στη διπλωματική αυτή μελετήθηκε η σύγχρονη έρευνα στο πεδίο της τρισδιάστατης όρασης υπολογιστών και συγκεκριμένα η τρισδιάστατη ανακατασκευή του ανθρώπινου σώματος, των χεριών και της έκφρασης. Επιπλέον, ερευνήθηκε το σύνθετο πρόβλημα της αναγνώρισης μη συνεχούς νοηματικής γλώσσας, και πως η τρισδιάστατή ανακατασκευή μπορεί να βοηθήσει στο πρόβλημα αυτό. Οι συνεισφορές της διπλωματικής αυτής είναι αρκετές και μπορούν να συνοψιστούν στις εξής:

- Προσφέραμε μια ενδελεχή και λεπτομερής βιβλιογραφική ανάλυση των πιο σύγχρονων και state-of-the-art δισδιάστατων και τρισδιάστατων μεθόδων για ανακατασκευή του ανθρώπου, στα τελευταία 5 χρόνια. Συγκεκριμένα, περιγράψαμε και εξηγήσαμε σε βάθος τη μεθοδολογία που υπάρχει πίσω από το διάσημο δισδιάστατο μοντέλο εξαγωγής ανθρώπινου σκελετού, το Openpose (2016-2019) [11, 12, 66, 77]. Ορίσαμε και εξηγήσαμε τα τρισδιάστατα παραμετρικά μοντέλα ου χρησιμοποιούνται για να περιγραφεί το ανθρώπινο σώμα, SMPL (2015) [48] και SMPL-X (2019) [56]. Επιπλέον, περιγράψαμε τις τεχνικές λεπτομέρειες πίσω από τις πιο ποιοτικές μεθόδους για εξαγωγή τρισδιάστατων παραμέτρων που περιγράφουν το ανθρώπινο σώμα, το πρόσωπο και τα χέρια μέσω μιας μόνο RGB εικόνας, δηλαδή των SMPL-ify (2016) [7], HMR (2018) [34], SMPLify-X (2019) [56] και ExPose (2020) [15].

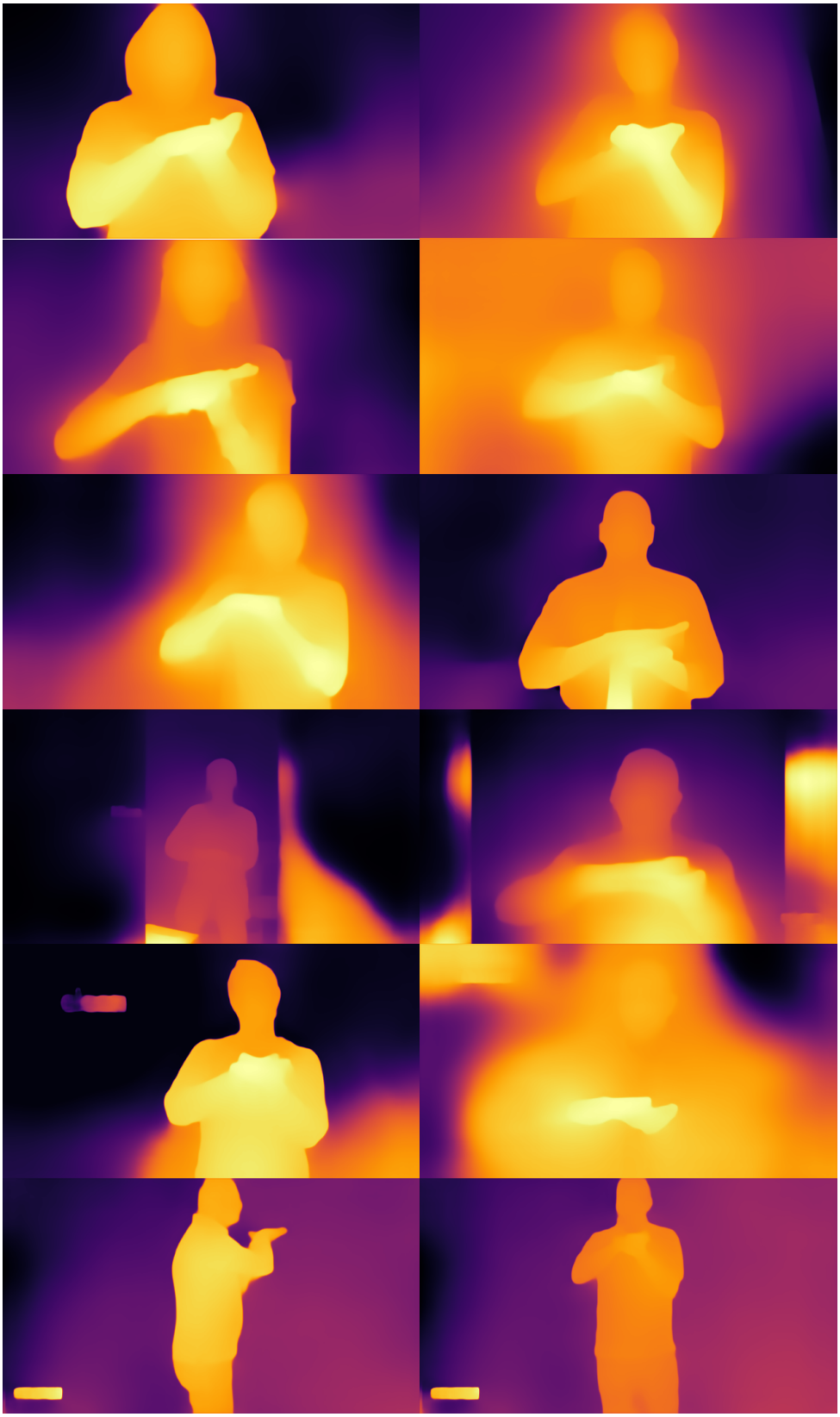

Figure 11: Προβλέψεις βάθους για χαρακτηριστικές εικόνες από την MS-ASL βάση που αποτυπώνουν το ίδιο νόημα “καθαρό”.

- Προσφέραμε μια παρόμοια βιβλιογραφική ανάλυση των πιο σημαντικών βάσεων για νοηματική γλώσσα καθώς και τις state-of-the-art μεθόδους για την αντιμετώπιση του προβλήματος αναγνώρισης της νοηματικής. Αναλύσαμε σε βάθος την MS-ASL βάση [74], και τις κορυφαίες τεχνικές που χρησιμοποιήθηκαν για να επιτευχθεί υψηλή επίδοση στο πρόβλημα της αναγνώρισης μη συνεχούς νοηματικής γλώσσας. Περιγράψαμε δύο από τις πιο βασικές ελληνικές βάσεις νοηματικής γλώσσας, τις Greek Sign Language Lemmas Dataset [43**?**] και Greek Sign Language Dataset [1]. Στο τομέα της συνεχούς νοηματικής γλώσσας παρουσιάσαμε τις δύο πιο επιφανείς βάσεις για αυτό το πρόβλημα, τις RWTH-PHOENIX-Weather 2014 dataset [38] και RWTH-PHOENIX-Weather 2014 Translation dataset [10], μαζί με τις state-of-the-art μεθόδους στο τομέα της αναγνώρισης νοηματικής γλώσσας των τελευταίων ετών.

- Αξιοποιήσαμε την Greek Sign Language Lemmas Dataset (GSLL) για τα πειράματα μας, την οποία οργανώσαμε εξ αρχής και διαθέσαμε δημόσια και περαιτέρω πειραματισμούς από την ερευνητική κοινότητα. Η GSLL βάση μαζί με στατιστικές λεπτομέρειες και οδηγίες χρήσης είναι διαθέσιμη στο https://robotics.ntua.gr/gsll-dataset.

- Εφαρμόσαμε τις 3D μεθόδους ανακατασκευής σώματος, προσώπου και χεριών στο πρόβλημα της αναγνώρισης μη συνεχούς νοηματικής γλώσσας, πετυχαίνοντας κορυφαία αποτελέσματα και ξεπερνώντας όλες τις υπόλοιπες γνωστές μεθόδους. Συγκεκριμένα, αξιοποιήσαμε το SMPL-X, ένα σύγχρονο παραμετρικό μοντέλο που επιτρέπει την εξαγωγή αρθρώσεων για το τρισδιάστατο σώμα, τις χειρονομίες και τις εκφράσεις του προσώπου από μια RGB εικόνα. Χρησιμοποιήσαμε αυτό το ολιστικό 3D εργαλείο για αναγνώριση νοηματικής γλώσσας, δείχνοντας ότι οδηγεί σε υψηλά ποσοστά αναγνώρισης σε σύγκριση με τα τρισδιάστατα συνελικτικά δίκτυα που παίρνουν σαν είσοδο εικόνες και την οπτική τους ροή, αλλά και από Recurrent νευρωνικά δίκτυα που παίρνουν σαν είσοδο δισδιάστατο σκελετό του Openpose.

- Επιπλέον, πραγματοποιήσαμε μια ablation έρευνα, που δείχνει την σημασία του να υπάρχουν διαθέσιμα και τα τρία κανάλια πληροφορίας, συγκεκριμένα, οι εκφράσεις του προσώπου, το σχήμα των χεριών αλλά και ολόκληρο το σώμα, για να επιτευχθεί η βέλτιστη αναγνώριση νοηματικής γλώσσας. Συγκεκριμένα, εκπαιδεύσαμε τρία διαφορετικά μοντέλα, παραλείποντάς τα χαρακτηριστικά του προσώπου, τα χαρακτηριστικά του σώματος και τα χαρακτηριστικά των χεριών αντίστοιχα. Δείχνουμε ότι κάθε μέρος του σώματος παίζει σημαντικό ρόλο στην επίτευξη μέγιστου αποτελέσματος στο πρόβλημα της αναγνώρισης νοηματικής γλώσσας.

- Συγκρίνουμε ευθέως δύο από τις σύγχρονες μεθόδους για τρισδιάστατη ανακατασκευή, το SMPLify-X και το ExPoSE, σε χρόνο εκτέλεσης αλλά και σε εκφραστικότητα. Αυτό σημαίνει, ότι τεστάρουμε και τις μεθόδους στο ίδιο πρόβλημα της αναγνώρισης μη συνεχούς νοηματικής γλώσσας για να ελέγξουμε την επάρκεια τους. Παράλληλα, παρέχουμε εικόνες και για τις δύο ανακατασκευές για ποιοτική σύγκριση, ενώ σχολιάζουμε το χρόνο εκτέλεσης τους. Τέλος, αξιοποιούμε το ExPose, στην ίσως πιο δύσκολη διαθέσιμη βάση για αναγνώριση μη συνεχούς νοηματικής γλώσσας, ανοίγοντας δρόμο για μελλοντική έρευνα.

- Πειραματιστήκαμε με μεθόδους εκτίμησης βάθους και στη συνέχεια αξιοποιήσαμε το πιο επιτυχές μοντέλο για να ενισχύσουμε το μοντέλο αναγνώρισης στο πρόβλημα της αναγνώρισης νοηματικής γλώσσας. Εκπαιδεύσαμε μοντέλα χρησιμοποιώντας μόνο

πληροφορία βάθους, συνδυαζόμενη πληροφορία βάθους και RGB εικόνας, και τέλος συνδυαζόμενης πληροφορίας βάθους και SMPL-X παραμέτρων.

Κλείνοντας, η διπλωματική αυτή ανοίγει μονοπάτια στο κόσμο της τρισδιάστατης ανακατασκευής και της αναγνώρισης νοηματικής γλώσσας. Το πρώτο μπορεί να αξιοποιηθεί με πολλούς τρόπους ώστε να βελτιώσει τις σημερινές μεθόδους που υπάρχουν για το δεύτερο, ενώ αποτελεί και ένα εξαιρετικό εργαλείο για άλλα προβλήματα, στις μέρες μας.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Computer Vision & Machine Learning

1.1.1 From Feature Extraction to Deep Learning

The main task of Computer Vision is to extract as much information possible from an image into a compact descriptor. These feature representations are then fed into a machine learning approach, like Support Vector Machines [18]. Until nowadays, Computer Vision was focused on finding sophisticated ways to extract such qualitative features from images. Most of the time features used to be some statistical properties or shape descriptors, or texture descriptors. For the method to be able to generalize successfully, features should be highly discriminative. Discriminative representation was enhanced by unsupervised learning such as Gaussian Mixture Models (GMMs), Principal Component Analysis (PCA), or manifold embedding. One of the most prominent methodologies that was exploited for a vast variety of computer vision tasks involved the detection of interest points over different scales and the extraction of local descriptors around these points [6, 50]. As local descriptors, gradient orientation was encoded into fixed-sized histograms. Each image was then represented by a set of different local descriptors. Classification applications required a supervised learning step, like SVMs, and a fixed-sized representation for the image. To this end, the set of local descriptors was organized into a histogram based on the Bag of Visual Words approach [45].

Until recently, meticulously crafting the feature extraction step, was the most important aspect of every Computer Vision task, while SVMs and Neural Networks were only employed at the very last step. Nowadays though, with the advance in computational capabilities of state-of-the-art hardware equipment, like GPUs, and the abundance of labeled data, neural network utilization has suddenly risen for every possible computer vision task during the last decade. For example, existing resources have enabled the efficient training of end-to-end deep neural networks, consisting from hundreds to million layers, like the very famous AlexNet [44] (2012). With the advance of the latter, the computer vision research community shifted towards using Deep Neural Networks (DNNs) as optimal feature extractors. In fact, the vast majority of the recent continuously expanding computer vision literature relies on DNNs for almost every possible sub-task. Since deep learning will also serve as the core component for a major portion of this thesis, we briefly analyze several important aspects of building and optimizing neural networks.

1.1.2 History of Neural Networks

McCulloch and Pitts were the first to introduce the concept of neural networks in 1943. The McCulloch-Pitts neuron is also known as linear threshold gate and consisted of a linear transform topped by a step activation function. In 1958, Rosenblatt proposed an enhancement over the McCulloch-Pitts Neuron, called the perceptron, introducing the idea of trainable weights along with an appropriate training algorithm for binary classification. Multiple neurons stacked to build a multi-layer feed-forward network were considered, in order to extend the neuron concept to classify non-linearly separable classes, like the famous XOR problem. Intermediate layers between input and output were referred to as hidden layers. The Perceptron and a multi-layer network example are depicted in Figures 1.1 (a) and (b), respectively.

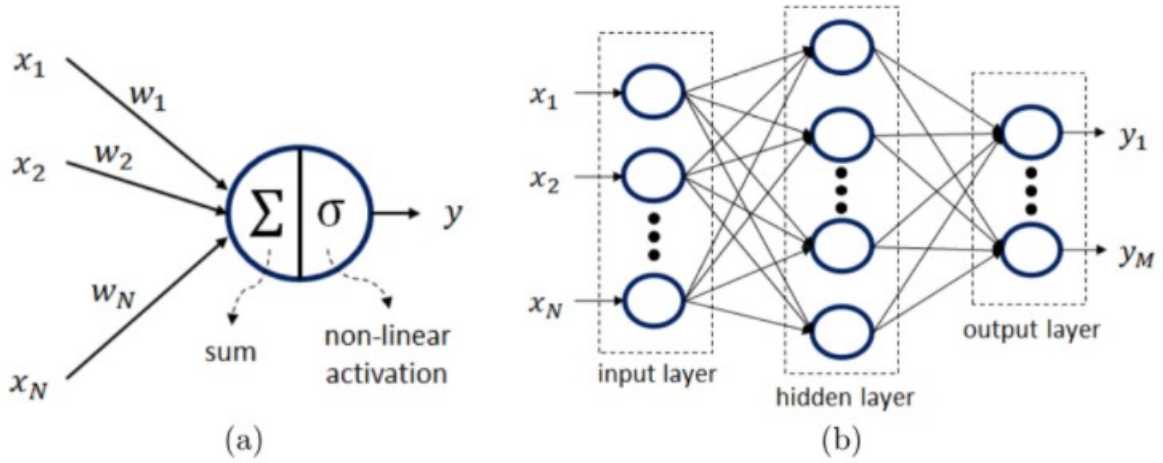

Figure 1.1: Visalization of (a) the Perceptron Neuron, consisted of a dot product and a non-linear activation, and (b) an example of multi-layer feed-forward neural network, consisted of multiple neurons and 3 layers: input, one hidden, and output.

Since the formulation of multi-layer feed-forward networks, leaps have been made towards complex structures of neurons, leading to two types of neural networks that will emerge multiple times in this thesis: Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs).

1.1.3 Convolutional Neural Networks

Since spatial context should be captured, image-related problems had always been challenging. One important factor aspect of neural networks is the introduction of Convolutional Neural Networks (CNNs), which can efficiently handle images and deduce spatial information from them. Before that, filtering was performed by convolution with a handcrafted kernel, designed to capture specific patterns like edges. Convolutional neural networks work by using trainable filters which can generate discriminative feature maps, optimized for each given task. CNNs revolutionized the computer vision field, pushing aside sub-optimal handcrafted features. Given $\mathbf{X}, \mathbf{Y}$, the input and output 3D tensors respectively, and $\mathbf{W}$ the 4D kernelled weight tensor, the convolutional layers perform the

operation $\mathbf{Y} = \mathbf{X} * \mathbf{W}$. which is defined as follows:

$$\mathbf{Y}[m] = \sum_{n=1}^{C_{in}} \mathbf{X}[n] * \mathbf{W}[m, n], m = 1, \ldots, C_{out} \tag{1.1}$$

$$\mathbf{Y} \in \mathbb{R}^{C_{out} \times H \times W}, \mathbf{X} \in \mathbb{R}^{C_{in} \times H \times W}, \mathbf{W} \in \mathbb{R}^{C_{in} \times C_{out} \times k_H \times k_W} \tag{1.2}$$

The spatial dimensions $H \times W$ and $k_h \times k_W$ correspond to the feature map and the kernel size, respectively, while C_in and C_out correspond to the number of 2D feature maps on the input and the output of convolution. Layers close to the input generate low-level features, such as edges, while layers close to the output generate high-level features of complex shape and texture. For example, the final layers of a CNN can generate a nose or eyes for face detection. Finally, stacking layers in successive order is not a necessary architectural requirement, while in fact, the majority of recent architectures contains complex information flows of multiple paths, like GoogleNet [70], ResNet [25], DenseNet [27] and so on.

1.1.4 Recurrent Neural Networks

Except from images, sequence modeling is a standout area of research with notable results in Speech Recognition, Sign Language Recognition, and Natural Language Processing. Since typical neural networks similar to CNNs cannot model a sequence of data, an alternation is needed. To this end, Recurrent Neural Networks (RNNs) were introduced. Consider the $\{\mathbf{x}_i\}$ segments that form the input sequence of an RNN and the $\{\mathbf{h}_i\}$ segments that form the hidden state. We have the following recurrent formulation:

$$\mathbf{h}_i = \sigma(f_h(\mathbf{h}_{i-1}) + f_x(\mathbf{x}_i)) \tag{1.3}$$

where $\sigma()$ is a nonlinear activation function and $f_h(), f_x()$ are the linear transformation functions for the hidden state and input respectively. Since these transformations are linear they can be formulated by a weight matrix and a bias. For instance, $f_h(\mathbf{x}) = \mathbf{W}_h\mathbf{x} + \mathbf{b}_h$. Each step of the recurrent formulation of Equation 1.3 shares the same weights. Hence, the unrolled version of RNN can be viewed as a typical neural network with shared weights. From the appearance of the first RNNs, more complicated variants have been constructed:

- Multi-layer Recurrent Neural Network: The output sequence of the first RNN is fed as an input sequence to the second RNN and so on. Each layer is described by the recurrent relation of Equation 1.3 and uses a pair of shared weight tensors $(\mathbf{W}_x^k, \mathbf{W}_h^k)$, where k is the layer's identifier.

- Bidirectional Recurrent Neural Network; Typical RNNs learn representations from previous time steps. However, often it is helpful to incorporate information from future steps. Bidirectional RNNs combine both information flows, rightward and leftward at each step.

- Long-Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) Networks; GRU Networks [16] contain gated recurrent unit which have an update gate that decides how much of information from the hidden state should be let through, and a reset gate that decides on how much information from the hidden state should be

discarded. Long Short Term Memory (LSTM) Networks [26] aim to retain information from distant time steps without vanishing it through time. Both units are preferred in the majority of recent sequence-related applications.

1.1.5 Categories of Layers

Convolutional Layer:

Neural networks, on their very basis, work on a simple idea; to alternate between linear transformation and non-linearities to build complex functions. The linear transformation can be:

- A fully connected layer which is a simple linear projection performed by a 2D tensor
- A convolutional layer which performs a convolution operation for every input/output channel pair resulting in a 4D tensor.

Activation Layer:

After every linear transformation, a non-linear function is following. Typical activation functions are tanh, sigmoid, ReLU, and others. These non-linearities contribute to the representational capabilities of neural networks.

Batch Normalization:

The concept of batch normalization [29] is to constraint the input and output of layers over a specific range of values. For that to happen, the running mean value is computed along with the standard deviation, updated at each batch by a momentum scheme. The input is then normalized to approximate a normal distribution of zero-mean and deviation of one. Batch normalization helps the convergence of the network's optimization by avoiding extreme values which may affect the gradients.

Dropout:

To solve one of the biggest problems that emerged through the use of neural networks, namely overfitting, data augmentation, and random noise were introduced. Specifically, since the network sometimes tends to learn exclusively the training dataset without the ability to generalize, a form of noise was introduced; the random zeroing of channels, otherwise known as dropout [69]. Dropout essentially assists the creation of multiple "paths" of information through different channels and avoids correlating a neuron with a specific input sample, thus enhancing generalization.

1.1.6 Optimization

Finally, we describe the way neural networks are trained. To train a neural network, with respect to the weights of the model, the following steps are required:

- Define a training set of input samples x_i and output targets y_i.
- Define a loss function $L(\hat{y}_i, y_i)$ which quantifies the proximity of the prediction $\hat{y}_i$ to the requested target y_i. Loss function plays a crucial role in the successful training of the neural network.

- Select an optimizing algorithm to perform $\theta^* = argmin_\theta \sum_i L(f_\theta(x_i), y_i)$, where θ is the set of all the weights comprising the network. No analytic solution exists for the aforementioned optimizing scheme, due to its complexity. Thus, iterative gradient-based algorithms are employed to gradually minimize the overall loss, like the gradient-descent method.

As said before, the loss function is critical for the effectiveness of the trained model. First of all, the loss function should reflect the task's goal. For instance, mean squared error (MSE) should be used for regression problems and Cross-Entropy (CE) for classification tasks. Losses can be complex and consist of multiple terms when considering multi-task or advanced problems. Secondly, the loss function should be differentiable since a gradient-based optimization scheme is used.

The gradients computation is performed layer-wise, starting from the loss function and moving backward to the input layer, when training a deep neural network. This method is called backpropagation. Computing the gradient of each parameter with respect to the loss function relies on the chain rule. Iterative gradient-based optimizations are summarized through computing gradients and updating weights. These two steps are performed iteratively until convergence is guaranteed. The convergence implies that the gradient would be almost zero, which translates to detecting an optimum. Nonetheless, this scheme cannot guarantee that this discovered optimum, is in fact, global.

After computing the gradient score of the objective function $J(theta)$ with respect to the parameters θ, the update of the weights is fundamentally done by the rule: $\theta \leftarrow \theta - \eta \times \nabla_\theta J(\theta)$, where η controls the convergence rate. The Stochastic Gradient Descent (SGD) optimization algorithm is used to eliminate the impractical method of calculating the gradients over the entire dataset. SGD performs a parameter update for each training sample, while an entire iteration of every dataset sample is referred to as the epoch.

1.2 Human Body

1.2.1 Body, Face and Hands

Body

The human body is the structure of a human being. It is composed of many different types of cells that together create tissues and subsequently organ systems. They ensure homeostasis and the viability of the human body. It comprises a head, neck, trunk (which includes the thorax and abdomen), arms and hands, legs, and feet.

The study of the human body involves anatomy, physiology, histology, and embryology. The body varies anatomically in known ways. Physiology focuses on the systems and organs of the human body and their functions. Many systems and mechanisms interact in order to maintain homeostasis, with safe levels of substances such as sugar and oxygen in the blood. The body is studied by health professionals, physiologists, anatomists, and by artists to assist them in their work.

Face

The face is the front of an animal's head that features three of the head's sense organs, the eyes, nose, and mouth, and through which animals express many of their emotions. The face is crucial for human identity, and damage such as scarring or developmental

Figure 1.2: Unveiled human body. Illustration of the main skeletal muscles constitutive of the human body in the anatomical reference posture. Around 600 muscles put in motion the various articulations composing the human skeleton. Image from https://en.wikipedia.org/wiki/Human_body.

deformities affects the psyche adversely. The front of the human head is called the face. It includes several distinct areas, of which the main features are:

- The forehead, comprising the skin beneath the hairline, bordered laterally by the temples and inferiorly by eyebrows and ears
- The eyes, sitting in the orbit and protected by eyelids and eyelashes
- The distinctive human nose shape, nostrils, and nasal septum
- The cheeks, covering the maxilla and mandibula (or jaw), the extremity of which is the chin
- The mouth, with the upper lip divided by the philtrum, sometimes revealing the teeth
- Facial appearance is vital for human recognition and communication. Facial muscles in humans allow the expression of emotions.

The face is itself a highly sensitive region of the human body and its expression may change when the brain is stimulated by any of the many human senses, such as touch, temperature, smell, taste, hearing, movement, hunger, or visual stimuli.

Faces are essential to expressing emotion, consciously or unconsciously. A frown denotes disapproval; a smile usually means someone is pleased. Being able to read emotion

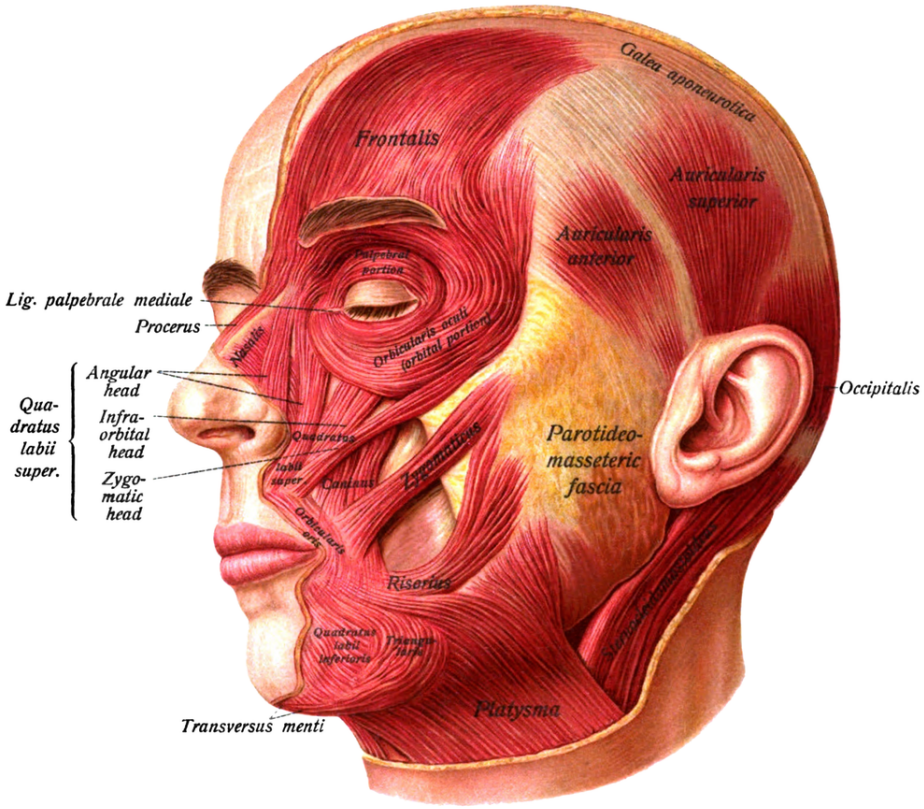

Figure 1.3: Ventrolateral aspect of the face with skin removed, showing muscles of the face. Image from https://en.wikipedia.org/wiki/Face.

in another's face is "the fundamental basis for empathy and the ability to interpret a person's reactions and predict the probability of ensuing behaviors". One study used the Multimodal Emotion Recognition Test to attempt to determine how to measure emotion. This research aimed at using a measuring device to accomplish what people do so easily every day: read emotion in a face. The muscles of the face play a prominent role in the expression of emotion, and vary among different individuals, giving rise to additional diversity in expression and facial features.

People are also relatively good at determining if a smile is real or fake. A recent study looked at individuals judging forced and genuine smiles. While young and elderly participants equally could tell the difference for smiling young people, the "older adult participants outperformed young adult participants in distinguishing between posed and spontaneous smiles". This suggests that with experience and age, we become more accurate at perceiving true emotions across various age groups.

Hands

A hand is a prehensile, multi-fingered appendage located at the end of the forearm or forelimb of primates such as humans, chimpanzees, monkeys, and lemurs. A few other vertebrates such as the koala (which has two opposable thumbs on each "hand" and fingerprints extremely similar to human fingerprints) are often described as having "hands" instead of paws on their front limbs. The raccoon is usually described as having "hands" though opposable thumbs are lacking.

The human hand normally has five digits: four fingers plus one thumb; these are often

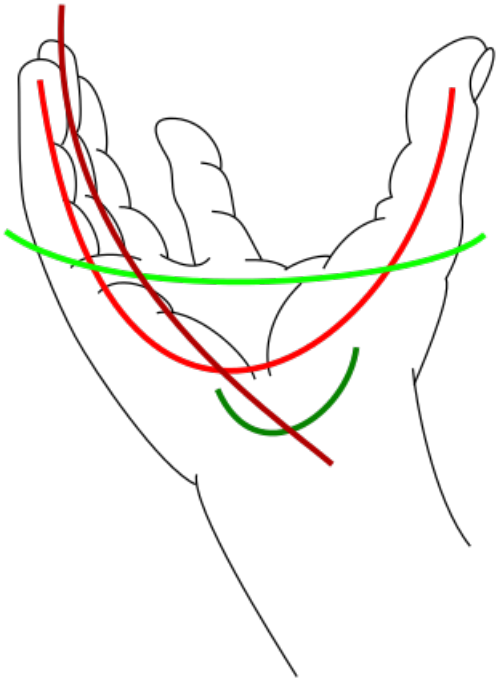

Figure 1.4: Arches of the hand. Red: one of the oblique arches. Brown: one of the longitudinal arches of the digits. Dark green: transverse carpal arch. Light green: transverse metacarpal arch. Image from https://en.wikipedia.org/wiki/Hand.

referred to collectively as five fingers, however, whereby the thumb is included as one of the fingers. It has 27 bones, not including the sesamoid bone, the number of which varies among people, 14 of which are the phalanges (proximal, intermediate, and distal) of the fingers and thumb. The metacarpal bones connect the fingers and the carpal bones of the wrist. Each human hand has five metacarpals and eight carpal bones.

Fingers contain some of the densest areas of nerve endings in the body and are the richest source of tactile feedback. They also have the greatest positioning capability of the body; thus, the sense of touch is intimately associated with hands. Like other paired organs (eyes, feet, legs) each hand is dominantly controlled by the opposing brain hemisphere, so that handedness—the preferred hand choice for single-handed activities such as writing with a pencil, reflects individual brain functioning.

Among humans, the hands play an important function in body language and sign language. Likewise, the ten digits of two hands and the twelve phalanges of four fingers (touchable by the thumb) have given rise to number systems and calculation techniques.

1.3 Sign Language

1.3.1 History

Sign languages have been around much longer than most people think. They existed in ancient Greece and even before recorded history. Next, we offer some perspective on how prolific sign language really is, by diving into the long and colorful history of how signs—and ASL in particular—came to be.

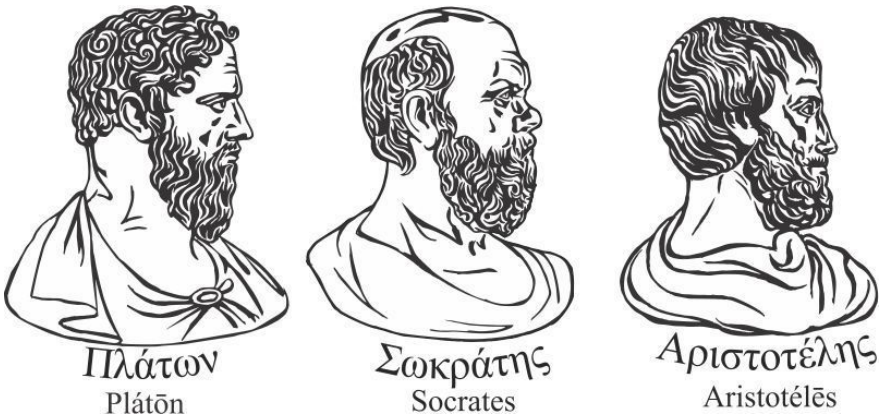

Figure 1.5: The Greek philosophers Plato, Socrates, and Aristotle were the first people in history to write about sign language and deaf members of their society. Image from https://aslblog.goreact.com/the-history-of-sign-language.

Earliest Sign Languages: No one knows exactly when sign language first appeared, but many sources agree that using hands to communicate has been around just as long as spoken language. And these early signing systems were the direct result of humans needing a new way to interact. Researchers believe that hunters on the open plains used signs to communicate with each other from great distances. Because of the lack of visual obstruction in a plains environment, the sign was the most obvious way to communicate without scaring off the animals they were hunting.

The ancient Great Plains Native Americans also developed a complex signing system. It's unclear what exactly the system was for, but many different theories exist. A popular one is that sign made intertribal trade possible. To overcome language barriers, the natives developed a standardized system of hand gestures to negotiate with tribes that did not speak their language—including European expeditioners. Multiple accounts of Columbus landing in the Americas claim that the natives communicated with his crew through sign.

Greek Philosophers: It is impossible to know exactly when and where the first deaf person tried out sign, but it is known that the first written record of sign language came from Ancient Greece. In the fifth century B.C., the philosopher Plato wrote the dialogue Cratylus. In it, he recorded Socrates saying, "If we had neither voice nor tongue, and yet wished to manifest things to one another, should we not, like those which are at present mute, endeavor to signify our meaning by our hands, head, and other parts of the body?" Apparently, ancient Greeks who could not speak did indeed have a rudimentary sign language to go about their daily lives.

Later Plato's student Aristotle became the first person ever to record a claim about deaf people—and unfortunately, it was not a good one. He believed that being able to hear speech was the only way people could learn. So according to Aristotle, it was completely impossible to educate deaf people. Even though there was not a shred of factual evidence to support his claim, Aristotle's theory caught hold and was widely believed for the next 2000 years throughout the world.

And the results were not pretty. During this era of history, deaf people were viewed as lesser humans who could not legally hold property. They could not get married because society was afraid that deafness was a hereditary trait that would be passed on to future generations. Deaf people were often denied citizenship and even religious rights. And though deafness was regarded as a shameful disability, any form of sign was ostracized and discouraged, making it nearly impossible for these people to communicate freely.

Figure 1.6: The first fingerspelling systems in history emerged in sixteenth-century Spain and Italy. Image from https://aslblog.goreact.com/the-history-of-sign-language.

Scholars of this period genuinely believed that deaf people could not learn, but some teachers still tried. In 685 A.D. the Archbishop of York, John Beverly, famously taught a deaf boy to speak. But instead of seeing this accomplishment as proof that Aristotle was wrong, thinkers of the era deemed this act as divine. The archbishop was later canonized for performing the miracle, but people still believed that the only way deafness could be "overcome" was to speak the same language as the general population.

Teachers in Italy and Spain: In the sixteenth century, philosophers and teachers finally started questioning Aristotle's claim that people who could not hear could not be educated. An Italian physician and mathematician named Girolamo Cardano (also known as Gerolamo or Geronimo) was the first voice to challenge Aristotle's long-standing assumption.

Cardano claimed that hearing was not necessary for a person to understand ideas and even started developing his own code of hand gestures. He believed that one could use written words matched with symbols of what they represented to communicate with deaf students. Although his code was never widely adopted, he did use his methods to teach his own deaf son. And Cardano's theories greatly influenced other leaders and thinkers of the time.

Around the same time as Cardano (about 1570), a Spanish monk named Pedro Ponce de Leon started educating his own deaf students—the sons of Spanish nobles. Because they were deaf, these young men were ineligible to inherit property. Leon taught them to read, write, and speak so they could claim the family fortunes that rightly belonged to them. And his efforts were successful.

Both Cardano and Leon inspired another Spanish monk named Juan Pablo de Bonet to take the biggest step in early sign language history. After developing his own methods of educating deaf pupils, Bonet published the first book on sign language in 1620. In it, he included his own manual alphabet of handshapes representing sounds. This was the first published system of fingerspelling in history.

Even though these early systems were designed to teach deaf people how to speak

Figure 1.7: By the 1700s, a standardized sign language—Old French Sign Language—already existed in Paris. L'Eppe added to this system at his school. Image from https://aslblog.goreact.com/the-history-of-sign-language.

other languages, Bonet's book was still a revolutionary landmark in the development of sign language as an officially recognized form of communication. His book sparked interest across Europe in educating deaf students, but it was not until the mid-1700s that the next groundbreaking achievement in sign language development took place.

French Sign Language Revolution: The French Deaf community already used a common sign language in Paris, one that had developed organically over centuries. L'Eppe added to this Old French Sign Language system by creating a series of hand signals to replace the sounds of the alphabet. As he taught the twins, l'Eppe uncovered a breakthrough in deaf education: that deaf people learn visually all the same things that other people learn by hearing. Deaf and mute people already had a language that was every bit as powerful and expressive as spoken French, and the key to educating them was training them to communicate with their hands instead of their voices.

In 1760 l'Eppe founded the first free public sign language school in the world, funded by his own inheritance. The school was called Institution Nationale des Sourds-Muets à Paris (The Royal Institution of Deaf-Mutes). As the French signing system and l'Eppe's methods of teaching continued to develop, deaf people from all over France flocked to his school. Even officials from other countries started to take notice. The emperor of Austria and the empress of Russia both sent teachers to learn l'Eppe's teaching style, and his influence eventually led to the creation of twenty-one schools total in France and many other countries.

Of course, l'Eppe wasn't the only influential sign language teacher of this time period. In England, Thomas Braidwood was establishing the Braidwood's Academy for the Deaf and Dumb around the exact same time that l'Eppe's school opened in France. Braidwood taught his pupils using a unique two-handed method of sign language, and he was pivotal in developing the same British Sign Language used today in the United Kingdom.

But not all the teachers of the time were accepting of signs. Samuel Heinicke started the first German school for the deaf in 1778, but unlike l'Eppe, Heinicke was a staunch oralist. He falsely believed that the primary function of education for deaf children should be to develop their spoken language skills so they could fully integrate into hearing society. This is the one area where l'Eppe's influence stood out among the other European teachers.

L'Eppe truly was the first "manualist" teacher, the first leader of deaf education who realized that sign language was the way deaf people should be communicating, and not just

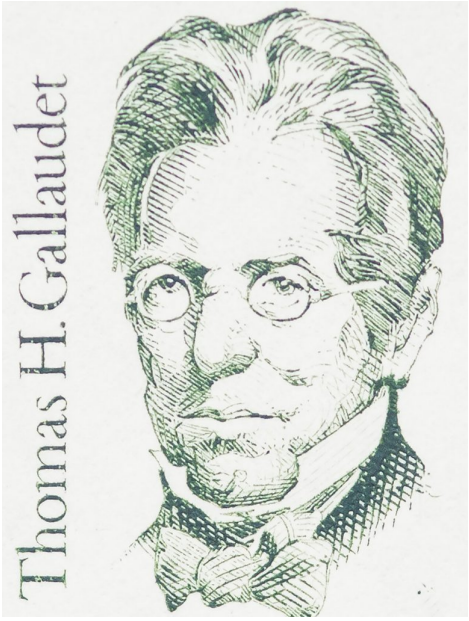

Figure 1.8: Thomas H. Gallaudet, the founder of the American School for the Deaf and the namesake of Gallaudet University. Image from https://aslblog.goreact.com/the-history-of-sign-language.

as a vehicle to help them speak oral languages. Aside from perpetuating the importance of sign, l'Eppe's unique background in theology and law also made him a valuable ally for deaf rights in both religion and the courtroom. He was one of the first people in history to publicly assert that deaf people deserved to be treated as fully functioning human beings with something meaningful to contribute to society, even if they spoke a different language. It's little wonder that today l'Eppe is known as the "Father of the Deaf."

The Great Gallaudet: Thomas Hopkins Gallaudet was a Yale graduate and an ordained clergyman in Hartford, Connecticut. He dreamed of becoming a professional minister, but his path took a different turn in 1814 when he met nine-year-old Alice Cogswell.

She was the deaf daughter of Gallaudet's neighbor Dr. Mason Fitch Cogswell. Gallaudet befriended Alice when he saw that the other children were not playing with her, and he began teaching her the names of objects by drawing pictures and words in the dirt. Right from the beginning of their friendship, Gallaudet was amazed by Alice's intelligence, personality, and enthusiasm to learn. He did not realize it at the time, but this relationship with this little girl was going to change Gallaudet's life forever—and the lives of millions of future deaf Americans too.

Dr. Cogswell was delighted to see his daughter's progress and convinced Gallaudet that he should learn more about educating deaf children. Perhaps even start a school. As a prominent member of Connecticut society, Dr. Cogswell used his connections to raise enough money to send Gallaudet to Europe to study established methods of deaf education. The funds were raised in just one afternoon, and soon Gallaudet was on a ship bound for England.

He hoped to be trained at one of the Braidwood schools for the deaf in England and Scotland, but the Braidwoods turned out to be far from welcoming. They were not in a hurry to give up their family sign and lip-reading methods without compensation. And Gallaudet wasn't convinced that their teaching methods were the best option for

educating deaf children anyway.

A discouraged Gallaudet parted ways with the Braidwoods, but shortly thereafter he met Abbe Roch-Ambroise Curcurron Sicard, l'Eppe's successor at the Paris school for the deaf. Sicard just happened to be visiting England during Gallaudet's trip and was giving lectures on deaf education along with two of his deaf assistants, Jean Massieu and Laurent Clerc. When Gallaudet introduced himself and explained his vision of establishing a school for the deaf in America, Sicard gladly invited him back to Paris to learn the French method of deaf education.

Gallaudet liked what he saw in Paris. He studied French sign with great enthusiasm, but he was quickly running out of money and needed to return home. Unsure if he could really start an American school all on his own, Gallaudet convinced the young Laurent Clerc to return with him to Hartford so they could start the school together. During the long sea voyage across the Atlantic, Gallaudet taught Clerc English and Clerc taught Gallaudet how to sign.

Sign Language today: Sign language is now recognized as the native communication and education method for Deaf people. No one knows exactly how many sign languages exist around the world today, but there are unique signing methods in just about every country on the globe. Sign language is now recognized as the native communication and education method for Deaf people. Many countries still do not have strong support for deaf education, and plenty still haven't recognized sign as an official language. But there's no doubt that sign has developed into a fully-fledged and beautiful language of its own right that has connected deaf people all around the world and impacted the lives of individuals everywhere.

1.3.2 Sign Language Alphabet

Memorizing the American Sign Language alphabet (also known as the American Manual Alphabet) is the first step when learning American Sign Language and most new sign language students rely on fingerspelling from the ASL alphabet when they don't know the sign for something.

Grammatically, fingerspelling is used in ASL for signing proper nouns (people's names, brand names, book and movie titles, and city and state names). So, it is recommended that sign language students don't fingerspell a word they don't know. Instead, we suggest trying to use signs you do know to describe the word or use gestures. If all else fails, though, go ahead and fingerspell it.

1.4 Contributions & Thesis Structure

This Diploma Thesis discusses the contemporary research field of 3D Computer Vision, namely the 3D Reconstruction of the human facial expression, body structure, and hand gesture. Moreover, this thesis investigates the complex task of Isolated and Continuous Sign Language Recognition, and how the former field can help. There is a plethora of contributions that this thesis offers. First of all, we offer a very detailed bibliographic analysis of the most contemporary and state-of-the-art 2D and 3D methods for the human reconstruction of the last 5 years. Next, a similar bibliographic analysis is followed for the most important sign language datasets and the state-of-the-art methods confronting the task of recognizing sign language. Moreover, during this thesis, we exploited the Greek Sign Language Lemmas Dataset for our experiments, which we re-organized and

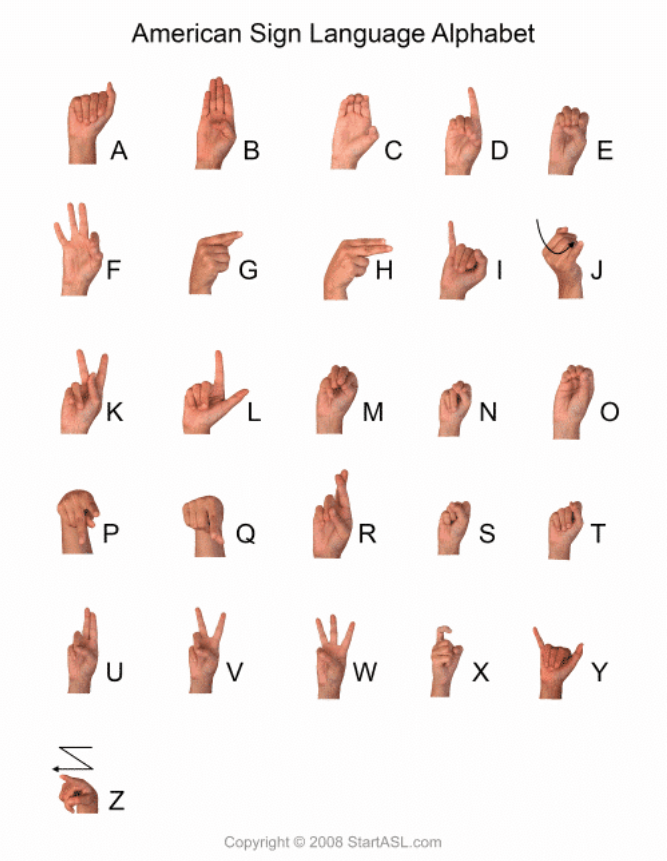

Figure 1.9: The American Sign Language Alphabet. Image from https://www.startasl.com/american-sign-language-alphabet.

made publicly available for further experimentation. Next, we applied 3D body, face, and hands reconstruction methods on the isolated SLR task, achieving top results and surpassing all other currently known methods. Furthermore, we conducted an ablation study, showing the importance of having all three channels of information; namely facial expression, hands shape, and body structure, for successfully recognizing Sign Language. We compared the two most recent 3D reconstruction methods, i.e. SMPL-X (CVPR 2019) [56] and ExPose (ECCV 2020) [15], in runtime and expressiveness, as well, while exploited the latter in further experiments in the MS-ASL [74] dataset. Finally, we experimented with depth estimation methods, as an additive channel to further increase accuracy to the aforementioned methods. To conclude, this thesis opens a path to the world of 3D Reconstruction and Sign Language Recognition. The former can be exploited in many ways to improve the current methods for the latter, while it is currently an exceptional tool for other tasks as well, nowadays, as it is highlighted in the Future Work section at the end of this thesis.

This thesis has the following structure. Section 1 offered an introduction to the topic of machine learning and computer vision. Moreover, it offers a brief discussion over the topic of the human body, as well as the history, alphabet, and vocabulary of the Sign Language.

Section 2 consists of an analytical bibliographical review of the state-of-the-art methods for 2D and 3D body, face, and hands reconstruction from 2015 onwards. Specifically, we present the most famous method for extracting the 2D skeleton body pose, hands, and facial characteristics of a human from a single RGB, namely Openpose. Next, we move to

the most important 3D models for parametrizing the human body (SMPL and SMPL-X) and describe them in detail, while we analytically present the methods for extracting such parameters from a single RGB frame (SMPLify, HMR, SMPLify-X, ExPose).

Section 3 offers a presentation of the most important datasets (MS-ASL, GSLL) for the task of the isolated sign language recognition, together with the state-of-the-art methods which manage to achieve top accuracy in this task. Furthermore, we present the most important datasets for continuous sign language recognition as well (PHOENIX-WEATHER, PHOENIX-WEATHER-TRANSLATION), and some of the most recent methods for confronting this problem.

In Section 4, we analytically present our methodology and experiments. In specific, we present the 3D tools we employ, as well as the pipeline which was used to deal with the problem of Sign Language Recognition. After presenting, the experimental setup, we proceed with the experimental results and an ablation study. Moreover, we discuss a comparison between two state-of-the-art 3D methods for body, face, and hands reconstruction, namely SMPLify-X and ExPose, as well as some experiments with the latter on the MS-ASL dataset. Finally, we present one more experiment that has to do with depth estimation of sign language videos, which is used as an enrichment piece to the 3D information granted from SMPL-X.

To conclude, Section 5 presents some future directions that can be followed as a continuation of this thesis and the main contributions of the thesis are being restated.

Chapter 2

Body Reconstruction

Body Reconstruction is the procedure of creating a parametric model that can accurately represent a wide variety of body shapes in natural human poses through a set of meaningful parameters. Hence, body reconstruction can be considered as a mapping between two spaces; $\mathbb{R}^n \rightarrow \mathbb{R}^m$, m $\ll$ n, where $\mathbb{R}^n$ is the space of a single RGB frame, while $\mathbb{R}^m$ is the space of a set of features which represent the human body. This procedure can be extended to cover the reconstruction of the hands and the facial characteristics, as well, while it should be noted that much more detail is needed to cover these parts, due to their smaller size compared to the body and the variety of expressions they can depict. The two main categories of body, face, and hands reconstruction are the 2D Reconstruction and the 3D Reconstruction.

2.1 2D Body Reconstruction

The main task in 2D body reconstruction is the pose estimation of a human in an RGB image. Specifically, the target is to locate a set of parameters or keypoints that can efficiently describe the human body, face, hands, and feet. In Computer Vision, 2D Body representation and reconstruction can be useful in numerous applications. In the next subsection, we present plausibly the most known and accurate method for real-time 2D Pose Estimation until now.

2.1.1 OpenPose

The proposed OpenPose method [11, 12, 66, 77], is probably the most acknowledged, accurate, and fast method for real-time Multi-Person 2D Pose Estimation until now. This real-time method is based on its bottom-up approach using Part Affinity Fields (PAFs) instead of the detection-based approach in other works. While it is the first bottom-up presentation of association score via PAFs, OpenPose achieves a running time that is invariant to the number of people visible in the image, rendering it a perfect method for real-time applications.

The overall procedure followed by OpenPose is shown in Figure 2.1. Given a single RGB image, the method feeds this image in a baseline VGG-like network to extract a feature map, which is given as input to a multi-stage CNN network, the architecture of which can be seen in Figure 2.2. The multi-stage CNN is responsible for generating a set of Part Confidence Maps and a set of Part Affinity Fields. Finally, a greedy algorithm combines the Confidence Maps and the Part Affinity Fields to obtain the 2D pose for

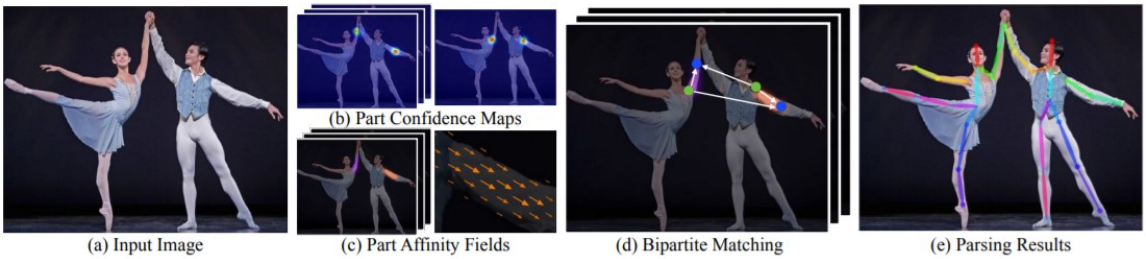

Figure 2.1: The overall presented pipeline. (a) OpenPose feeds the entire image to a CNN to jointly predict (b) confidence maps for body part detection and (c) PAFs for part association. (d) The parsing step performs a set of bipartite matchings to associate body part candidates. (e) The final assembly into full body pose for all people in the image. Image from [11].

each person in the RGB image. Confidence Map is a 2D representation of the belief that a particular body part can be located in any given pixel. So, each map corresponds to a joint and has the same size as the input RGB frame. A Part Affinity Field is a set of flow fields that encodes unstructured pairwise relationships between body parts. If, for example, a pixel is on a limb, then that pixel is represented by a 2D unit vector from the start joint to the end joint.

The first step of the multi-stage CNN is to computer the PAFs from the feature maps of the base network, namely $\mathbf{F}$. Let ϕ^1 be the CNN at the first stage of training. Then

$$L^1 = \phi^1(\mathbf{F}) \tag{2.1}$$

This procedure is repeated T_p times in order for the PAFs to be refined. Hence, if ϕ^t is the CNN at the stage t, and L^{t-1} the previous PAFs, then

$$L^t = \phi^t(\mathbf{F}, L^{t-1}), 2 \leq t \leq T_p \tag{2.2}$$

After T_p iterations, this process must be repeated for the confidence maps detection, given again the the baseline feature map $\mathbf{F}$ and the most updated PAFs prediction. This process is repeated for T_C iterations. Hence, letting ρ^t be the the CNN at the state t, then

$$S^{T_p} = \rho^t(\mathbf{F}, L^{T_p}), t = T_p \tag{2.3}$$

$$S^t = \rho^t(\mathbf{F}, L^{T_p}, S^{t-1}), T_p < t \leq T_p + T_C \tag{2.4}$$

The final Part Affinity Fields (PAFs) $\mathbf{L}$, and the confidence maps $\mathbf{S}$ are then processed by the greedy algorithm.

The parsing method contains three steps that can be described as follows: Step 1: Using the confidence maps, find all joints locations. Step 2: Using the part affinity fields and joints found in Step 1, find which joints go together to form limbs (body parts). Step 3: Associate limbs that belong to the same person to get the final list of human poses. A brief explanation of each step is followed for the completeness of the OpenPose algorithm.

Step 1: This step gets as input the confidence maps and the up-sampling scale, which is the difference in height and width between the initial RGB image and the confidence maps. The output of this step is a joints_list, which is a list of joint locations, where each item is a list of peaks $(x, y, probability)$. The algorithm follows the following procedure. For each joint, it gets the corresponding 2D heat-map for the joint in confidence maps and finds the peaks by thresholding the 2D heat-map. Next, for each peak, it takes the

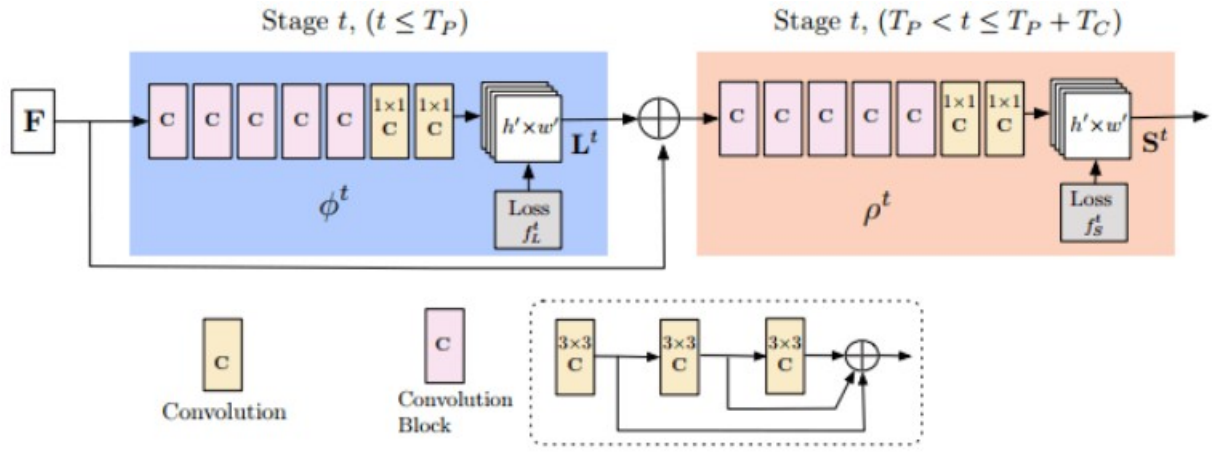

Figure 2.2: The proposed OpenPose architecture of the multi-stage CNN. The first set of stages predicts Part Affinity Fields ($\mathbf{L}^t$), while the last set predict Part Confidence Maps ($\mathbf{S}^t$). The corresponding image features are concatenated for each subsequent stage. Image from [11].

patch around the peak, and it scales it up according to the up-sampling scale. Afterward, it finds the maximum peak location and adds it to the list of peaks of the joint.

Step 2: This step gets as input the joints_list from the first step, the Part affinity fields (PAFs), the up-sampling scale, which is the difference in height and width between the input image and PAFs map, and finally the number of intermediate points between the source and destination joints. The output of this step is the connected_limbs, which is a list of connected limbs, where each item is a list of all limbs of that type found in the form (the id of the source joint, the id of the target joint, score of how good the connection is). The algorithm follows the following procedure. First, scale up the PAFs to the input size according to the up-sampling scale. Next, for each limb type i.e. right wrist elbow, get all the source and destination joint peaks, while if at least one of them is 0, skip this limb. Then, create a list to store all limb connection candidates and for each source peak and target peak, find the direction vector by subtracting the former from the latter. Normalize the vector into a unit vector, get PAFs values at each intermediate point and calculate the score of the current limb connection by averaging the PAFs values. Add the limb connections to the limb connection candidates while adding a score to penalize the long-distance limb. For each connection candidate, add the connection to the final list if the source and the destination are not selected for any connection.

Step 3: This step gets as input the joints_list from Step 1 and the connected_limbs from Step 2 and gives as output the poses, which is a list of human poses for each person in the RGB frame. Each item of this list contains the joint locations for that person. The algorithm follows the following procedure. For each limb type and for each connection of that type find the persons that are associated with either joint of the current connection. In case there is no person, create a new person with the current connection. If there is one person, then add the connection to that person. Finally, if there are two persons, merge these two persons into 1 and add the connection. Remove any person with very few joints.

This real-time Multi-Person 2D Pose Estimation method has numerous application from Action Recognition [51, 35], Robotic Visual Servoing [30] and Sign Language Recognition [74]. The OpenPose version that can jointly detect human body, hands, and facial keypoints on single images returns a total of 137 keypoints in the format of (x, y, p) where

Figure 2.3: OpenPose jointly detects human body, hands, facial and feet keypoints from a single RGB image containing viewpoint and appearance variation, occlusion, crowding, contact, and other common imaging artifacts. Image from [11].

(x, y) is the position of the keypoint and p the confidence level of the keypoint. In specific, OpenPose returns 25 keypoints for the 2D body pose, 21 keypoints for the 2D left hand, 21 keypoints for the 2D right hand, and 70 keypoints for the 2D face expression. Figure 2.3 shows results containing viewpoint and appearance variation, occlusion, crowding, contact, and other common imaging artifacts. Finally, Figure 2.4 shows an example frame with the produced 2D skeleton keypoints.

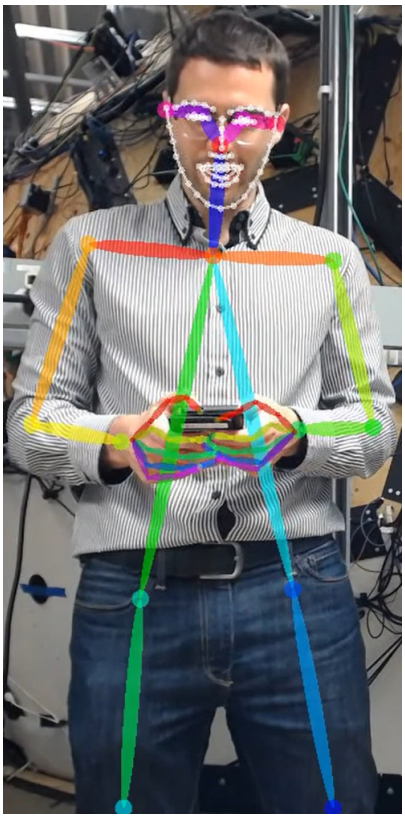

Figure 2.4: Example frame of the 2D skeleton produced by OpenPose showing the specific keypoints. Image from [11].

2.2 3D Reconstruction Methods

In computer vision and computer graphics, 3D reconstruction is the process of capturing the shape and appearance of real objects. One of the main goals of 3D Computer Vision, when it comes to 3D Body Reconstruction, is the creation of a realistic 3D model that captures the human body shape and pose dependent shape variation and accurately represents a wide variety of body shapes in natural human poses. Furthermore, a big task of 3D human reconstruction is to efficiently extract these realistic body models from single RGB images accurately and fast. In the next subsections, we present some of the most famous methods of 3D Body Modelling and 3D Body Reconstruction of the last 5 years.

2.2.1 SMPL & SMPL-ify

2.2.1.1 SMPL model

SMPL [48] is one of the first realistic 3D models of the human body that is based on skinning and blend shapes and is learned from thousands of 3D body scans. SMPL is a learned model of human body shape and pose-dependent shape variation that is more accurate than previous models and its compatible with existing graphics pipelines. This Skinned Multi-Person Linear model (SMPL) is a skinned vertex-based model that accurately represents a wide variety of body shapes in natural human poses. The parameters of the model are learned from data including the rest pose template, blend weights, pose-dependent blend shapes, identity-dependent blend shapes, and a regressor from vertices to joint locations. Unlike previous models, the pose-dependent blend shapes are a linear function of the elements of the pose rotation matrices. This simple formulation enables training the entire model from a relatively large number of aligned 3D meshes of different people in different poses.

2.2.1.2 SMPL-ify algorithm

SMPL-ify [7] is one of the very first methods for automatically estimating the 3D pose of the human body and its 3D shape from a single unconstrained image. To do so, the SMPL-ify method predicts the 2D body joint locations using DeepCut [59] and then fits the aforementioned statistical body shape model SMPL to these joints. This is done through the minimization of a sophisticated objective function which penalizes the error between the projected 3D model joints and the detected 2D ones. Qualitative results of the SMPL-ify method are shown in Figure 2.5. The SMPL-ify method can be divided into three main sub-tasks. First, the CNN-based DeepCut prediction of the 3D body joint locations takes place, while next, the body surface is approximated by a set of "capsules" where each one has a radius and an axis length. Finally, an objective function is carefully constructed for it to be minimized. These steps are briefly described below.

2D Body Joints and 3D Body modeling: A single RGB image is given as input to the DeepCut CNN to predict 2D body joints, J_{est}. The model provides for each joint i a confidence value, w_i. The body model, according to SMPL, is defined as a function $M(\beta, \theta, \gamma)$ where β, θ and γ stand for the shape, pose and translation respectively. The function M returns a triangulated surface with 6890 vertices. The shape parameters β are coefficients of low-dimensional shape space, learned from a training set of thousands of registered scans. The pose parameters θ represent the axis-angel representation of the

Figure 2.5: Two examples from the 3D pose and shape estimation using the SMPL-ify method. The original image is shown at the left, the fitted model at the middle and the 3D model rendered from a different viewpoint is shown at the right of the figure. Image from [7].

relative rotation between parts. If $J(\beta)$ denotes the function that predicts 3D skeleton joints locations from body shape, then those joints can be put in arbitrary poses by applying a global rigid transformation induced by pose θ denoted as $R_\theta(J(\beta)_i)$ for joint i.

Bodies Approximation with Capsules: One of the main challenges of 3D pose estimation is handling the interpenetration between body parts. Apparently, the SMPL model detects and prevents interpenetration which on the other hand, is extremely expensive to compute for non-convex and complex surfaces like the body. Using proxy geometries to compute collisions is a much more efficient method which is followed by the SMPL-ify system by approximating the body surface as a set of "capsules", whereas each has a specific radius and axis length. Two examples are shown in Figure 2.6

Objective Function and Optimization: The objective function used for minimization for the 3D pose and shape to be fitted to the CNN-detected 2D joints is the sum of five independent error terms;

$$E(\beta,\theta) = E_J(\beta,\theta;K,J_{est}) + \lambda_\theta E_\theta(\theta) + \lambda_\alpha E_\alpha(\theta) + \lambda_{sp} E_{sp}(\theta;\beta) + \lambda_\beta E_\beta(\beta) \tag{2.5}$$

where K are the camera parameters and λ_θ, λ_α, λ_{sp} and λ_β are scalar weights. The first error function, is called the joint-based data term which penalizes the weighted 2D distance between projected SMPL joints and estimated joints, J_{est}. Namely:

$$E_J\left(\beta,\theta;K,J_{est}\right) = \sum_{\text{joint } i} w_i \rho\left(\Pi_K\left(R_\theta\left(J(\beta)_i\right)\right) - J_{est,i}\right) \tag{2.6}$$

where Π_K is the projection from 3D to 2D induced by a camera with parameters K. A differentiable German-McClure penalty function [24] ρ is used for dealing with noisy estimates. Each joint i contributes by a weight w_i which is the confidence produced by DeepCut and when it comes for occluded joints, this value is usually low and the pose is driven by the pose priors. The next error function penalizes unnatural bending of elbows and knees and is given by the formula:

$$E_a(\theta) = \sum_i \exp\left(\theta_i\right) \tag{2.7}$$

where i sums over pose parameters (rotations). The exponential helps the strong penalization of positive bending which is unnatural, while negative and zero bendings ($\theta_i \leq 0$) are not heavily penalized. Next, given the significant variation of poses, it is of vital importance for the method, to represent the multi-modal nature of the data. This is done

Figure 2.6: Body shape approximation using capsules for two subjects. The original shape is shown at the left, the approximated shape with capsules is shown at the middle, while the capsules reposed are shown at the right. Yellow point clouds represent actual vertices of the model that is approximated. Image from [7].

by fitting SMPL to the CMU marker data, using MoSh [49] and then fitting a mixture of Gaussians to approximately 1 million poses. Thus

$$E_\theta(\theta) = -\log \sum_j \left(g_j \mathcal{N}\left(\theta; \mu_{\theta,j}, \Sigma_{\theta,j}\right)\right) \approx \min_j \left(-\log\left(c g_j \mathcal{N}\left(\theta; \mu_{\theta,j}, \Sigma_{\theta,j}\right)\right)\right) \tag{2.8}$$

where c is a positive constant and g_j are the mixture model weights of N=8 Gaussians. For the interpenetration error term, the capsule's volume are simplified into spheres with centers $C(\theta, \beta)$ along the capsule axis and radius $r(\beta)$ in order to relate it to the intersection volume between "incompatible" capsules. Considering a 3D isotropic Gaussian with $\sigma(\beta) = \frac{r(\beta)}{3}$ for each sphere, the penalty is defined as a scaled version of the integral of the product of Gaussians corresponding to "incompatible" parts:

$$E_{sp}(\theta; \beta) = \sum_i \sum_{j \in I(i)} \exp\left(\frac{\|C_i(\theta, \beta) - C_j(\theta, \beta)\|^2}{\sigma_i^2(\beta) + \sigma_j^2(\beta)}\right) \tag{2.9}$$

where the summation is over all the spheres$i$and I(i) are the spheres that are incompatible with i. Finally, a shape prion $E_\beta(\beta)$ is defined as

$$E_\beta(\beta) = \beta^T \Sigma_\beta^{-1} \beta \tag{2.10}$$

where Σ_β^{-1} is a diagonal matrix with the squared singular values estimated via the Principal Component Analysis from the shapes in the SMPL training set.

For the optimization, the camera translation, γ is initialized by assuming the person is standing parallel to the image plane. This estimation is further refined by minimizing E_J over the torso joints alone with respect to the camera translation and body orientation. The best technique for avoiding local minima is starting with high values of λ_θ and λ_β and gradually decreasing them. In case the person in the frame is captured in a side view, two initializations are attempted; one as described above and one with the orientation rotated by 180 degrees. The one with the lowest E_J is picked. The Equation 2.5 is minimized using Powell's dogleg method [52].

The optimization method takes almost one minute for a single RGB frame. SMPL and SMPL-ify gained great research attraction in the last few years, while many more contemporary methods tried to improve the technique, either when it comes to faster optimization and more detailed representation or when it comes to adding the 3D reconstruction of facial expression and hands. More examples of the SMPL and SMPL-ify method are shown in Figure 2.7.

Figure 2.7: Each sub-image shows the original image with the 2D joints fit by the CNN. To the right of that is the estimated 3D pose and shape and the model seen from another view. Image from [7].

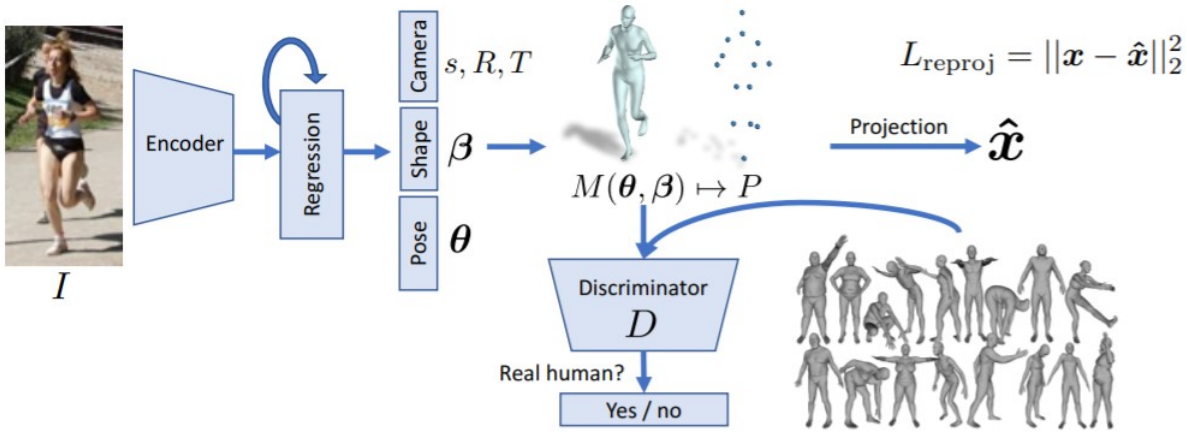

Figure 2.8: Overview of the proposed framework. A given RGB image I is passed through a convolutional encoder, features of which are given as input to an iterative 3D regression module that infers the latent 3D representation of the human by minimizing the joint reprojection error. A discriminator D is also exploited to tell if the 3D parameters come from a real human shape and pose. Image from [34].

2.2.2 Human Mesh Recovery (HMR)

While SMPLify offers a very qualitative and detailed reconstruction of the human body from a single RGB image, the optimization technique exploited takes up to a whole minute for each frame, rendering this method inappropriate for real-time applications. In 2018, Angjoo Kanazawa et al. [34] proposed an end-to-end method for the recovery of the human shape and pose named Human Mesh Recovery (HMR). This framework reconstructs a full 3D mesh from a single RGB image, running in real-time while showing competitive results compared to other reconstruction methods. Figure 2.8 shows the overview of the proposed framework.

This method used the Skinned Multi-Person Linear model (SMPL) which is described in Section 2.2.1.1 to encode the 3D mesh of the human body. SMPL encodes the 3D Body Representation through shape, which shows how each person varies in body proportions, height, and weight, and through pose, which shows how the 3D surface deforms with articulation. The shape β is given by the first 10 coefficients of a PCA shape space while the pose θ is given by relative 3D rotation of 23 joints in axis-angle representation. Moreover, the 3D keypoints used for reprojection error $X(\theta, \beta)$ are obtained by linear regression from the final mesh vertices. The weak-perspective camera model is employed and the solution yields the global rotation $R \in \mathbb{R}^{3\times3}$, the translation $t \in \mathbb{R}^2$, and the scale

Figure 2.9: Human Mesh Recovery (HMR) qualitative results using end-to-end adversarial learning of human pose and shape. The first two rows show results from the HMR model trained with some 2D-to-3D supervision, while the bottom row shows results from a model that is trained in a fully weakly-supervised manner without using any paired 2D-to-3D supervision. The full 3D body is inferred even in cases of occlusions and truncations. Head and limb orientations are captured as well. Image from [34].

$s \in \mathbb{R}$. Therefore, the 3D reconstruction of a human body is expressed through an 85 dimensional vector $\Theta = \{\theta, \beta, R, t, s\}$ and the projection of $X(\theta, \beta)$, given an orthographic projection Π is

$$\hat{\mathbf{x}} = s\Pi(RX(\theta, \beta)) + t \tag{2.11}$$

The iterative 3D Regression with feedback is responsible for producing Θ given an image encoding ϕ while aiming to minimize the joint reprojection error

$$L_{\text{reproj}} = \Sigma_i \left\| v_i \left(\mathbf{x}_i - \hat{\mathbf{x}}_i \right) \right\|_1 \tag{2.12}$$

where v_i is the visibility for each of the K joints (1 if visible, 0 if not), $\hat{\mathbf{x}}$ as defined in Equation 2.11 and $\mathbf{x_i} \in \mathbb{R}^{2\times K}$ is the ith ground truth 2D joints. In order to successfully regress θ through the 3D Regression, the authors regress θ in an iterative error feedback loop, where progressive changes are made recurrently to the current estimate. Specifically, given the image features ϕ and the current parameters Θ_t the regression module produces the residual $\Delta\theta_t$, and then the current estimate Θ_{t+1} is updated by adding the residual to the current estimate. The mean $\bar{\Theta}$ is chosen as an initial estimate Θ_0.

To eliminate the chance of an anthropometrically implausible 3D body or a body with gross self-intersections to minimize the reprojection loss, a discriminator network D is exploited to tell whether SMPL parameters correspond to a real body or not. This is mentioned as an adversarial prior since D acts as a data-driver prior that guides the 3D inference. Figure 2.9 shows qualitative results for the HMR method.

2.2.3 SMPL-X & SMPLify-X

A much more detailed and qualitative approach appeared in 2019 from G. Pavlakos and V. Choutas et al. [56] who not only improved the SMPL model by proposing a new, unified, 3D model of the human body SMPL-X but improved as well the SMPL-ify

approach for reconstructing 3D Hands, Face, and Body from a single monocular image. SMPL-X, through thousands of 3D scans, is trained to fit fully articulated hands and an expressive face. Furthermore, SMPlify-X is an improved version of the SMPlify method in a way that 2D features for face, hands, and feet are used to fit the full SMPL-X model, a new more accurate, and faster interpenetration penalty is defined and a new neural network pose prior is being trained.

Unified model SMPL-X: SMPL-X uses vertex-based linear blend skinning with learned corrective blend shape while having $N = 10,475$ vertices and $K = 54$ joints, including joints for the neck, jaw, eyeballs, and fingers. To better facilitate hands and face, the pose parameters θ are decomposed into θ_f for the jaw joint, θ_h for the finger joints, and θ_b for the remaining body joints. The shape parameters for body, face, and hands are noted as usual, with β, while the facial expression parameters with ψ. Hence:

$$M(\beta, \theta, \psi) = W(T_p(\beta, \theta, \psi), J(\beta), \theta, \mathcal{W}) \tag{2.13}$$

$$T_p(\beta, \theta, \psi) = \bar{(T)} + B_S(\beta; \mathcal{S}) + B_E(\psi, \mathcal{E}) + B_P(\theta; \mathcal{P}) \tag{2.14}$$

In the aforementioned equation, $B_S(\beta; \mathcal{S}) = \sum_{n=1}^{|\beta|} \beta_n \mathcal{S}_n$ is the shape blend shape function given the linear shape coefficients β and the orthonormal principle components of vertex displacements capturing shape variations due to different person identity $\mathcal{S}_n$. Next, $B_P(\theta; \mathcal{P}) = \sum_{n=1}^{9K} \left(R_n(\theta) - R_n\left(\theta^*\right)\right) \mathcal{P}_n$ is the pose blend shape function which adds corrective vertex displacements to the template mesh $\bar{T}$ given a mapping function $\mathcal{R}$ from the pose vector θ to a vector of concatenated part-relative rotation matrices, the n^{th} element of $\mathcal{R}(\theta), \mathcal{R}_n(\theta)$, the pose vector of the rest pose θ^* and the orthonormal principle components of vertex displacements $\mathcal{P}_n$. Finally, $B_E(\psi, \mathcal{E}) = \sum_{n=1}^{|\psi|} \psi_n \mathcal{E}$ is the expression blend shape function, given the principle components capturing variations $\mathcal{E}$ and the PCA coefficients ψ. The 3D joint locations J vary between differenct shapes and are a function of body shape according to $J(\beta) = \mathcal{J}(\bar{T} + B_S(\beta; \mathcal{S}))$, where $\mathcal{J}$ is a sparse linear regressor of 3D joint locations from mesh vertices. A linear blend skinning function W rotates the vertices in T_p, around the joints $J(\beta)$ smoothed by blend weights $\mathcal{W}$. The template is fitted into four datasets of 3D human scans to estimate the shape $\{\mathcal{S}\}$ and body pose $\{\mathcal{W}, \mathcal{P}, \mathcal{J}\}$ space parameters, while the hand and face parameters are leveraged from MANO [64] and FLAME [47] which have learned the pose space and pose corrective blendshapes for the hands through 1500 hand scans, and the expression space $\mathcal{E}$ for the face from 3800 head scans respectively.

SMPLify-X: Similarly to the SMPL-ify method described in Section 2.2.1.2, to fit SMPL-X to a single RGB image the authors solve an optimization problem by minimizing the following function

$$E(\beta, \theta, \psi) = E_J + \lambda_{\theta_b} E_{\theta_b} + \lambda_{\theta_f} E_{\theta_f} + \lambda_{m_h} E_{m_h} + \lambda_a E_a + \lambda_\beta E_\beta + \lambda_\mathcal{E} E_\mathcal{E} + \lambda_\mathcal{C} E_\mathcal{C} \tag{2.15}$$

The data term E_J a re-projection loss is exploited for minimizing the weighted robust distance between the estimated 2D joints, J_{est}, and the 2D projection of the corresponding posed 3D SMPL-X joints $R_\theta(J(\beta))_i$ for each joint i, and is given by

$$E_J(\beta, \theta\, K, J_{est}) = \sum_{\text{joint } i} \gamma_i \omega_i \rho \left(\Pi_K \left(R_\theta(J(\beta))_i\right) - J_{\text{est},i}\right) \tag{2.16}$$

given the the 3D to 2D projection with intrinsic camera parameters K, Π_K. It is important to mention that the 2D detection not only for the body but for the hands, face and feet

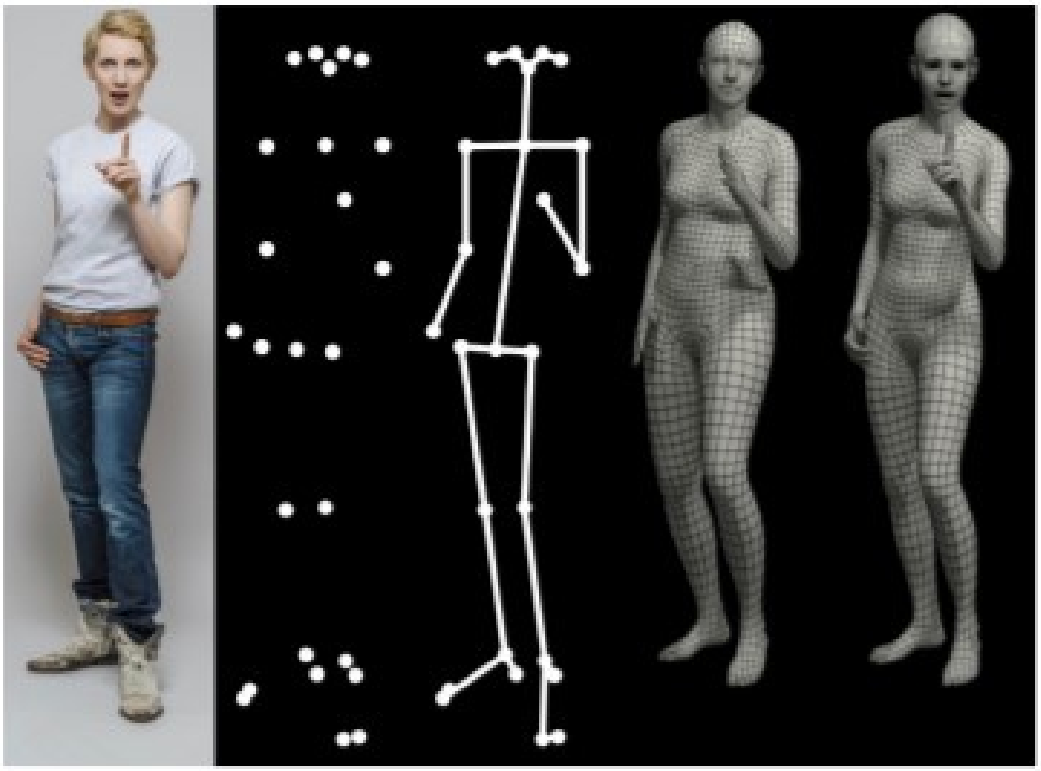

Figure 2.10: SMPL-X model and SMPLify-X method: The major joints of the body are not sufficient to represent body pose, hand pose, and facial expression, all together. This approach estimates a detailed and expressive 3D model from a single RGB image. From left to right; RGB image, major joints, 2D skeleton (using OpenPose), SMPL (female) and SMPL-X (female). Image from [56].

keypoints as well, are made with OpenPose which is described in Section 2.1.1. The terms $E_{m_h}(m_h), E_{\theta_f}(\theta_f), E_\beta(\beta) and E_\mathcal{E}(\psi)$ are L_2 priors for the hand pose, facial pose, body shape and facial expressions, penalizing deviation for the neutral state. The $E_a(\theta_b) = \sum_{i \in (elbows, knees)} exp(\theta_i)$, similar to SMPLify in Equation 2.7, while $E_{\theta_\beta}(\theta_\beta)$ is a VAE-based body pose prior and $E_C(\theta_{b,h,f}, \beta)$ is an interpenetration penalty. Finally, λ denotes optimization weights while an annealing scheme is followed.

Collision penalizer: While SMPL-ify penalizes penetrations through a collision model which is based on an ensemble of capsules, this is only an approximation of the human body. Since SMPL-X models the fingers and the face as well, a more detailed collision model is required. For that, a list of colliding triangles $\mathcal{C}$ are detected by employing Bounding Volume Hierarchies (BVH) [71] and local conic 3D distance fields Ψ defined by the triangles $\mathcal{C}$ and their normals n are computed. Then the penetrations can be penalized by the depth of the intrusion, computed by the position in the distance field. Given two colliding triangles f_s and f_t, the intrusion is bi-directional; the vertices v_t of f_t are the intruders in the distance field of Ψ_{f_s} of the receiver triangle f_s and are penalized by $\Psi_{f_s}(v_t)$ and vice-versa. Hence, the collision term E_C is defined as

$$E_\mathcal{C}(\theta) = \sum_{(f_s(\theta), f_t(\theta)) \in \mathcal{C}} \left\{ \sum_{v_s \in f_s} \| -\Psi_{f_t}(v_s)\, n_s \|^2 + \sum_{v_t \in f_t} \| -\Psi_{f_s}(v_t)\, n_t \|^2 \right\} \tag{2.17}$$

Some qualitative results of the SMPL-ify method and the SMPL-X model can be seen in Figure 2.11.

Figure 2.11: Qualitative results of SMPL-X for in-the-wild images. Gray color depicts the gender-specific model for confident gender detections, while the blue color depicts the gender-neutral model that is used when the gender classifier is uncertain. Image from [56].

2.2.4 ExPose

While the SMPL-X model seems to be the most detailed and accurate model for representing the human body and SMPify-X the most efficient algorithm for extracting SMPL-X features from a single RGB image, the time constraint sets a major drawback for real-time applications. As mentioned earlier, the SMPLify-X method requires approximately a minute per image due to the minimization of the Equation 2.15. In ECCV 2020, Vasileios Choutas, Georgios Pavlakos et al. [15] proposed ExPose, which stands for EXporessive POse and Shape rEgression. In contrast to SMPL-X that not only is slow due to the optimization-based method but also requires 2D keypoints as input, ExPose directly regresses the body, face, and hands in SMPL-X format, from an RGB image in almost real-time. While the HMR method was fast due to its regression technique, the estimation of hands and face was poor due to the downscaling caused by the neural network. ExPose exploits body-driven attention for these regions to extract higher-resolution crops from the original image while feeding them to dedicated refinement modules with part-specific knowledge from existing face- and hand-only datasets.

3D Body Representation: This work chooses to represent the human body through SMPL-X, which is described in detail in the previous section. The expression parameters $\beta \in \mathbb{R}^{10}$ and the expression parameters $\psi \in \mathbb{R}^{10}$ are described by 10 coefficients from the corresponding PCA spaces. The pose vector $\theta \in \mathbb{R}^{J \times D}$ models the articulation of the limbs, the hands, and the face, where D is the rotation representation, here chosen as 6, which describes the relative rotation of the $J = 53$ major joints, including 22 main body joints, 1 for the jaw and 15 joints per hand for the finger. Hence, the posed joints are denoted with $X(\theta, \beta)$ and the final set of SMPL-X parameters is the vector $\Theta = \{\beta, \theta, \psi\} \in \mathbb{R}^{338}$.

Body-driven Attention: Let I be the full resolution image and $T_b \in \mathbb{R}^{2\times3}$ an affine transformation used for extracting a bounding box of the body I_b. Then, the body crop I_b is fed into a neural network g similar to the HMR technique to produce the parameters Θ_b, the camera scale s_b, and the translation t_b. The recovered posed joints X are projected

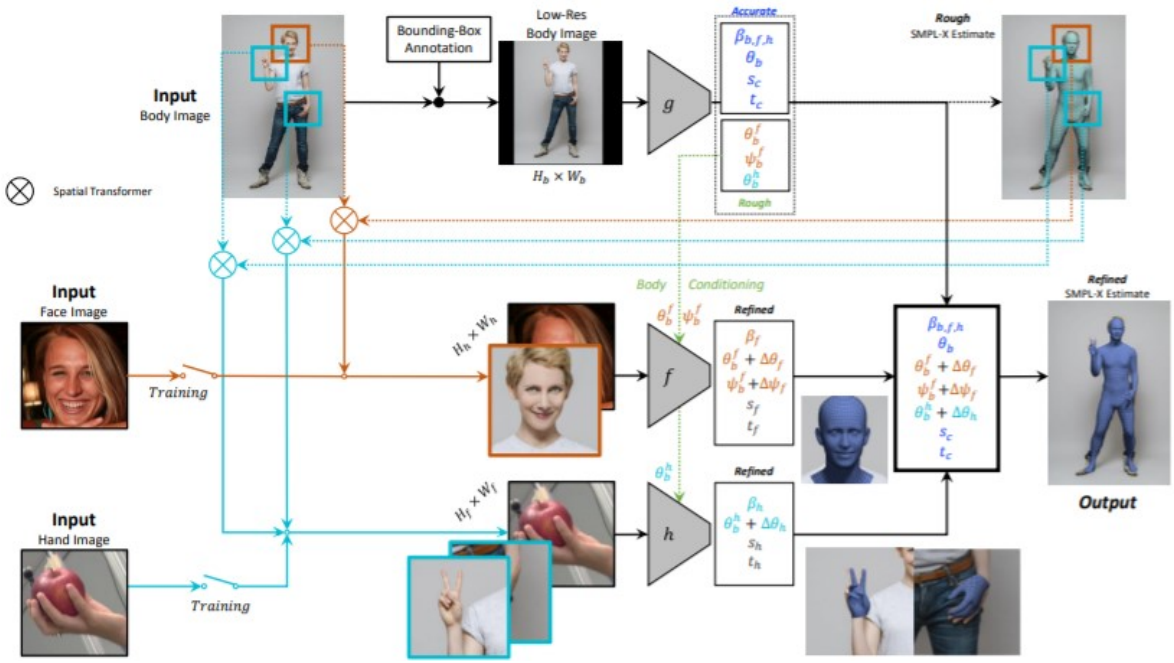

Figure 2.12: The body-driven attention method proposed by ExPose. A body image is extracted using a bounding box and fed to a neural network $g(\cdot)$, that predict body pose θ_b, hand pose $theta_h$, facial pose $theta_f$, shape β, expression ψ, camera scale s and translation t. The face and hands are extracted from the original resolution image using bilinear interpolation and then are fed to part specific sub-networks $f(\cdot)$ and $h(\cdot)$ to produce the final estimates. The part specific networks receive hand and face only data for extra supervision. Image from [15].

on the image $x = s(\Pi(X) + t)$, given an orthographic projection Π. Next, new affine transformations $T_h, T_f \in \mathbb{R}^{2\times3}$ are used to extract higher resolution hand and faces images using spatial transformers (ST): [31]

$$I_h = ST(I; T_h), I_f = ST(I; T_f) \tag{2.18}$$

Similar to I_b, the hand and face images are fed to a hand network h and a face network f, to refine the respective parameter predictions. The hand parameters θ_h include the orientation of the wrist θ^{wrist} and finger articulation $\theta^{fingers}$. The face parameters contain the expression coefficients ψ_f and the facial pose θ_f. The refinement of the parameters of the body network is done by predicting offsets for each of the parameters and conditioning the part specific networks on the corresponding body parameters:

$$\left[\Delta\theta^{\text{wrist}}, \Delta\theta^{\text{fingers}}\right] = h\left(I_h; \theta_b^{\text{wrist}}, \theta_b^{\text{fingers}}\right), [\Delta\theta_{\text{f}}, \Delta\psi] = f\left(I_f; \theta_b^{\text{f}}, \psi_b\right) \tag{2.19}$$

where $\theta_b^{wrist}, \theta_b^{fingers}, \theta_b^f, \psi_b$ are the wrist pose, finger pose, facial pose and expression predicted by $g(\cdot)$ respectively. The predicted 3D meshes are aligned to their respective images I_h and I_f through a set of weak-perspective camera parameters $\{s_h, t_h\}$ and $\{s_f, t_f\}$produced by the hand and head sub-networks. The final predictions are equal to:

$$\boldsymbol{\theta}_h = \left[\boldsymbol{\theta}^{\text{wrist}}, \boldsymbol{\theta}^{\text{fingers}}\right] = \left[\boldsymbol{\theta}_b^{\text{wrist}}, \boldsymbol{\theta}_b^{\text{fingers}}\right] + \left[\Delta\boldsymbol{\theta}_{\text{wrist}}, \Delta\boldsymbol{\theta}_{\text{fingers}}\right] \tag{2.20}$$

$$[\boldsymbol{\psi}, \boldsymbol{\theta}_f] = \left[\boldsymbol{\psi}_b, \boldsymbol{\theta}_b^f\right] + [\Delta\boldsymbol{\psi}, \Delta\boldsymbol{\theta}_f] \tag{2.21}$$

Figure 2.13: Qualitative results of the ExPose method. The raw RGB image is shown on the left. The naive regression from a single body image is shown in the middle, which fails to capture detailed finger articulation and facial expressions. The ExPose results are shown at the right. Note that due to the attention mechanism, ExPose is able to recover details and produce results of similar quality as SMPLify-X, while being 200 times faster. Image from [15].

Thus, the full resolution of the image is being utilized while the network has also the ability to leverage hand- and face-only information to supplement the training of the hand and face sub-network. Figure 2.12 shows the prediction module in detail.

Loss function: To train the model, the authors combine a body, face, and hands loss function, namely:

$$L = L_{body} + L_{hands} + L_{face} + L_h + L_f \tag{2.22}$$

The body network is trained using a 2D re-projection loss, 3D joints loss and a loss of the parameters Θ. Hence, $L_body = L_{reproj} + L_{3Djoints} + L_{SMPL-X}$ where:

$$L_{\text{3D Joints}} + L_{\text{SMPL-X}} = \sum_{j=1}^{J} \left\| \hat{\boldsymbol{X}}_j - \boldsymbol{X}_j \right\|_1 + \|\{\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\psi}}\} - \{\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}\}\|_2^2 \tag{2.23}$$

$$L_{\text{reproj}} = \sum_{j=1}^{J} v_j \left\| \hat{\boldsymbol{x}}_j - \boldsymbol{x}_j \right\|_1 \tag{2.24}$$

where hat terms denote ground-truth quantities and u_j is a binary variable denoting visibility of each of the J joints. For the hand and head only data a re-projection loss is

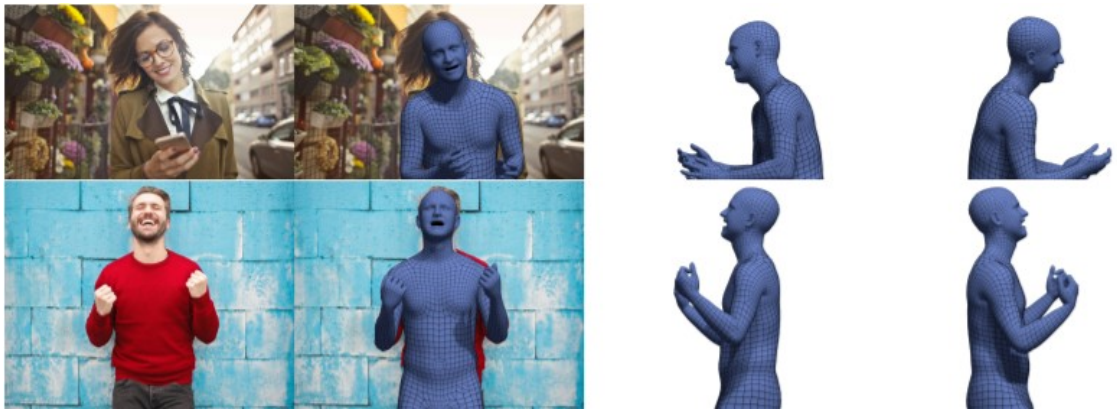

Figure 2.14: Further qualitative results for the ExPose method. The image shows the input image, the ExPose prediction overlayed on the image and rendering from different viewpoints. Image from [15].

employed, using only the subset of joints of each part, and parameter losses:

$$L_{\text{hand}} = L_{\text{reproj}} + \left\| \left\{ \hat{\boldsymbol{\beta}}_h, \hat{\boldsymbol{\theta}}_h \right\} - \{ \boldsymbol{\beta}_h, \boldsymbol{\theta}_h \} \right\|_2^2 \tag{2.25}$$

$$L_{\text{face}} = L_{\text{reproj}} + \left\| \left\{ \hat{\boldsymbol{\beta}}_f, \hat{\boldsymbol{\theta}}_f, \hat{\boldsymbol{\psi}}_f \right\} - \{ \boldsymbol{\beta}_f, \boldsymbol{\theta}_f, \boldsymbol{\psi}_f \} \right\|_2^2 \tag{2.26}$$

Due to the fact that fingers and facial landmarks have a much smaller magnitude compared to those of body joints, an extra penalty is used for these. A 2D re-projection loss L_h and L_f is applied in the hand and face image coordinate space using the affine transformation T_h and T_f respectively.

$$L_h = \sum_{j \in \text{ Hand}} v_j \left\| T_h T_b^{-1} \left(\hat{\boldsymbol{x}}_j - \boldsymbol{x}_j \right) \right\|_1, L_f = \sum_{j \in \text{Face}} v_j \left\| T_f T_b^{-1} \left(\hat{\boldsymbol{x}}_j - \boldsymbol{x}_j \right) \right\|_1 \tag{2.27}$$

Figures 2.13 and 2.14 show qualitative results of the ExPose method.

Chapter 3

Sign Language Recognition

With the term "Sign Language" we refer to a language that employs signs made with the hands and other movements, including facial expressions and postures of the body, used primarily by people who are deaf. Sign Languages are languages that use visual-manual modality to convey meaning and are expressed through manual articulations in combination with non-manual elements. Sign Languages do have their own grammar and lexicon, they are not universal and they are not mutually intelligible with each other. Humans are able, due to their natural ability, to identify continuous and isolated sign language after they have been trained to identify and understand it. Unfortunately, that is not the case with computers since Sign Language Recognition (SLR) is considered a very hard task due to the need of combining information from three different channels; face, body and hands. Sign Language Recognition is mainly divided into two main tasks; isolated sign language recognition and continuous sign language recognition.

Sign languages are natural languages communicable purely by vision via sequences of time-varying 3D shapes. They serve for communication in the Deaf communities, as well as among deaf and hearing people if the latter learn to sign. They also serve as inspiration and/or models for building sets of gestures for human-computer communication or interaction. They convey information and meaning via spatio-temporal visual patterns, which are formed by manual (handshapes) and non-manual cues (facial expressions and upper body motion). Computer-based processing and recognition of sign videos is also broadly related to vision-based human-computer and human-robot interaction using gesture recognition.

While significant progress exists in the field of automatic sign language recognition from the computer vision and pattern recognition fields, e.g. see [75, 54, 3, 72, 53] and the references therein, it still remains a quite challenging task especially for continuous sign language. In addition to signs having a complex multi-cue 4D space-time structure, the difficulty in their automatic recognition is also due to the large variability with respect to inter-signer or intra-signer variations of signing while expressing the same concept-word. Due to the above variability, instead of recognizing each sign as a whole 'visual word', a more efficient approach (inspired by speech recognition) is to decompose signs into *subunits*, resembling the phonemes of speech, and recognize them as a specific sequence of subunits by using some statistical model, e.g. via Hidden Markov Models (HMMs).

Clearly, the subunits approach performs much better on large vocabularies and continuous language; further, the subunits are reusable and help with signer adaptation. In lack of a lexicon, a computational technique to find such subunits is *data-driven*, i.e. perform unsupervised clustering on a large database and use the cluster centroids as subunits. This performs well in several instances, especially when the subunits are pre-classified

and statistically modeled based on visual features into dynamic vs. static, as done in [72]. A further improved performance accompanied with phonetic interpretability may be obtained if the chosen subunits are also based on the phonetic structure of a sign, as for example by incorporating the Posture-Detention-Transition-Steady Shift (PDTS) system [32] of phonetic labels. In [60] the phonetic information provided by the PDTS transcriptions of sign videos was combined with the automatically extracted visual features to create statistically trained *phonetic subunits* and a corresponding lexicon, which were then used for optimally aligning (via Viterbi decoding) the data with the phonetic labels and hence providing the missing temporal segmentation, as well as better sign recognition. Vogler and Metaxis in [75] present a novel framework to ASL recognition that aspires to being a solution to the scalability problems. It is based on breaking down the signs into their phonemes and modeling them with parallel hidden Markov models. Finally, another paper that exploits hidden Markov models for spatiotemporal inputs in the sign language recognition problem is [13]. The proposed approach deals with temporal and spatial aspects of the spatiotemporal domain in a discriminative as well as coupling manner. Self Organizing Maps (SOM) model the spatial aspect of the problem and Markov models its temporal counterpart. Incorporation of adjacency, both in training and classification, enhances the overall architecture with robustness and adaptability.

While information and meaning in sign languages are mainly conveyed by moving handshapes, they are also conveyed in part by non-manual cues such as facial expressions. These expressions can be visually modeled by deformable models that encode both geometric shape and brightness texture information. Deformable masks provided by active appearance models (AAMs) [17] can successfully help with detecting and tracking several types of informative events in frames from a sign sequence, e.g. eye blinking, as done in [5]. AAMs [65] have also significantly boosted the performance of handshape recognition in sign language videos.

With the advancement of deep neural networks, much progress has been made in independent and continuous SLR, with the use of CNN and LSTM networks. Koller et al. in [40] used a pretrained GoogleNet CNN architecture followed by 2 Bidirectional LSTM layers to achieve, the currently minimum, 26.8% word error rate in the RWTH-PHOENIX-Weather 2014 continuous sign language dataset [38]. In [74], Joze and Koller have experimented with different deep learning methods in independent SLR, like the I3D [14] that consists of a plethora of Conv3D layers and inception modules, or the hierarchical co-occurrence network (HCN) [46] for body key-points. An integral component of our approach is the use of a recently introduced parametric body model, SMPL-X [56] that can jointly model the body, the hands and the face of the person. With the exception of Adam [33], this is the only available model that can jointly capture these three channels of information. Previous statistical models, focus only the body (e.g., SCAPE [4] and SMPL [48]), or add hands, but still miss the facial expression (e.g., SMPL+H [64]), which is crucial for the task of sign language recognition. Conveniently, SMPL-X is also accompanied by a method that allows us to reconstruct the model parameters for a person from a single image. The method is called SMPLify-X and is based on the SMPLify approach by Bogo et al. [7].

3.1 Isolated Sign Language Recognition

3.1.1 Introduction

The term isolated sign language recognition is used to describe the task of recognizing a single sign which is not in the context of a complete sentence. More specifically, the computer's task is to identify a specific sign among a set of signs, which is depicted by a signer in a single video. Many datasets have been created for this task and a variety of methods have been exploited to increase the accuracy. In the next subsections we present some of the main isolated sign language datasets along with their methods for achieving high accuracy.

3.1.2 The MS-ASL Dataset

One of the most famous datasets for isolated sign language recognition is the MS-ASL dataset proposed by V. Joze and O. Koller in BMVC 2019 [74]. They propose the first real-life large-scale sign language data set comprising over 25,000 annotated videos, which they thoroughly evaluate with state-of-the-art methods from sign and related action recognition. Unlike the current state-of-the-art, the data set allows to investigate the generalization to unseen individuals (signer-independent test) in a realistic setting with over 200 signers. Previous work mostly deals with limited vocabulary tasks, while in this paper, the authors cover a large class count of 1000 signs in challenging and unconstrained real-life recording conditions. They further propose I3D, known from video classifications, as a powerful and suitable architecture for sign language recognition, outperforming the current state-of-the-art by a large margin. The data set is publicly available to the community. Some characteristics of the 4 proposed subsets of the MS-ASL data set is shown in the next Table. Moreover, Figure 3.1 illustrates a histogram of the duration of the 25,513 video samples of signs after the manual touch-up. To highlight the diversity of this dataset, Figure 3.2 shows a set of frames that represent the exact same sign "nice".

			Number of Videos			Duration	Videos per class		
Data set	Class	Subjects	Train	Validation	Test	Total	[hours:min]	Min	Mean
ASL100	100	189	3789	1190	757	5736	5 : 33	47	57.4
ASL200	200	196	6319	2041	1359	9719	9 : 31	34	48.6
ASL500	500	222	11401	3702	2720	17823	17 : 19	20	35.6
ASL1000	1000	222	16054	5287	4172	25513	24 : 39	11	25.5

Next, we present the state-of-the-art methods used to solve the problem of isolated sign language recognition. Isolated sign language recognition can be considered similar to action recognition or gesture detection as it is a video classification task for a human being. Three main categories or combinations of them can be considered to confront this challenging task.

- Exploit the RGB image using 2D convolution on it and do a recurrent network on top of that.
- Extract body joints in the form of skeleton for the signer, using 2D reconstruction methods.
- Using 3D convolution or 3D reconstruction features.

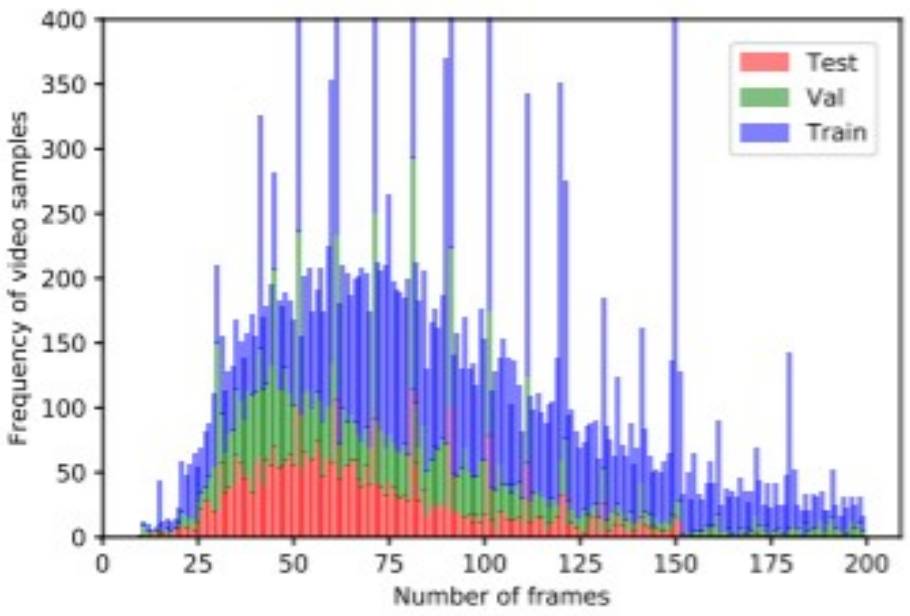

Figure 3.1: Histogram of frame numbers for ASL1000 video samples. Image from [74].

2D-CNN: Extraction of features is achieved from each frame of the video independently using 2D convolutional layers. Next LSTM layer [26] was used on the top of 2D convolutional networks, which records the temporal ordering and long range dependencies by encoding the states. VGG16 [67] network was used as the convolutional network and a single LSTM layer of size 256 with batch normalization was used as the recurrent network. This method is refereed to as **VGG-LSTM**. Another famous method implemented for continuous sign language by O. Koller et al. in CVPR 2017 [40] was tested in this dataset which used GoogleNets [70] as the 2D-CNN followed by 2 bi-directional LSTM layers and 3-state HMM. This method is reported as [40] and it will be further discussed later on in this Chapter.

Body Keypoints: Extracting 2D skeleton keypoints of the signer is another method that can be exploited for isolated sign language recognition. A work that covers hand and face keypoints along with the classical skeleton [66] is used for these experiments. 137 keypoint are extracted in total, where is keypoints is in the form of $(x, y, confidence)$. Hence, since each video contains 64 frames, a neural network with input of $64 \times 137 \times 3$ is needed. Figure 3.3 illustrates the extracted 137 body keypoints for a set of frames from a video sample of MS-ASL. The network exploied in this scenario is the hierarchical co-occurrence network (HCN) [46] which originally used 15 joints. The input of the extended network is 137 body keypoints as well as per frame difference of them. The network included three layers of 2D convolution on top of each input as well as two extra 2D convolution layers after the concatenation of two paths. The architecture can be seen in Figure 3.4. This method is refereed to as **HCN**.

3D-CNN: In the last few years, 3D convolutional networks have shown promising perfromance in action recognition tasks including two of the most famous models, C3D [73] and I3D [14] networks. While C3D did not converge for any of the experiments performed, the I3D network was trained successfully. It contains several 3D convolutional layers followed by 3D max-pooling layers and inflated Inception-V1 submodules. The architecture can be seen in Figure 3.5. This experiment is referred to as **I3D**.

The aforementioned methods were trained on four different MS-ASL subsets (ASL100,

Figure 3.2: Characteristic frames from the MS-ASL dataset depicting the exact same sign "clean" / "nice".

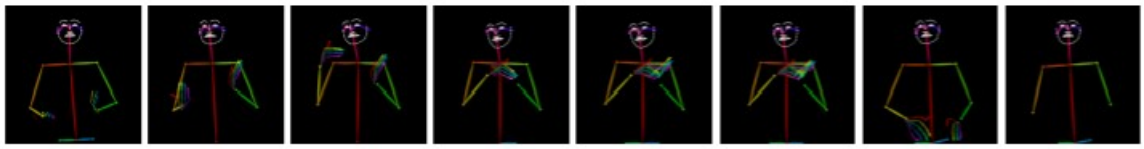

Figure 3.3: Extracted 137 body keypoints for a video sample from MS-ASL. Image from [74].

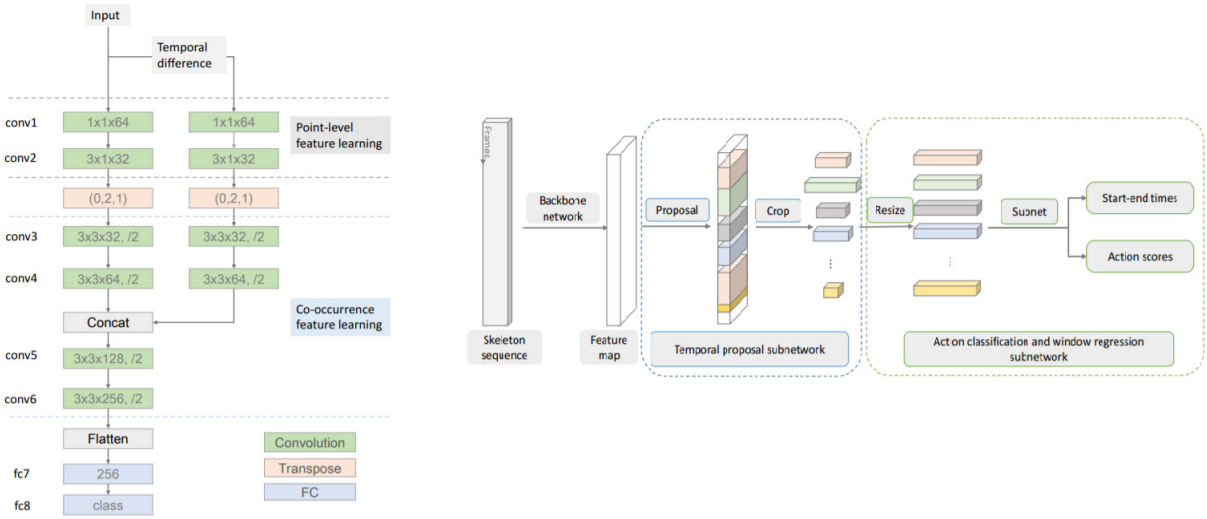

Figure 3.4: Overview of the proposed hierarchical co-occurrence network: The temporal action detection framework. The backbone network is described in the left. Two subnetworks are designed for temporal proposal segmentation and action classification respectively. Image from [46].

ASL200, ASL500, AS1000). Table **??** reports the results for the average per class accuracy. The experimental results suggest that this dataset is very difficult for 2D-CNN network or that one LSTM layer can not propagate the recurrent information well. Re-Sign [40] method which report as state-of-the-art in some continuous-sign-language datasets, does not manage to achieve top result in the challenging MS-ASL dataset. Next, the body keypoints based approach with the HCN network is doing relatively better compared to 2D-CNN but there is room for improvement due to the network's simplicity and the simplicity in the keypoint extraction. Finally, the state-of-the-art method for action and gesture recognition, for the last few years, seems to perform equally good in the task of sign language recognition as well.

3.1.3 The Greek Sign Language Lemmas Dataset (GSLL)

The Greek Sign Language Lemmas Dataset [43, 72], or GSLL Dataset for short, is an isolated sign language dataset, the development of which was supported by the EU research project Dicta-Sign. The GSLL dataset contains 347 different signs/classes signed by two signers; Kostas and Olga (male and female). Each sign is repeated from 5 to 17 times. These 347 classes are recorded through a total of 3,464 videos containing 161,050 frames. Four examples frames are shown in Figure 3.6. Moreover, Table 3.1 shows some statistics for the dataset and its respective subsets. The indicative suggested splitting in train, dev and test set was used in the experiments of [43].

Some preliminary experiments were conducted using this dataset by A. Kratimenos

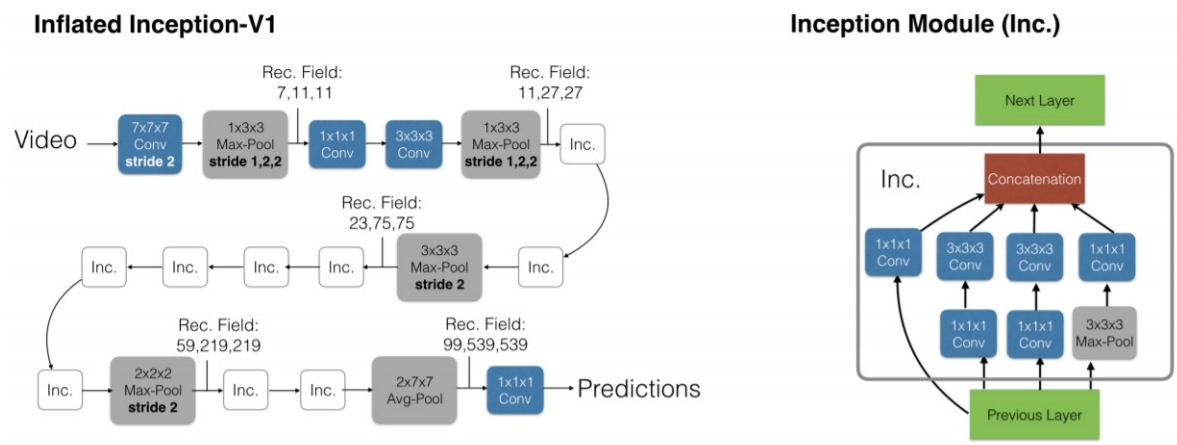

Figure 3.5: Overview of the proposed I3D network: The Two-Stream Inflated 3D ConvNet (I3D) that is based on 2D ConvNet inflation: filters and pooling kernels of very deep image classification ConvNets are expanded into 3D, making it possible to learn seamless spatio-temporal feature extractors from video while leveraging successful ImageNet architecture designs and even their parameters. Image from [14].

GSLL Subset	Videos	Frames	TrainSet	DevSet	TestSet
50 classes	538	22808	318	106	114
100 classes	1038	45437	618	206	214
200 classes	2038	92599	1218	406	414
300 classes	3038	140771	1818	606	614
347 classes	3464	161050	2066	695	703

Table 3.1: Statistics for the Greek Sign Language Lemmas Dataset and its respective subsets. Indicative suggested splitting in train, dev and test set used in the experiments of [43].

et al. in ECCVW 2020 [42]. We present the training methodology and the results that led to further experimentation in [43]. The results of [43] will be discussed analytically in Chapter 4.

Training Methodology: We do not intervene on the length of each feature sequence, resulting in various lengths from 10 to 300 frames per sign. Next, we present the methods with which we confront this problem. **Raw Image**: We reshape each frame in a 175×175 array and normalize its pixels to $[0, 1]$. We feed our images' sequence in a Conv3D-LSTM model, the structure of which is similar to [41], alongside with a VGG16-LSTM model which is initialized with Imagenet weights and is followed by a Global Average 2D Pooling layer. **Openpose**: We extract 411 parameters for each frame and feed the sequence in an RNN consisting of one Bi-LSTM layer of 256 units and a Dense layer for classifying, after applying standard scaling to our features. **SMPL-X**: Due to SMPL-X ability to interpret the 3D structure of the body in detail, we strongly believe that this method will provide key features for action and sign recognition. In this case, we extract 88 features per frame and follow the same procedure as with the Openpose.

According to Table 3.2, Openpose and SMPL-X models, which consist of 1.4 and 0.7 million parameters respectively, outperform the Conv3D-LSTM and VGG16-LSTM model, which consist of 43 and 15 million parameters respectively. This can be attributed to the fact that the former two eliminate the redundant information from each frame, keeping only the essential key-points. VGG16, in specific, fails to converge and reduce its

Figure 3.6: Characteristic frames the Greek Sign Language Lemmas Dataset from both signers.

Method \ GSLL Subset	Subset 50	Subset 100	Subset 200	Subset 300
Raw Image	88.59%	84.58%	71.98%	55.37%
Openpose features	96.49%	94.39%	93.24%	91.86%
SMPL-X features	**96.52%**	**95.87%**	**95.41%**	**95.28%**

Table 3.2: Comparison of the three representations for sign classification: i) Raw RGB images ii) Openpose 2D skeleton keypoints and iii) SMPL-X parameters.

loss, achieving an accuracy below 10% for all classes. This does not come as a surprise to us since Joze and Koller in [74] have trained a VGG16-LSTM model for the MS-ASL dataset which achieved 13.33% for the ASL100 Subset and just 1.47% for the ASL500 Subset. GSLL Dataset is characterised by a very uniform environment between each sign and each signer (only two signers in front of a blue cloth). The MS-ASL dataset for instance, consists of 222 distinct signers where each signer performs in a completely altered environment. We strongly believe that Openpose and mainly SMPL-X will by far outperform convolutional models in these datasets, which simulate more accurately the real world. Finally, SMPL-X seems to outperform the features produced by Openpose especially with the increase of different signs, dictating that a more detailed and qualitative representation of the human body is needed for the Sign Language Recognition task. While varying and more complex signs are being added to the train set, Openpose fails to convey the small details that differentiate these signs, while SMPL-X holds its accuracy almost fixed.

Figure 3.7: Examples frames from the GSL dataset. Image from [1].

Model	Results
GoogLeNet+TConvs [21]	86.03%
3D-ResNet [61]	86.23%
I3D [14]	**89.74%**

Table 3.3: Results in the GSL isolated dataset using three state-of-the-art methods for sign language recognition.

3.1.4 The Greek Sign Language (GSL) Dataset

The Greek Sign Language (GSL) [1] is a large-scale RGB+D dataset, suitable for Sign Language Recognition (SLR) and Sign Language Translation (SLT). The video captures are conducted using an Intel RealSense D435 RGB+D camera at a rate of 30 fps. Both the RGB and the depth streams are acquired in the same spatial resolution of 848×480 pixels. To increase variability in the videos, the camera position and orientation is slightly altered within subsequent recordings. Seven different signers are employed to perform 5 individual and commonly met scenarios in different public services. The average length of each scenario is twenty sentences.

The dataset contains 10,290 sentence instances, 40,785 gloss instances, 310 unique glosses (vocabulary size) and 331 unique sentences, with 4.23 glosses per sentence on average. Each signer is asked to perform the pre-defined dialogues five consecutive times. In all cases, the simulation considers a deaf person communicating with a single public service employee. The involved signer performs the sequence of glosses of both agents in the discussion. For the annotation of each gloss sequence, GSL linguistic experts are involved. The given annotations are at individual gloss and gloss sequence level. A translation of the gloss sentences to spoken Greek is also provided. Figure 3.7 shows some key frames from the GSL dataset.

This dataset contains continuous and isolated information for both tasks in sign language recognition. For the isolated part, the validation set consists of 2,231 gloss instances, the test set 3,500, while the remaining 34,995 are used for training. All 310 unique glosses are seen in the training set.

In table 3.3, quantitative results are reported for the isolated setup. Classification accuracy is reported in percentage. It can be seen that 3D baseline methods achieve higher gloss recognition rate than 2D ones. I3D+BLSTM clearly outperforms other architectures in this setup, by a margin of 2.2%. I3D+BLSTM and 3D-ResNet+BLSTM

were pretrained on Kinetics, which explains their superiority in performance as they contain motion priors. The 3D CNN models achieve satisfactory results in datasets created under laboratory conditions, like GSL, yet in challenging scenarios, I3D+BLSTM clearly outperforms 3D-ResNet+BLSTM.

3.2 Continuous Sign Language Recognition

3.2.1 Introduction

The term **continuous sign language recognition** is used to describe the task of recognizing a complete sentence signed by a signer in a single video. The task of continuous SLR is indeed a more complete and harder problem than the isolated SLR similar to the fact that it is harder to detect speech instead of individual words in the NLP field. Many methods have been exploited to solve this problem, including the ones in the isolated task, while HMM's, Expectation Maximization [40], LSTM's and most recently transformers [9] have been used to cover the continuity aspect of the problem. Many datasets are available for Continuous Sign Language Recognition while the most famous are the RWTH-PHOENIX-Weather 2014 Dataset [38], the RWTH-PHOENIX-Weather 2014-T Dataset [10] and the SIGNUM Dataset [76]. In the next subsections we present some of the main continuous sign language datasets along with their methods for achieving high accuracy.

3.2.2 RWTH-PHOENIX-Weather 2014

Over a period of three years (2009 - 2011) the daily news and weather forecast airings of the German public tv-station PHOENIX featuring sign language interpretation have been recorded. Currently, only the weather forecasts of a subset of 386 editions have been transcribed using gloss notation. The transcriptions have been carried out by deaf and hard-of-hearing native speakers of German sign language. Additionally, the spoken German weather forecast has been transcribed in a semi-automatic fashion using the RASR speech recognition system. Moreover, an additional translation of the glosses into spoken German has been created to capture allowable translation variability.

The signing is recorded by a stationary color camera placed in front of the sign language interpreters. Interpreters wear dark clothes in front of an artificial grey background with color transition. All recorded videos are at 25 frames per second and the size of the frames is 210 by 260 pixels. Each frame shows the interpreter box only.

The RWTH-PHOENIX-Weather 2014 Dataset contains a total of 5672 train, 540 validation and 629 test videos in Continuous Sign Language Recognition performed by 9 different signers in a smooth and uniform background. The state-of-the-art results are shown in Table 3.4 below, in Word Error Rate (the lower the better). Some example figures are shown in 3.8

Next, we briefly describe one of the state-of-the-art methods for the RWTX-PHOENIX-Weather 2014 dataset, namely the one by O. Koller et al. in CVPR 2017 [40]. The authors proposed a pretrained CNN - 2 Bi-directional LSTM network followed by a hybrid HMM, while proposing an iterative re-alignment approach using the Expectation - Maximization algorithm, to deal with the weak annotated nature of the continuous sign language recognition. Specifically, this work presents an iterative re-alignment approach applicable to visual sequence labelling tasks such as gesture recognition, activity recognition and

Author	WER Dev (%)	WER Test (%)
Koller, Ney and Bowden CVPR 2016 [39]	47.1	45.1
Koller, Zargaran, Ney and Bowden, BMVC 2016 [53]	38.3	38.8
Camgoz, Hadfield, Koller and Bowden, ICCV 2017 [8]	40.8	40.7
Cui, Liu and Zhang, CVPR 2017 [20]	39.4	38.7
Huan, Zhou, Zhang, Li and Li, AAAI 2018 [28]	–	38.3
Koller, Zargaran and Ney, CVPR 2017 [40]	27.1	26.8
Koller, Camgoz, Ney, and Bowden, TPAMI 2019 [36]	**26.0**	**26.0**

Table 3.4: Published State of the Art Continuous Sign Language Recognition Results on RWTH-PHOENIX-Weather 2014 Multisigner (in Word Error Rate, the lower the better)

continuous sign language recognition. Instead of relying to frame-to-frame labeling, that in most of the cases is not available or they are noisy, the authors propose an algorithm that treats the provided training labels as weak labels and refines the label-to-image alignment on-the-fly n a weakly supervised fashion. Next, using the series of frames and sequence-level labels, a depp recurrent CNN-BLSTM network is exploited for training end-to-end. The resulting deep neural network is embedded into an HMM, which corrects the frame labels and continuously improves the performance in several alignments. The whole end-to-end architecture is shown in Figure 3.9 (left).

The basic idea of the iterative re-alignment algorithm relies on Expectation Maximisation (EM) [22]. The algorithm is initialised with a provided frame labelling or a frame-state-alignment generated by standard CNN training. Then, iteratively, a maximisation step is performed, which corresponds to fitting the CNN-LSTM model to the data and then an expectation step is performed, in which the previously trained model is embedded in a hybrid HMM recognition. As depicted in Figure 3.9 (right), after each successful re-alignment the following iteration of CNN-LSTM training benefits from the new frame-state labels and it also uses the previous iteration's model weight for initialization.

The best reported results on the RWTH-PHOENIX-Weather 2014 dataset is reported by O. Koller et al. in TPAMI 2019 [36] and will be described in the next subsection since it reports the best results for that dataset as well.

3.2.3 RWTH-PHOENIX-Weather 2014 Translation

The RWTH-PHOENIX-Weather 2014 T provides spoken language translations and gloss level annotations for German Sign Language videos of weather broadcasts. The dataset contains over 0.95 million frames with more than 67,000 signs from a sign vocabulary of more than 1,000 and 99,000 words from a German vocabulary of more than 2,800. Similar to the RWTH-PHOENIX-Weather 2014 dataset, the signing in the translation dataset is recorded by a stationary color camera placed in front of the sign language interpreters. Interpreters wear dark clothes in front of an artificial grey background with color transition. All recorded videos are at 25 frames per second and the size of the frames is 210 by 260 pixels. Each frame shows the interpreter box only.

While this dataset is mostly used for the task of Sign Language Translation, which this thesis does not cover, some works have used this dataset to test their methods for recognition as well. Table 3.5 shows some of the state-of-the-art results in this task when it comes to the recognition task. Next, we present the two state-of-the-art methods proposed for the task of continuous sign language recognition in the RWTH-PHOENIX-Weather 2014 Translation dataset.

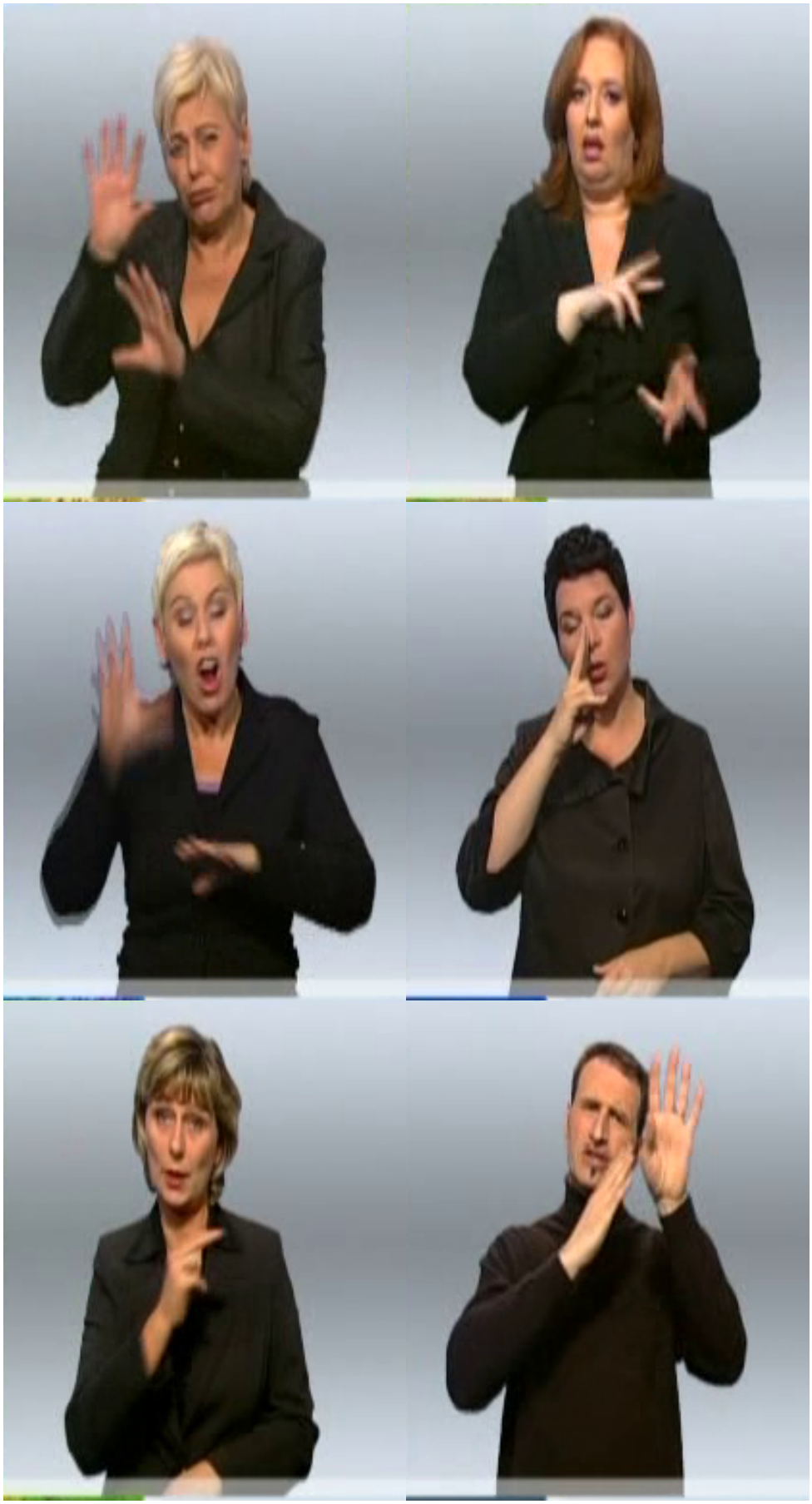

Figure 3.8: Characteristic frames from the RWTH-PHOENIX-Weather 2014 dataset.

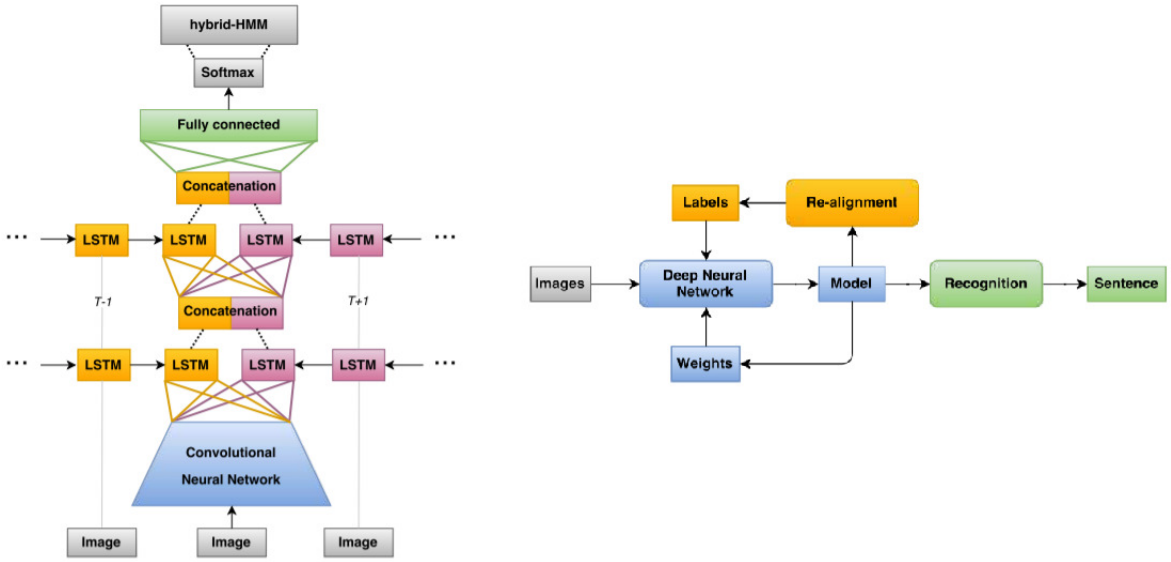

Figure 3.9: Left: End-to-end CNN-LSTM architectures with two BLSTM layers. Right: Overview of iterative re-alignment algorithm used to refine the training labels. Image from [40].

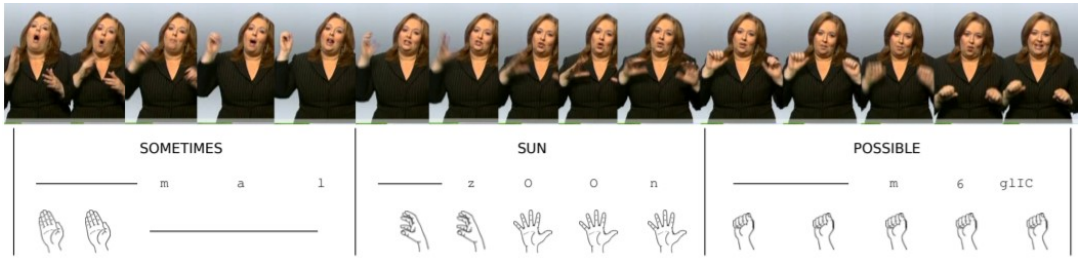

Figure 3.10: Example showing from top to bottom: the a video segment of continuous sign language and the three aligned streams: the sign glosses, the mouth shapes described by phonemes and the hand shapes. Vertical bars illustrate the synchronisation constraints across all streams, horizontal bars represent the garbage class. Image from [37].

O. Koller et al. in [36] present a new approach to the field of weak supervised learning in the video domain, which exploits sequence constraints within each independent stream and combines them by explicitly imposing synchronisation points to make use of parallelism that all sub-problems share. This is done with multi-stream HMMs while adding intermediate synchronisation constraints among the streams. They embed powerful CNN-LSTM models in each HMM stream following the hybrid approach. This allows the discovery of attributes which on their own lack sufficient discriminative power to be identified. An example of a video segment of continuous sign language and the three aligned streams is shown in Figure 3.10.

Furthermore, rather than constraining the input of expert networks by error prone preprocessing (e.g. tracking and cropping the mouth for lip reading), multiple loss functions

Author	WER Dev (%)	WER Test (%)
Camgoz, Koller, Hafield and Bowden, CVPR 2020[9]	24.61	24.49
Koller, Camgoz, Ney, and Bowden, TPAMI 2019 [36]	**22.1**	**24.1**

Table 3.5: Published State of the Art Continuous Sign Language Recognition Results on RWTH-PHOENIX-Weather 2014 Translation (in Word Error Rate, the lower the better)

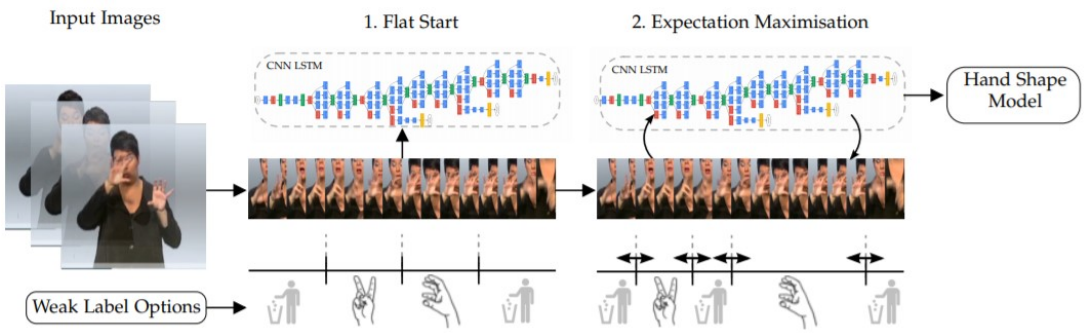

Figure 3.11: Single CNN-HMM Stream. Showing initialisation and iterative label and temporal segmentation refinement in an expectation maximisation fashion. We first linearly partition the input stream (1. Flat Start), train a CNN-LSTM model and use this model to re-estimate a new segmentation. Image from [37].

are added with weakly learnt labels. As such, preprocessing is dispensed and powerful mouth and hand shape classifier is learned directly from full images. As a result, the proposed hybrid multi-stream CNN-LSTM HMM achieves significantly faster convergence as opposed to standard single stream methods. Figure 3.11 shows the single CNN-HMM stream, while Figure 3.12 shows the multi-stream CNN-HMM with synchronisation at the sign end. Specifically, it shows how to incorporate sequential parallelism in the learning. To do so, the expectation step in the Expectation-Maximization algorithm [22] is modified to incorporate synchronisation constraints in the HMM that estimates the Viterbi alignment. Each stream is modelled in a hybrid fashion where a CNN-LSTM estimates the HMM emission probabilities of its stream symbols. The HMM has independent streams that can evolve freely. But, synchronisation points between the stream are introduced, which can only be reached by all stream at the same time. They do not resemble standard HMM states as they do not emit any symbols, but they recombine the posterior of all independent streams into a single posterior probability. The exact way this recombination is implemented is a design choice and will be a weighted sum in this paper's case. To sum up, each stream is a separate CNN-LSTM and during modelling (maximisation step) all streams have access to the input images which can be the same or different for each stream.

Finally, we will briefly describe the N. Camgoz et al. in CVPR 2020 results [9], which although they do not manage to surpass the results of O. Koller et al. in [36], they exploit a very contemporary tool that may become extremely useful in the next few years, namely the transformers. A transformer is a deep learning model that adopts the mechanism of attention, weighing the influence of different parts of the input data. It is used primarily in the field of natural language processing (NLP). It also has applications in tasks such as video understanding. Like recurrent neural networks (RNNs), transformers are designed to handle sequential input data, such as natural language, for tasks such as translation, text summarization and continuous sign language recognition. However, unlike RNNs, transformers do not require that the sequential data be processed in order. Rather, the attention operation provides context for any position in the input sequence. For example, if the input data is a natural language sentence, the transformer does not need to process the beginning of the sentence before the end. Rather, it identifies the context that confers meaning to a word in the sentence. Due to this feature, the transformer allows for much more parallelization than RNNs and therefore reduces training times.

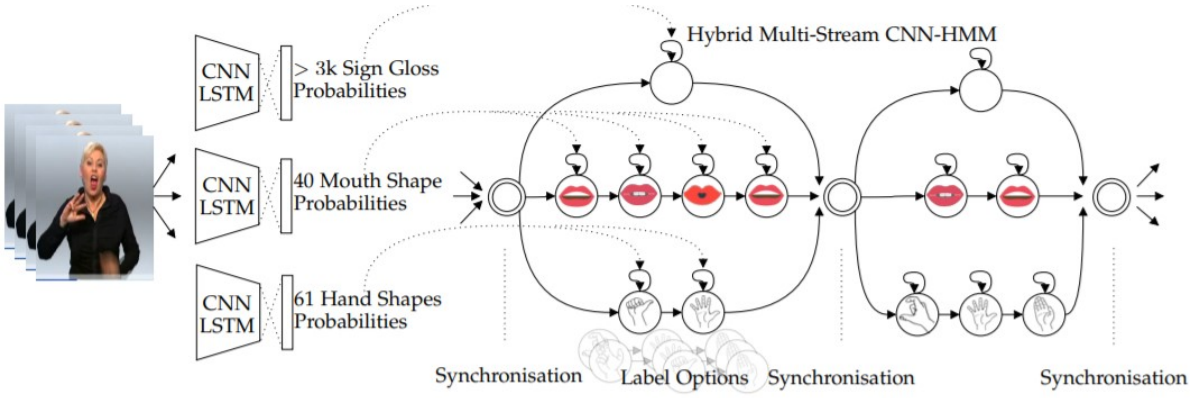

Figure 3.12: Multi-stream (3-stream) CNN-HMM with synchronisation at the sign end. Three independent CNN-LSTM models are trained on the same full frame input, while having different loss functions yielding classifiers for sign-gloss, mouth & hand shape modalities. In a hybrid multi-stream HMM framework the networks model HMM emission probabilities. All streams can evolve different in time, but have to recombine at the sign ends which have been chosen as synchronisation points. The HMM is used to re-estimate the frame labelling, improving the modelling in several EM iterations. Image from [37].

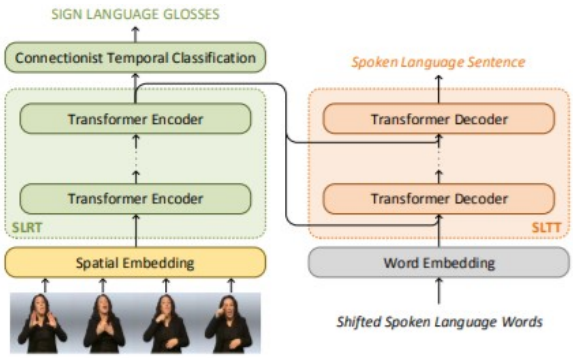

Figure 3.13: An overview of the end-to-end Sign Language Recognition and Translation approach using transformers. Image from [9].

N. Camgoz et al. [9] introduce a novel transformer based architecture that jointly learns continuous sign language recognition ad translation while being trainable in an end-to-end manner. This is achieved by using a Connectionist Temporal Classification (CTC) loss to bind the recognition and translation problems into a single unified architecture. This joint approach does not require any ground-truth timing information, simultaneously solving two co-dependant sequence-to-sequence learning problems and leads to significant performance gains. Figure 3.13 gives an overview of the aforementioned description. Finally, Figure 3.14 shows an overview of a single layers Sign Language Transformer. To help the translation networks with sign language understanding and to achieve continuous sign language recognition, a Sign Language Recognition Transformer (SLRT) is introduced, an encoder transformer model trained using a CTC loss [2], to predict sign gloss sequences. SLRT takes spatial embeddings extracted from sign videos and learns spatio-temporal representations. These representations are then fed to the Sign Language Translation Transformer (SLTT), an autoregressive transformer decoder model, which is trained to predict one word at a time to generate the corresponding spoken language sentence.

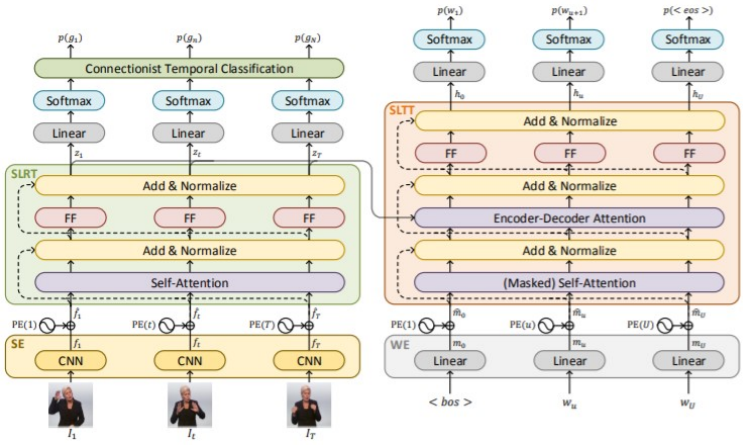

Figure 3.14: A detailed overview of a single layered Sign Language Transformer.
(SE: Spatial Embedding, WE: Word Embedding , PE: Positional Encoding, FF: Feed Forward).
Image from [9].

Chapter 4

3D Body Reconstruction for Sign Language Recognition

In this chapter, we describe how the aforementioned tool, namely SMPL-X [56], can become extremely useful for the sign language recognition task. We provide the technique through which we create features from SMPL-X, while some problematic cases of the SMPL-ify algorithm are also mentioned. Next, we define the experimental setup, namely all the model architectures exploited, the other feature extraction methods used for comparison, as well as the training parameters. Finally, we present the results of our experiments and evaluate the techniques.

4.1 From SMPL-X to SLR

As mentioned earlier, SMPL-X is a 3D model of human body pose, hand pose, and facial expression able to facilitate the analysis of human actions, interactions, and emotions. Based on this fact, it is reasonable to exploit such a tool in a task that requires a detailed depiction of the human body, face, and hands, namely the task of sign language recognition.

4.1.1 Using SMPL-X for feature creation

Although SMPL-X can reconstruct with very high accuracy the person in a specific sequence, the ultimate goal is to recognize the signs for each image sequence. The main insight is that the low dimensional parametric representation of SMPL-X should capture the majority of the information that is transmitted during a sign, i.e., the body pose, the hand pose, and the facial expressions. This should make it a very effective intermediate representation for sign language recognition. In specific, SMPL-X takes each frame of a video and reconstructs the 3D body, face, and hands of the signer conveying it in 88 parameters. Hence, each video, or equivalently each sign is converted to a sequence of vectors of length 88. This sequence of SMPL-X parameters across the frames of a sign can then be used as input to a classifier to classify the sign to one of the corresponding categories. Figure 4.2 shows the procedure followed to produce the sequence of these vectors, given a sign video.

Figure 2 provides a set of raw images and their SMPL-X reconstruction so one can observe the qualitative results provided by the SMPL-X model. Indeed, this model can adequately reconstruct both hand shape and facial expressions apart from body structure

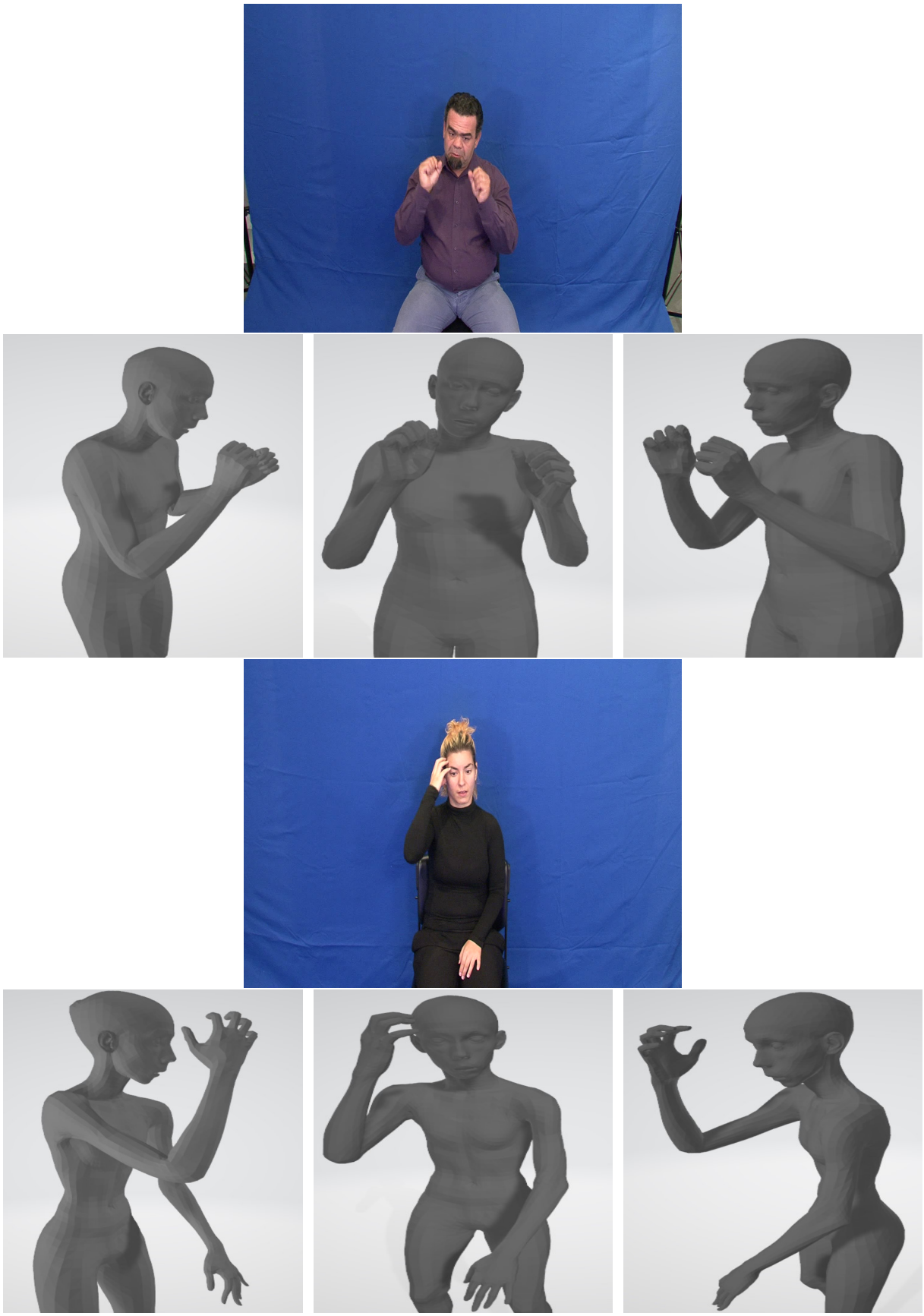

Figure 4.1: Example figure of the 3D Body, Face and Hands Reconstruction produced by SMPLify-X for the raw RGB frame at the top, viewed from different angles.

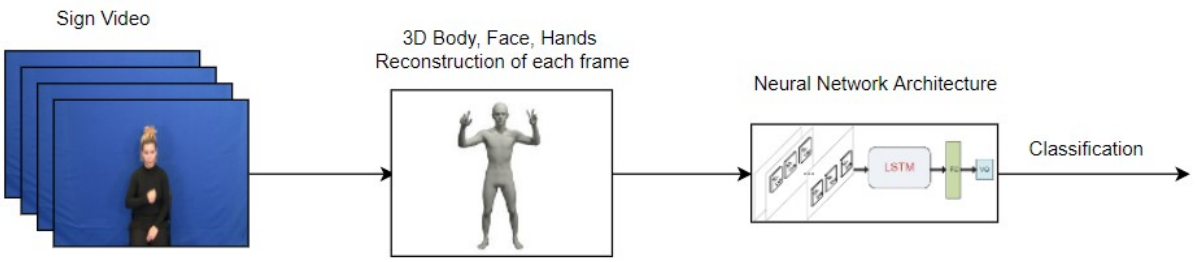

Figure 4.2: The pipeline used to produce a sequence of features from a sign video, to be used for classification.

in the sign language context, a fact that will be a key feature in a successful recognition schema. Images are being taken from different datasets (Greek Sign Language Lemmas [72, 43], MS-ASL [74], Weather-Phoenix-2013 [38]) depicting different people with varying background to emphasize the ability of SMPL-X to efficiently reconstruct the 3D representation of a person in various conditions.

4.1.2 Problematic Cases and Execution Times

While SMPL-X offers numerous traits useful not only for the task of sign language recognition but generally in the action recognition area as well, it comes with some drawbacks as well. First of all, compared to other 3D methods for reconstructing the human body, face, and hands, like HMR [34] or ExPose [15], is much slower. Due to the optimization procedure, the reconstruction from a single RGB image takes around one minute with the use of a common GPU. This fact limits this method from scaling up to real-time applications. On the other hand, SMPL-X and SMPLify-X is plausibly the most detailed and accurate reconstruction method. Figure 4.3 shows the comparison in time and expressiveness between HMR, ExPose, and SMPLify-X.

Moreover, the SMPLify-X algorithm as described in Section 2.2.3 uses OpenPose [11, 12, 66, 77] 2D skeleton keypoints for initialization. Apparently, when the hips are missing from the image, which is usually the case in the task of Sign Language Recognition, the initialization fails, and SMPLify-X cannot minimize its sophisticated loss function. This results in the construction of "monsters", instead of the detailed and qualitative 3D representation of the human body. Figure 4.4 demonstrates some examples where SMPLify-X fails to successfully reconstruct the human body, face, and hands according to the RGB frame.

4.2 Experimental Setup

Since our goal is to test the ability of the proposed method to adequately extract 3D hand, face, and body features, we limit our approach to non-continuous sign language recognition. Continuous SLR contains a syntactic and linguistic structure that is beyond the focus of this work. This means that we exclude from our experimentation datasets like RWTH-PHOENIX-Weather 2014 [38] and SIGNUM [76] that consist of full sentences. Instead, we focus on the Greek Sign Language Lemmas Dataset (GSLL) [1], which is described in detail in Section 3.1.3 [72, 43] which proved to be ideal for our experiments.

[1]The dataset can be found in: https://robotics.ntua.gr/gsll-dataset

Figure 4.3: Comparison between three state-of-the-art 3D Reconstruction methods for body, face and hands, namely HMR, SMPLify-X and ExPose.

Figure 4.4: Failure cases of SMPLify-X due to OpenPose initialization, when hips are missing from the RGB image. Image from [15] Supp. Material.

The MS-ASL [74] dataset consists of 222 signers and extremely varying backgrounds, which makes it challenging for Conv3D networks to converge. To make a more fair comparison between 3D reconstruction and 3D convolutional networks, we choose the GSLL dataset which consists of only two signers and 347 different signs (classes) in almost 3500 videos and a steady blue background cloth. We revise Table 4.1 which provides more details for the dataset and our selected subsets.

Independent sign language recognition can be considered a task that is similar to

GSLL Subset	Videos	Frames	TrainSet	DevSet	TestSet
50 classes	538	22808	318	106	114
100 classes	1038	45437	618	206	214
200 classes	2038	92599	1218	406	414
300 classes	3038	140771	1818	606	614
347 classes	3464	161050	2066	695	703

Table 4.1: Statistics for the Greek Sign Language Lemmas Dataset and its respective subsets. Indicative suggested splitting in train, dev and test set used in the experiments of [43]

action recognition. Thus we expect similar techniques to work well on SLR too. We decided to not intervene on the length of the features' sequence. Thus our features vary in sequence length from a minimum of 10 frames to a maximum of 300. Next, we present the methods with which we confront this problem.

Openpose: We extract 411 parameters for each frame and feed the sequence in an RNN consisting of one Bi-LSTM layer of 256 units and a Dense layer for classifying, after applying standard scaling to our features. We believe that providing a recurrent network with these features will eliminate any redundant information (e.g background, clothes, lighting) that a raw image contains.

Raw Image and Optical flow: A 3D state-of-the-art method for action recognition and signing is the I3D network [74, 14]. We reshape each frame to a 175×175 array and normalize its pixels to $[0, 1]$. We feed raw images to a VGG16-LSTM model as well which is initialized with Imagenet [23] weights, for further experimentation. Figure 4.5 shows the architecture described, in more detail.

SMPL-X: Due to its ability to interpret the structure of the body in detail, we strongly believe that this method will provide key features for this task. Moreover, SMPL-X requires Openpose parameters to extract its features, hence we assume that the latter provides more qualitative and deeper features than the former. Moreover, SMPL-X provides 3D information, in comparison to Openpose which results in 2D only keypoints, so the extracted features should be strictly more informative. This method extracts 88 features per frame, creating a (length of sequence) $\times$ 88 array for each sequence, which is being standard scaled as in the Openpose experiments. Similar to Openpose, we employ the same neural network architecture not only because both experiments treat the same form

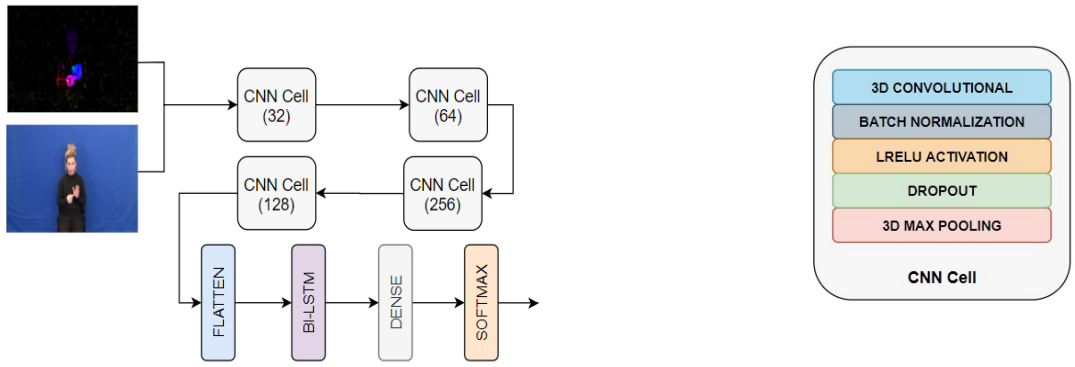

Figure 4.5: The architecture used for the Convolutional I3D-type model. On the left is the proposed architecture with the 3D CNN Cells followed by one Bidirectional LSTM layer. On the right are the interior layers of each 3D CNN cell.

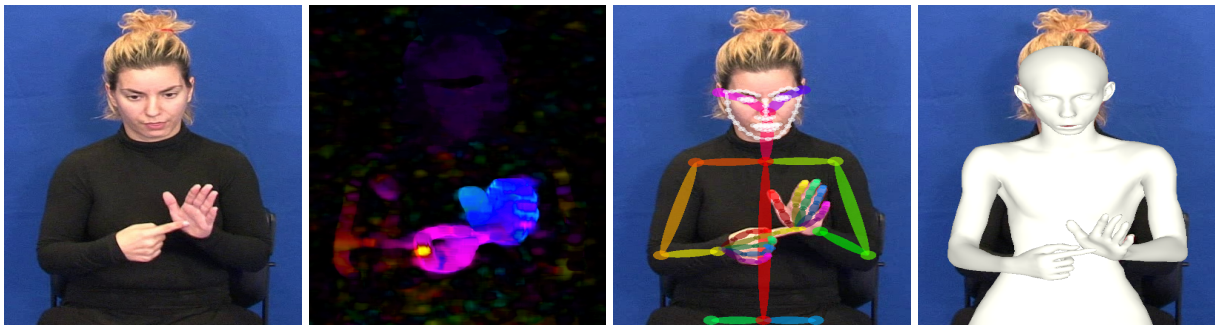

Figure 4.6: i) First image: Raw RGB frame, ii) Second Image: Optical flow of a frame, iii) Third Image: Openpose 2D Skeleton, iv) Fourth image: 3D Body Reconstruction produced by SMPL-X.

of features, but primarily so that we can directly compare the two methods independently of the type of architecture.

We train all networks using categorical cross-entropy loss. SGD is used to optimize the loss function, with an initial learning rate of 0.0001 and 10% decay rate per epoch, while the batch size is set to 1, due to varying sequence length. We perform Learning Rate Reduction and Early Stopping by monitoring the validation loss with patience of 3 and 5 epochs respectively. Figure 4.6 shows an example raw RGB frame with its optical flow, the 2D skeleton pose produced by OpenPose, and the 3D SMPL-X reconstruction.

4.3 Experimental Evaluation

4.3.1 Experimental Results

According to Table 4.2, Openpose and SMPL-X models, which consist of 1.6 and 0.9 million parameters respectively, outperform the Conv3D-LSTM and VGG16-LSTM model, which consist of 43 and 15 million parameters respectively. This can be attributed to the fact that the former two eliminate the redundant information from each frame, keeping only the essential body features. Specifically, VGG16 fails to converge and reduce its loss, achieving an accuracy below 10% for all classes. This does not come as a surprise to us since Joze and Koller in [74] have trained a VGG16-LSTM model for the MS-ASL dataset which achieved 13.33% for the ASL100 Subset and just 1.47% for the ASL500 Subset. As mentioned earlier, GSLL Dataset is characterized by a uniform environment between each sign and each signer (two signers in front of blue cloth). The MS-ASL dataset consists of 222 distinct signers where each signer performs in a completely altered environment. We strongly believe that Openpose and mainly SMPL-X will by far outperform convolutional models in these datasets, which simulate more accurately the real world. Finally, SMPL-X seems to outperform the features produced by Openpose especially with the increase of different signs, dictating that a more detailed and qualita-

Method \ GSLL Subset	Subset 50	Subset 100	Subset 200	Subset 300	Full Dataset	Parameters
3D RGB & Optical Flow Images	90.41%	86.85%	80.79%	71.36%	65.95%	43.41 million
2D Openpose Skeleton	96.49%	94.39%	93.24%	91.86%	88.59%	1.55 million
3D SMPL-X Reconstruction	**96.52%**	**95.87%**	**95.41%**	**95.28%**	**94.77%**	0.88 million

Table 4.2: Comparison of the three methods for training: i) Raw RGB images and their Optical Flow ii) Openpose skeleton key-points and iii) 3D Body Reconstruction key-points.

Parameters	Openpose	SMPL-X
All	**88.59**%	**94.77**%
Without Face	88.34%	93.19%
Without Hands	70.20%	89.58%
Without Body	84.21%	85.02%

Table 4.3: Experiments with subset of features produced by Openpose and SMPL-X.

tive representation of the human body is needed for the Sign Language Recognition task. While varying and more complex signs are being added to the train set, Openpose fails to convey the small details that differentiate these signs, while SMPL-X holds its accuracy almost fixed.

4.3.2 Ablation Study

To further examine the features produced by SMPL-X, we experiment with a combination of a subset of features produced by it. Specifically, as mentioned in Section 3, the SMPL-X method produces a total of 88 features, 10 for shape parameters, 3 for global orientation, 24 for left and right-hand pose, 3 for jaw pose, 6 for left and right eye pose, 10 for expression and 32 for the body pose. Moreover, it is widely known that sign language does not depend solely on gestures but fundamentally on body pose and facial expressions as well. To demonstrate this fact, we proceed to a couple of more experiments. First, we remove all information that comes from facial expressions (jaw pose, left and right eye pose, and expression) and train the model again with a total of 69 features. Secondly, we only remove the body pose information and train the model with a total of 50 features. Finally, for the sake of completeness, we remove the left and right-hand pose and train the model with a total of 64 parameters. We conduct the same experiments for Openpose by separating pose keypoints (75 parameters), face keypoints (210 parameters), and left and right-hand keypoints (126 parameters). Table 4.3 sums up the results from all the aforementioned experiments.

First of all, we can see that omitting any of these three channels indeed reduces the accuracy of our model. In fact, we expect the omission of facial characteristics to affect even more the accuracy in the continuous sign language where the face plays a crucial role in expressing the intensity of a word. For example, “rain” and “snow” have the exact same hand configurations, whereas only the mouth shape changes. Furthermore, we observe that removing hand information in SMPL-X is less harmful than removing body pose. That can be attributed to the fact that when few and simple signs are available, the sign can be mainly conveyed through the movement of the arms while the hands often remain straight. Nonetheless, both hands and body structure (chiefly due to arms) are of vital importance for SLR while at the same time, omitting facial expression affects the model’s performance. On the other hand, Openpose due to the fact that it has very few parameters for the body, i.e. only 75 out of 411, is much more destructive to remove hands features than the body.

4.4 Further Experiments

4.4.1 SMPL-X vs ExPose

As described earlier, SMPL-X is maybe the most qualitative way for reconstructing 3D body structure, facial expressions, and hand gestures from a single RGB image. Apparently, it is not the fastest, though. Specifically, the approximately one-minute optimization algorithm per RGB image renders SMPLify-X incapable of real-time applications. ExPose, on the other hand, which was published in 2020, maintains the quality of the reconstruction using body-driven attention as described in 2.2.4, and performs in almost real-time with the use of a common GPU. Next, we compare qualitatively, the two methods by reconstructing the 3D representation for both SMPL-X and ExPose, for some given frames from Sign Language Datasets.

Furthermore, we compare these two methods in the task of the Isolated Sign Language Recognition, applying the same procedure described in Sections 4.1 and 4.2. Table 4.4 shows the accuracy of SMPL-X and ExPose for different subsets of the GSLL dataset.

Method \ GSLL Subset	Subset 50	Subset 100	Subset 200	Subset 300	Parameters
SMPLify-X	96.49%	94.39%	93.24%	91.86%	0.88 million
ExPose	**99.20%**	**99.46%**	**97.74%**	**96.47%**	1.59 million

Table 4.4: Comparison of the the two 3D methods for reconstructing body, face and hands. i) SMPLify-X ii) ExPose

We observe, that ExPose helps the network recognize better the differences between each sign video. Despite the fact that ExPose, in order to achieve almost real-time reconstruction, falls short of expressiveness, it can adequately decode the details of the human body, face, and hands and in fact better than SMPLify-X. Moreover, ExPose features make the neural network converge much faster in comparison to SMPLify-X. The comparison between those two state-of-the-art methods seems to offer intriguing research information and should be investigated in bigger datasets, as well.

4.4.2 ExPose on the MS-ASL

We further examine the performance of ExPose, in one of the most challenging datasets available for Isolated Sign Language Recognition, namely the MS-ASL dataset. We quickly remind the state-of-the-art results on this dataset for different methods, as described in Section 3.1.2, while adding our results using ExPose. We train a simple Recurrent Neural Network with only one LSTM layer and present the results in Table 4.5. Our experiments are limited only to a small subset of the dataset. We can see that the neural network initialized with ExPose features achieves an astonishing 37.39% accuracy. This method should be compared with the HCN network that was described in Section 3.1.2 and is a complicated recurrent neural network that is fed with Openpose skeleton keypoints. We expect that ExPose combined with HCN will by far surpass the current HCN and Re-sign method, while on a much bigger subset of the MS-ASL subset, where the I3D diverges fast, ExPose might stand out as the best method.

Figure 4.7: Qualitative results of the MiDas Depth Estimation model. Image from [63].

4.4.3 Depth Channel

4.4.3.1 MiDaS: Depth Estimation

MiDaS is a robust monocular depth estimation model which is expected to perform across diverse environments, developed by R. Ranftl et al. in [63]. This work develops novel loss functions that are invariant to the major sources of incompatibility between datasets including unknown and inconsistent scale and baselines. These losses enable training on data that was acquired with diverse sensing modalities such as stereo cameras, laser scanners, and structured light sensors. Qualitative examples are shown in Figure 4.7

R. Ranftl et al. improved MiDaS model by proposing Vision Transformers for Dense Prediction in [62]. Dense vision transformers is an architecture that leverages vision transformers in place of convolutional networks as a backbone for dense prediction tasks. Tokens are being assembled from various stages of the vision transformer into image-like representations at various resolutions which are progressively combined into full-resolution predictions using a convolutional decoder. The transformer backbone processes representations at a constant and relatively high resolution and has a global receptive field at every stage. These properties allow the dense vision transformer to provide finer-grained and more globally coherent predictions when compared to fully convolutional networks. For the task of monocular depth estimation, dense vision transformers achieve a 28% improvement in relative performance when compared to the state-of-the-art fully-convolutional network MiDaS model in [63]. A qualitative comparison between these two is shown in Figure 4.8.

Method \ MS-ASL Subset	Subset 100	Subset 200	Subset 500	Subset 1000
Naive Classifier	0.99%	0.50%	0.21%	0.11%
VGG+LSTM [21, 19]	13.33%	7.56%	1.47%	-
Naive Classifier	0.99%	0.50%	0.21%	0.11%
HCN [46]	46.08%	35.85%	21.45%	15.49%
Re-sign [40]	45.45%	43.22%	27.94%	14.69%
I3D [14]	**81.76%**	**81.97%**	**72.50%**	**57.69%**
ExPose [15]	37.39%	-	-	-

Table 4.5: Comparison of the the ExPose method using a simple LSTM RNN with the state-of-the-art methods for this dataset.

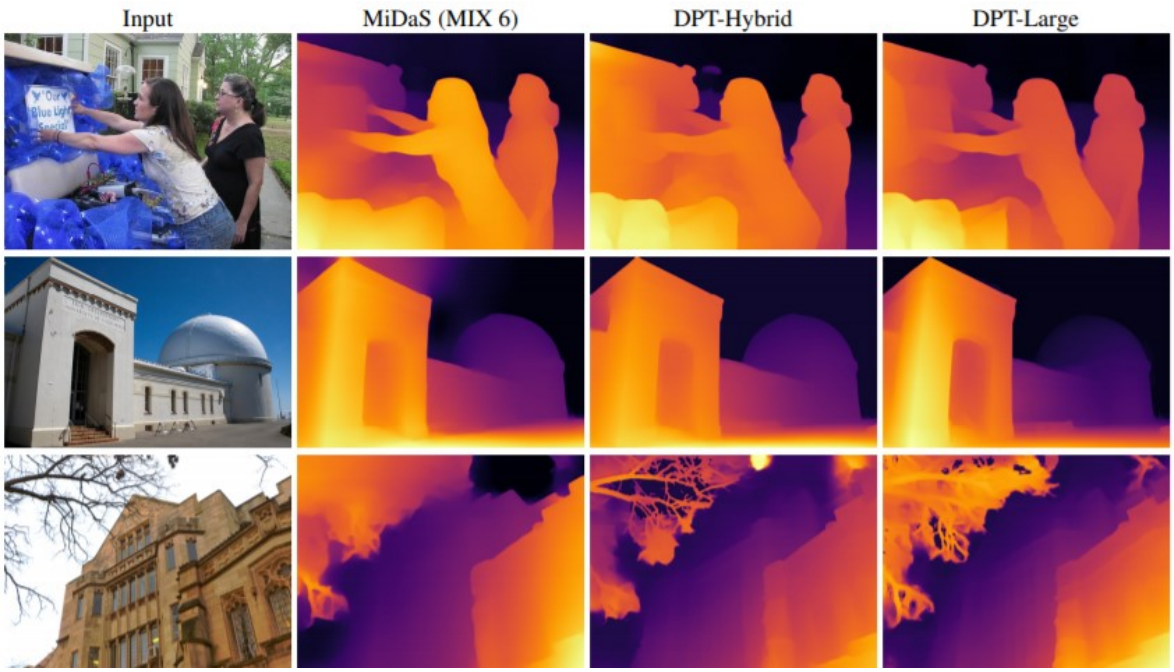

Figure 4.8: Qualitative comparison between the fully-convolutional network MiDaS and the Depth Vision Transformer in the task of monocular depth estimation. Image from [62].

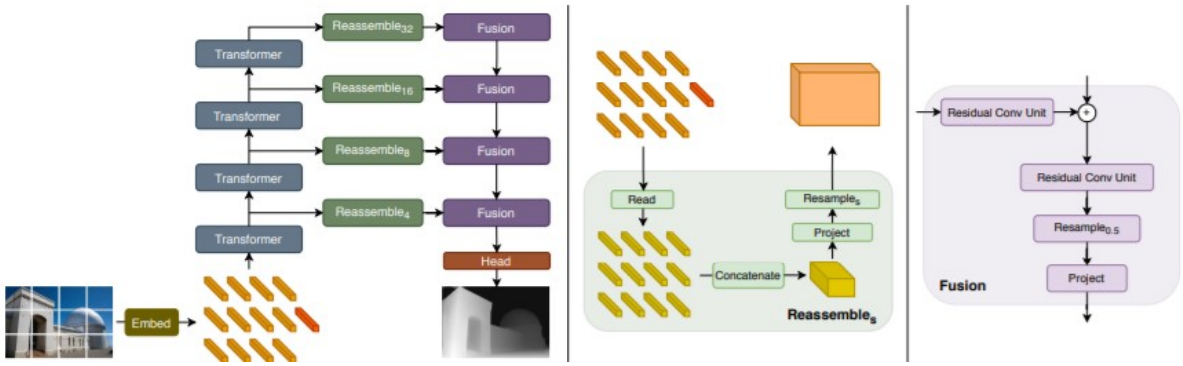

Figure 4.9: The depth vision transformer architecture used for the task of monocular depth estimation. Image from [62].

The architecture of the depth vision transformer is shown in Figure 4.9. On the left is the architecture overview. The input image is transformed into tokens (orange) either by extracting non-overlapping patches followed by a linear projection of their flattened representation (DPT-Base and DPT-Large) or by applying a ResNet-50 feature extractor (DPT-Hybrid). The image embedding is augmented with positional embedding and a patch-independent readout token (red) is added. The tokens are passed through multiple transformer stages. The tokens are reassembled from different stages into an image-like representation at multiple resolutions (green). Fusion modules (purple) progressively fuse and upsample the representations to generate a fine-grained prediction. At the center is the overview of the $Reassemble_S$ operation. Tokens are assembled into feature maps with $\frac{1}{s}$ the spatial resolution of the input image. At right the fusion blocks combine features using residual convolutional units and upsample the feature maps.

For our experiments, we use the most contemporary and successful model for monocular depth estimation, namely the MiDaS model that used depth vision transformers.

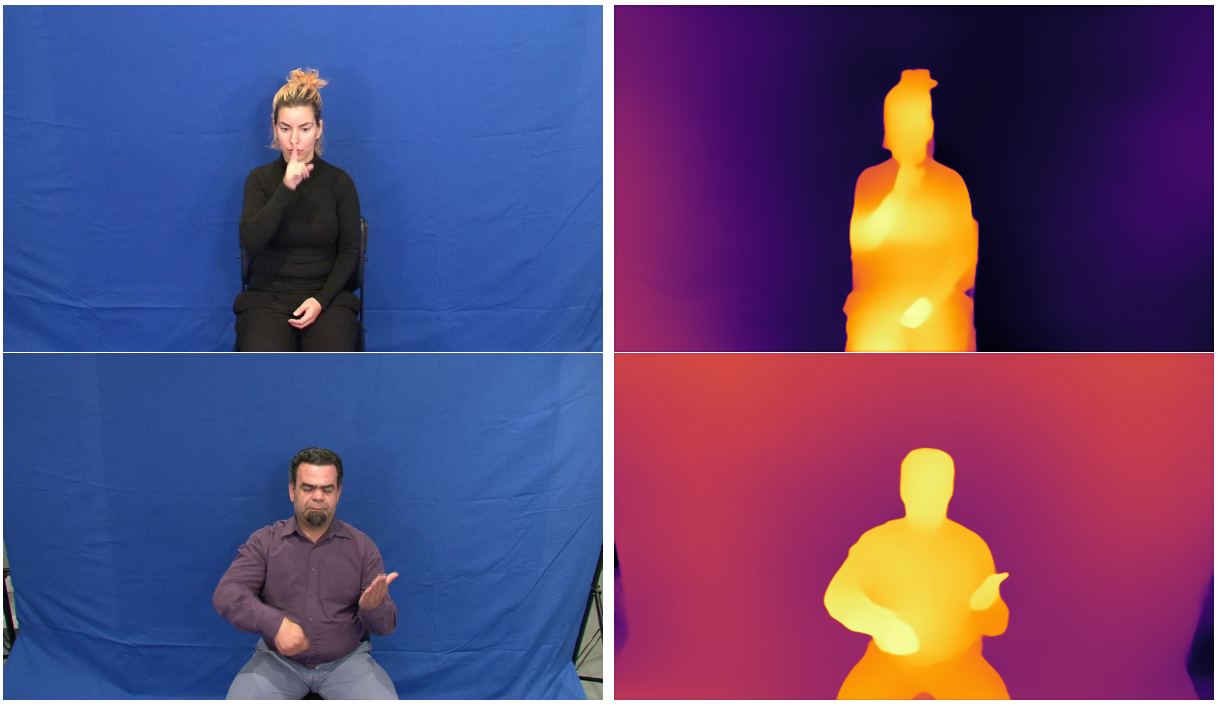

Figure 4.10: Example frames from the GSLL dataset and their MiDaS Depth Estimation.

4.4.3.2 Experiments with Depth Channel

We believe that the Depth information can increase the performance of the network when used as a secondary channel. Since the Greek Sign Language Lemmas Dataset does not contain depth information, we used the aforementioned tool MiDaS for depth estimation. Two example frames of the dataset with their depth estimation using MiDaS are shown in Figure 4.10.

To test the depth information, we train a model using only that information. Next, we train a two-channel CNN-LSTM model using the architecture in Figure 4.2 and finally we combine the SMPL-X information with Depth information. The results are shown in Table 4.6. We see that training a neural network with only depth information, is not viable for the task of Sign Language Recognition, in comparison to other tasks, where using only the depth can yield a good accuracy. The RGB results have been reported before. When combining the information of RGB images with their depth, the neural network slightly deteriorates since it cannot handle the extra information and eliminate the redundant one. Experimentations with the MS-ASL dataset show that the depth channel worsens the results even more while some examples of the depth prediction are shown in Figure 4.11

Method \ GSLL Subset	Subset 50	Subset 100	Subset 200	Subset 300
Only Depth	11.56%	9.91%	9.1%	6.50%
RGB	**88.59%**	**84.58%**	**71.98%**	**55.37%**
RGB + Depth	85.81%	82.59%	70.01%	52.10%

Table 4.6: Experimental results for the three methods of training: i) using only depth information, ii) combining raw RGB images and their depth information and iii) combining SMPL-X features and Depth Information

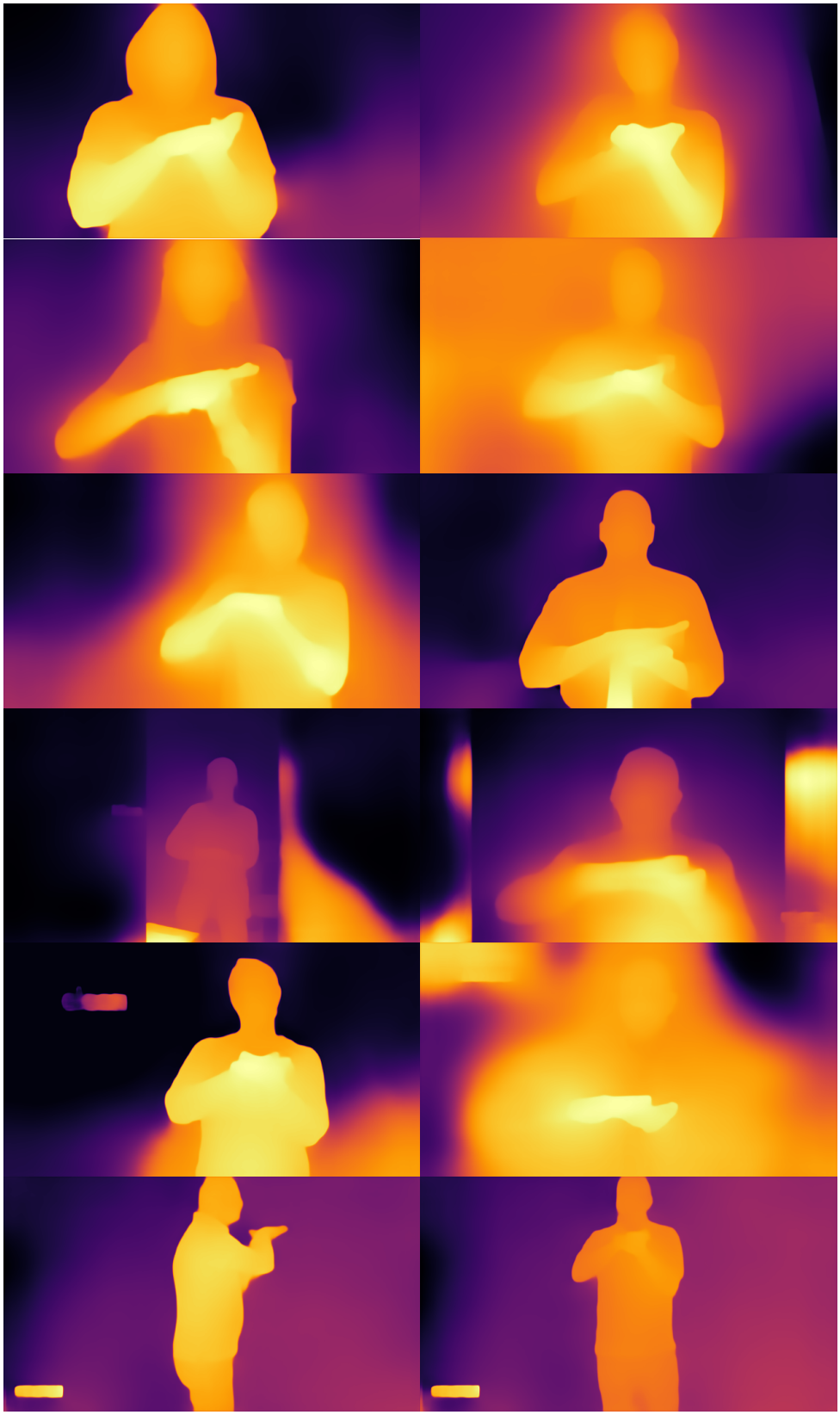

Figure 4.11: Depth Estimation of characteristic frames from the MS-ASL dataset depicting the exact same sign "clean"/"nice" shown in Figure 3.2

Chapter 5

Conclusions

5.1 Future Directions

5.1.1 Using SLR to improve 3D Reconstruction

In the previous chapter, it was made clear that 3D Reconstruction Methods can significantly help in the task of Sign Language Recognition. An interesting future direction, though, is to consider how sign language recognition can be used to increase the qualitative accuracy of the 3D body, face, and hand reconstruction itself. Thus, sign Language Recognition will act as feedback to the 3D Reconstruction. To briefly expand this idea, consider one sign performed N times by a signer. Hence, N videos of the same sign are available and equivalently with the "same" 3D body reconstruction. Each video $V_1, ..., V_N$ consists of F_i frames where $i = 1, ..., N$. Since each frame is represented by 88 SMPL-X parameters, each video V_i consists of $F_i \times 88$ features.

The first task is to find the outlier video, which is the 3D body reconstruction that differs more from the other $N - 1$ reconstructions. Finding an outlier video, of course, is not an easy task. The first method that we can consider is to train a recurrent neural network with 2 bidirectional LSTM layers, in multiple signs. Let these two layers consist of 128 and 64 units. Thus, each video V_i is converted from an $F_i \times 88$ array to a 256 vector, to a 128 vector, and finally to a $|C|$ vector for classifying, where $|C|$ is the number of signs. Then, we extract the features from the penultimate layer to achieve a representation of a video from an $F_i \times 88$ array to a 128-length vector. So, the task now is to find the outlier array between N 128-length arrays. This can be achieved, with the use of a 128D Gaussian Distribution. The data will be fed into the Gaussian and the μ, Σ will be calculated. Finally, the probability of each vector to belong in the distribution is calculated, and the one with the smallest is considered the outlier. A second approach can be exploited, as well. Each video is split into its corresponding frames, creating $F_1 + ... + F_N = F_total$ frames of 88 features each. The main goal is to find a Gaussian Mixture Model (GMM) for all these frames. We expect, that each cluster will convey one hand sub-gesture (i.e hands resting on legs, hands rising, hands turning around, and so on). Using the Expectation Maximization algorithm we fit the F_{total} – 88D vectors in the GMM and compute μ, Σ for each. Then, for each frame, the probability to belong in the GMM is calculated, while the probability of a video is computed as:

$$P(V_i) = P(frame_1) \cdot ... \cdot (frame_{F_i}) \tag{5.1}$$

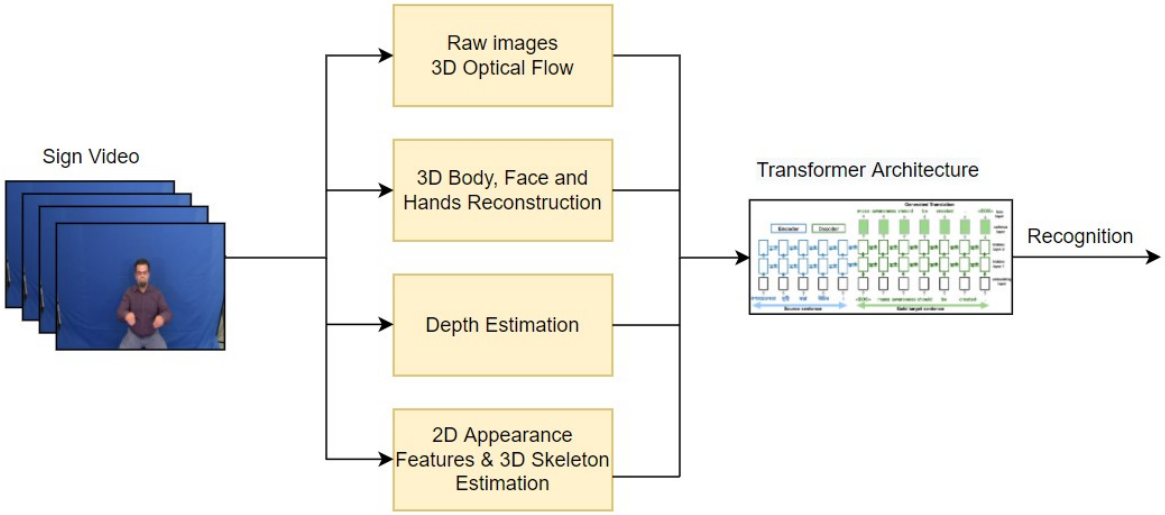

Figure 5.1: A block diagram of the method proposed for the task of continuous sign language recognition. From each video, the raw frames and their optical flow are being used, the 3D information from hand, body, and face are being extracted, the depth channel is being estimated, and finally, the 2D Appearance Features and the 3D Skeleton are being estimated, as well. All these features are Incorporated in an advanced transformer architecture used for continuous sign language recognition.

5.1.2 3D Body Reconstruction for Continuous Sign Language Recognition

The whole Diploma Thesis deals with the task of Isolated Sign Language Recognition, which is an easier one compared to the more generalized problem of Continuous Sign Language Recognition. Many tools have been exploited for this task, like HMM's or transformers which incorporate linguistic features with computer vision techniques to form complete sentences. These techniques have been discussed in detail, in Chapter 3. While, SMPL-X and other 3D body, face, and hands reconstruction methods seem to be strong tools for such a complicated task, combining more channels of information within a very strong neural network. Next, we briefly describe the features that can be combined all together to decipher the complex visual task of a signer signing a complete sentence using their whole body, facial expression, and hand gestures.

Raw images & Optical Flow: The raw image, and its optical flow could never be missing from a strong convolutional neural network. Many tasks until today, have shown that raw RGB frames through sophisticated neural networks achieve great results in continuous sign language recognition [40], and hence, the information taken from this should not be neglected. 3D Optical flow shall be exploited, as well. **Depth Channel**: As shown in the previous Chapter, the Depth Channel indeed helps to increase the accuracy in recognizing signs, and thus it should be included too. **3D Reconstruction**: The outstanding importance of combing 3D information is being highlighted throughout the whole Diploma Thesis. Using state-of-the-art methods like SMPL-X or ExPose is a must to achieve higher accuracy in this task. **2D Appearance Features from Hands and Face & 3D Skeleton**: There are many contemporary methods for extracting 2D appearance features not only for the skeleton but for the hands and face, as well. Moreover, in [55] a 3D hand skeleton is being estimated through 2D features and deep learning techniques, information that play a role in recognition.

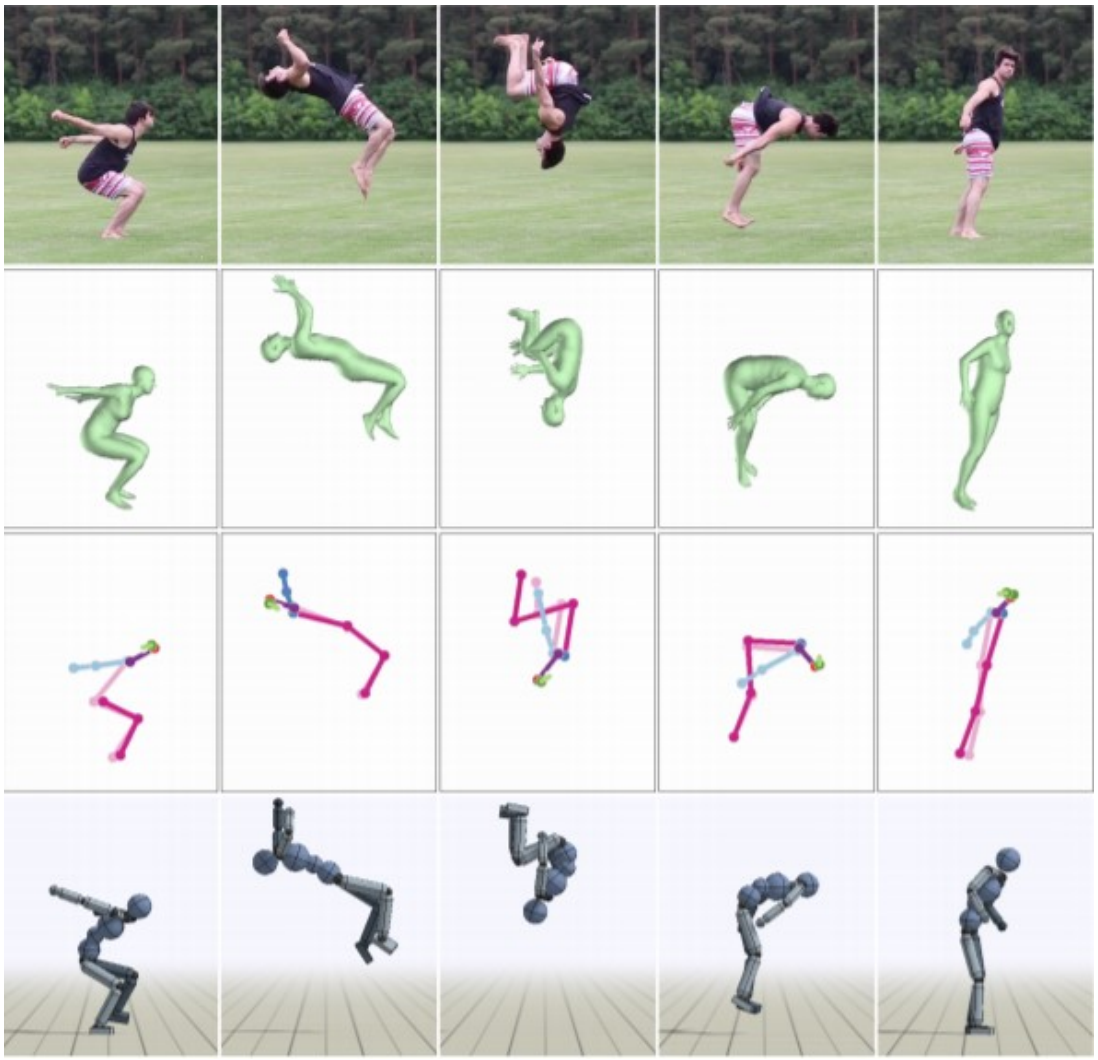

Figure 5.2: Comparison of the motions generated by different stages of the pipeline for backflip A. Top-to-Bottom: Input video clip, 3D pose estimator, 2D pose estimator, a simulated character. Image from [58].

Finally, to incorporate the time domain and linguistic characteristics, transformers could be exploited. As described in Chapter 3, the contemporary architecture used in [9], altered to combine all the aforementioned channels of information, should lead to state-of-the-art results in the field of continuous sign language recognition. All these are being summarized in a block diagram in Figure 5.1

5.1.3 3D Reconstruction for Other Tasks

In this thesis, we decided to highlight the importance of creating robust, fast, and qualitative systems for 3D Human Reconstruction, through the task of Isolated and Continuous Sign Language Recognition. Nonetheless, 3D Computer Vision and 3D Reconstruction is an extremely hot topic nowadays, which is applied to numerous tasks. We will briefly mention two paths that 3D reconstruction can be useful when applied.

Autonomous Driving: A self-driving car, also known as an autonomous vehicle (AV or auto), driver-less car, or robo-car is a vehicle that is capable of sensing its environment and moving safely with little or no human input. To be successful, self-driving cars

combine a variety of sensors to perceive their surroundings, such as radar, sonar, GPS, odometry, and cameras. With the excessive development of Computer Vision, cameras play a crucial role in similar tasks, since their recordings can be interpreted deeper and more accurately. 3D Reconstruction of the driver or the passengers can play a crucial role in deciphering the intentions of the former when it comes to the driving decisions, or the safety of the latter when they are crossing the road. 3D Reconstruction though does not stop on humans. Objects are being reconstructed with great detail, as well, and thus, reconstructing, key objects of the driving environment (i.e traffic lights, signs, and so on) will lead to a more successful autonomous-driving task.

Reinforcement Learning: Reinforcement learning (RL) is an area of machine learning concerned with how intelligent agents ought to take actions in an environment to maximize the notion of cumulative reward. Reinforcement learning is one of three basic machine learning paradigms, alongside supervised learning and unsupervised learning. Reinforcement learning is being used more and more in the last few years, primarily in robotic applications. Deep reinforcement learning, is when the agent combines optical information apart from solely behaving through a reward function. Deep RL algorithms are able to take in very large inputs (e.g. every pixel rendered to the screen in a video game) and decide what actions to perform to optimize an objective (eg. maximizing the game score). Deep reinforcement learning has been used for a diverse set of applications including but not limited to robotics, video games, natural language processing, computer vision, education, transportation, finance, and healthcare. According to [58] and Figure 5.2 using 3D body, face, and hands reconstruction, helps the agent to assimilate information faster and easier from videos and images. Hence, combining 3D Reconstruction with Deep Reinforcement Learning might be a very prosperous research path.

5.2 Contributions

This Diploma Thesis discussed the contemporary research field of 3D Computer Vision, namely the 3D Reconstruction of the human facial expression, body structure, and hand gesture. Moreover, this thesis investigated the complex task of Isolated and Continuous Sign Language Recognition, and how the former field can help. There is a plethora of contributions that this thesis has offered:

- We offered a very detailed bibliographic analysis of the most contemporary and state-of-the-art 2D and 3D methods for human reconstruction of the last 5 years. In specific, we described and explained in detail the methodology behind the most famous 2D skeleton-based model, namely Openpose (2016-2019) [11, 12, 66, 77]. We defined and explained the 3D parametric models used to describe the human body; SMPL (2015) [48] and SMPL-X (2019) [56]. Next, we described the technical work behind the most qualitative methods for extracting the 3D parameters that describe the human body, face and hands from a single RGB image; SMPL-ify (2016) [7], HMR (2018) [34], SMPLify-X (2019) [56] and ExPose (2020) [15].

- We offered a similar bibliographic analysis for the most important sign language datasets and the state-of-the-art methods for confronting the task of recognizing sign language. We deeply elaborated on the MS-ASL dataset [74], and on the state-of-the-art methods exploited to achieve high performance on the task of isolated sign language recognition. We discussed two of the main Greek sign language

datasets, namely the Greek Sign Language Lemmas Dataset [43, 72] and the Greek Sign Language Dataset [1]. On the continuous aspect, we presented the two most prominent datasets for the task of continuous sign language recognition, namely the RWTH-PHOENIX-Weather 2014 dataset [38] and the RWTH-PHOENIX-Weather 2014 Translation dataset [10], along with the state-of-the-art methods in the field of Sign Language Recognition of the last few years.

- We exploited the Greek Sign Language Lemmas Dataset (GSLL) for our experiments, which we re-organized and made publicly available for further experimentation. The GSLL dataset along with statistical details and instruction can be found in https://robotics.ntua.gr/gsll-dataset.

- We applied 3D body, face, and hands reconstruction methods on the task of isolated sign language recognition, achieving top results and surpassing all other currently known methods. Specifically, we employed SMPL-X, a contemporary parametric model that enables joint extraction of 3D body shape, face and hands information from a single image. We use this holistic 3D reconstruction for SLR, demonstrating that it leads to higher accuracy than recognition from raw RGB images and their optical flow fed into the state-of-the-art I3D-type network for3D action recognition and from 2D Openpose skeletons fed into a Recurrent Neural Network.

- Furthermore, we conducted an ablation study, showing the importance of having all three channels of information; namely facial expression, hands shape, and body structure, for successfully recognizing Sign Language. In specific, we trained three different models, by omitting the facial expression information, the body information, and the hand gesture information respectively. We show that each part plays an important role in optimal results in SLR.

- We directly compared the two most recent 3D reconstruction methods, namely SMPLify-X and ExPose, in run-time efficiency and expressiveness. That means, that we tested both methods in the same task of isolated sign language recognition to check their efficiency. In parallel, we provided images for both reconstructions for qualitative comparison while discussing their run-time complexity. Finally, we exploited the ExPose method, perhaps in the most challenging dataset for isolated SLR, opening roads for future exploitation.

- We experimented with Depth Estimation methods and then exploited the most successful one to enhance our recognition models for the task of Sign Language Recognition. We trained models using only depth information, only RGB information and finally we combined information of raw RGB frames and depth information.

To conclude, this thesis opens a path to the world of 3D Reconstruction and Sign Language Recognition. The former can be exploited in many ways to improve the current methods for the latter, while it is currently an exceptional tool for other tasks as well, nowadays.

Appendices

5.3 List of Publications

Bibliography

[1] Kratimenos, A. and Pavlakos, G. and Maragos P., "3D Hands, Face and Body Extraction for Sign Language Recognition", in Proc. European Conference on Computer Vision Workshops (ECCVW), 2020.

Abstract: For the problem of Sign Language Recognition (SLR), the majority of the information is included in three main channels; hand gestures, facial expression and body pose. While many state-of-the-art works have managed to deeply elaborate on these features independently, to the best of our knowledge, no work has adequately combined all these three information channels, particularly in 3D, to efficiently recognize Sign Language. In this work, we employ SMPL-X, a contemporary parametric model that enables joint extraction of 3D body shape, face and hands information from a single image. We use this holistic 3D reconstruction for SLR, demonstrating that it leads to higher accuracy than recognition from raw RGB images, or 2D skeletons. Simultaneously, we demonstrate the importance of combining the information from all three channels, to achieve the best recognition results.

[2] Kratimenos, A. and Pavlakos, G. and Maragos P., "Independent Sign Language Recognition with 3D Body, Hands, and Face Reconstruction", in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020.

Abstract: Independent Sign Language Recognition is a complex visual recognition problem that combines several challenging tasks of Computer Vision due to the necessity to exploit and fuse information from hand gestures, body features and facial expressions. While many state-of-the-art works have managed to deeply elaborate on these features independently, to the best of our knowledge, no work has adequately combined all three information channels to efficiently recognize Sign Language. In this work, we employ SMPL-X, a contemporary parametric model that enables joint extraction of 3D body shape, face and hands information from a single image. We use this holistic 3D reconstruction for SLR, demonstrating that it leads to higher accuracy than recognition from raw RGB images and their optical flow fed into the state-of-the-art I3D-type network for 3D action recognition and from 2D Openpose skeletons fed into a Recurrent Neural Network. Finally, a set of experiments on the body, face and hand features showed that neglecting any of these, significantly

reduces the classification accuracy, proving the importance of jointly modeling body shape, facial expression and hand pose for Sign Language Recognition

[3] Kratimenos*, A. and Avramidis*, K.and Garoufis, C. and Zlatintsi, A. and Maragos, P."Augmentation Methods on Monophonic Audio for Instrument Classification in Polyphonic Music", in Proc. European Signal Processing Conference (EUSIPCO), 2020

Abstract: Instrument classification is one of the fields in Music Information Retrieval (MIR) that has attracted a lot of research interest. However, the majority of that is dealing with monophonic music, while efforts on polyphonic material mainly focus on predominant instrument recognition. In this paper, we propose an approach for instrument classification in polyphonic music from purely monophonic data, that involves performing data augmentation by mixing different audio segments. A variety of data augmentation techniques focusing on different sonic aspects, such as overlaying audio segments of the same genre, as well as pitch and tempo-based synchronization, are explored. We utilize Convolutional Neural Networks for the classification task, comparing shallow to deep network architectures. We further investigate the usage of a combination of the above classifiers, each trained on a single augmented dataset. An ensemble of VGG-like classifiers, trained on non-augmented, pitch-synchronized, tempo-synchronized and genre-similar excerpts, respectively, yields the best results, achieving slightly above 80% in terms of label ranking average precision (LRAP) in the IRMAS test set.ruments in over 2300 testing tracks.

[4] Avramidis*, K. and Kratimenos*, A. and Garoufis, C. and Zlatintsi, A. and Maragos, P. "Deep Convolutional and Recurrent networks for polyphonic instrument classification from monophonic raw audio waveforms.", in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020.

Abstract: Sound Event Detection and Audio Classification tasks are traditionally addressed through time-frequency representations of audio signals such as spectrograms. However, the emergence of deep neural networks as efficient feature extractors has enabled the direct use of audio signals for classification purposes. In this paper, we attempt to recognize musical instruments in polyphonic audio by only feeding their raw waveforms into deep learning models. Various recurrent and convolutional architectures incorporating residual connections are examined and parameterized in order to build end-to-end classifiers with low computational cost and only minimal preprocessing. We obtain competitive classification scores and useful instrument-wise insight through the IRMAS test set, utilizing a parallel CNN-BiGRU model with multiple residual connections, while maintaining a significantly reduced number of trainable parameters.

Bibliography

[1] N. Adaloglou, T. Chatzis, I. Papastratis, A. Stergioulas, G. T. Papadopoulos, V. Zacharopoulou, G. Xydopoulos, K. Atzakas, D. Papazachariou, and P. Daras, "A comprehensive study on sign language recognition methods," *arXiv preprint arXiv:2007.12530*, 2020.

[2] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 1, no. 1, p. 99, 2018.

[3] U. Agris, J. Zieren, U. Canzler, B. Bauer, and K. F. Kraiss, "Recent developments in visual sign language recognition," *Universal Access in the Information Society*, vol. 6, pp. 323–362, 2008.

[4] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis, "SCAPE: Shape Completion and Animation of People," *ACM Trans. on Graphics*, vol. 24, p. 408–416, 2005.

[5] E. Antonakos, V. Pitsikalis, and P. Maragos, "Classification of extreme facial events in sign language videos," *EURASIP Journal on Image and Video Processing*, vol. 2014:14, 2014.

[6] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[7] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image," in *Proc. Eur. Conf. Computer Vision*, 2016.

[8] N. Camgoz, S. Hadfield, O. Koller, and R. Bowden, "Subunets: End-to-end hand shape and continuous sign language recognition," in *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[9] N. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Sign language transformers: Joint end-to-end sign language recognition and translation," in *Proc. IEEE Conf. Computer Vision & Pattern Recognition*, 2020.

[10] N. Camgöz, S. Hadfield, O. Koller, and R. B. Hermann Ney, "Neural sign language translation," in *Proc. IEEE Conf. Computer Vision & Pattern Recognition*, 2018.

[11] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 43, no. 1, pp. 172–186, 2019.

[12] Z. Cao, T. Simon, S. Wei, and Y. Sheikh, "Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields," in *Proc. IEEE Conf. Computer Vision & Pattern Recognition*, 2017.

[13] G. Caridakis, K. Karpouzis, A. Drosopoulos, and S. Kollias, "Non parametric, self organizing, scalable modeling of spatiotemporal inputs: The sign language paradigm," *Neural Networks*, vol. 36, pp. 157–166, 2012.

[14] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," in *Proc. IEEE Conf. Computer Vision & Pattern Recognition*, 2017.

[15] V. Choutas, G. Pavlakos, T. Bolkart, D. Tzionas, and M. J. Black, "Monocular expressive body regression through body-driven attention," in *European Conference on Computer Vision (ECCV)*, 2020.

[16] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *arXiv preprint arXiv:1412.3555*, 2014.

[17] T. Cootes, G. J. Edwards, and C. Taylor, "Active Appearance Models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[18] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[19] R. Cui, H. Liu, and C. Zhang, "Recurrent convolutional neural networks for continuous sign language recognition by staged optimization," in *Proc. IEEE Conf. Computer Vision & Pattern Recognition*, July 2017.

[20] ——, "Recurrent convolutional neural networks for continuous sign language recognition by staged optimization," in *Proc. IEEE Conf. Computer Vision & Pattern Recognition*, 2017.

[21] ——, "A deep neural framework for continuous sign language recognition by iterative training," *IEEE Transactions on Multimedia*, vol. 21, pp. 1880–1891, 2019.

[22] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society. Series B (methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[23] J. Deng, W. Dong, R. Socher, L. Li-Jia, L. Kai, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," *Proc. IEEE Conf. Computer Vision & Pattern Recognition*, 2009.

[24] S. Geman and D. McClure, "Statistical methods for tomographic image reconstruction," *Bulletin of the International Statistical Institute*, vol. 52, no. 4, pp. 5–21, 1987.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision & Pattern Recognition*, 2016.

[26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–180, 1997.

[27] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Computer Vision & Pattern Recognition*, 2017.

[28] J. Huang, W. Zhou, Q. Zhang, H. Li, and W. Li, "Video-based sign language recognition without temporal segmentation," in *Proc. of the AAAI Conference on Artificial Intelligence*, 2018.

[29] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. of the International Conference on Machine learning (ICML)*, 2015.

[30] M. J. Islam, J. Mo, and J. Sattar, "Robot-to-robot relative pose estimation using humans as markers," arXiv, 2020.

[31] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[32] R. E. Johnson and S. K. Liddell, "A Segmental Framework for Representing Signs Phonetically," *Sign Language Studies*, vol. 11, no. 3, 2011.

[33] H. Joo, T. Simon, and Y. Sheikh, "Total Capture: A 3D Deformation Model for Tracking Faces, Hands, and Bodies," in *Proc. IEEE Conf. Computer Vision & Pattern Recognition*, 2018.

[34] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Proc. IEEE Conf. Computer Vision & Pattern Recognition*, 2018.

[35] S. Kim, K. Yun, J. Park, and J. Choi, "Skeleton-based action recognition of people handling objects," in *Proc. IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019.

[36] O. Koller, N. Camgoz, H. Ney, and R. Bowden, "Weakly supervised learning with multi-stream cnn-lstm-hmms to discover sequential parallelism in sign language videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 9, p. 2306–2320, 2020.

[37] O. Koller, N. C. Camgoz, H. Ney, and R. Bowden, "Weakly Supervised Learning with Multi-Stream CNN-LSTM-HMMs to Discover Sequential Parallelism in Sign Language Videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 9, pp. 2306–2320, 2020.

[38] O. Koller, J. Forster, and H. Ney, "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers," *Computer Vision and Image Understanding*, vol. 141, pp. 108–125, 2015.

[39] O. Koller, H. Ney, and R. Bowden, "Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled," in *Proc. IEEE Conf. Computer Vision & Pattern Recognition*, 2016.

[40] O. Koller, S. Zargaran, and H. Ney, "Re-Sign: Re-Aligned End-to-End Sequence Modelling with Deep Recurrent CNN-HMMs," in *Proc. IEEE Conf. Computer Vision & Pattern Recognition*, 2017.

[41] A. Kratimenos, K. Avramidis, C. Garoufis, N. Zlatintski, and P. Maragos, "Augmentation methods on monophonic audio for instrument classification in polyphonic music," in *Proc. in European Signal Processing Conference (EUSIPCO)*, 2020.

[42] A. Kratimenos, G. Pavlakos, and P. Maragos, "3D Hands, Face and Body Extraction for Sign Language Recognition," in *Proc. 16th European Conference on Computer Vision, ACVR Workshop*, 2020.

[43] ——, "Independent Sign Language Recognition with 3D Body, Hands, and Face Reconstruction," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2021.

[44] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS)*, 2012.

[45] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Conf. Computer Vision & Pattern Recognition*, 2006.

[46] C. Li, Q. Zhong, D. Xie, and S. Pu, "Co-occurrence Feature Learning from Skeleton Data for Action Recognition and Detection with Hierarchical Aggregation," in *Proc. International Joint Conferences on Artificial Intelligence*, 2018.

[47] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4D scans," *ACM Transactions on Graphics*, vol. 36, no. 6, pp. 194:1–194:17, 2017.

[48] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, vol. 34, no. 6, pp. 248:1–248:16, 2015.

[49] M. M. Loper, N. Mahmood, and M. J. Black, "MoSh: Motion and shape capture from sparse markers," *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, vol. 33, no. 6, pp. 220:1–220:13, 2014.

[50] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Computer Vision and Image Understanding*, vol. 60, no. 2, pp. 91–110, 2004.

[51] D. C. Luvizon, D. Picard, and H. Tabia, "2D/3D Pose Estimation and Action Recognition Using Multitask Deep Learning," in *Proc. IEEE Conf. Computer Vision & Pattern Recognition*, 2018.

[52] J. Nocedal and S. Wright, "Numerical optimization," *Springer*, 2016.

[53] H. N. O. Koller, S. Zargaran and R. Bowden, "Deep Sign: Hybrid CNN-HMM for Continuous Sign Language Recognition," in *Proc. of the British Machine Vision Conference (BMVC)*, 2016.

[54] S. Ong and S. Ranganath, "Automatic Sign Language Analysis: A Survey and the Future beyond Lexical Meaning," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, pp. 873–891, 2005.

[55] M. Parelli, K. Papadimitriou, G. Potamianos, G. Pavlakos, and P. Maragos, "Exploiting 3D Hand Pose Estimation in Deep Learning-Based Sign Language Recognition from RGB Videos," in *Proc. 16th European Conference on Computer Vision, SLRTP Workshop*, 2020.

[56] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3D hands, face, and body from a single image," in *Proc. IEEE Conf. Computer Vision & Pattern Recognition*, 2019.

[57] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, "Expressive Body Capture: 3D Hands, Face, and Body From a Single Image," in *Proc. IEEE Conf. Computer Vision & Pattern Recognition*, 2019.

[58] X. B. Peng, A. Kanazawa, J. Malik, P. Abbeel, and S. Levine, "Sfv: Reinforcement learning of physical skills from videos," *ACM Trans. Graph.*, vol. 37, no. 6, 2018.

[59] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele, "Deepcut: Joint subset partition and labeling for multi person pose estimation," in *Proc. IEEE Conf. Computer Vision & Pattern Recognition*, 2016.

[60] V. Pitsikalis, S. Theodorakis, C. Vogler, and P. Maragos, "Advances in Phonetics-based Sub-Unit Modeling for Transcription Alignment and Sign Language Recognition," in *CVPR Workshop*, Colorado Springs, USA, jun 2011.

[61] J. Pu, W. Zhou, and H. Li, "Iterative alignment network for continuous sign language recognition," in *Proc. IEEE Conf. Computer Vision & Pattern Recognition*, 2019.

[62] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," *ArXiv preprint*, 2021.

[63] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

[64] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, vol. 36, no. 6, pp. 245:1–245:17, 2017.

[65] A. Roussos, S. Theodorakis, V. Pitsikalis, and P. Maragos, "Dynamic Affine-Invariant Shape-Appearance Handshape Features and Classification in Sign Language Videos," *J. Machine Learning Research*, vol. 14, pp. 1627–1663, 2013.

[66] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand Keypoint Detection in Single Images Using Multiview Bootstrapping," in *Proc. IEEE Conf. Computer Vision & Pattern Recognition*, 2017.

[67] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.

[68] K. Snoddon, "Wendy sandler diane lillo-martin, sign language and linguistic universals. cambridge: Cambridge university press, 2006. pp. xxi, 547. pb $45.00." *Language in Society*, vol. 37, 2008.

[69] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research (JMLR)*, vol. 15, no. 56, pp. 1929–1958, 2014.

[70] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Computer Vision & Pattern Recognition*, 2015.

[71] M. Teschner, S. Kimmerle, B. Heidelberger, G. Zachmann, L. Raghupathi, A. Fuhrmann, M. Cani, F. Faure, N. Magnenat-Thalmann, W. Strasser, and P. Volino, "Collision Detection for Deformable Objects," in *Eurographics State-of-the-Art Report (EG-STAR)*, 2004.

[72] S. Theodorakis, V. Pitsikalis, and P. Maragos, "DynamicStatic Unsupervised Sequentiality, Statistical Subunits and Lexicon for Sign Language Recognition," *Image and Vision Computing*, vol. 32, pp. 533–549, 2014.

[73] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[74] H. Vaezi Joze and O. Koller, "MS-ASL: A Large-Scale Data Set and Benchmark for Understanding American Sign Language," in *Proc. British Machine Vision Conference*, 2019.

[75] C. Vogler and D. Metaxas, "A Framework for Recognizing the Simultaneous Aspects of American Sign Language," *Computer Vision and Image Understanding*, vol. 81, no. 3, pp. 358–384, 2001.

[76] U. von Agris, M. Knorr, and K. Kraiss, "The significance of facial features for automatic sign language recognition," in *Proc. 8th IEEE Conf. Automatic Face & Gesture Recognition*, 2008.

[77] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional Pose Machines," in *Proc. IEEE Conf. Computer Vision & Pattern Recognition*, 2016.