NATIONAL TECHNICAL UNIVERSITY OF ATHENS
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING
DIVISION OF SIGNALS, CONTROL AND ROBOTICS

Deep Affective Bodily Expression Recognition in the Presence of Medical Face Masks

Diploma Thesis

of

Nikolaos Kegkeroglou

Supervisor : Petros Maragos
Professor NTUA
Co-supervisor : Panagiotis P. Filntisis
Postdoctoral Researcher NTUA

COMPUTER VISION, SPEECH COMMUNICATION AND SIGNAL PROCESSING GROUP
Athens, March 2023

[This page intentionally left blank]

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

Deep Affective Bodily Expression Recognition in the Presence of Medical Face Masks

Diploma Thesis

of

Nikolaos Kegkeroglou

Supervisor : Petros Maragos
Professor NTUA
Co-supervisor : Panagiotis P. Filntisis
Postdoctoral Researcher NTUA

Approved by the examination committee on 13th March 2023.

....	....	....
Petros Maragos	Alexandros Potamianos	Athanasios Rontogiannis
Professor NTUA	Associate Professor NTUA	Associate Professor NTUA

Athens, March 2023

....
NIKOLAOS KEGKEROGLOU
Graduate of Electrical and
Computer Engineering NTUA

The content of this thesis does not necessarily reflect the views of the Department, the Supervisor, or the committee that approved it.

to my parents for their love & selfless support

[This page intentionally left blank]

Abstract

Emotions prepare people to deal with important events, and thus, play a vital role in their various relationships and decision-making. The possible applications of an interface capable of assessing human emotional states, make Emotion Recognition an exciting research field. There are three major contenders for the role of a general model on emotion mechanisms. The most common one is the basic emotion or categorical approach, that assumes the existence of a small, fixed number of discrete emotions. Emotional information is conveyed by a wide range of multimodal cues, facial expressions being the most used by the research community. In this thesis we focus on expression from the body, motivated by the fact, that the COVID-19 pandemic has undoubtedly changed the standards and affected all aspects of our lives, especially social life. Nowadays people extensively wear face masks, as it is one of the essential means to prevent the transmission of the pandemic. As a result, emotional reading from face can be strongly irritated by the presence of a mask. The type of algorithm that we use to tackle the problem of Emotion Recognition is Deep Learning, as these type of methods have yielded excellent results, due to the massive amounts of digital data in combination with powerful processing hardware, and on most cases has outperformed conventional machine learning methods. In this thesis, we wish to conduct insightful studies and come to fruitful applicative conclusions, regarding the area of Affective Bodily Expression Recognition. We adopt a proven deep learning-based visual recognition model called Temporal Segment Network and perform an experimental study about the face occlusion effect, caused by a face mask, on emotion recognition performance. This is achieved, by creating a medical face mask application tool and using it on a children emotion database, named EmoReact. We compare results based on the input modality and show, that although performance drops considerably with face, with full body we observe little to no decrease. Also, incorporating the whole body into the input, gives superior results over the plain masked face cropped image. We enhance our model with some proven techniques and almost fully overcome the face mask consequences, regarding performance. Lastly, as an essential step towards making a real-world emotion recognition interface, we create a real-time setup of the model, present multiple input versions of it and study the face mask effect in-the-wild.

Keywords — Emotion Recognition, Deep Learning, COVID-19, Real-Time, Medical Face Mask, Affective Bodily Expression Recognition

[This page intentionally left blank]

Περίληψη

Τα συναισθήματα προετοιμάζουν τους ανθρώπους να αντιμετωπίσουν σημαντικά γεγονότα και επομένως, παίζουν ζωτικό ρόλο στις διάφορες σχέσεις και τις αποφάσεις τους. Οι πιθανές εφαρμογές μιας διεπαφής, που είναι ικανή να αποτιμήσει τα ανθρώπινα συναισθήματα, κάνουν την Αναγνώριση Συναισθήματος ένα συναρπαστικό ερευνητικό τομέα. Υπάρχουν τρία υποψήφια μοντέλα για τους μηχανισμούς συναισθήματος. Το πιο συχνό είναι το μοντέλο βασικών συναισθημάτων ή κατηγορικό, το οποίο υποθέτει την ύπαρξη ενός μικρού, σταθερού αριθμού από διακριτά συναισθήματα. Η συναισθηματική πληροφορία μεταφέρεται από ένα ευρύ φάσμα από πολυτροπικές μορφές, με τις εκφράσεις του προσώπου να είναι η πιο δημοφιλής ερευνητικά. Σε αυτή την εργασία εστιάζουμε στην έκφραση μέσω του σώματος, με κίνητρο το γεγονός ότι η πανδημία COVID-19 έχει αδιαμφησβήτητα αλλάξει τα δεδομένα και έχει επηρεάσει όλες τις πλευρές της ζωή μας, ειδικά της κοινωνικής. Σήμερα οι άνθρωποι φορούν εκτεταμένα μάσκες προσώπου, αφού είναι ένα από τα αναγκαία μέσα για τον περιορισμό της εξάπλωσης της πανδημίας. Αυτό ίσως έχει ως αποτέλεσμα, η αναγνώριση συναισθήματος μέσω του προσώπου να γίνει πολύ δυσκολότερη από την παρουσία μιας μάσκας. Ο τύπος αλγορίθμου που χρησιμοποιούμε για να προσεγγίσουμε το πρόβλημα είναι η Βαθιά Μάθηση, αφού αυτού του είδους οι μέθοδοι έχουν επιφέρει εξαιρετικά αποτελέσματα, εξαιτίας του τεράστιου όγκου ψηφιακών δεδομένων σε συνδυασμό με ισχυρές υπολογιστικές μονάδες, και στις περισσότερες περιπτώσεις έχουν ξεπεράσει σε επίδοση τις μεθόδους συμβατικής μηχανικής μάθησης. Σε αυτή την εργασία, επιθυμούμε να πραγματοποιήσουμε διορατικές μελέτες και να καταλήξουμε σε χρήσιμα και εφαρμόσιμα συμπεράσματα, σχετικά με την περιοχή της Αναγνώρισης Συναισθηματικών Εκφράσεων του Σώματος. Υιοθετούμε ένα αποδεδειγμένο μοντέλο οπτικής αναγνώρισης βασισμένο στη βαθιά μάθηση, και μελετάμε την επίδραση της απόκρυψης του προσώπου, που προκαλείται από μια μάσκα, στην επίδοση αναγνώρισης συναισθήματος. Δημιουργούμε ένα εργαλείο εφαρμογής ιατρικής μάσκας προσώπου και το χρησιμοποιούμε πάνω σε μία βάση δεδομένων με παιδιά, την EmoReact. Συγκρίνουμε τα αποτελέσματα των διαφορετικών μορφών πληροφορίας σαν είσοδο και δείχνουμε, ότι παρ΄όλο που με το πρόσωπο η επίδοση πέφτει σημαντικά, με το σώμα παρατηρούμε ελάχιστη έως καθόλου χειροτέρευση. Επίσης, συμπεριλαμβάνοντας ολόκληρο το σώμα στην είσοδο, πετυχαίνουμε καλύτερα αποτελέσματα απ' ότι μόνο με το πρόσωπο. Στη συνέχεια, ενισχύουμε το μοντέλο με αποδεδειγμένες τεχνικές και ξεπερνάμε σχεδόν πλήρως τις συνέπειες της μάσκας. Τέλος, ως αναγκαίο βήμα για να φτιάξουμε μια πραγματική διεπαφή αναγνώριση συναισθήματος, δημιουργούμε ενα μοντέλο σε περιβάλλον πραγματικού χρόνου, παρουσιάζοντας εκδοχές του και μελετώντας την επίδραση της μάσκας σε παραδείγματα πραγματικού κόσμου.

Λέξεις-κλειδιά — Αναγνώριση Συναισθήματος, Βαθιά Μάθηση, COVID-19,
Ιατρική Μάσκα Προσώπου, Πραγματικός Χρόνος,
Αναγνώριση Συναισθηματικών Εκφράσεων του Σώματος

[This page intentionally left blank]

Acknowledgements

To begin with, I would like to thank Prof. Petros Maragos for teaching the courses “Signals and Systems”, “Digital Signal Processing” and “Computer Vision”, that piqued my interest, and for giving me the opportunity to conduct my Diploma Thesis in the Computer Vision, Speech Communication and Signal Processing Laboratory (CVSP). Then, I would like to thank Dr. Panagiotis Filntisis for his valuable contribution, guidance and advice during this thesis. Finally, I thank the people close to me for their support throughout the entirety of my studies.

Nikolaos Kegkeroglou
March 2023

[This page intentionally left blank]

Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω τον Καθηγητή Πέτρο Μαραγκό για τη διδασκαλία των μαθημάτων “Σήματα και Συστήματα”, “Ψηφιακή Επεξεργασία Σήματος” και “Όραση Υπολογιστών”, που μου κέντρισαν το ενδιαφέρον, καθώς και για την ευκαιρία που μου έδωσε να εκπονήσω τη Διπλωματική μου Εργασία στο Εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων. Στη συνέχεια, θα ήθελα να ευχαριστήσω τον Δρ. Παναγιώτη Φιλντίση για την πολύτιμη συνεισφορά, την καθοδήγηση και τις συμβουλές του κατά τη διάρκεια αυτής της εργασίας. Τέλος, ευχαριστώ τους κοντινούς μου ανθρώπους για την υποστήριξή τους καθ’ όλη τη διάρκεια των σπουδών μου.

Νικόλαος Κεγκέρογλου
Μάρτιος 2023

[This page intentionally left blank]

Εκτεταμένη Περίληψη

1 Εισαγωγή

Σύμφωνα με τον Paul Ekman, τα συναισθήματα είναι μια διαδικασία, κατά την οποία οι άνθρωποι αισθάνονται ότι συμβαίνει κάτι σημαντικό σχετικά με την ευεξία τους, και ένα σύνολο ψυχολογικών αλλαγών και συναισθηματικών συμπεριφορών ξεκινά να χειρίζεται την κατάσταση. Με άλλα λόγια, τα συναισθήματα προετοιμάζουν τους ανθρώπους να αντιμετωπίσουν σημαντικά γεγονότα και επομένως, παίζουν ρόλο ζωτικής σημασίας στις διάφορες σχέσεις και τις αποφάσεις τους [1].

1.1 Συναισθηματική Υπολογιστική

Τα συναισθήματα έχουν παραδοσιακά μελετηθεί από ψυχολόγους εδώ και πάνω από έναν αιώνα, και έχουν κυριως συσχετιστεί με την αντίληψη του μυαλού [2] [3]. Μόνο πρόσφατα, με τη βοήθεια νέων αισθητήρων μέτρησης, οι υπολογιστές μπορούν να καταγράψουν και να επεξεργαστούν συναισθηματικά χαρακτηριστικά, πετυχαίνοντας αυτό που αποκαλείται Συναισθηματική Υπολογιστική [4]. Παρ' όλο που τα ανθρώπινα αισθήματα δεν μπορούν να προσεγγιστούν άμεσα, η Picard, μία πρωτοπόρος του πεδίου, αναφέρεται στην "αναγνώριση συναισθήματος" ή την "αναγνώριση συναισθηματικής κατάστασης", ως μέτρηση παρατηρήσιμων συναρτήσεων τέτοιων καταστάσεων. Οπότε άν οι παρατηρήσεις κάπου αντιστοιχούν με μεγάλη πιθανότητα σε ένα υποκείμενο συναίσθημα ή συνδυασμό συναισθημάτων, αυτές οι παρατηρήσεις μπορούν να χρησιμοποιηθούν για να εξαχθούν αυτές οι καταστάσεις [5].

1.2 Θεωρίες των Συναισθημάτων

Σύμφωνα με έρευνες στην ψυχολογία, υπάρχουν τρία υποψήφια μοντέλα για τους μηχανισμούς συναισθήματος: το μοντέλο βασικών συναισθημάτων ή κατηγορικό, το διαστατικό, και το συνδυαστικό. Μία εκτενής έρευνα για τα παραπάνω, μπορεί να βρεθεί στα [6] and [7]. Ακολούθως, περιγράφουμε τις βασικές αρχές τους και σημειώνουμε τα κύρια πλεονεκτήματα και μειονεκτήματά τους.

1.2.1 Κατηγορική Προσέγγιση

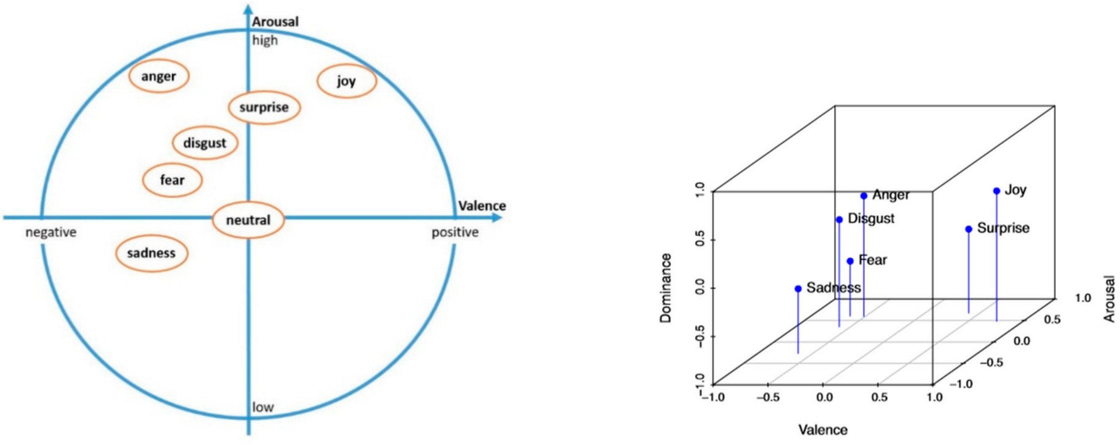
Η κατηγορική προσέγγιση βασίζεται στην ερμηνεία που έδωσε ο Tomkin [8] [9] στον ισχυρισμό του Darwin [3], που υποθέτει την ύπαρξη ενός μικρού, σταθερού αριθμού από διακριτά συναισθήματα. Αυτή η υπόθεση έχει υποστηριχτεί από ερευνητικό έργο του Paul Ekman και λοιπών [10] [11], στο οποίο διεξήγαγαν διάφορα πειράματα για την ανθρώπινη κρίση πάνω σε φωτογραφίες με επιτηδευμένα απεικονισμένη συμπεριφορά με

Κατηγορικό Μοντέλο Συναισθημάτων

το πρόσωπο και συμπέραναν ότι έξι βασικά συναισθήματα μπορούν να αναγνωριστούν καθολικά: χαρά (happiness), λύπη (sadness), έκπληξη (surprise), φόβος (fear), θυμός (anger), αηδία (disgust). [12] [13] (Σχήμα 1.1). Αυτή η θεωρία είναι η πιο συχνά χρησιμοποιούμενη στη βιβλιογραφία, παρά τη δυσκολία της να χειριστεί περίπλοκες συναισθηματικές καταστάσεις ή μίξη συναισθημάτων [14]. Έχει, επίσης, λάβει κριτική από τον James Russell, ο οποίος εξέθεσε κάποιες ασάφειές του [15] και πρότεινε, αντ' αυτού, ένα διαστατικό μοντέλο, το οποίο αναλύουμε παρακάτω.

1.2.2 Διαστατική Προσέγγιση

Τα μοντέλα συναισθημάτων με διαστατική προσέγγιση, υποθέτουν ότι οι συναισθηματικές καταστάσεις δεν είναι ανεξάρτητες μεταξύ τους, αλλά συνδέονται με συστηματικό τρόπο [16]. Σύμφωνα με έρευνες [17], η πλειονότητα της ποικιλίας των συναισθημάτων καλύπτεται από τρεις διαστάσεις: valence, arousal, dominance. Το valence αναφέρεται στο πόσο θετικό ή αρνητικό είναι το αναπαριστούμενο συναίσθημα και κειμένεται από άχαρα έως χαρούμενα αισθήματα. Το arousal αναφέρεται στο πόσο ενθουσιώδες ή απαθές είναι το συναίσθημα και κειμένεται από νύστα/βαρεμάρα έως μανιώδη ενθουσιασμό. Τέλος, το dominance αναφέρεται στο κατά πόσο το υποκείμενο έχει τον έλεγχο της κατάστασης [18]. Χρησιμοποιώντας αυτές τις διαστάσεις, μπορούμε να αναπαραστήσουμε το χώρο του συναισθήματος (Σχήμα 1.2). Παρ' όλο που αυτή η θεωρία έχει υπάρξει χρήσιμη σε εφαρμογές [19], έχει προκύψει ο ισχυρισμός, ότι συναισθήματα όπως η "σύγχηση" δεν μπορούν να απεικονιστούν με ξεκάθαρο τρόπο, ο "φόβος" και η "αηδία" γίνονται δυσδιάκριτα, και χάνεται πληροφορία όταν μειώνουμε το χώρο σε λίγες διαστάσεις.



V-A: 2D (credit: [20])

V-A-D: 3D (credit: [21])

Ο Χώρος του Διαστατικού Μοντέλου Συναισθημάτων

1.2.3 Συνδυαστική Προσέγγιση

Η ανάγκη για ένα σύστημα που αποφεύγει τους περιορισμούς της κατηγορικής και της διαστατικής προσέγγισης, οδήγησε τον Scherer και συναδέλφους του να προτείνουν μια άλλη προσέγγιση, βασισμένη στη θεωρία εκτίμησης [22]. Τα συνδυαστικά μοντέλα, όπως αποκαλούνται, εστιάζουν στη μεταβλητότητα των συναισθηματικών καταστάσεων, όπως προκύπτουν από τους διαφορετικούς τύπους μοτίβων εκτίμησης. Αυτή η προσέγγιση είναι η λιγότερο μελετημένη από τις τρεις, κυρίως λόγω της απαίτησης ενός περίπλοκου, πολυσυστατικού και εκλεπτυσμένου τρόπου μέτρησης των αλλαγών.

1.3 Τρόποι Έκφρασης Συναισθήματος

Η συναισθηματική πληροφορία μεταφέρεται από ένα ευρύ φάσμα από πολυτροπικές μορφές, τις εκφράσεις του προσώπου, την κίνηση και στάση του σώματος, την ομιλία και τη γλώσσα, και ούτω καθεξής. Παρ' όλ' αυτά, σε αυτή την εργασία εστιάζουμε στην έκφραση μέσω του σώματος.

1.3.1 Πρόσωπο

Οι εκφράσεις του προσώπου (Σχήμα 1.3) είναι ένας από τους πιο σημαντικούς τρόπους επικοινωνίας για τη μεταφορά πληροφορίας για τη συναισθηματική κατάσταση του ανθρώπου και είναι ο πιο μελετημένος από την ερευνητική κοινότητα για την αναγνώριση συναισθήματος. Οι περισσότερες παραδοσιακές μέθοδοι έχουν χρησιμοποιήσει "χειροποίητα" χαρακτηριστικά [23], ενώ πιο πρόσφατες προσεγγίσεις χρησιμοποιούν βαθιά μάθηση [24] (Υποενότητα 2.1).

1.3.2 Σώμα

Ενώ ερευνητικές εργασίες βασισμένες στις εκφράσεις του προσώπου βρίθουν, η αναγνώριση συναισθήματος μέσω της κίνησης και στάσης του σώματος παραμένει ένα λιγότερο εξερευνημένο αντικείμενο. Πρόσφατες έρευνες στη νευροβιολογία έχουν δείξει ότι η στάση και κίνηση του σώματος περιέχουν χρήσιμα χαρακτηριστικά για την αναγνώριση του ανθρώπινου συναισθήματος [25] [26]. Σε άλλα πειράματα, αποδείχθηκε ότι οι

(credit)

Εκφράσεις Προσώπου

Παίκτρια Τένις

Απόκρυψη Προσώπου

Σωματικές Εκφράσεις

εκφράσεις μέσω προσώπου και σώματος λειτουργούν συμπληρωματικά για την οπτική αντίληψη του συναισθήματος. Όταν ζητήθηκε στους συμμετέχοντες να προβλέψουν το νικητή ενός επαγγελματικού αγώνα τένις κοιτώντας σε απομονωμένες εικόνες προσώπου, απέδωσαν χειρότερα σε σύγκριση με όταν τους παρουσιάστηκαν οι ίδιες εικόνες που περιείχαν ολόκληρο το ανθρώπινο σώμα [27]. Μία ενδιαφέρουσα παρατήρηση είναι ότι όταν τα πρόσωπα και τα σώματα ανταλλάχθηκαν με τα αντίθετα συναισθήματα, οι συμμετέχοντες χρησιμοποίησαν το σώμα για να προβλέψουν το αποτέλεσμα, το οποίο υπονοεί ότι οι άνθρωποι αντιλαμβάνονται τη συναισθηματική πληροφορία εκφρασμένη με το σώμα ως περισσότερο διαγνωστική απ' ότι με το πρόσωπο (Σχήμα 1.4a). Επιπρόσθετα, η αναγνώριση σωματικών εκφράσεων είναι κρίσιμη όταν τα πρόσωπα δεν είναι διαθέσιμα. Η ορατότητα των χαρακτηριστικών του προσώπου ενός κοινωνικού συντρόφου δεν είναι εγγυημένη, για παράδειγμα όταν τα επίπεδα φωτισμού πέφτουν, άτομα φορούν αξεσουάρ όπως μάσκες (1.4b) (εκτενώς χρησιμοποιούμενες στις μέρες μας εξαιτίας της πανδημίας COVID-29 [28]), γυρνούν τα κεφάλια ή η απόσταση μεταξύ παρατηρητή και παρατηρούμενου προσώπου αυξάνεται [29]. Επειδή τα σώματα είναι μεγαλύτερα και πιο εκφραστικά από τα πρόσωπα σε αυτές τις καταστάσεις, οι παρατηρητές μπορούν αντ' αυτού να εντοπίσουν κοινωνική πληροφορία από τις σωματικές κινήσεις. Η σωματική έκφραση μπορεί επίσης να χρησιμοποιηθεί ως μια βοηθητική ροή πληροφορίας μαζί με το πρόσωπο [30], για να τη σωστή αποσαφήνιση της αντίστοιχης έκφρασης.

1.4 Εφαρμογές

Οι πιθανές εφαρμογές μιας διεπαφής ικανής να αποτιμήσει ανθρώπινες συναισθηματικές καταστάσεις είναι πολυάριθμες, μερικές από τις οποίες αναφέρονται παρακάτω.

1.4.1 Αλληλεπίδραση Ανθρώπου-Ρομπότ

Έρευνα των Reeves και Nass [32], έδειξε ότι οι άνθρωποι γενικά συμπεριφέρονται στους υπολογιστές όπως πιθανόν θα συμπεριφέρονταν σε άλλους ανθρώπους. Τα ρομπότ και τα συστήματα που είναι ικανά να αναγνωρίσουν, να ερμηνεύσουν και να επεξεργαστούν το ανθρώπινο συναίσθημα [33], θα μπορούσαν να ταιριάξουν καλά σε αυτό το σενάριο, κάνοντας την αλληλεπίδραση πιο επιδραστική και ευχάριστη. Ειδικά όταν αλληλεπιδρούν με παιδιά (Αλληλεπίδραση Ρομπότ-Παιδιού) [34], είναι ακόμη πιο κρίσιμο τα ρομπότ να έχουν εμπάθεια, λόγω των ιδιαιτεροτήτων των παιδιών [35].

1.4.2 Μάθηση Υποβοηθούμενη από Υπολογιστή

Η μάθηση είναι η υπέρτατη συναισθηματική εμπειρία [36]. Ένα στιγμιότυπο μάθησης μπορεί να ξεκινήσει με περιέργεια και έντονο ενδιαφέρον. Όσο όμως η δυσκολία αυξάνεται, μπορεί κανείς να βιώσει σύγχηση, θυμό ή αγχος, και τελικά να εγκαταλείψει τη μάθηση [37] Ένα ρομπότ-παιδαγωγός, που είναι σε θέση να εκτιμήσει τη συναισθηματική κατάσταση του μαθητευόμενου, μπορεί να ανταποκριθεί κατάλληλα και να δώσει ενθαρρυντικές συμβουλές. Υπάρχον έργο έχει δείξει ότι τα ρομπότ-παιδαγωγοί ενισχύουν τη διαδικασία της μάθησης, προσωπικοποιώντας τις ενθαρρυντικές τους στρατηγικές στη συναισθηματική συμπεριφορά του μαθητευόμενου [38] [39].

1.4.3 Φροντίδα της Υγείας

Οι διαταραχές ψυχικής υγείας, όπως η κατάθλιψη και η ψύχωση, είναι ανά τον κόσμο σε άνοδο. Τα συστήματα αναγνώρισης συναισθήματος μπορούν να αποτελέσουν μια αποδοτική στρατηγική για την πρόληψη, την παρακολούθηση και την περίθαλψη τέτοιων διαταραχών. Πρόσφατες εφαρμογές περιέχουν μια αυτόματη δομή για αναγνώριση συναισθήματος βασισμένη στη γλώσσα του σώματος για πρόβλεψη ψυχιατρικών συμπτωμάτων [40], καθώς και μία φορετή συσκευή που αναλύει την κατάσταση στρες και συναισθήματος του χρήστη και του περιβάλλοντός του, χρησιμοποιώντας αλγόριθμους αναγνώρισης συναισθήματος με βάση εκφράσεις προσώπου και σημάτων φυσιολογίας [41].

1.4.4 Ψηφιακή Ψυχαγωγία

Η αναγνώριση συναισθήματος έχει βρει εύφορο έδαφος και στην περιοχή της ψηφιακής ψυχαγωγίας, καθώς έχει προσφέρει εκτίμηση αισθητικής εμπειρίας χρηστών [42] σε βιντεοπαιχνίδια που χρησιμοποιούν ολόκληρο το σώμα.

1.5.3 Προκλήσεις Σωματικής Έκφρασης

Ποιότητα Επισημειώσεων Συνόλου Δεδομένων

Στην ανάλυση του προσώπου, οι εκφράσεις μπορούν να κωδικοποιηθούν με την κίνηση συγκεκριμένων μυών του προσώπου, χρησιμοποιώντας το Facial Action Coding System (FACS) [50], που αναπτύχθηκε από τους Ekman και Friesen. Αυτό το σύστημα είναι καθιερωμένο και ευρέως χρησιμοποιούμενο από την ερευνητική κοινότητα. Παρ' όλ' αυτά, δεν υπάρχει ένα αντίστοιχο σύστημα για τη σωματική έκφραση και τις κινήσεις του σώματος. Ο Dael και λοιποί πρότειναν το σύστημα κωδικοποίησης Body Action and Posture (BAP) [51], το οποίο διαχωρίζει τις σωματικές μονάδες δράσης από αυτές της στάσης, που ήταν μια αξιοσημείωτη προσπάθεια για ένα αξιόπιστο σύστημα κωδικοποίησης. Άλλες προσεγγίσεις μελέτησαν τη σχέση μεταξύ συναισθηματικών καταστάσεων και μιας υψηλού-επιπέδου (π.χ. επιτηδευμένες κινήσεις και στάσεις μπαλέτου) ή χαμηλού-επιπέδου (π.χ. κίνηση άρθρωσης) περιγραφή είτε κίνησης είτε στάσης [52]. Δυστυχώς, οι περισσότερες από αυτές έχουν βασιστεί σε ένα περιορισμένο σύνολο επιτηδευμένων σωματικών εκφράσεων και μας λείπουν ακόμη καθιερωμένα περιεκτικά πρωτόκολλα. Όπως έχει προταθεί [53], η ασυνέχεια των συναισθηματικών συνόλων δεδομένων που έχει προκύψει από πλυθοπορισμό υπάρχει λόγω της πιθανής αναξιοπιστίας των συμμετεχόντων που προσλαμβάνονται και της φυσικής μεταβλητότητας της αντίληψης των συναισθηματικών εκφράσεων των άλλων ανθρώπων. Ως αποτέλεσμα, η επισημείωση σωματικής έκφρασης γίνεται μία πρόκληση ακόμα και για ειδικούς του παιδιού και επομένως, η συμφωνία μεταξύ των επίσημειωτών είναι γενικά όχι πολύ υψηλή σε υπάρχοντα σύνολα δεδομένων.

Πολυπλοκότητα Σωματικής Πόζας

Ένα ανθρώπινο σώμα είναι μία εκλεπτυσμένη οντότητα με αρθρώσεις και άκρα, και περιέχει κινηματική δομή, αλλά και πληροφορία σχήματος. Ο ακριβής προσδιορισμός των συντεταγμένων των αρθρώσεων και η επεξεργασία ενός μεγάλου αριθμού από βαθμούς ελευθερίας είναι ένα από τα αρχαιότερα προβλήματα στην όραση υπολογιστών, εξαιτίας της πολυπλοκότητας των μοντέλων που σχετίζονται με την παρατήρηση της πόζας.

Μεταβλητότητα

Επιπλέον τεχνικές δυσκολίες που μπορεί να αντιμετωπίσουμε είναι το υψηλό επίπεδο ετερογένειας στην συμπεριφορά των ανθρώπων, το θορυβώδες φόντο, και συχνά οι μη αμελητέες διαφορές στην κλίμακα, η οπτική και άλλες ρυθμίσεις κάμερας. Επίσης, υπάρχει μεγάλη μεταβλητότητα που εισάγεται από τις παραλλαγές στην ανθρώπινη εμφάνιση λόγω του ρουχισμού, του σωματικού σχήματος, του μεγέθους και της κόμης. Επιπρόσθετα, αποτελέσματα ερευνών [54], Υποδεικνύουν την ανάγκη να θεωρήσουμε την κουλτούρα ως ένα συγκεκριμένο παράγοντα σχεδιασμού για ένα σύστημα αναγνώρισης σωματικών εκφράσεων. Ενώ υπάρχουν ομοιότητες μεταξύ των πολιτισμών, στον τρόπο με τον οποίο μεταφέρουν, αναγνωρίζουν, και αποτυπώνουν την συναισθηματική έννοια στην πόζα, υπάρχουν διαφορές. Τα αποτελέσματα τους δείχνουν ότι τα συναισθήματα είναι και καθολικά και συγκεκριμένα για κάθε πολιτισμό και όταν κάνουμε αξιολόγηση σε επίπεδο πολιτισμού, κάποιες κουλτούρες είναι πιο όμοιες μεταξύ τους από αλλες.

2.1 Βαθιά Μάθηση

Εισάγουμε κάποιες βασικές έννοιες της βαθιάς μάθησης, αφού είναι το είδος αλγόριθμο που θα χρησιμοποιήσουμε σ' αυτή την εργασία. Οι τεχνικές συμβατικής μηχανικής μάθησης [55] ήταν περιορισμένες στην ικανότητά τους να επεξεργαστούν φυσικά δεδομένα στην ωμή μορφή τους. Ο στόχος ήταν η σχεδίαση ενός εξαγωγέα χαρακτηριστικών που θα μετέτρεπε τα ωμά δεδομένα (όπως τις τιμές πίξελ μιας εικόνας) σε μία κατάλληλη εσωτερική αναπαράσταση ή ένα διάνυσμα χαρακτηριστικών από το οποίο ένα σύστημα μάθησης, συχνά ένας ταξινομητής, θα μπορούσε να εντοπίσει ή να κατηγοριοποιήσει μοτίβα στην είσοδο. Για δεκαετίες, η δημιουργία ενός συστήματος μηχανικής–μάθησης ή αναγνώρισης-προτύπων χρειαζόταν προσεκτική μηχανική και σημαντική εξειδίκευση. Η Βαθιά Μάθηση [56] επιτρέπει στα υπολογιστικά μοντέλα που συντίθενται από πολλαπλά υπολογιστικά επίπεδα να μαθαίνουν αναπαραστάσεις δεδομένων με πολλαπλά επίπεδα αφαιρετικότητας. Αυτές οι μέθοδοι έχουν βελτιώσει δραματικά τα σύγχρονα δεδομένα τεχνολογίας στην αναγνώρισης ομιλίας, την οπτική αναγνώριση αντικειμένων, τον εντοπισμό αντικειμένων και σε πολλά άλλα πεδία, όπως την ανακάλυψη φαρμάκων και τη γονιδιωματική.

Κατηγοριοποίηση

Η κατηγοριοποίηση είναι ένα πρόβλημα μηχανικής μάθησης, όπου ζητείται από το μοντέλο να προσδιορίσει σε ποια από τις k κατηγορίες ανήκει η είσοδος, με την οποία το τροφοδοτούμε. Για να λύσουμε αυτό το πρόβλημα, ο αλγόριθμος μάθησης συνήθως πρέπει να παράξει μια συνάρτηση $f : \mathbb{R}^n \to \{1, ..., k\}$. Όταν $y = f(x)$, το μοντέλο αναθέτει την είσοδο, που περιγράφεται από το διάνυσμα x, σε μια κατηγορία που προσδιορίζεται από τον κωδικό αριθμό y.

Επιβλεπόμενη Μάθηση

Η πιο συνήθης μορφή μηχανικής μάθησης, βαθιάς η μη, είναι η επιβλεπόμενη μάθηση. Φανταστείτε ότι θέλουμε να χτίσουμε ένα σύστημα που κατηγοριοποιεί εικόνες, οι οποίες περιέχουν, ας πούμε, ένα σπίτι, ένα αυτοκίνητο, έναν άνθρωπο ή ενα κατοικίδιο. Πρώτα, συλλέγουμε ένα μεγάλο σύνολο δεδομένων από εικόνες με σπίτια, αυτοκίνητα, ανθρώπους και κατοικίδια, κάθε μία από τις οποίες είναι επισημειωμένη με την κατηγορία στην οποία ανήκει. Κατά τη διάρκεια της εκπαίδευσης, τροφοδοτείται στο μηχάνημα μία εικόνα και εκείνο παράγει μία έξοδο στην μορφή ενός διανύσματος με σκορ, ένα για κάθε κατηγορία. Ιδανικά, θέλουμε η επιθυμητή κατηγορία να έχει το υψηλότερο σκορ από όλες, αλλά αυτό είναι απίθανο να συμβεί πριν την εκπαίδευση.

Συνάρτηση Κόστους

Η εκπαίδευση πραγματοποιείται υπολογίζοντας μία αντικειμενική συνάρτηση, η οποία μετράει το σφάλμα (ή την απόσταση) μεταξύ των σκορ της εξόδου και του επιθυμητού μοτίβου εξόδου. Ύστερα, το μηχάνημα τροποποιεί τις εσωτερικές παραμέτρους του για να μειώσει αυτό το σφάλμα. Αυτές οι τροποποιούμενες παράμετροι, βάρη όπως αποκαλούνται, είναι πραγματικοί αριθμοί που καθορίζουν την συνάρτηση εισόδου–εξόδου του μηχανήματος. Σε ένα τυπικό σύστημα, μπορεί να υπάρχουν εκατοντάδες εκατομμύρια από αυτά τα τροποποιούμενα βάρη και εκατοντάδες εκατομμύρια επισημειωμένα παραδείγματα με τα οποία εκπαιδεύεται το μηχάνημα.

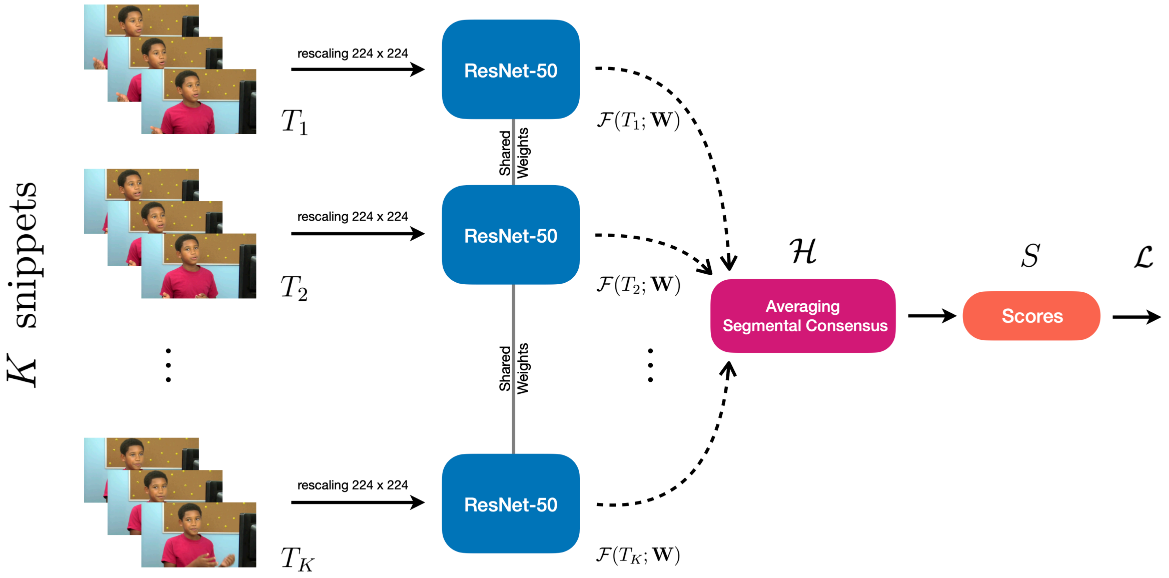
Σύνολο Δεδομένων

Το σύνολο δεδομένων από εικόνες που συλλέγεται, διαχωρίζεται στην συνέχεια σε τρία υποσύνολα: εκπαίδευσης, επαλήθευσης και αξιολόγησης. Το σύνολο εκπαίδευσης χρησιμοποιείται, για να τροποποιηθούν κατάλληλα τα βάρη ενός διαχωριστή βαθιάς μάθησης, κατά τη διάρκεια της διαδικασίας μάθησης. Το σύνολο επαλήθευσης χρησιμοποιείται για να ρυθμιστούν οι υπερπαράμετροι (π.χ. η αρχιτεκτονική), ενώ το σύνολο επαλήθευσης χρησιμοποιείται μόνο για να δούμε την επίδοση του μοντέλου.

3 Μοντέλο Οπτικής Αναγνώρισης Συναισθήματος

3.1 Βάση Δεδομένων EmoReact

Αν και έχει γίνει σημαντικός όγκος έρευνας στην αυτόματη αναγνώριση συναισθημάτων σε ενήλικες, η αναγνώριση συναισθημάτων σε παιδιά δεν είναι τόσο μελετημένη. Διαλέγουμε την EmoReact (Σχήμα 2.20) ως το σύνολο δεδομένων μας και επομένως μελετάμε



Αρχιτεκτονική Αρχικού Μοντέλου

ένα πρόβλημα αλληλεπίδρασης Ρομπότ-Παιδιού. Εκτός από την δυσκολία που βάζει η φύση του προβλήματος, μία ακόμη πρόκληση που αντιμετωπίζουμε είναι η ανισορροπία των δειγμάτων της βάσης (Σχήμα 3.1).

3.2 Αρχιτεκτονική

Οι πολύπλοκες δράσεις, όπως η έκφραση συναισθημάτων, αποτελούνται από πολλαπλά στάδια τα οποία εκτείνονται με κάποια χρονική διάρκεια και θα ήταν σημαντική απώλεια αν αποτυγχάναμε να τα αξιοποιήσουμε. Από την άλλη, όπως προαναφέραμε, κάθε συναίσθημα που εκφράζεται δεν είναι παρόν καθ' όλη τη διάρκεια του βίντεο εισόδου. Αυτά υποδεικνύουν ότι έχουμε ανάγκη για ένα τρόπο αποδοτικής καταγραφής γενικών χαρακτηριστικών. Υιοθετούμε την δομή Temporal Segment Network (TSN) [117], η οποία αντί να επεξεργάζεται ολόκληρο το βίντεο, καθώς και κάθε εικόνα ξεχωριστά, λειτουργεί πάνω σε μία σειρά από μικρής διάρκειας κομμάτια του, αραιά δειγματοληπτημένα από αυτό. Κάθε κομμάτι της σειράς θα παράξει τις δικές του πρώιμες προβλέψεις για τις κατηγορίες συναισθήματος και στη συνέχεια, θα βγει μία ομόφωνη προβλέψει ως το τελικό αποτέλεσμα. Επομένως, αυτή η δομή επιτρέπει στο μοντέλο να προσπελαύνει διάφορα κομμάτια του βίντεο, αλλά και να αντιμετωπίσει την αδυναμία της απλής αρχιτεκτονικής δικτύου να μοντελοποιήσει μακράς διαρκείας χρονική συσχέτιση, οπότε και είναι πιο πιθανό να παρατηρήσει το αντίστοιχο συναίσθημα. Έτσι αγνοεί πλεονάζουσα πληροφορία, οπότε αποφεύγει το overfitting και προσφέρει μία είδους επαύξηση δεδομένων, ενώ ταυτόχρονα, μειώνει τον υπολογιστικό φόρτο. Ως αρχιτεκτονική κορμού του δικτύου, υιοθετούμε την ResNet, η οποία υλοποιεί συνδέσεις μεταξύ επιπέδων, δηλαδή προωθεί την έξοδο κάποιων στην είσοδο κάποιων άλλων. Τα κύρια πλεονεκτήματά της είναι: (i) μειώνει την επίδραση του προβλήματος vanishing gradient, δημιουργώντας αυτή την απεικόνιση ταυτότητας, (ii) αυξάνει την ταχύτητα της εκπαίδευσης, και (iii) δεν αυξάνει τις παραμέτρους προς εκπαίδευση.

3.3 Μετρική Αξιολόγησης

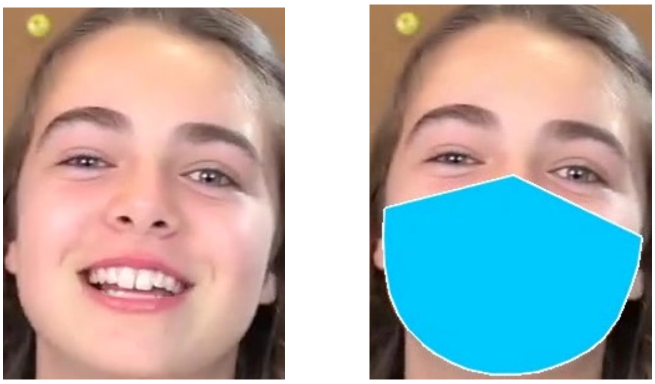
Η μόνη μετρική αξιολόγησης που έχει αποδειχθεί ότι είναι εύρωστη σε ανισόρροπα σύνολα δεδομένων είναι η Area Under the Curve of Receiver Operating Characteristic [118] (ROC AUC). Είναι ένα γράφημα του True Positive Rate (TPR) ή Recall, ως συνάρτηση του False Positive Rate (FPR) και δείχνει την απόδοση ενός μοντέλου κατηγοριοποιήσης σε όλα τα κατώφλια αποφάσης (Σχήμα 3.5).

4 Μελέτη Επίδρασης Ιατρικής Μάσκας Προσώπου

Η πανδημία COVID-19 έχει αδιαμφησβήτητα αλλάξει τα δεδομένα και έχει επηρεάσει όλες τις πλευρές της ζωή μας, ειδικά την κοινωνική. Σήμερα οι άνθρωποι φορούν εκτεταμένα μάσκες προσώπου, αφού είναι ένα από τα αναγκαία μέσα για τον περιορισμό της εξάπλωσης της πανδημίας. Αυτό ίσως έχει ως αποτέλεσμα, η αναγνώριση συναισθήματος μέσω του προσώπου να γίνει πολύ δυσκολότερη από την παρουσία μιας μάσκας. Με κίνητρο το γεγονός αυτό, δημιουργούμε ένα εργαλείο εφαρμογής ιατρικής μάσκας προσώπου και το χρησιμοποιούμε πάνω στην EmoReact, για να μελετήσουμε την επίδραση της απόκρυψης του προσώπου λόγω της μάσκας, στην επίδοση αναγνώρισης συναισθήματος. Για τον εντοπισμό της επιφάνειας του προσώπου και του σώματος σε πραγματικό χρόνο, χρησιμοποιούμε το Mediapipe (Υποενότητα 4.2).

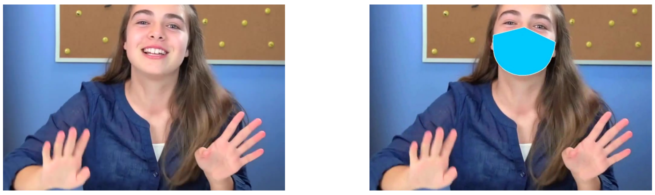
Αποτελέσματα Επίδρασης

Με πρώτη ματιά, παρατηρούμε ότι, με είσοδο το πρόσωπο, η επίδοση πέφτει σημαντικά (≈ 3-4%). Αυτό είναι ένα αποτέλεσμα που περιμέναμε, καθώς η μάσκα καλύπτει την πλειονότητα του προσώπου, συμπεριλαμβανομένου του πιο εκφραστικού χαρακτηριστικού του, του προσώπου.



Αποτελέσματα Επίδρασης Μάσκας με Είσοδο το Πρόσωπο

Κομμάτια	**ROC AUC**		**Επίδραση**
	Αρχική	Μάσκα	
1	0.755	0.728	−2.7%
3	0.769	0.733	−3.6%
5	0.767	0.732	−3.5%
10	0.770	0.741	−2.9%



Αποτελέσματα Επίδρασης Μάσκας με Είσοδο Ολόκληρο το Σώμα

Κομμάτια	**ROC AUC**		**Επίδραση**
	Αρχική	Μάσκα	
1	0.752	0.752	-
3	0.759	0.758	−0.1%
5	0.758	0.754	−0.4%
10	0.761	0.759	−0.2%

Διαισθητικά, αν κανείς προσπαθήσει να προβλέψει τα συναισθήματα που εκφράζονται στις δυο εικόνες, υποψιαζόμαστε ότι θα είχε καλύτερη τύχη, χωρίς την παρουσία της μάσκας. Κοιτώντας τα αποτελέσμα με είσοδο ολόκληρο το σώμα, η πρώτη και πιο σημαντική παρατήρηση που κάνουμε, είναι ότι η χειροτέρευση της επίδοσης είναι πολύ μικρή έως καθόλου (0-0.4%). Αυτά τα αποτελέσματα υποδεικνύουν ότι το μοντέλο μπορεί να εκμεταλλευτεί τη σωματική πληροφορία με τέτοιο τρόπο, ώστε ακόμη και με την εφαρμογή μιας μάσκας προσώπου, και επομένως την απώλεια πληροφορίας, χάνει μόνο ένα μηδαμινό ποσοστό επίδοσης.

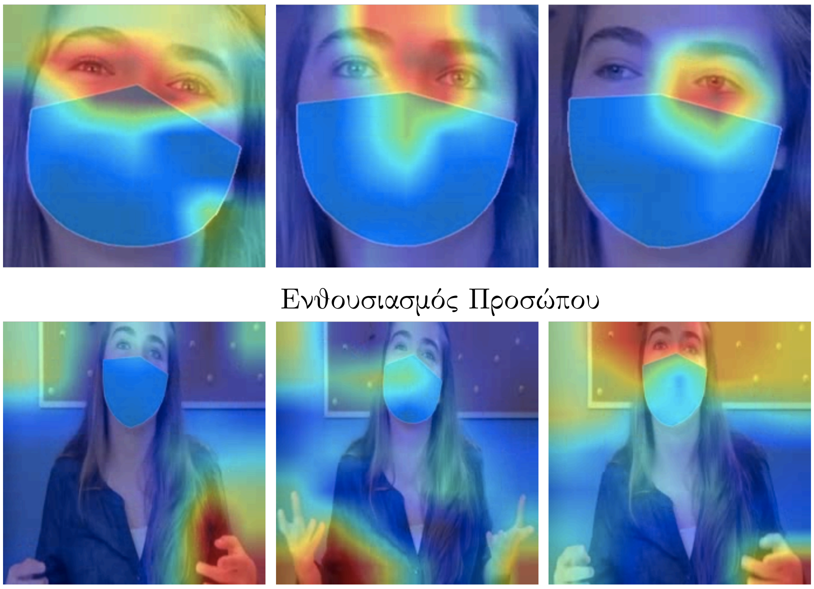
Συγκρίνουμε την επίδοση του μοντέλου με είσοδο το πρόσωπο και με είσοδο όλο το σώμα, όταν έχει εφαρμοστεί η μάσκα, και δείχνουμε ότι συμπεριλαμβάνοντας ολόκληρο το σώμα στην είσοδο, πετυχαίνουμε καλύτερα αποτελέσματα απ' ότι με μόνο το πρόσωπο. Το προφανές συμπέρασμα που βγάζουμε είναι πως η καλύτερη επιλογή σε περιπτώσεις απόκρυψης του προσώπου είναι η κατευθυνση προς την αναγνώριση σωματικών εκφράσεων.

Σύγκριση Αποτελεσμάτων Προσώπου και Ολόκληρου Σώματος με Μάσκα

Κομμάτια	**ROC AUC**		**Επίδραση**
	Πρόσωπο	Ολόκληρο Σώμα	
1	0.728	0.752	+2.4%
3	0.733	0.758	+2.5%
5	0.732	0.754	+2.2%
10	0.741	**0.759**	+1.8%

Οπτική Επεξήγηση

Για να έχουμε μια καλύτερη κατανόηση της επίδρασης της μάσκας, χρησιμοποιούμε μια τεχνική για την παραγωγή οπτικών επεξηγήσεων για τις προβλέψεις. Θέλουμε να εξερευνήσουμε που εστιάζει το μοντέλο στην εικόνα εισόδου και πως η συμπεριφορά του διαφέρει για διαφορετικές κατηγορίες συναισθημάτων. Η μέθοδος που διαλέγουμε λέγεται Gradient-weighted Class Activation Mapping (Grad-CAM) [130], η οποία χρησι-μοποιεί τις παραγώγους μιας κατηγορίας συναισθήματος, που ρέουν μέχρι το τελευταίο επίπεδο του δικτύου για να παράγει ένα τραχύ χάρτη εντοπισμού, τονίζοντας τις σημαντι-κές περιοχές της εικόνας για να γίνει η πρόβλεψη. Παρέχουμε κάποια παραδείγματα, όπου η ένταση, και άρα η εστίαση του μοντέλου, αυξάνεται από το μπλε στο κόκκινο χρώμα. Βλέπουμε ότι για το πρόσωπο, το μοντέλο εστιάζει στο άνω μέρος του, πιθανώς χρησιμοποιώντας χαρακτηριστικά όπως τα μάτια, τα φρύδια και το μέτωπο. Γενικά, το μοντέλο καταφέρνει να αγνοήσει θορυβώδη χαρακτηριστικά, όπως η μάσκα και το φόντο. Μιλήσαμε επίσης νωρίτερα και για τη μεγάλη μεταβλητότητα που εισάγεται από τις παραλλαγές στην ανθρώπινη εμφάνιση, τις οποίες δυσκολίες το μοντέλο ξεπερνά και εστιάζει σε εκφραστικά χαρακτηριστικά. Τέλος, σε παράδειγμα του σώματος, βλέπουμε ότι το μοντέλο εστιάζει ταυτόχρονα και στο σώμα και στο πρόσωπο, κάνοντας μίξη πληροφορίας διαφορετικών τρόπων έκφρασης σε ένα και μοναδικό stream RGB εικόνας.



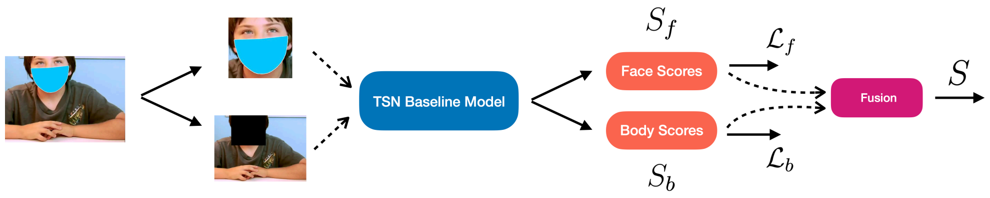
Ενθουσιασμός Προσώπου

Ενθουσιασμός Σώματος

Περιοχές Απόφασης Ενθουσιασμού

5 Ενίσχυση Μοντέλου

Τα αποτελέσματα του αρχικού μοντέλου είναι μια βάση, πάνω στην οποία χτίζουμε και επιδιώκουμε βελτιώσεις, με τελικό στόχο να πετύχουμε την επίδοση που έχει το μοντέλο, με την αρχική είσοδο, δηλαδή χωρίς μάσκα. Η πρώτη τεχνική που αξιοποιούμε είναι το Temporal Shift Module (TSM) [131], ένα γενικό και αποδοτικό εργαλείο,



Σύστημα Μίξης Σκορ Τρόπων Έκφρασης

που απολαμβάνει ταυτόχρονα υψηλή αποδοτικότητα και επίδοση. Μπορεί να εισαχθεί σε δίκτυα για να προσδώσει χρονική μοντελοποίηση με κανένα παραπάνω υπολογισμό και καμία παράμετρο προς εκπαίδευση. Το TSM μεταθέτει ένα μέρος των καναλιών του τανυστή εισόδου, κατά τη χρονική διάσταση, και προς τα μελλοντικά και προς τα παρελθοντικά frame (Σχήμα 5.1). Δεν μεταθέτει ολα τα κανάλια, λόγω αποφυγής μεταφοράς μεγάλου όγκου δεδομένων, άρα και δέσμευσης πόρων. Επίσης, για να διατηρηθεί η χωρική αντίληψη του μοντέλου, που χάνεται με τη μετάθεση, τοποθετούμε το TSM μέσα στο κλαδί του ResNet, ώστε να διατηρηθεί η απεικόνιση ταυτότητας. Από τη σύγκριση των αποτελεσμάτων, παρατηρούμε ότι η τεχνική αυτή αυξάνει ελάχιστα την επίδοση και καταλήγουμε στο συμπέρασμα ότι η μάθηση χωρικών χαρακτηριστικών είναι πιο σημαντική για μια συναισθηματική έκφραση, ενώ η χρονική δομή παίζει συμπληρωματικό ρόλο. Για περαιτέρω ενίσχυση του μοντέλου, εκμεταλλευόμαστε την πληροφορία του προσώπου και του σώματος ξεχωριστά. Προτείνουμε να διαχωρίσουμε τα χαρακτηριστικά τους, ώστε να αποφύγουμε πιθανή σύγχηση πληροφοριών από τους διαφορετικούς τρόπους έκφρασης. Ο κορμός του μοντέλου παραμένει ίδιος, αλλά τώρα επεξεργάζεται την εικόνα προσώπου, και την εικόνα σώματος με σβησμένο το πρόσωπο, σε δύο ξεχωριστές τροφοδοτήσεις. Αφού παραχθούν τα σκορ S_f και S_b από το πρόσωπο και το σώμα αντίστοιχα, γίνεται μίξη για να αποκτήσουμε τα τελικά σκορ S. Πειραματιζόμαστε με δύο συναρτήσεις μίξης: μέγιστο και μέσο. Η κατάλληλη συνάρτηση μίξης φαίνεται να είναι το μέσο, καθώς το μέγιστο δίνει σχετικά κακά αποτελέσματα, που ίσως οφείλονται σε αποκλεισμό σωστών αρνητικών προβλέψεων του ενός τρόπου έκφρασης, από λανθασμένων θετικών προβλέψεων του άλλου. Συνδυάζοντας και τις δύο τεχνικές, δηλαδή TSM με 1/4 μερική μετάθεση και τη μέθοδο μίξης, πετυχαίνουμε το αποτέλεσμα 0.768 ROC AUC, δηλαδή παρόμοιο με το 0.769 που είχαμε χωρίς μάσκα.

6 Αναγνώριση Πραγματικού Χρόνου

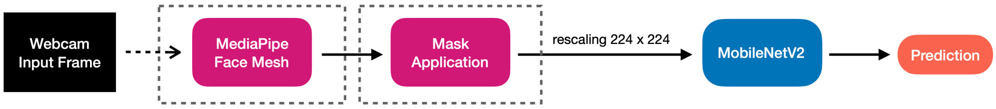
Ένα αναγκαίο βήμα για να φτιάξουμε μια πραγματική διεπαφή αναγνώριση συναισθήματος είναι να την κάνουμε να τρέχει σε πραγματικό χρόνο. Έτσι, δημιουργούμε ενα μοντέλο σε περιβάλλον πραγματικού χρόνου με δύο εκδοχές, είτε με είσοδο το πρόσωπο, είτε το σώμα. Έως τώρα κάναμε τα πειράματά μας στην EmoReact, που είναι μια βάση με μικρό αριθμό δειγμάτων. Έτσι, το μοντέλο μας δεν έχει πλήρη εικόνα του πραγματικού κόσμου, οπότε δε θα έτρεχε αποτελεσματικά σε ένα τέτοιο σενάριο. Έτσι, εκπαιδεύουμε το μοντέλο σε μία μεγαλύτερη βάση με ένα πιο ευρύ φάσμα από δείγματα. Για το πρόσωπο χρησιμοποιούμε την AffectNet [132], που είναι με διαφορά η μεγαλύτερη βάση εκφράσεων προσώπου με περισσότερα από ένα εκατομμύριο δείγματα. Για το σώμα, χρησιμοποιούμε την BoLD [100], μια μεγάλη βάση με βίντεο σε σενάρια πραγματικού κόσμου, που περιέχει δέκα χιλιάδες βίντεο με ανθρώπους εκφράζοντας συναισθήματα, κυρίως μέσω σωματικών κινήσεων. Εκτός από τα σύνολα δεδομένων, αλλάζουμε και την αρχιτεκτονική κορμού του δικτύου. Η αναγνώριση πραγματικού χρόνου απαιτεί ελαφρά δίκτυα, ώστε ο χρόνος απόκρισης να είναι χαμηλός. Το ResNet-50 που χρησιμοποιούμε, παρ' όλα τα πλεονεκτήματα που έχει, είναι υπολογιστικά βαρύ και δε μπορεί να τρέξει αποδοτικά. Γι' αυτό το αντικαθιστούμε με το MobileNetV2 [126], που έχει σχεδιαστεί ειδικά για εφαρμογές με περιορισμένους υπολογιστικούς πόρους. Στον παρακάτω πίνακα φαίνεται η σύγκριση των δύο αρχιτεκτονικών κορμού και παρακάτω η γραμμή επεξεργασίας του online demo. Αρχικά, η εικόνα από την κάμερα προωθείται στο εργαλείο Face Mesh του MediaPipe, για να εξαχθούν τα σημεία του προσώπου και να περικοπεί το πρόσωπο. Στη συνέχεια, εφαρμόζεται προαιρετικά η ιατρική μάσκα προσώπου και η εικόνα αλλάζει κλίμακα, ώστε να ταιριάζει στις απαιτήσεις του μοντέλου. Το μοντέλο μας είναι ένα απλό MobileNetV2, που παράγει την πρόβλεψη του εκφρασμένου συναισθήματος. Στην περίπτωση του σώματος, παραλείπεται το βήμα του Face Mesh και αντ' αυτού, τοποθετούμε την κάμερα και το χρήστη στην επιθυμητή απόσταση, για να παρέχουμε πληροφορία με το σώμα.

Σύγκριση Αποτελεσμάτων Διαφόρων Συνδυασμών

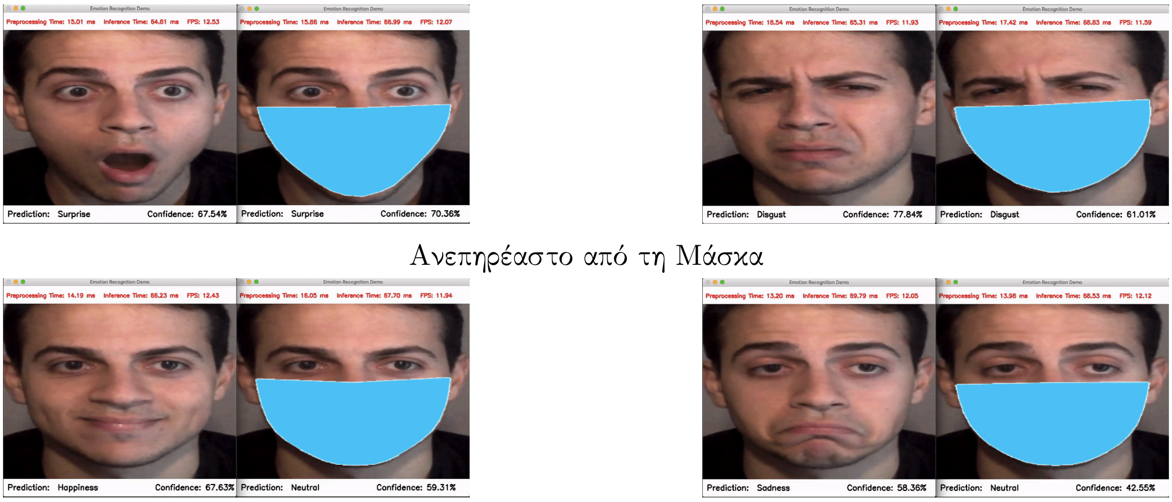
Μοντέλο	**Είσοδος - 3 Κομμάτια**	**Μετάθεση**	**Συνάρτηση**	**ROC AUC**
TSN	Πρόσωπο με Μάσκα	-	-	0.733
TSN	Σκέτο Σώμα	-	-	0.736
TSN	Ολόκληρο Σώμα με Μάσκα	-	-	0.758
TSM	Ολόκληρο Σώμα με Μάσκα	1/8	-	0.762
		1/4	-	0.763
TSN	Μίξη	-	Μέγιστο	0.758
		-	Μέσο	0.758
TSM	Μίξη	1/8	Μέγιστο	0.729
			Μέσο	0.767
		1/4	Μέγιστο	0.731
			Μέσο	**0.768**
TSN	Πρόσωπο χωρίς Μάσκα	-	-	0.769

Σύγκριση Κορμών για Αναγνώριση Συναισθήματος σε Πραγματικό Χρόνο

Κορμός	**Παράμετροι**	**Απόκριση**	**FPS**	**Ακρίβεια (%)**	
				ImageNet	AffectNet
ResNet-50	25.0M	170ms	∼6	77.6	57.9
MobileNetV2	3.5M	60ms	∼17	71.9	57.2



Γραμμή Επεξεργασίας Online Demo



```
Preprocessing Time: 15.01 ms   Inference Time: 64.81 ms   FPS: 12.53
Prediction:  Surprise          Confidence: 67.54%

Preprocessing Time: 15.86 ms   Inference Time: 66.99 ms   FPS: 12.07
Prediction:  Surprise          Confidence: 70.36%

Preprocessing Time: 18.54 ms   Inference Time: 65.31 ms   FPS: 11.93
Prediction:  Disgust           Confidence: 77.84%

Preprocessing Time: 17.42 ms   Inference Time: 68.83 ms   FPS: 11.59
Prediction:  Disgust           Confidence: 61.01%
```

Ανεπηρέαστο από τη Μάσκα

```
Preprocessing Time: 14.19 ms   Inference Time: 66.23 ms   FPS: 12.43
Prediction:  Happiness         Confidence: 67.63%

Preprocessing Time: 16.05 ms   Inference Time: 67.70 ms   FPS: 11.94
Prediction:  Neutral           Confidence: 59.31%

Preprocessing Time: 13.20 ms   Inference Time: 69.79 ms   FPS: 12.05
Prediction:  Sadness           Confidence: 58.36%

Preprocessing Time: 13.98 ms   Inference Time: 68.53 ms   FPS: 12.12
Prediction:  Neutral           Confidence: 42.55%
```

Επηρεασμένο από τη Μάσκα

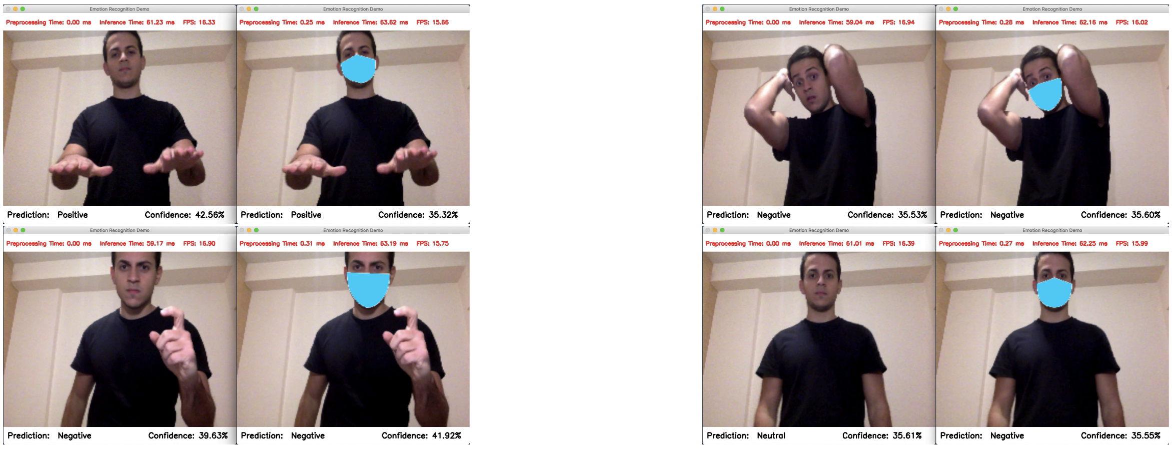
Παραδείγματα Αναγνώρισης Συναισθήματος σε Πραγματικό Χρόνο μέσω Προσώπου

Πρόσωπο

Η εκδοχή του demo με πρόσωπο κατηγοριοποιεί μεταξύ των 8 συναισθημάτων της AffectNet. Παρακάτω, παρουσιάζουμε κάποια παραδείγματα αναγνώρισης σε πραγματικό χρόνο σε άγνωστα δείγματα, όπου συγκρίνουμε τις προβλέψεις μεταξύ αρχικής και εισόδου με μάσκα προσώπου. Ενώ το μοντέλο προβλέπει σωστά κάποια συναισθήματα, παρατηρούμε ότι είναι επιρρεπές σε λάθη όταν μπαίνει η μάσκα.

Σώμα

Η εκδοχή του demo με σώμα κατηγοριοποιεί μεταξύ των 26 κλάσεων της BoLD, η οποία περιέχει πολύ περίπλοκα συναισθήματα. Σε αυτή την περίπτωση, θα ήταν πολύ δύσκολο για το μοντέλο να χειριστεί ένα τόσο μεγάλο αριθμό από κατηγορίες, οπότε τις απεικονίζουμε στις 3 γενικές κατηγορίες: Θετικό, Ουδέτερο, Αρνητικό. Παρακάτω, παρουσιάζουμε κάποια παραδείγματα για το body demo και παρατηρούμε ότι παραμένει ανεπηρέαστο από τη μάσκα, στην εργασία της αναγνώρισης.



```
Preprocessing Time: 0.00 ms   Inference Time: 61.23 ms   FPS: 16.33
Prediction:  Positive         Confidence: 42.56%

Preprocessing Time: 0.25 ms   Inference Time: 63.62 ms   FPS: 15.66
Prediction:  Positive         Confidence: 35.32%

Preprocessing Time: 0.00 ms   Inference Time: 59.04 ms   FPS: 16.94
Prediction:  Negative         Confidence: 35.53%

Preprocessing Time: 0.28 ms   Inference Time: 62.16 ms   FPS: 16.02
Prediction:  Negative         Confidence: 35.60%

Preprocessing Time: 0.00 ms   Inference Time: 59.17 ms   FPS: 16.90
Prediction:  Negative         Confidence: 39.63%

Preprocessing Time: 0.31 ms   Inference Time: 63.19 ms   FPS: 15.75
Prediction:  Negative         Confidence: 41.92%

Preprocessing Time: 0.00 ms   Inference Time: 61.01 ms   FPS: 16.39
Prediction:  Neutral          Confidence: 35.61%

Preprocessing Time: 0.27 ms   Inference Time: 62.25 ms   FPS: 15.99
Prediction:  Negative         Confidence: 35.55%
```

Παραδείγματα Αναγνώρισης Συναισθήματος σε Πραγματικό Χρόνο μέσω Σώματος

Contents

List of Acronyms

CRI	Child-Robot Interaction
CNN	Convolutional Neural Network
ABER	Affective Bodily Expression Recognition
STGCN	Spatial-Temporal Graph Convolutional Network
ResNet	Residual Network
TSN	Temporal Segments Network
ROC AUC	Area Under the Curve of Receiver Operating Characteristic
TSM	Temporal Shift Module

[This page intentionally left blank]

List of Figures

List of Tables

[This page intentionally left blank]

Chapter 1

Introduction

According to Paul Ekman, emotions are a process, in which people sense that something important to their welfare is occuring, and a set of psychological changes and emotional behaviors begins to deal with the situation. In other words, emotions prepare people to deal with important events, and thus, play a vital role in their various relationships and decision-making [1].

1.1 Affective Computing

Emotion or affect has traditionally been studied by psychologists, since over a century ago, and has mainly been associated with the mind's perception [2] [3]. Only recently, with the aid of new measuring sensors, computers can capture and process affect features, achieving what is called Affective Computing [4]. While people's feelings cannot be directly accessed, Picard, one of the pioneers in the area, refers to "emotion recognition" or "recognition of affective state", as measuring observable functions of such states. Hence, if one's observations correspond with high probability to an underlying emotion or combination of emotions, these observations may be used to infer the states themselves [5].

1.2 Theories of Emotion

According to the research in psychology, there are three major contenders for the role of a general model on emotion mechanisms: the basic emotion or categorical, the dimensional, and the componential appraisal models. An extensive review of the above, can be found in [6] and [7]. Subsequently, we describe their basic principles and highlight their main advantages and drawbacks.

1.2.1 Categorical Approach

The categorical approach is based on Tomkins's interpretation [8] [9] of Darwin's account [3] that assumes the existence of a small, fixed number of discrete emotions. This assumption has been supported by research work of Paul Ekman et al. [10] [11], in which various experiments were conducted on human judgement of still photographs of intentionally displayed facial behavior and concluded that six basic

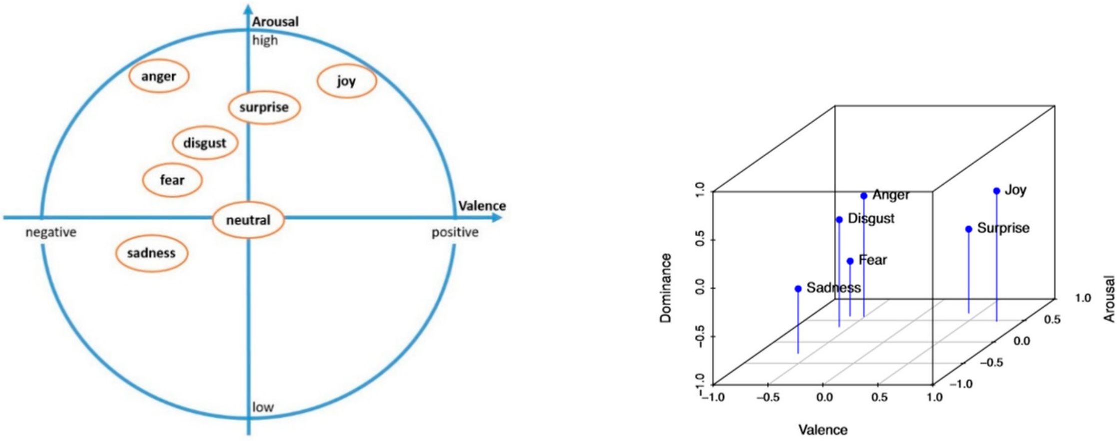
happiness
sadness
surprise
fear
anger
disgust

Figure 1.1: The Categorical Emotion Model

emotions can be recognized universally: happiness, sadness, surprise, fear, anger and disgust [12] [13] (Fig. 1.1). This theory has been the most commonly adopted in the literature, despite its struggle handling complex affective state or blended emotions [14]. It has, also, received criticism by James Russell, who exposed some of its uncertainties [15] and proposed a dimensional model, instead, which we analyze below.

1.2.2 Dimensional Approach

Emotion models with a dimensional approach, assume that affectives states are not independent from one another, but are related in a systematic manner [16]. According to research work [17], the majority of affect variability is covered by three dimensions: valence, arousal and dominance. Valence refers to how positive or negative the represented emotion is, ranging from unpleasant to pleasant feelings. Arousal refers to how excited or apathetic the emotion is, ranging from sleepiness/boredom to frantic excitement. Lastly, dominance refers to the subject's control level of the situation [18]. Using these dimensions, we can represent the space of emotion (Fig. 1.2). Although this theory has been shown useful in applications [19], it has been claimed, that emotions like "confusion" cannot be clearly depicted, "fear" and "disgust" become indistinguishable, and there is loss of information when reducing the emotion space in low dimensions.



(a) V-A: 2D (credit: [20])

(b) V-A-D: 3D (credit: [21])

Figure 1.2: The Dimensional Emotion Model Space

1.2.3 Componential Approach

The need of a system that avoids the restrictions of the categorical and the dimensional approach, has led Scherer and colleagues to propose another approach, based on appraisal theory [22]. Componential models, as they are called, focus on the variability of different emotional states, as produced by different types of appraisal patterns. This approach has been the least investigated of the three, mainly due to its require of complex, multicomponential and sophisticated measurements of change.

1.3 Emotion Modalities

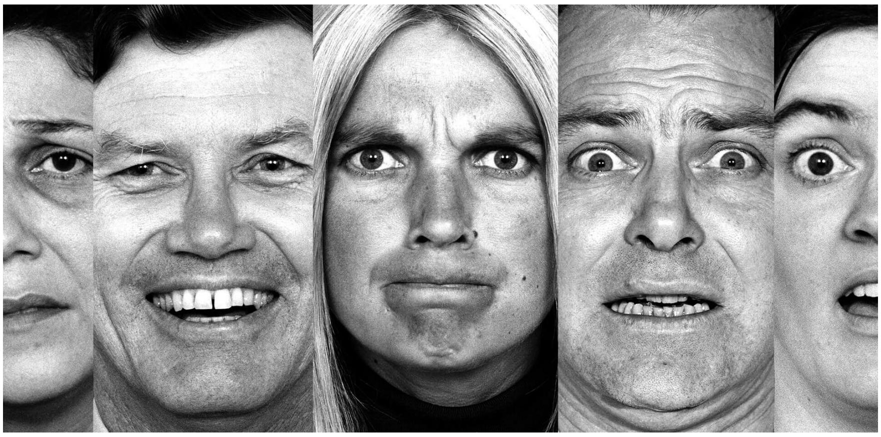
Emotional information is conveyed by a wide range of multimodal cues, including facial expression, body movement and posture, speech and language, and so forth. In this thesis, however, we will focus on emotion recognition from bodily expression.

1.3.1 Face

Facial expressions (Fig. 1.3) are one of the most important forms of communication used to convey information about one's emotional state and have been the most commonly used modality for emotion recognition, by the research community. Most traditional methods have used handcrafted features [23], whereas more recent approaches use deep learning [24] (see 2.1).

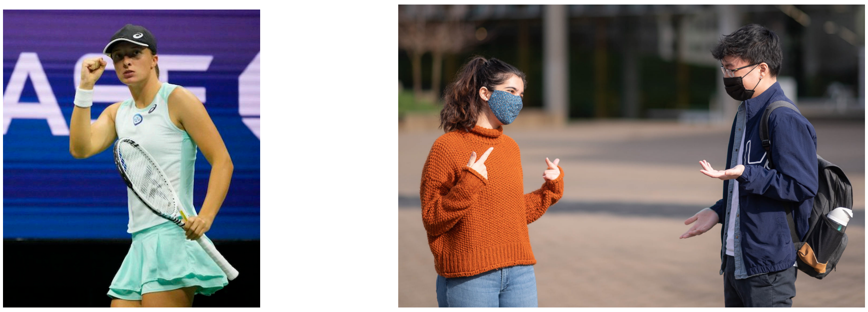
1.3.2 Body

While works based on facial expressions abound, recognizing affect from body movement and posture remains a less explored topic. Recent studies in neurobiology have shown that body posture and movement contain useful features for recognizing human affect [25] [26]. In other experiments, it was shown that facial and bodily expressions work complementary for visual perception of emotion. When the participants were asked to predict the winner of a professional tennis game by looking at



(credit)

Figure 1.3: Facial Expressions



(a) Winning Tennis Player

(b) Face Occlusion

Figure 1.4: Bodily Expressions

isolated face images, they performed worse than when the whole images containing the whole bodies were presented [27]. An interesting point, is that when the faces and the bodies were paired oppositely, the participants still used the body to predict the outcome, which suggests that humans perceive bodily expressed emotional information as more diagnostic than facial (Fig. 1.4a). Furthermore, bodily expression recognition is crucial when facial cues are not available. The visibility of a social partner's facial features is not guaranteed, for example, when light levels decrease, individuals wear accessories like masks (Fig. 1.4b) (extensively used nowadays due to COVID-19 pandemic [28]), turn their heads or distance between the observer and an observed face increases [29]. Because bodies are bigger and more expressive than faces in those situations, observers can detect social information from bodily motions, instead. Bodily expression can also be used as an auxiliary stream of information besides the face [30], to correctly disambiguate the corresponding facial expression.

1.3.3 Speech

Speech is the most natural method of communication between humans. The field of automatic speech recognition has tremendously progressed with valuable research contributions, including statistical methods like hidden markov models, finite-state transducer-based systems like Kaldi, as well as deep neural networks. However, human-computer interaction through speech still does not feel natural, due to the computer's inability of recognizing human affect. This has introduced a new research field, namely speech emotion recognition, which aims at extracting the emotional state of the speaker. A detailed survey can be found in [31].

1.4 Applications

The possible applications of an interface capable of assessing human emotional states are numerous, some of which are mentioned below.

1.4.1 Human-Robot Interaction

Research by Reeves and Nass [32], has shown that humans generally treat computers as they might treat other people. Robots and systems that are able to recognize, interpret and process human affect [33], are arguably well suited to this, making the interaction more effective and pleasant. Especially when interacting with children (Child-Robot Interaction - CRI) [34], it is even more crucial to have empathic robots, due to the children's particularities [35].

1.4.2 Computer-Assisted Learning

Learning is the quintessential emotional experience [36]. A learning episode might begin with curiosity and fascination. But as its difficulty increases, one may experience confusion, frustration or anxiety, and thus, may abandon learning [37]. A tutoring agent, who is able to estimate the learner's affective state, can respond appropriately and give encouraging suggestions. Existing work has shown that robot tutors enhance learning, by personalizing their motivational strategies to the student's emotional behavior [38] [39].

1.4.3 Health Care

Mental health disorders, like depression and psychoses, are on the rise across the world. Emotion recognition systems can be an effective strategy for preventing, monitoring and treating such disorders. Recent applications include an automated framework for body language based emotion recognition for psychiatric symptom prediction [40] and a wearable device that analyzes stress status and emotions of the user and their environment, using facial and physiological signal emotion recognition algorithms [41].

1.4.4 Digital Entertainment

Recognition of affect has also found fertile ground in the area of digital entertainment, allowing for aesthetic experience estimation [42] in full-body computer games.

1.5 Bodily Expression

1.5.1 Body Form vs Movement

According to studies [43] [44], there are two separate pathways in the brain for recognizing biological information, one for form information (i.e., the description of the configuration of a stance) and one for motion information. The study of Lange and Lappe [45] claims that *"...a model that analyses global form information and then integrates the form information temporally"* can best explain results from psychophysical experiments of biological motion perception. They argue that information about the temporal development of the movement is only used if necessary to resolve inconsistencies and if it is essential to the type of task. This argument is also supported by previous research work [46], indicating that form information can be instrumental in the recognition of biological motion. In Fig. 1.5, we show some sample frames from the Geneva Multimodal Emotion Portrayals (GEMEP) corpus dataset [47], which may be considered as characteristic body forms for the corresponding emotions.

However, a recent study by Atkinson et al. [48], determined that both form and motion signals are assessed for affect perception from the body. Specifically, the authors concluded that motion signals can be sufficient for recognizing basic emotions, but that recognition accuracy is significantly impaired when the form information is disrupted by inverting and reversing the clip. In Fig. 1.6, we show some consecutive frames from body movements that express joy and anger. One can notice it is easier to recognize emotion from a sequence of frames (form & movement), compared to looking at a single frame (form). Summarizing, both form and motion information is useful and important for Affective Bodily Expression Recognition.

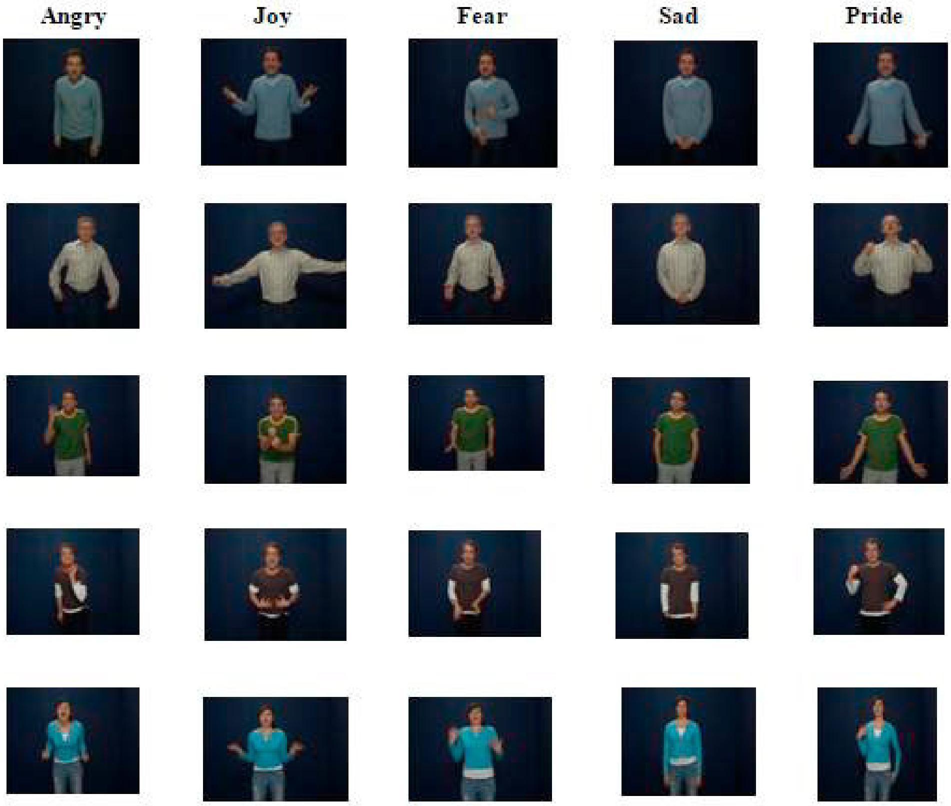
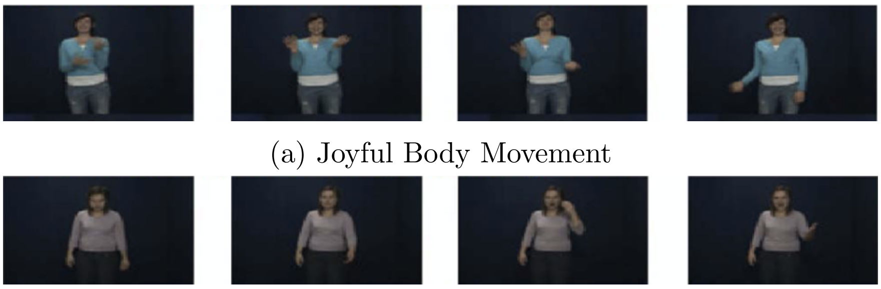


Figure 1.5: Body Forms



(a) Joyful Body Movement

(b) Angry Body Movement

Figure 1.6: Body Movements

1.5.2 Hand Expression

Research exploring emotion recognition from the body tends to refer to “the body” as a whole entity. This is surprising given the important role that hands play in our daily activities, as they are the way we interact with the world and could be differentially contributing to emotion recognition. In recent research work [49], they investigated the influence of hands on emotion recognition from the body and was shown that removing the hands, significantly reduced recognition accuracy for fearful and angry body stimuli, which suggest the hands may play a key role in Affective Bodily Expression Recognition.

1.5.3 Challenges

Dataset Annotation Quality

In facial analysis, facial expressions can be encoded with movement of individual muscles of the face, using the Facial Action Coding System (FACS) [50], developed by Ekman and Friesen. This system has been established and widely used by the research community. However, there is no such gold standard for bodily expression and body movements. In [51], Dael et al. proposed the Body Action and Posture (BAP) coding system, that discriminated body action units from body posture units, which was a remarkable attempt for a reliable coding system. Other approaches have investigated the relationship between affective states and a high-level (e.g. acted ballet movements and postures) or low-level (e.g. joint rotation) description of either movement or form [52]. Unfortunately, most of the above have relied on a limited set of acted body expressions and we still lack comprehensive protocols. As suggested in [53], inconsistency of crowdsourced affective datasets exists due to the possible untrustworthiness of recruited participants and the natural variability of humans perceiving others’ affective expressions. As a result, annotating bodily expression becomes challenging even for experts of the field and thus, inter-annotator agreement is generally not high in existing datasets.

Body Pose Complexity

A human body is a sophisticated entity with joints and limbs, and contains body kinematic structure, as well as body shape information. Identifying fine-grained joint coordinates and working with a large number of degrees of freedom is one of the longest-lasting problems in computer vision because of the complexity of the models that relate observation with pose.

Variations

Additional technical challenges that we may face are the high level of heterogeneity in people’s behaviors, the noisy background and the often substantial differences in scale, perspective and other camera configurations. There is also large variability introduced by differences in people’s appearance due to clothing, body shape, size and hairstyles. Furthermore, results of studies [54], indicate a need for considering culture as one specific design factor for a bodily expression recognition system. While some similarities do exist among cultures in how they convey, recognize, and attribute emotional meaning to posture, there are also some differences. Their results show that emotions are both universal and culturally specific and while evaluating cultural dimensions, some cultures are more similar than others.

1.6 Thesis Structure

The thesis consists of the following chapters:

- In Chapter 2, we study the existing literature work on emotion recognition pipelines, starting with human detection, moving to body pose estimation, and finally representation building and recognition. We study both traditional and modern methods, as it is important to understand the evolution of the area.

- In Chapter 3, we present the visual emotion recognition model used to tackle this challenging topic, discussing its structure and benefits. We also address the challenges of our dataset, that are label imbalance and child-robot interaction, that are taken into account in the model's various configurations.

- In Chapter 4, we perform an experimental study about the face occlusion effect, caused by a face mask, on emotion recognition performance. We talk about the motivation of this study, the tools used and their implementation, as well as the insights that our results suggest.

- In Chapter 5, we adopt temporal modeling and modality fusion techniques to enhance our model and aspire to overcome the face mask consequences, regarding performance. Various comparisons are presented, depending on the input modality and the combination of methods.

- In Chapter 6, we create a real-time setup of the model and evaluate it on a real-world scenario. We expand our dataset domains on large facial and bodily affection databases, adopt a lightweight low-inference time network architecture, present multiple input versions of a demo and study the face mask effect in-the-wild.

- In Chapter 7, we summarize the conclusions and insights that came out of this thesis, to be further utilized and exploited, and also suggest future work that was not covered.

Chapter 2

Background Work

In this section, we will study the main components of an Affective Bodily Expression Recognition (ABER) system. An important preparation step, which influences all the subsequent design decisions for such an automatic pipeline is the determination of the appropriate modelling of input (human body) and targets (emotion). Depending on the type of the model that has been chosen, either a publicly accessible database can be utilized, or a new one needs to be created. Similarly, other elements of the system need to be selected and configured, such that they are compatible with each other, and overall, comprise an integrous effective system. Regardless of the foregoing differences between various types of ABER systems, the common first step is human detection, i.e. to subtract the background from every frame which represents a human presenting a gesture. The second step is detection of the human body pose in order to reduce irrelevant variation of data caused by posture. The final part of the pipeline consists in building an appropriate representation of the data and applying a learning technique (usually classification or regression) to map this representation to the targets.
Two general approaches of the above stages can be discriminated in the literature: handcrafted feature methods and deep learning based methods. In recent years, deep learning methods have been very popular due to the massive amounts of digital data in combination with powerful processing hardware, yielded excellent results and on most cases outperformed non-deep state-of-the-art methods. However, they usually suffer from high computational need and lower performance speed.

2.1 Deep Learning

In this section, some basic concepts of deep learning are introduced, as it is the type of algorithm that is used in this thesis. Conventional machine learning techniques [55] were limited in their ability to process natural data in their raw form. The goal was to design a feature extractor that transformed the raw data (such as the pixel values of an image) into a suitable internal representation or feature vector from which the learning subsystem, often a classifier, could detect or classify patterns in the input. For decades, constructing a machine-learning or pattern-recognition system, required careful engineering and considerable domain expertise. Deep Learning [56] allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. These methods have dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection and many other domains such as drug discovery and genomics.

2.1.1 Machine Learning

Classification

Classification is a machine learning task, where the model is asked to specify which of k categories some input belongs to. To solve this task, the learning algorithm is usually asked to produce a function $f : \mathbb{R}^n \rightarrow \{1, ..., k\}$. When $y = f(x)$, the model assigns an input described by vector x, to a category identified by numeric code y.

Supervised learning

The most common form of machine learning, deep or not, is supervised learning. Imagine that we want to build a system that can classify images as containing, say, a house, a car, a person or a pet. We first collect a large dataset of images of houses, cars, people and pets, each labelled with its category. During training, the machine is shown an image and produces an output in the form of a vector of scores, one for each category. We want the desired category to have the highest score of all categories, but this is unlikely to happen before training.

Loss Function

Training is done by computing an objective function that measures the error (or distance) between the output scores and the desired pattern of scores. The machine then modifies its internal adjustable parameters to reduce this error. These adjustable parameters, called weights, are real numbers that define the input–output function of the machine. In a typical system, there may be hundreds of millions of these adjustable weights, and hundreds of millions of labelled examples with which to train the machine.

Dataset

The dataset of images that is collected, is then divided into three subsets of training, validation and test. The training dataset is used to fit the weights of a deep learning classifier during the learning process. The validation set is used to tune the hyperparameters (i.e. the architecture), while the test set is used only to assess the performance of the model.

2.1.2 Convolutional Neural Network

Convolutional Neural Networks (CNN) are designed to process data that come in the form of multiple arrays, for example a colour image composed of three 2D arrays containing pixel intensities in the three colour channels. A typical architecture (Fig. 2.1) consists of repetitions of a stack of several convolution layers and a pooling layer, followed by one or more fully connected layers.

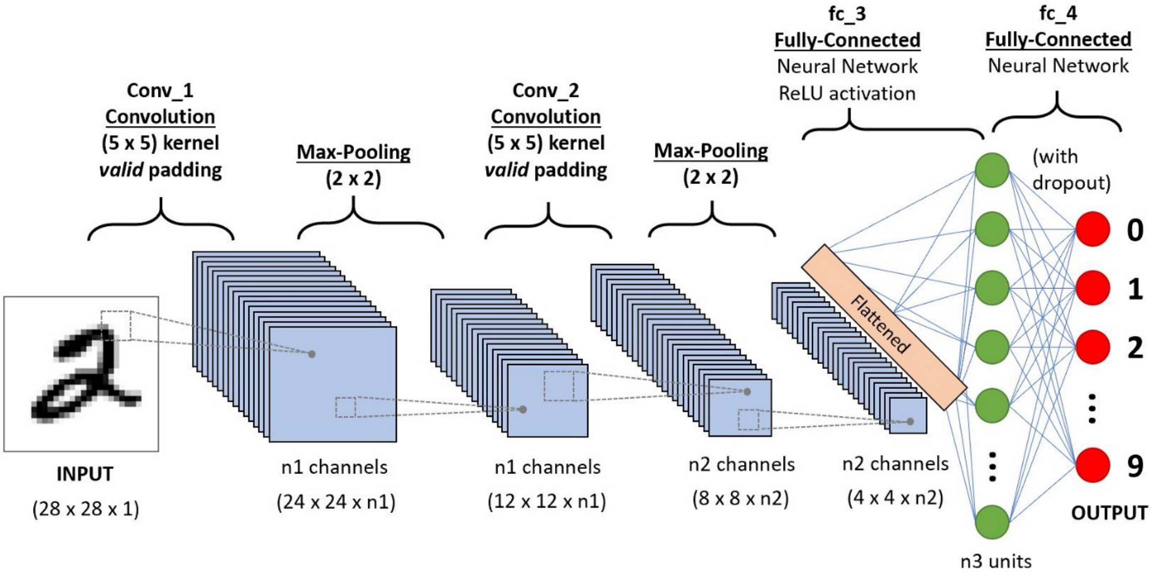


Figure 2.1: Convolutional Neural Network Architecture

- Convolutional Layer: A convolution layer is a fundamental component of the CNN architecture that performs feature extraction, which typically consists of a combination of linear and nonlinear operations, i.e., convolution operation and activation function.
- Pooling Layer: A pooling layer provides a typical downsampling operation that reduces the in-plane dimensionality of the feature maps in order to introduce a translation invariance to small shifts and distortions, and decrease the number of subsequent learnable parameters. It is of note that there is no learnable parameter in any of the pooling layers, whereas filter size, stride, and padding are hyperparameters in pooling operations, similar to convolution operations.
- Fully Connected Layer: The output feature maps of the final convolution or pooling layer is typically flattened, i.e., transformed into a one-dimensional (1D) array of numbers (or vector), and connected to one or more fully connected layers, also known as dense layers, in which every input is connected to every output by a learnable weight. Once the features extracted by the convolution layers and downsampled by the pooling layers are created, they are mapped by a subset of fully connected layers to the final outputs of the network, such as the probabilities for each class in classification tasks. The final fully connected layer typically has the same number of output nodes as the number of classes. Each fully connected layer is followed by a nonlinear function, such as ReLU.

2.1.3 Other Components

Backpropagation

When we use a deep neural network to accept an input x and produce an output $f(x)$, information flows forward through the network. The inputs x provide the initial information that then propagates up to the hidden units at each layer and finally produces $f(x)$. This is called forward propagation. During training, forward propagation can continue onward until it produces a scalar loss $\mathcal{L}$. The back-propagation algorithm [57], often simply called backprop, allows the information from the loss to then flow backwards through the network, in order to compute the gradient of the loss with respect to the weights. This procedure is nothing more than a practical application of the chain rule for derivatives.

Optimizer

After computing the gradient, the model modifies/updates its weights to reduce the error and increase performance on the desired task. This is done by using gradient descent [58], an iterative algorithm, that starts from a random point of the loss function and travels down its slope in steps until it reaches the lowest point of that function. The parameters are updated in each iteration, also called epoch, by subtracting its gradient multiplied by a hyperparameter called learning rate, which heavily influences the convergence of the algorithm and must be chosen wisely. Lastly, momentum is an extension to the gradient descent optimization algorithm that allows the search to build inertia in a direction in the search space and overcome the oscillations of noisy gradients and coast across flat spots of the search space.

Regularization

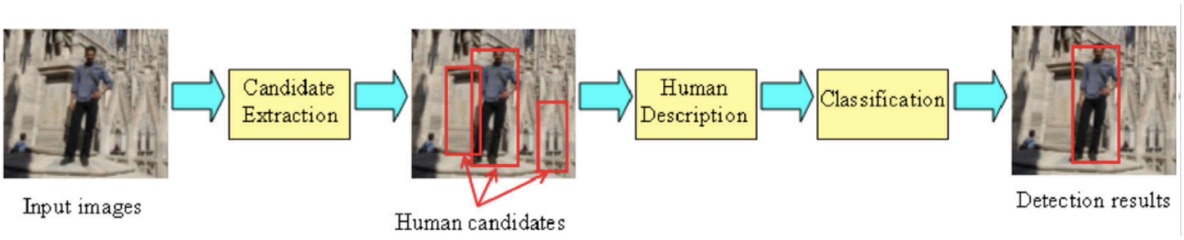
A central problem in deep learning is how to make a model perform well not just on the training set, but also on the test set. Many strategies used are explicitly designed to reduce the test error, possibly at the expense of increased training error, known collectively as regularization [59], and they have shown to improve the generalization performance. That means performing better on the test set, at the expense of performing worse at the training set. An effective regularizer is one that makes a profitable trade, reducing variance significantly while not overly increasing the bias.

2.2 Human Detection

Human detection in images usually consists in determining rectangular bounding boxes that enclose humans. A general framework for human detection is extracting candidate regions that are potentially covered by human objects, describing the extracted regions, classifying/verifying the regions as human or non-human, and postprocessing or adjusting the size of those regions (Fig. 2.2). Modern techniques might not exactly follow this modularization, either by jointly learning representation and classification or by directly proposing detection regions from input.

2.2.1 Candidate Region Extraction

There are a number of ways to extract the candidate regions. A common approach, called window-based detection, extracts windows at various scales and positions, without any prior knowledge of the size and location of the human object, classifies



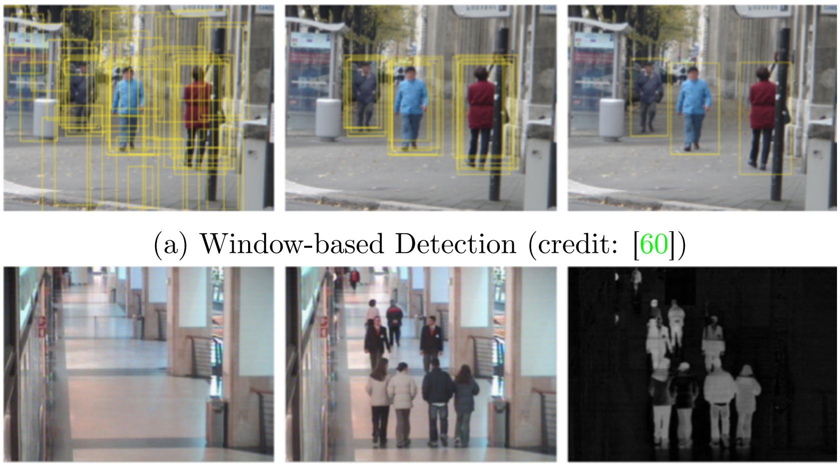
(credit: [60])

Figure 2.2: Framework for Human Detection

human windows, and then merges overlapping ones, e.g. with non-maximal suppression [61], to finally extract human candidate windows (Fig. 2.3a). When the input to the detection system is a video sequence, a well-known technique, namely, background subtraction can be used to obtain human candidates. In particular, moving objects are segregated from the background by calculating the difference between the current image and a reference background in a pixel-wise fashion (Fig. 2.3b).

2.2.2 Human Description

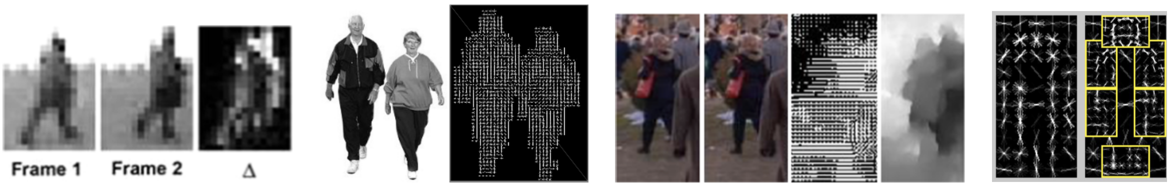
While candidate region extraction is useful to enhance the efficiency of the detection by limiting the search space of human objects, the description of the human objects also plays a key factor in the effectiveness and robustness of human detection. Pioneering work was proposed by Viola et al. [62], who built an efficient moving person detector for videos, exploiting both motion and appearance information (Fig. 2.4a). Following a method previously applied to face detection [63], it employs a cascade structure for efficient detection, and utilizing AdaBoost [64] for automatic feature selection. In [65], Dalal and Triggs introduced the method of Histograms of Oriented Gradients (HOG), which is based on evaluating well-normalized local histograms of image gradient orientations in a dense grid and showed substantial gains over intensity based features [66]. Local object appearance and shape can often be



(a) Window-based Detection (credit: [60])

(b) Background Subtraction (credit: [60])

Figure 2.3: Candidate Region Extraction Methods



(a) [62] (b) [65] (c) [67] (d) [68]

Figure 2.4: Examples of Human Description

characterized rather well by the distribution of local intensity gradients or edge directions (Fig. 2.4b). For the use of optical flows (Fig. 2.4c), histogram of flows (HOF) was proposed in [67]. Earlier works assumed no prior knowledge over the structure of the human body. Arguably, one of the most important contributions in this direction was the Deformable Part Models (DPM) [68], due to its robustness in dealing with articulation. A DPM is a set of parts and connections (Fig. 2.4d), which relate to a geometry prior. Local appearance is easier to model than global appearance and training data can be shared across deformations.

2.2.3 Human Candidate Classification

Once the human descriptors are extracted from the candidate regions, the classification step is invoked to classify the candidate regions as human or non-human. In [69], Mikolajczyk et. al developed a probabilistic method of robust part detectors (Fig. 2.5), in which the classification of a human object was performed using a naive Bayesian classifier based on the part's locations and classification scores. Apart from generative methods, Support Vector Machines (SVM) [70] are often used to classify the human and non-human descriptors, by maximizing the margin between these two classes. Satpathy et al. [71] observed that the human class often distributes in a small space surrounded by the non-human class in the feature space, which happens due to the diversity of the non-human class (Fig. 2.6). Consequently, a linear SVM may not be able to discriminate the two classes, so a hyper quadratic one was used.

Deep Neural Networks (DNN) have shown their potential in human detection [60]. In [72], various combinations of body parts were represented by nodes in a deep belief network [73]. In [74], sub tasks of human detection such as feature selection, object description, occlusion handling were organised into different layers of a deep convolutional neural network and the parameters of each layer were jointly learned through the network. In [75], the mixtures of parts were encoded in a deep architecture called switchable deep network, which was able to infer (and switch to) the most appropriate mode of the mixtures and its robustness was verified by trying different feature types.

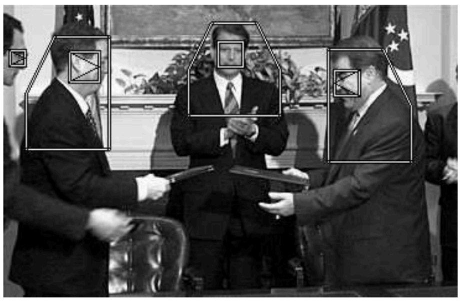


Figure 2.5: Part Detector Example (credit: [69])

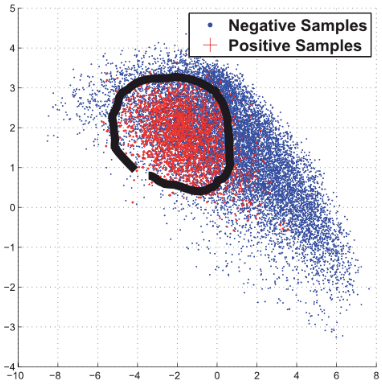


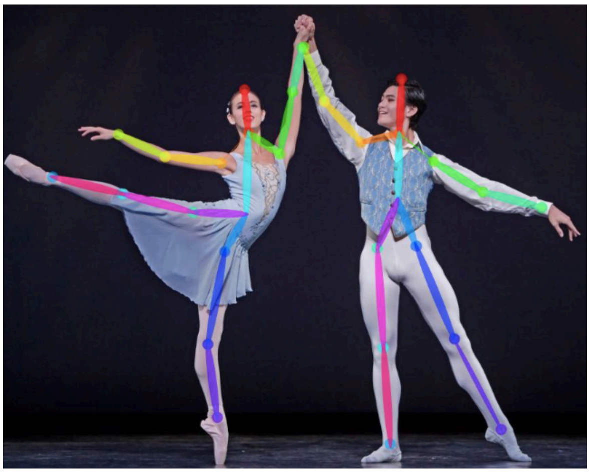
Figure 2.6: Human Features Distribution (credit: [71])

2.3 Body Pose Estimation

Once background has been subtracted, the next step in our ABER pipeline is human body pose estimation, which aims to automatically locate the human body parts from images or videos. One important aspect of body pose estimation is human body modeling, in order to represent keypoints and features extracted from input data. For example, most methods use an N-joints rigid kinematic model to describe and infer human body pose, and render 2D poses (Fig. 2.7).

2.3.1 Handcrafted Feature Methods

The change in human posture is provided by the deformation of the body structure. Complex body deformation will lead to very complex changes of the human body shape. Instead, body deformations could be regarded as a set of local part deformations, such as rotation, scale and translation. The common element of the developed methods toward this direction are: (i) division of the body into several parts, (ii) use of star/tree structure, considering the position relationship and co-occurence relation between parts, and (iii) attribution of each part to different weight to understand different poses [77]. Early works include the Pictorial Structures (PS) model [78] and its improvements [79] [80], in which an object is regarded as the connection between parts. We mentioned earlier that the DPM [68] is used for human pose estimation. However, methods based on DPM often use the simple appearance model, such as HOG features and require the use of background subtraction, making them insufficient in complex scenes. To address this problem, Andriluka et al. [81] proposed a generic approach for human pose estimation, based on the PS model . A more recent approach is the Flexible Mixtures-of-Parts model (FMP) [82], which captures contextual co-occurrence relations between parts, augmenting standard spring models that encode spatial relations. Different from the DPM and PS model, in FMP, the deformable model uses tree structure rather than star structure [77], and also is rotation and scale invariant. Although these traditional approaches showed great results in determining the accurate locations of body parts, the field experienced a considerable change with the emergence of deep neural networks.



(credit: [76])

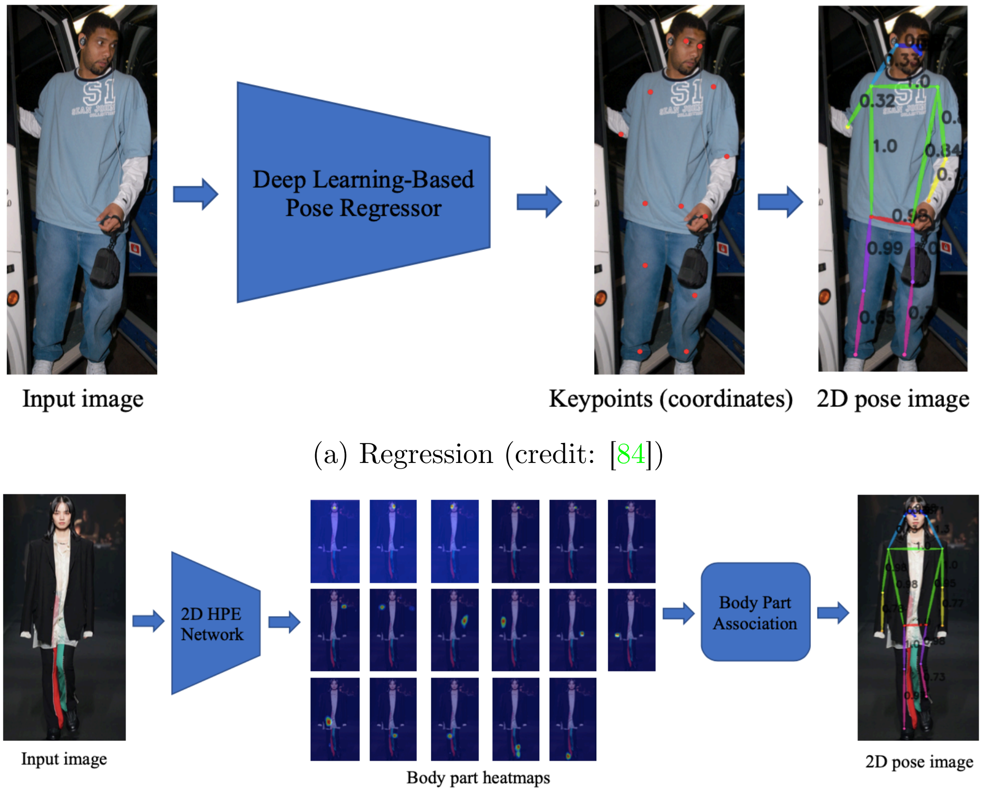
Figure 2.7: Kinematic Human Body Modeling

2.3.2 Deep Learning Methods

In general, there are two categories for body pose estimation pipelines that employ deep learning techniques: regression methods and body part detection methods [83] [84]. Regression methods apply an end-to-end framework to learn a mapping from the input image to body joints or parameters of human body models (Fig. 2.8a), while the goal of body part detection methods is to predict approximate locations of body parts and joints, which are normally supervised by heatmaps representation (Fig. 2.8b).

Regression methods
Regression methods apply an end-to-end framework to learn a mapping from the input image to body joints or parameters of human body models. Starting with DeepPose [85], Toshev and Szegedy proposed a cascaded deep neural network regressor to predict human keypoints directly. However, it is difficult to learn mapping directly from feature maps without other procedures. Carreira et [86] used a self-correcting model. By feeding back error predictions, the predicted keypoint locations are refined progressively. Sun et. al [87] proposed a structure-aware approach called "compositional pose regression". Unlike other related works, this approach re-parameterizes pose representation using bones instead of joints, which is more primitive, stable, and easier to learn. Long-range interactions between bones are encoded by a compositional loss function.

Body part detection methods
Body part detection methods tackle pose estimation as a heatmap prediction problem. The goal is to estimate K heatmaps $\{H_1, H_2, ..., H_K\}$ for a total of K keypoints. The pixel value $H_i(x, y)$ in each keypoint heatmap indicates the probability that the keypoint lies in the position (x, y). The target heatmap is generated by a 2D Gaussian centered at the ground-truth joint location and networks are trained by minimizing the discrepancy (e.g. MSE) between the predicted and the target heatmaps. Compared with joint coordinates, heatmaps provide richer supervision information by



(a) Regression (credit: [84])

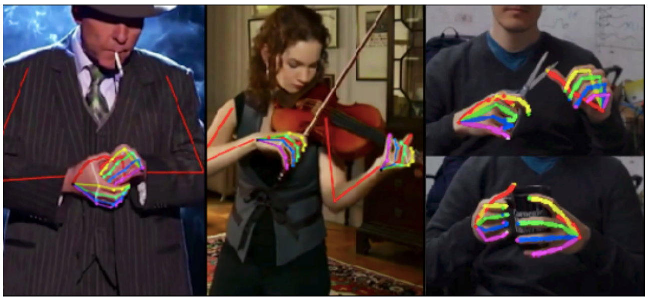
(b) Body Part Detection (credit: [84])

Figure 2.8: Body Pose Estimation Methods

preserving the spatial location information to facilitate the training of convolutional networks. Wei et al. [88] introduced a convolutional networks-based sequential framework with multiple stages, producing increasingly refined predictions. Newell et al. [89] proposed, an encoder-decoder network, named Stacked Hourglass, that consists of consecutive steps of pooling and upsampling layers to capture information at every scale. Based on the above network and with the emergence of Generative Adversarial Networks (GANs) [90], Chen et al. [91] constructed a structure-aware conditional adversarial network, which contains an hourglass [89] network-based pose generator and two discriminators to discriminate against reasonable body poses from unreasonable ones.

Real-time pose estimation

The aforementioned top-down approaches, whose runtime is proportional to the number of people in the image, have high computational cost. Cao et. al [76] presented an efficient method for multi-person pose estimation with state-of-the-art performance, which consisted of two convolutional network branches to jointly predict heatmaps and part affinity fields and lastly bipartite matching of the joints. Bazarevsky et al. [92] adopted a combined heatmap, offset, and regression approach. They use an encoder-decoder network architecture to predict heatmaps for all joints based on [89], followed by another encoder that regresses directly to the coordinates of all joints. The heatmap branch can be discarded, allowing for on-device real-time body pose detection.



(credit: [93])

Figure 2.9: Hand Keypoint Detection Examples

2.3.3 Hand Keypoint Detection

We mentioned in the previous section, that hands may play a key role in ABER. Hand keypoint detection has been an active research topic in the industry, but the vast majority of methods required specialized hardware, such as depth sensors or powerful processors. However, in [93] Simon et al. have achieved realtime 2D hand detection with accuracy comparable to those methods (Fig. 2.9).

2.4 Representation Building and Emotion Recognition

2.4.1 Handcrafted Feature Methods

The large majority of the handcrafted feature methods developed to recognize emotion from body gestures use geometrical representations. A great part of these methods build simple static or dynamic features related to the coordinates of either joints of kinematic models or of parts of the body like head, hands or torso. Some of the most used expressive features are: (i) displacements [94], where the feature vector consists of displacement measures between a frame with the neutral expression and one where the expression is at its apex, (ii) hands shape and palm orientation [95], (iii) motion cues [96] [97] [98] [99], from simple features, such as velocity and acceleration, to more advanced descriptors:

- Smoothness/Jerkiness, which refers to the value of high-order derivatives, describes the smoothness of a movement and is used to specify the arousal level communicated by a move.
- Curvature: $k = \frac{\dot{x}\cdot\ddot{y} - \dot{y}\cdot\ddot{x}}{(\dot{x}^2+\dot{y}^2)^{3/2}}$ measures the rate at which a tangent vector turns as a trajectory bends (e.g., a hand trajectory following the contour of a small circle will bend sharply, and hence will have higher curvature; by contrast, a hand trajectory following a straight line will have zero curvature).
- Quantity of Motion: $\text{QoM} = \text{Area}\left(\sum_{i=0}^{n} \text{Silhouette}[t-i] - \text{Silhouette}[t] \right)$ can be considered as an overall measure of the amount of detected motion, involving velocity and force.

- Contraction Index (CI) is a measure of the degree of contraction and expansion of the body (i.e. the space it occupies) and is calculated using the minimum rectangle surrounding the body. The use of space in terms of judged expansiveness or spatial extension of movements have also been regarded as another relevant indicator for distinguishing between active and passive emotions.

A further processing step consists of analyzing the temporal profiles of the above expressive features in order to get information on their temporal dynamics. Some dynamic features used are: maximum, minimum, mean, standard deviation, gesture duration, maximum/main peak duration and number of maxima.

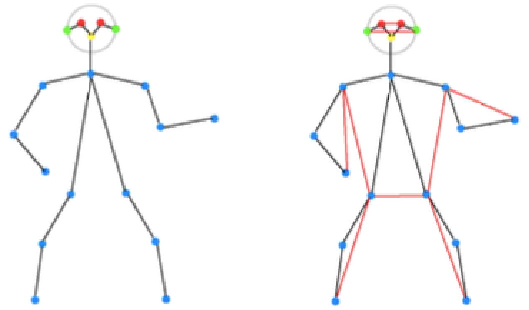
A more organised approach [100] is to follow the Laban notation, originally proposed by Rudolf Laban [101]. Laban Movement Analysis (LMA) uses four components to record human body movements: body, effort, shape, and space. Body category represents structural and physical characteristics of the human body movements. It describes which body parts are moving, which parts are connected, which parts are influenced by others, and general statements about body organization. Effort category describes inherent intention of a movement. Shape describes static body shapes, the way the body interacts with something, the way the body changes toward some point in space, and the way the torso changes in shape to support movements in the rest of the body.

Let $p_i^t \in \mathbb{R}^2$ denote the coordinate of the i-th joint at the t-th frame.

The first part of features in LMA, *body component*, captures the pose configuration. For feet-hip, feet, hands-shoulder, hands and centroid-pelvis, the distance between the specified joints are computed frame by frame (Fig. 2.10a).

The second part of features in LMA, *effort component*, captures body motion characteristics. Based on the pose, joints velocity v_i^t, acceleration a_i^t and jerk j_i^t are computed as:

$$v_i^t = \frac{p_i^{t+\tau} - p_i^t}{\tau}, \; a_i^t = \frac{v_i^{t+\tau} - v_i^t}{\tau}, \; j_i^t = \frac{a_i^{t+\tau} - a_i^t}{\tau}$$



(a) Natural Skeleton (b) Extra Limbs in red

(credit: [100])

Figure 2.10: Illustration of the Human Skeleton for LMA

Furthermore, angles, angular velocity and angular acceleration between each pair of limbs (Fig. 2.10b) as calculated for each pose:

$$\theta^t(i,j,m,n) = \arccos\left(\frac{(p_i^t - p_j^t)\cdot(p_m^t - p_n^t)}{||p_i^t - p_j^t||\cdot||p_m^t - p_n^t||}\right)$$

$$\omega_k^t = \frac{\theta^{t+\tau}(i,j,m,n) - \theta^t(i,j,m,n)}{\tau}$$

$$a_k^t = \frac{\omega^{t+\tau}(i,j,m,n) - \omega^t(i,j,m,n)}{\tau}$$

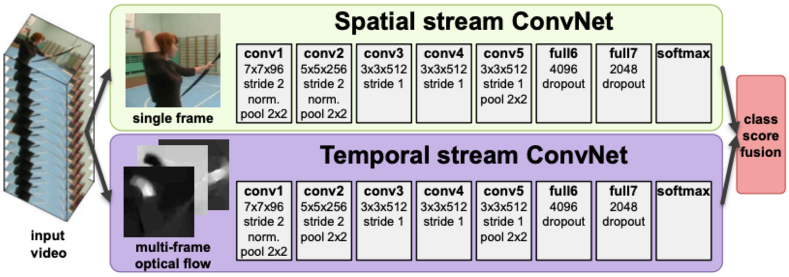
for a frame window τ.

The third part of features in LMA, *shape component*, captures body shape. For upper, lower, left side, right side and whole body, the area of bounding box that contains the specified joints is used to approximate volume.

Finally, all features are summarized by their basic statistics: maximum, minimum, mean and standard deviation over time. The significance of the LMA features was tested in [100], showing that they are strongly correlated with arousal. However, they seem to be correlated to a much lesser degree with valence, dominance and categorical emotions.

2.4.2 Deep Learning Methods

In order to address Deep Affective Bodily Expression Recognition (DABER), it would be wise to study proven background work on human action recognition [102] [103], as a more general case of emotion recognition. Current state-of-the-art results of human action recognition are achieved by two-stream network-based deep-learning methods [104]. One stream takes static images with the form of 2D joint coordinates as input (spatial stream), as the static appearance by itself is a useful clue, since some actions are strongly associated with particular objects. The other stream takes stacked optical flow between video frames as input (temporal stream), which is shown to effectively encode motion and significantly improve accuracy, when fused with the spatial stream. (Fig. 2.11).



(credit: [104])

Figure 2.11: Two-stream Convolutional Neural Network Architecture

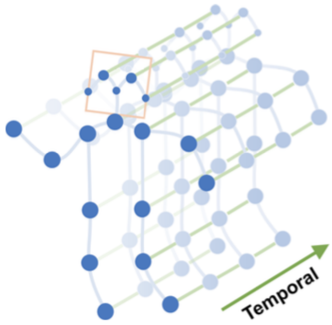
Yan et al. [105] showed that human body joints can be represented by a graph with their natural connectivity and considering the time dimension, a skeleton sequence can be represented by a spatio-temporal graph (Fig. 2.12). The intra-body edges between body joints are defined based on the natural connection, while the inter-frame edges connect the same joints between consecutive frames. This spatio-temporal graph is used as input to a Spatial-Temporal Graph Convolutional Network (STGCN). The STGCN implements an extension of a regular CNN, by redefining the sampling function, the weighting function and the spatial graph convolution, as well as modeling the spatial temporality, in order to fit these operations on a graph.

Instead of the neighboring pixels of a vertex, the sampling function can be defined on the vertices, that are at most D edges away from that vertex, which constitute the neighbor set. The weighting function can be defined as a mapping of subsets of the neighbor set, to their subset labels, which represents the weight. Several strategies of this "partitioning" have been proposed. The spatial graph convolution can now be easily described using the previous functions, and lastly, the spatial temporality can be defined by extending the concept of neighborhood to also include temporally connected vertices.

Besides action recognition, these types of networks can also be utilized for the more specific task of emotion recognition.

Considering an input video with a person expressing emotion, the expressions that correspond to each emotion are not present throughout the video, but usually only during shorter periods. A method that effectively captures the general pattern of the features during the temporal sequence is to use global average temporal pooling, over the body joint input sequence. In [30], Filntisis et al. developed a multi-cue affect recognition method, comparing different implementations for the ABER branch, including a regular DNN (Fig. 2.13), which actually achieved the best performance.

Pikoulis et al. [106] proposed a multi-stream network ensemble for visual emotion recognition in the wild, with a two-stream convolutional net, as well as an STGCN-based stream (Fig. 2.14). Regarding the modality of bodily expression, the network



(credit: [105])

Figure 2.12: Body Spatio-Temporal Graph

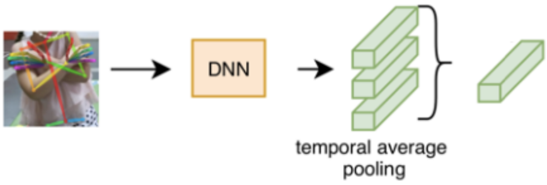


Figure 2.13: Temporal Average Pooling (credit: [30])

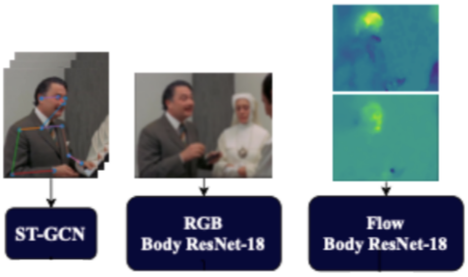
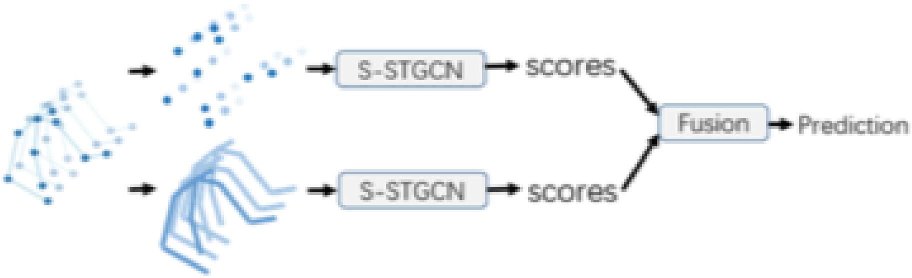


Figure 2.14: Bodily Expression Branches (credit: [106])

processes: (i) the body crops of each frame instance through the RGB stream, (ii) stacked optical flow through the Flow stream, and (iii) skeleton joints through the STGCN stream.

Because the original STGCN is based on the fixed topology of the body joints graph, which may not be appropriate for emotion recognition tasks, Shi et al. [107] proposed a self-attention enhanced spatial graph convolutional layer, to not only include local neighborhood vertices in the process of message passing, but also other vertices with high relevance of information. This layer's architecture is inspired by the work of Vaswani et al. [108] on attention mechanisms. Along with a temporal convolutional and some functional layers, they constitute the basic block of the Self-Attention Enhanced Spatial Temporal Graph Convolutional Network (S-STGCN).

In addition to the joint positions, the bone information representing the lengths and orientations of the human bones (2.15), has also been proved to be useful for skeleton-based action recognition [109] and may also play an important role in emotion recognition tasks. For this reason, Shi et al. [107] proposed a two-stream architecture, where the joint and the bone data are fed into two S-STGCNs, each stream is trained respectively and the output tensors are fused to predict emotion labels.



(credit: [107])

Figure 2.15: Bone information additional stream

2.5 Datasets

Below, we list some datasets that can be used to train and evaluate a DABER system.

GEMEP

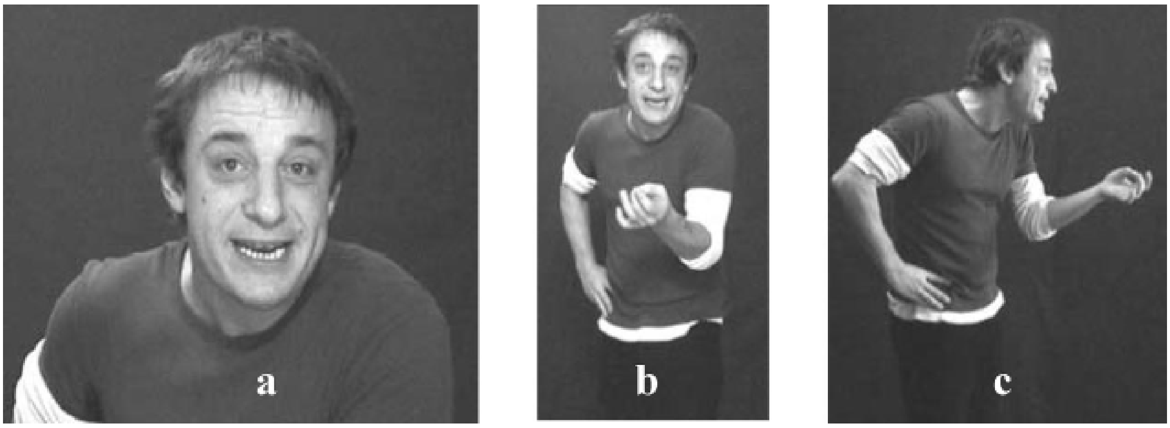
Bänziger et al. [47] introduced the GEMEP corpus, which included 10 professional theater actors performing non-spontaneous expressions. The expressions were recorded with three digital cameras. The first was used to zoom in on the facial expressions and head orientations of the expressers, the other recorded a zoomed out view, which includes body postures and gestures and the third camera recorded body postures and gestures from a right profile view (Fig. 2.16) for a total of 5,040 raw videos. The annotations include emotion categories for low/neutral/high values of arousal and both for positive and negative valence, for a total of 18 categories.

BoLD

Luo et al. [100] introduced a large scale in-the-wild video dataset, named the Body Language Dataset corpus (BoLD), annotated with categorical and continuous emotions. It consists of of 9,876 video clips of humans expressing emotion, primarily through body movements (Fig. 2.17). Each clip can contain more than one character, yielding a total of 13,239 annotations, split into a training, validation, and test set. The dataset has been annotated by crowdsourcing employing two widely accepted categorizations of emotion. The first one is the categorical annotation with a total of 26 labels first used in [110], by collecting and processing an extensive affective vocabulary and the second annotation regards the continuous emotional dimensions of the VAD [16].

FABO

Gunes and Piccardi [111] created a bimodal database consisting of representative samples of human multi-modal expressive behavior, that combine face and body



(credit: [47])

Figure 2.16: GEMEP Corpus Samples

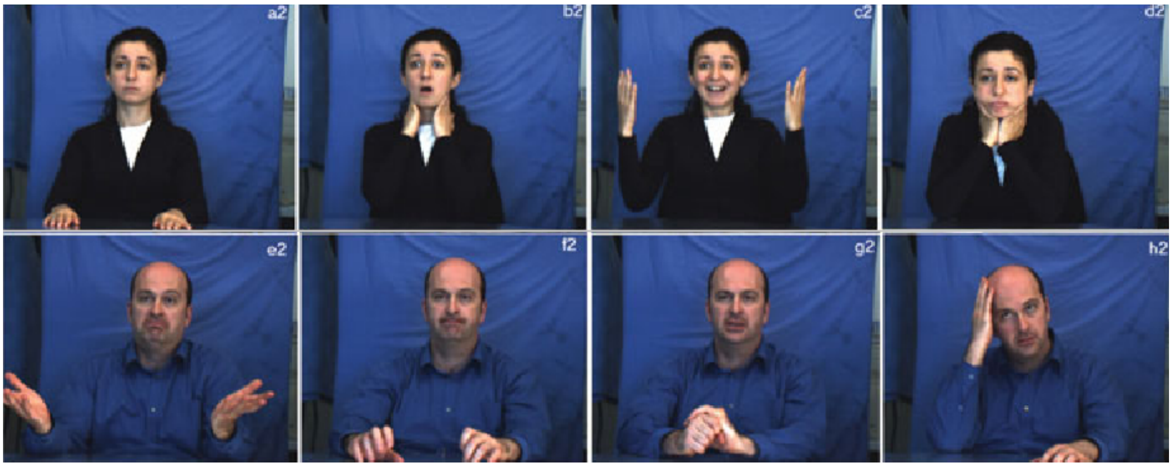
(credit: [100])

Figure 2.17: BoLD Samples

(mainly hand) expressions recorded simultaneously (Fig. 2.18). It was one of the first readily available databases combining affective face and body information in a genuine bimodal manner, which could be used for automatic analysis of human nonverbal affective behavior. Its size is approximately 9GB after compressing an 1-hour long recording, while the emotion labels include the 6 basic emotions.

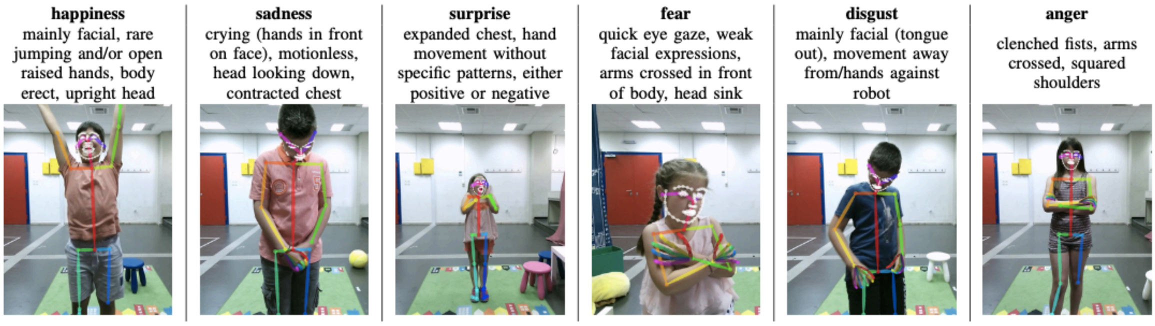
BRED

Filntisis et al. [30] collected a challenging dataset for evaluating emotion recognition systems, named the BabyRobot Emotion Database (BRED), which includes multimodal recordings of children interacting with two different robots (Fig. 2.19). The emotions included are the 6 basic emotions: Anger, Happiness, Fear, Sadness, Disgust, and Surprise. Each recording is hierarchically annotated, by first using only the face or body, and then both to decide for the expressed emotion, resulting in 3 labels for each video. Specifically for the body labels, the inter-annotator agreement is very high.



(credit: [111])

Figure 2.18: FABO Dataset Samples

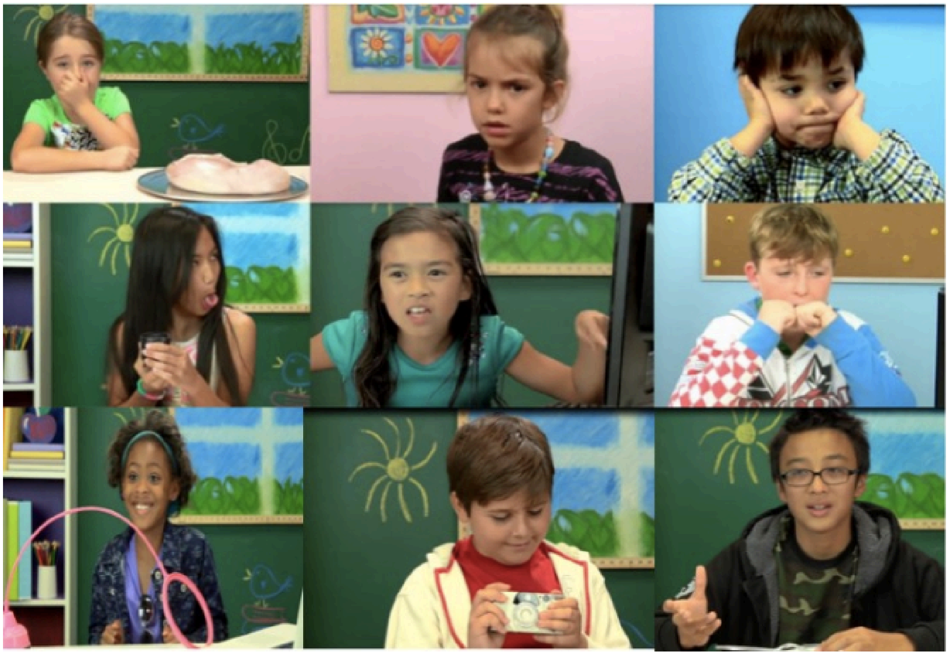

happiness
mainly facial, rare jumping and/or open raised hands, body erect, upright head
sadness
crying (hands in front on face), motionless, head looking down, contracted chest
surprise
expanded chest, hand movement without specific patterns, either positive or negative
fear
quick eye gaze, weak facial expressions, arms crossed in front of body, head sink
disgust
mainly facial (tongue out), movement away from/hands against robot
anger
clenched fists, arms crossed, squared shoulders

(credit: [30])

Figure 2.19: BRED Samples

EmoReact

The EmoReact dataset [112] contains videos of 63 children, aged between 4 and 14, expressing emotions while discussing different topics. Its the largest dataset of its kind both in size of data and number of annotated emotion labels. The videos are collected from the YouTube channel React (Fig. 2.20) and features multi-label annotations on eight different categorical emotions: Curiosity, Uncertainty, Excitement, Happiness, Surprise, Disgust, Fear, Frustration. The co-occurrence of emotions has been analyzed, showing that curiosity has mostly appeared with uncertainty, surprise and fear. Curiosity has been defined as a need, thirst or desire for knowledge about something, which can mean the curious person is uncertain about some aspects about a topic. Discovering knowledge about that topic can be surprising, especially if it is contradictory to one's previous beliefs, and can cause happiness.



(credit: [112])

Figure 2.20: EmoReact Dataset Samples

[This page intentionally left blank]

Chapter 3

TSN-Based Visual Emotion Recognition Model

In this chapter, we present the visual emotion recognition model, that will be used to tackle the challenging topic of DABER and provide the experimental results. It is based on literature work [113] [114] [115] that has shown proven results in the area. We discuss its structure and benefits and also address the challenges of our dataset, which are label imbalance and child-robot interaction, and are taken into account in the model's various configurations.

3.1 Database

3.1.1 Child-Robot Interaction

Although there has been a considerable amount of research on automatic emotion recognition in adults, emotion recognition in children has been understudied. We choose EmoReact (see 2.4) as our dataset and therefore, study a CRI problem. This is more challenging as children tend to fidget and move around more than adults, leading to more self-occlusions and non-frontal head poses. Children's behavior and natural characteristics, articulation, spontaneity and body height differ, so perception systems need to be specifically trained, to be able to tackle CRI problems. This difference, along with the lack of children-related big data for training recognition algorithms turn usual recognition tasks into a seriously challenging problem.

3.1.2 EmoReact Imbalance

EmoReact is an imbalanced dataset, which means it includes an unequal number of videos for each emotion label (Fig. 3.1), and is something that we must address in our upcoming configuration choices. Furthermore, by observing the training set we possess, we can argue that some emotions (Fear, Frustration, Disgust) are expressed in a pretty low number of samples. This results in possible lack of diversity and less ease to generalize well across unseen individuals, introducing an extra degree of difficulty to our problem.

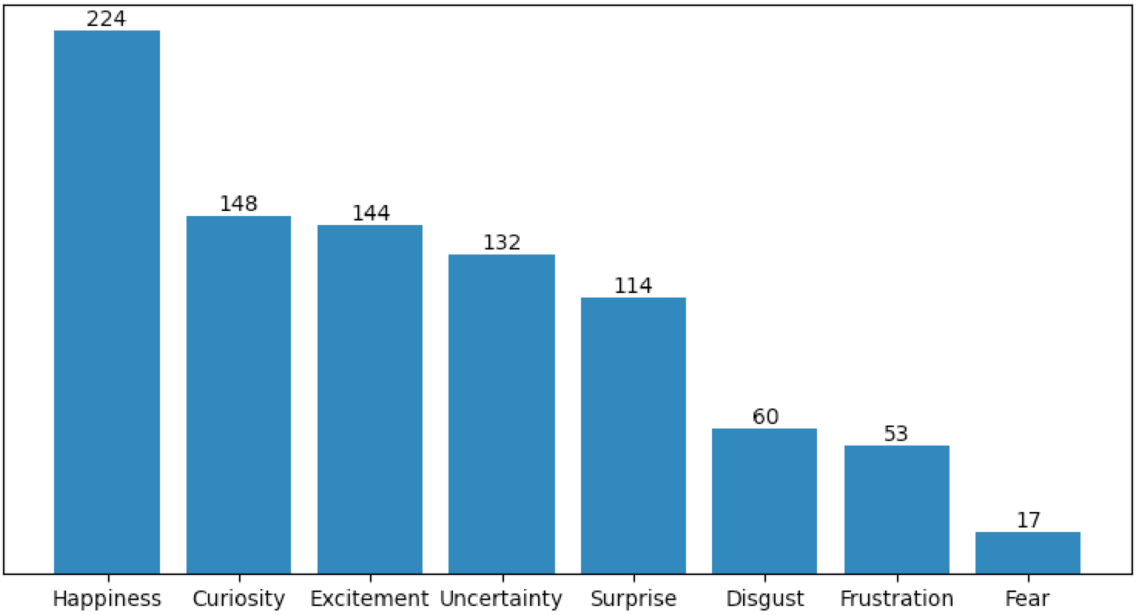


Figure 3.1: EmoReact Training Set Imbalance

3.2 Backbone Architecture

As our backbone network, we choose the most commonly used architecture for action & emotion recognition visual streams, which is the Residual Network [116] (ResNet). ResNets implement shortcut connections, i.e forwarding the output of certain layers (Fig. 3.2). Stacking layers should not degrade the network performance, because we could simply stack identity mappings upon the current network, and the resulting architecture would perform the same. This indicates that the deeper model should not produce a training error higher than its shallower counterparts. Authors also hypothesize that letting the stacked layers fit a residual mapping is easier than letting them directly fit the desired underlying mapping. Its main advantages are: (i) reducing the effect of the vanishing gradient problem that regular CNNs suffer from, by performing an identity mapping, (ii) accelerating the speed of training, and (iii) no addition of extra trainable parameters.

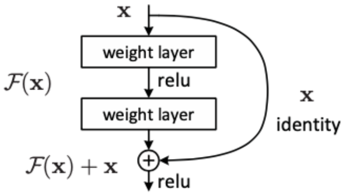

x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 3.2: ResNet Building Block

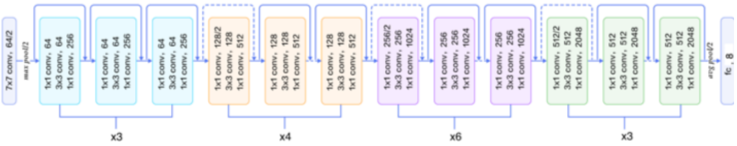


Figure 3.3: ResNet-50 Backbone

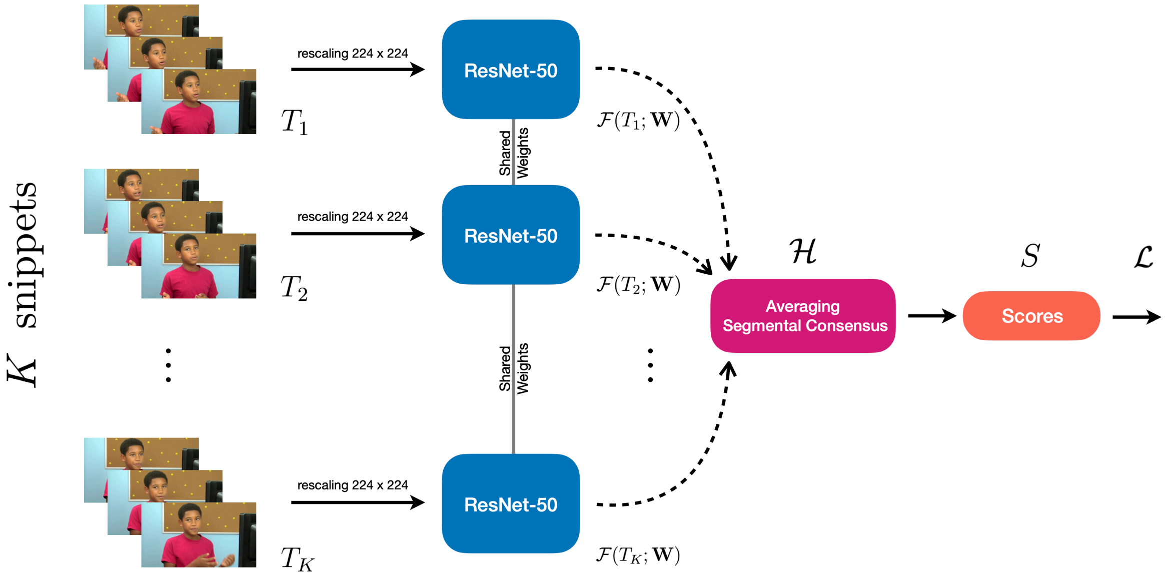


Figure 3.4: TSN-Based Model Architecture

3.3 Temporal Segment Network

3.3.1 Feature Capturing

Complex actions, like emotional expressions, comprise multiple stages spanning over a period of time and it would be quite a loss failing to utilize them. On the other hand, as we previously mentioned, each expressed emotion is not present throughout a whole input video. These facts, indicate that we are in need of effective general feature capturing. While the plain CNN architecture considers the whole input sequence, as well as each frame in the video separately, the Temporal Segment Network (TSN) framework [117] operates on a sequence of short snippets sparsely sampled from the entire video. Each snippet in this sequence will produce its own preliminary prediction of the emotion classes and then, a consensus among the snippets will be derived as the video-level prediction. Therefore, it allows the network to access several parts of the video, but also tackles the inability of the former to model long-range temporal structure, thus, being more likely to observe the corresponding expression.

3.3.2 Method

The overall architecture of our model is shown in Fig. 3.4. Formally, given a video V, we divide it into K segments $\{S_1, S_2, ..., S_K\}$ of equal duration M, and transform them into a sequence of snippets $(T_1, T_2, ..., T_K)$. Each snippet T_k is produced, by randomly sampling $N < M$ consecutive frames from its corresponding segment S_k. Finally, scores of the different snippets are fused using the segmental consensus function $\mathcal{H}$, that is applied on the representations of all different snippets to obtain the final scores S:

$$S = \text{TSN}(T_1, T_2, ..., T_K) = \mathcal{H}(\mathcal{F}(T_1; \mathbf{W}), \mathcal{F}(T_2; \mathbf{W}), ..., \mathcal{F}(T_K; \mathbf{W}))$$

Here $\mathcal{F}(T_k; \mathbf{W})$ denotes the function representing the application of a CNN with parameters $\mathbf{W}$ on the snippet T_k. The consensus function $\mathcal{H}$ we will use is simple averaging, and subsequently, the obtained scores S are fed to a loss function $\mathcal{L}$ to perform the training step.

3.3.3 Benefits

This framework offers several benefits to emotion recognition. Compared to processing the entire video, the sampling process ignores redundant information in consecutive video frames, helping avoid overfitting and offering a type of data augmentation. Simultaneously, it reduces computational costs of training, as it does not consider the entire input video chunk. Lastly, we equip our TSN model with the advantages mentioned in Sec. 3.2, by choosing a ResNet with 50 layers (ResNet-50) (Fig. 3.3) as our TSN backbone network architecture. We note, that before feeding the network with the input, we rescale EmoReact's RGB images from full resolution to (224,224,3).

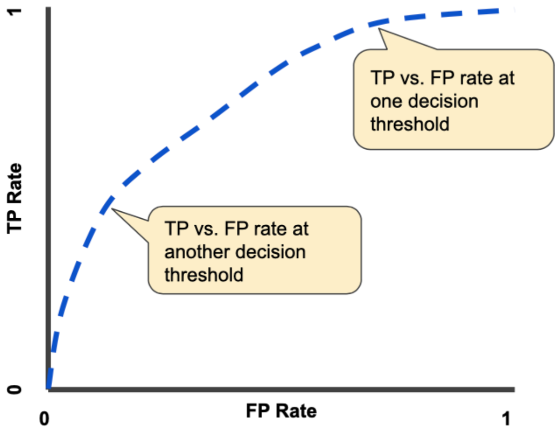
3.4 Model & Training Configurations

3.4.1 Evaluation Metric

Following prior work, the only evaluation metric that has been shown to be robust to imbalanced datasets is the Area Under the Curve of Receiver Operating Characteristic [118] (ROC AUC). It is a graph of the True Positive Rate (TPR), also known as Recall, as the function of the False Positive Rate (FPR), and shows the performance of a classification model at all decision thresholds (Fig. 3.5).

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad , \quad \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

where TP, TN, FP, FN: correctly/falsely predicted positive/negative labels.



(credit)

Figure 3.5: Receiver Operating Characteristic curve

The scores S, of size 8 per sample, are then averaged to obtain a single overall performance metric. They are treated as a collection of 8 binary problems, one for each class. There are then a number of ways to average binary metric calculations across the set of classes. One way is to simply calculate the mean of the binary metrics, giving equal weight to each class. However, the assumption that all classes are equally important is often untrue, such that macro-averaging will over-emphasize the typically low performance on an infrequent class. In order to take the label imbalance of our dataset into account (Fig. 3.1), we give each sample-class pair an equal contribution to the overall metric. Rather than summing the metric per class, we sum the dividends and divisors that make up the per-class metrics to calculate an overall quotient. This metric is called Unbalanced Average ROC AUC, and intuitively, it tells how much the model is capable of distinguishing between the classes. To calculate it per class, we measure the entire two-dimensional area underneath the entire ROC curve from (0,0) to (1,1). Looking at Fig. 3.5, there is a trade-off, between more true negatives as we increase the threshold, but at the expense of more false positives. Compared to metrics such as accuracy, Hamming loss, or F1 score, ROC doesn't require optimizing a threshold for each label.

3.4.2 Loss Function

Since our task is binary multi-label classification, which means to assign variable number of emotion labels to the input video, the loss function we choose is Binary Cross-Entropy (BCE). For multiple classes it is defined as below:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{ij}\log\left(p_{ij}\right) + \left(1-y_{ij}\right)\log\left(1-p_{ij}\right)\right)$$

where N, M denote the number of samples and classes, and y_{ij}, p_{ij} denote the label and probability prediction of sample i for class j respectively, after suppressing the scores S to [0,1] with a sigmoid function.

Our choice is justified by the several benefits of this loss function. It can be shown that cross-entropy depends on the relative errors and not on the absolute errors; thus it gives the same weight to small and large values [119]. A major advantage is that it diverges if one of the outputs converges to the wrong extreme, hence the gradient descent reacts fast. On the other hand, other functions may approach a constant in this case, resulting in gradient descent wandering on a plateau, even though the error may not be small. Furthermore, it satisfies the conditions of the well-formed functions [120].

3.4.3 Optimizer

As our optimizer, we select Stochastic Gradient Descent (SGD), which is faster than plain gradient descent. SGD randomly picks a number of samples, called batch, from the whole set at each iteration to reduce the computations enormously. Because SGD has trouble navigating areas around local optima, we use a method called Momentum [121] that helps dampen oscillations. This is done by introducing a term $\gamma = 0.9$, which adds a fraction of the update vector of the past time step to the current update vector. Intuitively, we give "momentum" to dimensions, whose gradients point in the same directions, and slow down dimensions that change directions, gaining faster convergence and reduced oscillation. All models are trained for 60 epochs, with a batch size of 8, starting with a learning rate of 1e-2, which is then reduced by a factor of 10 at 20 and 40 epoch milestones. For evaluation, we select the epoch with the best validation ROC AUC and apply the corresponding network on the test set, to finally report the best overall performance achieved.

3.4.4 Regularization

To prevent overfitting, we use the L2 regularization technique, also known as weight decay, because it penalizes weights according to their L2 norm. The objective function of minimizing the prediction loss on the training data is replaced with the new objective function, minimizing the sum of the prediction loss and the penalty term. It involves adding a term to the objective function that is proportional to the sum of the squares of the weights. This is how the final loss function looks like using the weight decay technique:

$$\mathcal{L} = \mathcal{L}_{\text{BCE}} + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

where $\lambda = 5e - 4$, is the weight decay factor, penalizing the weight vector $\mathbf{w}$ when it grows too large. This encourages the model to learn simpler functions that are less likely to overfit the training data.

Chapter 4

Medical Face Mask Effect Study

In this chapter, an experimental study about the effect of a face mask on emotion recognition is performed. We apply a medical face mask on the EmoReact children's faces and compare emotion recognition results to when the faces are visible. We examine the case of when the mask is applied to the image of the whole body, as well as only the face, and compare modality performance. Finally, we use some visual explanation techniques to see on what features of the input our model focuses and come to conclusions about the mask effect.

4.1 Motivation

The COVID-19 pandemic has undoubtedly changed the standards and affected all aspects of our lives, especially social life. It has fostered a pervasive use of medical face masks, making a serious impact on facial communication. Studies investigated how the presence of a face mask affects emotion recognition accuracy and have revealed that it diminishes the people's ability to accurately categorize a facial expression [122] [123] [124]. In order to address this problem, it is natural to turn towards bodily expression, as the main cue to recognize affect.

4.2 MediaPipe

MediaPipe offers unified Machine Learning (ML) solutions for live and streaming media, that work across Android, iOS, desktop/cloud, web and IoT. It supports built-in fast ML inference and processing accelerated even on common hardware. MediaPipe is also free and open source, since the framework and solutions are both under Apache 2.0, allowing full extensibility and customizability. The datasets used to train these solutions contain images, that were captured in a real-world lighting, noise, and motion environment, on a diverse set of smartphone cameras, both front and back-facing. A very interesting and important factor is that dataset samples were grouped into several evenly distributed geographic subregions, in order to take cultural fairness into account, and gave promising results in fairness evaluation.

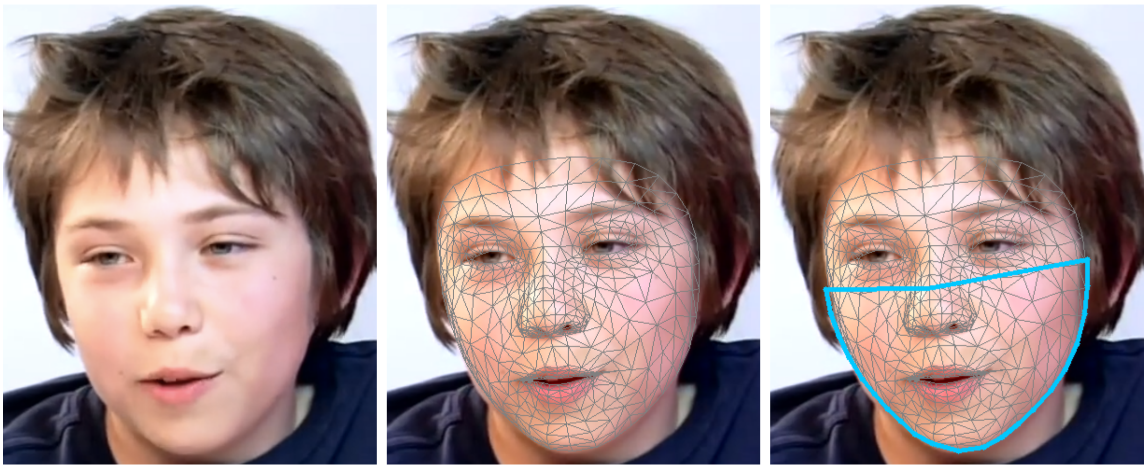
4.2.1 Mask Application

The mask is applied by tracking the facial surface geometry, using Google's MediaPipe Face Mesh [125]. This tool is an end-to-end CNN-based model for inferring an approximate 3D mesh representation of a human face from single camera input. Its architecture is similar to MobileNetV2 [126] with customized blocks for real-time performance. It uses a relatively dense mesh model of 468 vertices and is well-suited for face-based augmented reality effects. It also provides a face detection flag, indicating the likelihood of the face being present in the input image. We track the 2D coordinates of the right and left jawline vertices, starting from just below the eyes until the chin, and one extra vertex for the nose, in order to form a polygon that is finally filled to represent the mask (Fig. 4.1). The jawlines for the mask are created by tracking the edge x-axis vertices and accordingly selecting among several jawline candidates, that we manually created for this particular face mesh model [1]. In Fig. 4.2, we display several samples of EmoReact after the application of the mask and showcase our tool's robustness to face orientation.

4.2.2 Body Detection

In order to incorporate bodily expressions, we need a way to track the human body. Google's MediaPipe Holistic combines human pose and hand tracking tools in a semantically consistent end-to-end solution, is tailored for real-time environments and demonstrate real-time inference speed on mobile GPUs with high prediction quality. BlazePose [127] is a lightweight CNN MobileNetV2-like architecture for human pose estimation, that produces 33 body keypoints for a single person, including main body joints, as well as eyes and nose, using both heatmaps and regression to keypoint coordinates. Detection confidence is provided using two flags for each keypoint, one for the probability of the joint being located within the frame, and one for the probability it is not occluded by another bigger body part or another object. MediaPipe

[1] The code for the mask application tool is publicly available at: https://github.com/nkegke/medical-face-mask-applier



(a) Original Image (b) Face Mesh Tracking (c) Mask Polygon

Figure 4.1: Mask Application Steps

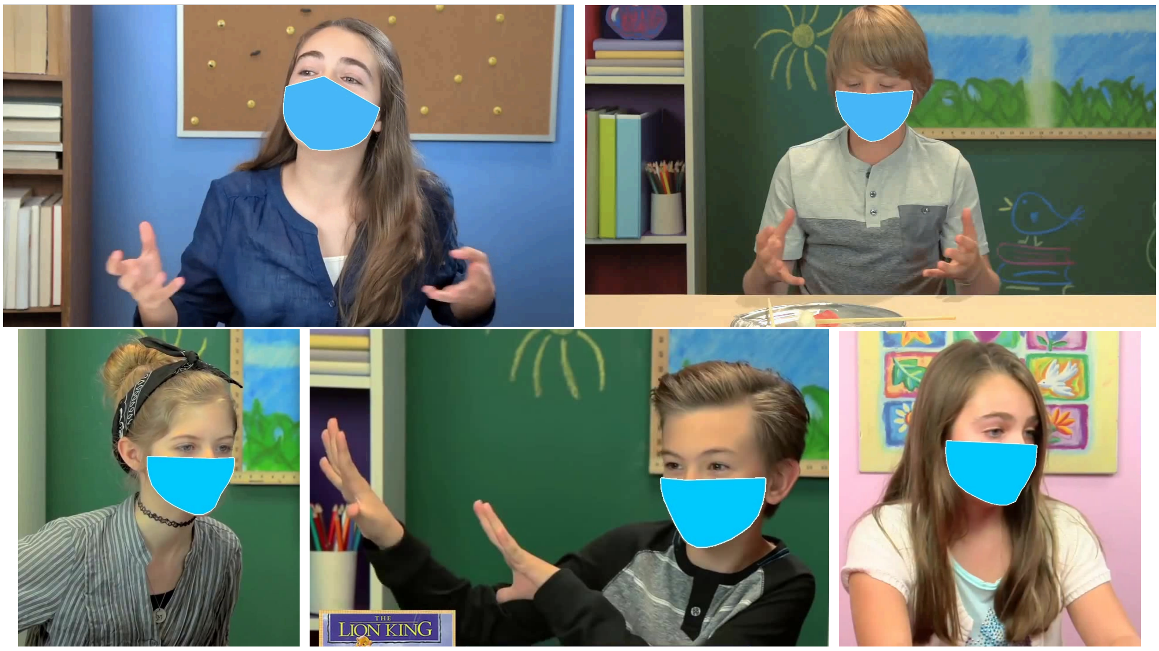
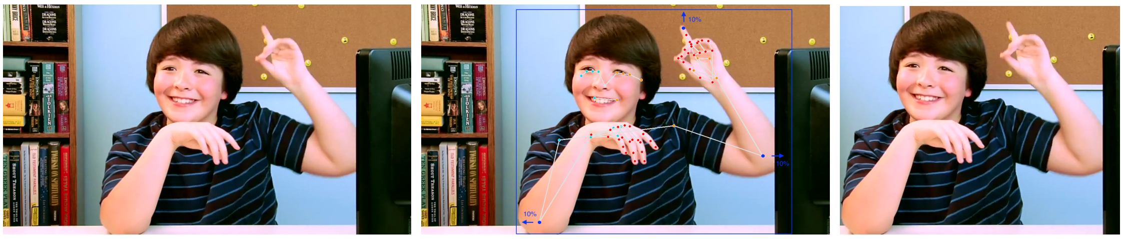


Figure 4.2: EmoReact Mask Samples

Hands [128] is a pipeline that predicts hand skeleton from single RGB camera, consisting of a single-shot palm detector, followed by a hand landmark regression model. We combine keypoints tracked by both tools and create a bounding box with the edge points, expanded by a factor of 10% at each respective dimension, which is then cropped as the input image. Fig. 4.3 demonstrates the process we just described and 4.3c is the input that is given to our model, where most of the background noise is removed and full body information dominates the image crop.

4.3 Mask Effect Results

We study the effect of applying the mask onto the children's faces, by comparing the emotion recognition results of our model when the input is the default, versus when it is masked. Apart from the tools we described in the previous subsections, we also extract the visual face features for face cropping, using OpenFace [129], an open source facial behavior analysis toolkit.



(a) Original Image (b) Body Pose Tracking (c) Body Crop

Figure 4.3: Body Cropping Steps

4.3.1 Performance vs Speed Trade-off

In Table 4.1, we perform an ablation study on the number of segments and consequently the number of snippets, which are used during the TSN training, by considering 4 different values: 1, 3, 5 and 10. By increasing the number of segments, we significantly increase computational load, and therefore inference time. On the other hand, when we provide the model with multiple parts of the video, it might achieve better performance. The numbers reported stand for training with a single RTX 2080 GPU, but one can use multiple ones and increase batch size proportionally for faster training.

Table 4.1: TSN Training Computational Load

Segments	**Time per Epoch (sec.)**	
	Training	Validation
1	6	4
3	14	10
5	23	16
10	32	23

4.3.2 Mask Effect on Face Input

In Table 4.2, we report results on face input. At first sight, performance drops considerably (≈ 3-4%). This is a result we expected, as the mask covers the majority of the face, including the most expressive facial feature, the mouth. Intuitively, if one would try to predict the emotions expressed in those two images, we sense that they would have a better chance without the presence of the mask. Regarding the number of segments used, performance peaks at 10, but increasing it above 3 does

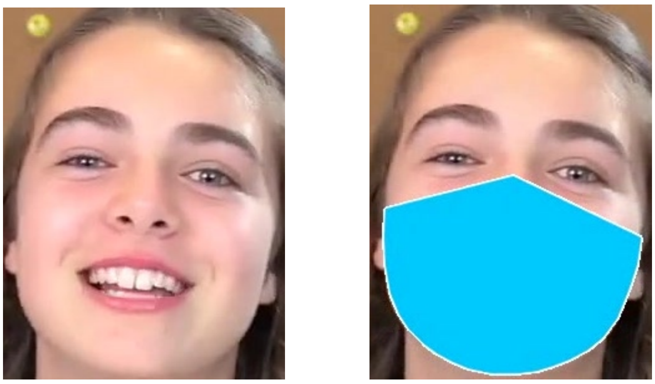


Table 4.2: Mask Effect Results on Face Input

Segments	**ROC AUC**		**Performance**
	Default	Mask	
1	0.755	0.728	−2.7%
3	0.769	0.733	−3.6%
5	0.767	0.732	−3.5%
10	0.770	0.741	−2.9%

not result in significant performance difference. This means that feeding the model with more than 3 parts of the video, does not necessarily make it model temporal structure better. Because we have to work with the aforementioned trade-off of performance versus speed, and showed that even a small number of segments can achieve satisfactory performance, a balanced option would be 3 segments, as it is 3 times faster than 10. The same pattern occurs for masked body results as well, which we examine hereupon.

4.3.3 Mask Effect on Full Body Input

Looking at Table 4.3, which shows results on full body input, the first and most important observation we make, is that performance decrease is very little to none (0-0.4%). These results suggest that the model can exploit body information in such a way, that even with the application of a face mask, and consequently face information loss, it only suffers minimal performance drop. We also note the same pattern of performance with the masked face input results, which is increasing performance as complexity goes up. However, performance increase from 3 to 10 segments is minimal (0.1%), which again suggests working towards the speed side of the trade-off.

4.3.4 Masked Face vs Full Body Results

We compare model performance with masked face versus masked full body crop input and show that incorporating the whole body into the input, gives superior results over face crop. With black we highlight the best overall result, whereas the blue highlighted model is our suggestion as the optimal trade-off. The obvious conclusion is that moving towards bodily expression recognition is our best option, when the face is occluded. However, this is only a baseline result, which we could build on and pursue improvements (see Chap. 5).

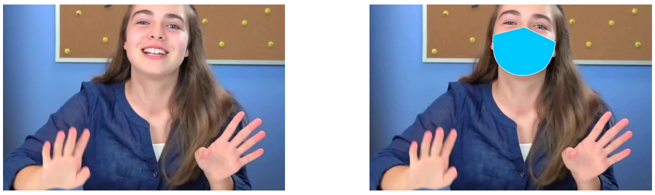


Table 4.3: Mask Effect Results on Full Body Input

Segments	**ROC AUC**		**Performance**
	Default	Mask	
1	0.752	0.752	-
3	0.759	0.758	−0.1%
5	0.758	0.754	−0.4%
10	0.761	0.759	−0.2%

Table 4.4: Masked Input Result Comparison

Segments	**ROC AUC**		**Performance**
	Face	Full Body	
1	0.728	0.752	+2.4%
3	0.733	**0.758**	+2.5%
5	0.732	0.754	+2.2%
10	0.741	**0.759**	+1.8%

4.4 Visual Explanation

To have a better understanding of the mask effect on performance, we utilize a technique for producing visual explanations for predictions. We wish to explore where our CNN model focuses in the input image and how its behaviour varies for the different emotion category targets.

Method

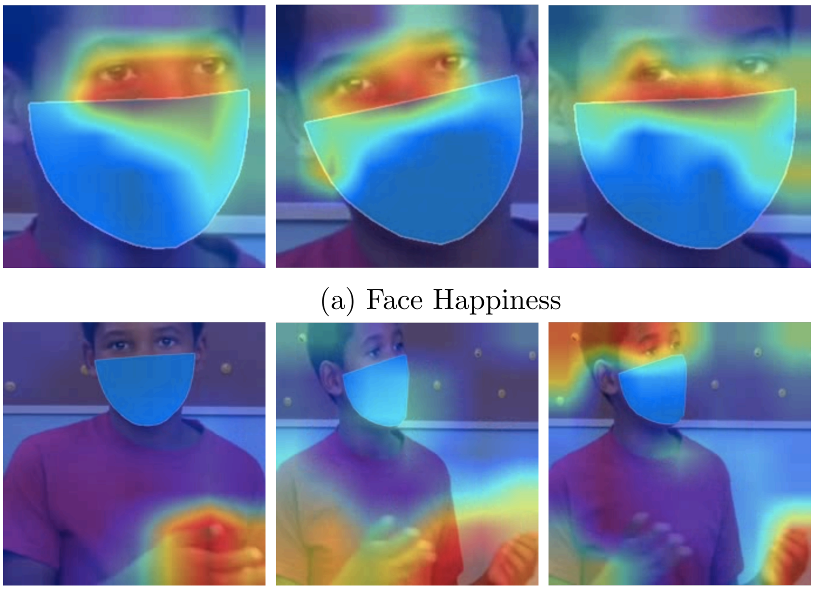
The method we choose is the Gradient-weighted Class Activation Mapping (Grad-CAM) [130], which uses the gradients of an emotion category target in a classification network, flowing into the final convolutional layer to produce a coarse localization map, highlighting the important regions in the image for predicting the concept. The way it works, is that is assigns importance values to each neuron of the output layer for a particular decision of interest. To obtain the class-discriminative localization map $L^{C}_{\text{Grad-CAM}} \in \mathbb{R}^{u \times v}$, it performs the following computations:

$$L^{C}_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_{k} \alpha^{c}_{k} A^{k}\right), \qquad \alpha^{c}_{k} = \frac{1}{Z}\sum_{i}\sum_{j}\frac{\partial y^{c}}{\partial A^{k}_{ij}}$$

where u, v denote width and height of the map, c denotes the class, y^{c} denotes the class score, A^{k} denotes layer k feature map and α^{c}_{k} denotes neuron importance weights. More details on the implementation can be read in [130].

Examples

We provide some example frames of the activation mapping, where intensity, and therefore model focus, increases from blue to red. Starting from the face examples, we can see that the model focuses on the upper part of the face. The facial features that could be utilized are the eyes, the eyebrows and the forehead. Intuitively, the emotion categories that can be expressed by those features are:

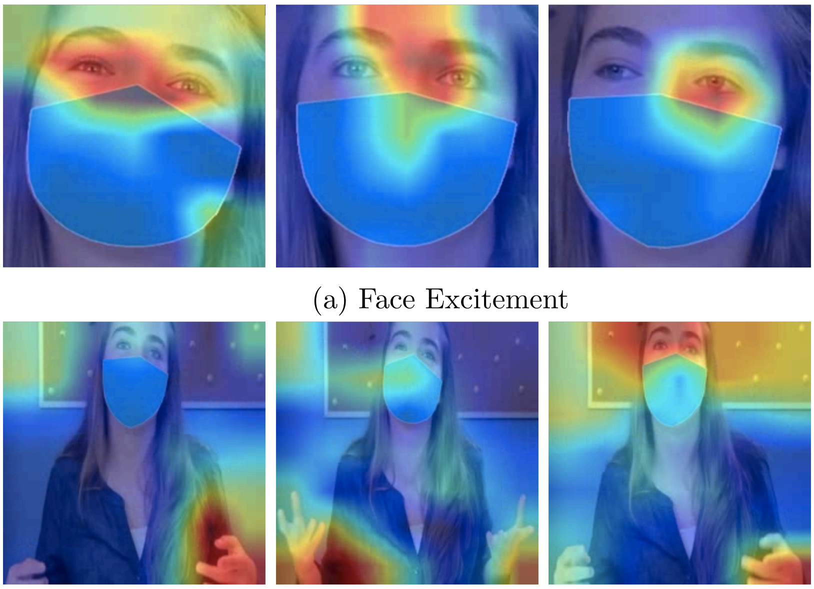


(a) Face Happiness

(b) Body Happiness

Figure 4.4: Happiness Decision Regions

- Excitement: raised eyebrows, eyes wide open (Fig. 4.5a)
- Curiosity: eyes half-closed (Fig. 4.6a)
- Frustration: frowning (Fig. 4.7a)



(a) Face Excitement

(b) Body Excitement

Figure 4.5: Excitement Decision Regions

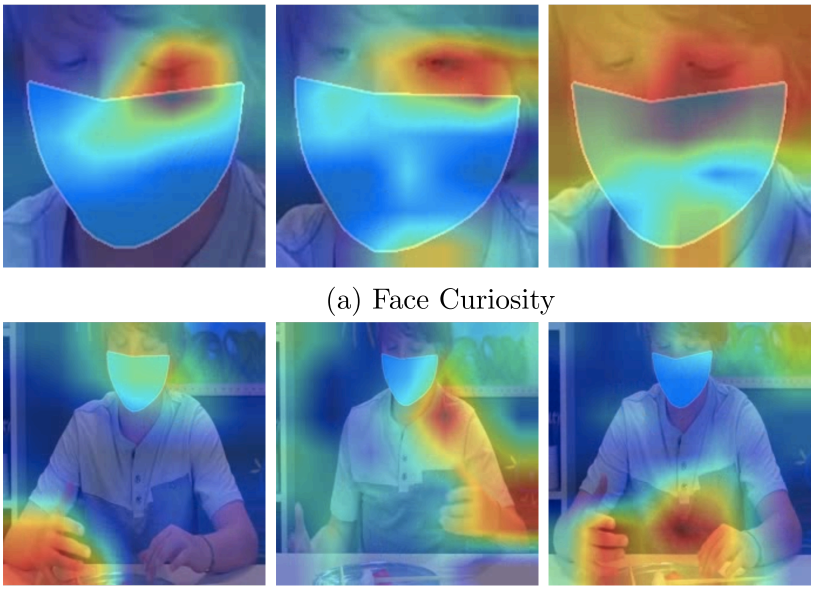
Happiness is not really an emotion that is conventionally recognized by those features, as most of us think of just a simple smile as the expression of happiness. However, we observe that the model is still able to focus on the eyes (Fig. 4.4a) and recognize happiness.

Regarding the body examples, the hands are visible and provide information that is utilized by the model. On emotion target level, the bodily expressed features are:

- Happiness: calm hands (Fig. 4.4b)
- Excitement: hands wide open, arms away from the body (Fig. 4.5b)
- Curiosity: hands investigating, shrugged shoulder (Fig. 4.6b)
- Frustration: fist next to face (Fig. 4.7b)

One interesting mapping is the right-most face frustration example (Fig. 4.7b), where the face crop includes the hand closed in a fist, and the model localizes and incorporates it as an expressive feature.

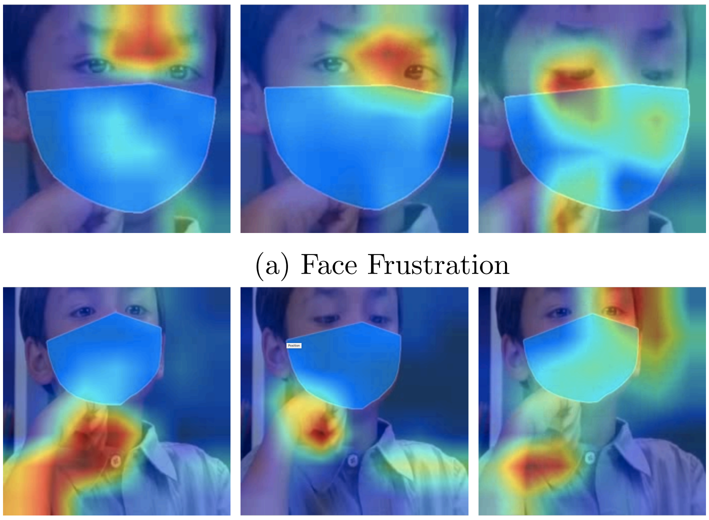
Overall, the model is able to ignore noisy features, like the mask and the background. It is crucial to note, that the background is considered noise in this dataset, as the videos were recorded in a directed setup and it can be the same for different reaction topics. We also talked earlier (see 1.5.3) about the large variability introduced by differences in people's appearance due to clothing, body shape, size and hairstyles. According to these frame examples, the model is able to overcome these difficulties and focus on the expressive features. Lastly, in some body examples, we observe, that the model focuses not only the body, but also on the face, fusing different modality information in a single RGB stream to make predictions.



(a) Face Curiosity

(b) Body Curiosity

Figure 4.6: Curiosity Decision Regions



(a) Face Frustration

(b) Body Frustration

Figure 4.7: Frustration Decision Regions

Below (Fig. 4.8), we report per emotion ROC AUC and compare face versus full body input performance. Full body outperforms face in all emotions, except for Excitement and Frustration. This could be translated as these two emotions being expressed more by facial than bodily features from the children involved. For Fear, performance is a lot higher with full body compared to face, which intuitively makes sense as children tend to utilize their body more to express fear [30]. Another conclusion we could come up to is that for some emotion pairs, like Curiosity-Uncertainty or Excitement-Surprise, which intuitively are quite similar to each other, performance is lower for each emotion individually, because it is harder for the model to distinguish them.

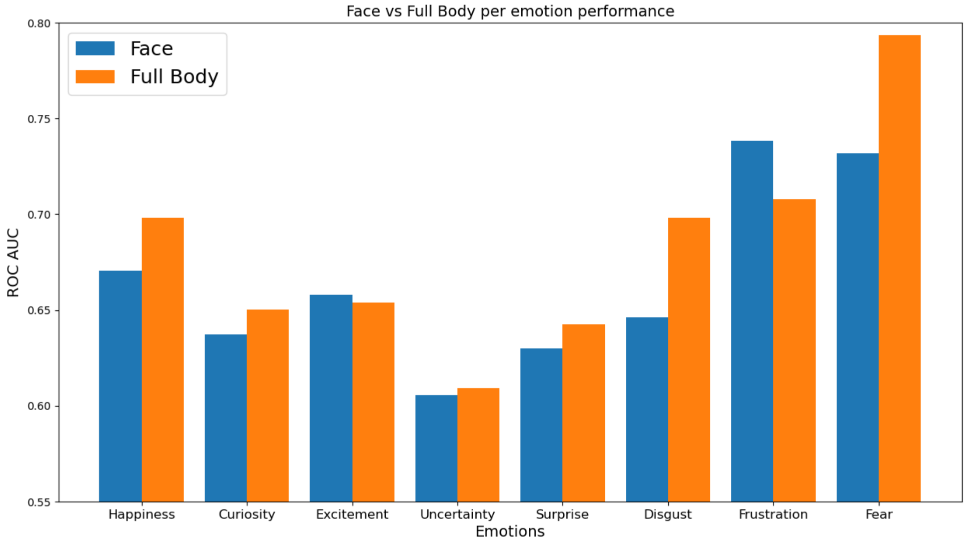


Figure 4.8: Input Modality per Emotion Performance

4.5 Mask Effect Conclusions

Summarizing our experimental study, the results that we reported help us come to numerous conclusions. In the presence of a face mask, emotion recognition performance from just the face drops considerably and urges us to incorporate the body modality. By providing the full body image, our model can sustain its performance to the same level and outperform the masked face. Also, our visualizations prove that a single RGB stream network can learn both from facial expressive features, like the eyes, the eyebrows and the forehead, as well as bodily, such as posture, arms, hands and shoulders. It is also able to ignore noisy features, e.g. the mask, the clothes and the background. Regarding the performance for each emotion, our results can be intuitively validated, as discussed. Moving to the next section, we will use 3 segments of the video, as they seem sufficient for the model to form temporal structure and achieve satisfactory performance.

Chapter 5

Model Enhancement

In this chapter, we aspire to fully overcome the consequences of the face mask. We enhance our previous model with some proven techniques and look to achieve performance results as good as with the default/unmasked input.

5.1 Temporal Shift Module

The Temporal Shift Module (TSM) [131] is a generic and effective technique that enjoys both high efficiency and high performance. It can be inserted into CNNs to achieve temporal modeling at zero computation and zero parameters. Concretely, an activation in a video model can be represented as $A \in \mathbb{R}^{N \times C \times T \times H \times W}$, where N is the batch size, C is the number of channels, T is the temporal dimension, H and W are the spatial resolutions. In Fig. 5.1, each distinct tensor is of dimensions $N \times H \times W$. TSM shifts part of the channels C along the temporal dimension T, both forward and backward; thus facilitate information exchanged among neighboring frames. Finally, edge tensors are either truncated if there is no previous frame, or padded with zeros if there is no future frame in the temporal segment. We experiment with shifting 1/8 and 1/4 of the number of channels, which are the proposed fractions.

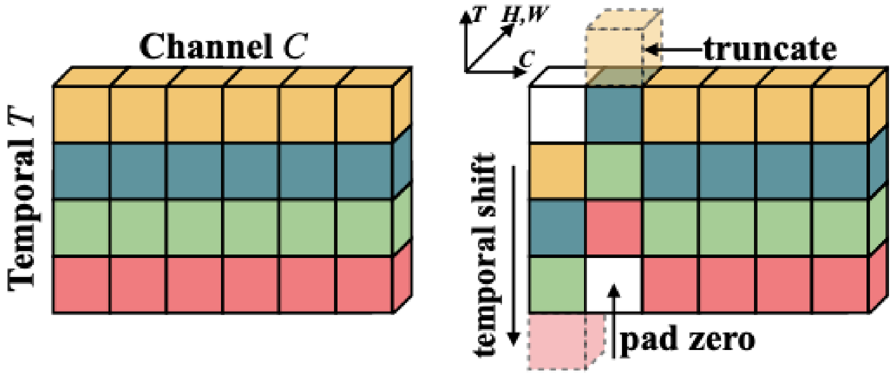
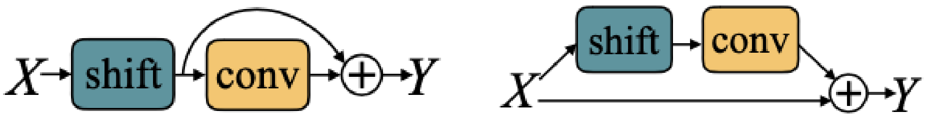


Figure 5.1: Temporal Shift Module



(a) In-place Shift

(b) Residual Shift

Figure 5.2: Module Placement

5.1.1 Partial Shift

The option to use partial shift, and not simply shifting all of the channels to neighboring frames, is justified by the drawbacks of the latter technique. While it enjoys no computation, it involves data movement which increases the memory footprint and inference latency on hardware. To make matters worse, this effect is exacerbated by the large activation map size (5D tensor) of working with videos. This is making the overall inference slow and is something that should be avoided.

5.1.2 Residual Shift

By shifting part of the channels to neighboring frames, the information contained in the channels is no longer accessible for the current frame, which may harm the spatial modeling ability of the CNN backbone. Therefore, we need to balance the model capacity for spatial feature learning and temporal feature learning. Instead of inserting TSM as In-place Shift, which means before each convolutional layer or residual block (Fig. 5.2a), we place it inside the residual branch as Residual Shift (Fig. 5.2b). This way, we can address the degraded spatial feature learning problem, as all the information in the original activation is still accessible after temporal shift through identity mapping.

5.1.3 TSM vs TSN

We insert the TSM technique on our existing TSN model, by implementing the aforementioned partial residual shift on several backbone blocks. Until now, the TSN model processed only one of the N consecutive frames of each snippet. Therefore, we were heavily based on spatial structure. With the TSM module, we attempt to exploit temporality, by mingling information among neighboring snippet frames and expect performance improvement, along with unnoticeable increase in training time.

5.1.4 TSM Results

In Table 5.1, a clear observation that can be made, is that inserting TSM improves performance very slightly. Either by shifting 1/4 or 1/8 of the channels, the difference is minimal. We come to the conclusion, that spatial feature learning plays a more important role for an emotional expression, while temporal structure is complementary.

Table 5.1: TSN vs TSM Model Performance Comparison

Model	**Input - 3 Segments**	**Shift**	**ROC AUC**
TSN	Masked Face	-	0.733
TSN	Masked Full Body	-	0.758
TSM	Masked Full Body	1/8	0.762
		1/4	**0.763**

5.2 Modality Fusion

For further model enhancement, we are looking to take advantage of the face and body information separately, by fusing the preliminary modality prediction scores with a late fusion scheme.

5.2.1 Fusion Method

Currently, our baseline model processes the full body crop, which also includes the masked face, as a single RGB input image, which can lead to irrelevant information confusion. The proposed method is separating the face and body features, in order to avoid the aforementioned issue. Our core model remains as is, but now processes the face crop, and the plain body crop with the corresponding face area blacked out, in two separate forward passes (Fig. 5.3). Though this way the computational load is increased, basically doubled, we do not add any extra trainable parameters to the model. After producing the scores S_f and S_b from face and plain body respectively, we use a late fusion scheme to obtain the final scores S. Finally, the final loss $\mathcal{L}$ is simply the summation of the individual modality losses $\mathcal{L}_f$ and $\mathcal{L}_b$, while all other model configurations remain the same as described in Chap. 3 .

5.2.2 Fusion Results

In Table 5.2, we report fusion results after experimenting with two different aggregation functions: maximum and average. We also present an extra row of the plain body input, the performance of which is expected to be on the same scale with the masked face. A first observation we can make is, that using maximum as the aggregation function gives poor results, as it is actually outperformed by the plain body crop method. That might happen, because we are utilizing different modality information with a single input and wrong positive predictions (false positives) from

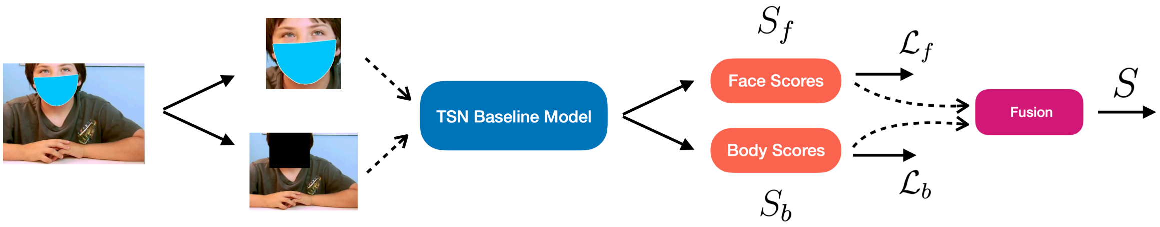


Figure 5.3: Modality Score Late Fusion Scheme

one modality are canceling out possible correct negative predictions (true negatives) from the other. On the other hand, averaging seems a choice that blends well, as it clearly improves performance. Intuitively, it makes sense to have a balanced consensus between the input modalities, as the way the various emotions are expressed, generally differ. For reference, the last row reports the balanced ROC AUC result from [112], where features extracted from the OpenFace framework are used with an SVM classifier, which our TSM Fusion method clearly outperforms.

Table 5.2: TSN Fusion Scheme Performance Comparison

Input - 3 Segments	**Aggregation**	**ROC AUC**
Masked Face	-	0.733
Plain Body	-	0.736
Fusion	Maximum	0.724
	Average	**0.764**

5.3 Method Combination

In Table 5.3, we report results when combining the TSM and fusion techniques. It seems that when utilizing both, the same conclusions as earlier apply. That means, TSM seems to give slight temporal modeling ability to the model and the fusion method results suggest that it effectively takes advantage of the face and body information separately, and possibly avoids irrelevant information confusion. The best overall performance is 0.768 ROC AUC and is achieved by the averaging fusion method, when using TSM with 1/4 partial shift. Compared to 0.769, which is the best face result achieved with no mask applied, reported in Table 4.2, we almost fully overcome face information loss and achieve similar performance. For reference, the last row reports the balanced ROC AUC result from [112], where features extracted from the OpenFace framework are used with an SVM classifier, which our TSM Fusion method clearly outperforms.

Table 5.3: Method Combination Performance Results

Model	**Input - 3 Segments**	**Shift**	**Aggregation**	**ROC AUC**	
				Unbalanced	Balanced
TSN	Masked Face	-	-	0.733	-
TSN	Masked Full Body	-	-	0.758	-
TSM	Fusion	1/8	Max.	0.729	-
			Avg.	0.767	-
		1/4	Max.	0.731	-
			Avg.	**0.768**	**0.696**
TSN	Unmasked Face	-	-	0.769	0.698
[112]		-	-	-	0.620

Chapter 6

Real-Time Online Emotion Recognition

In this chapter, we create a real-time setup of the model. This is an essential step towards making a real-world emotion recognition interface. We will present two different versions, depending on the input modality used, whether it will be face or body.

6.1 Domain Expansion

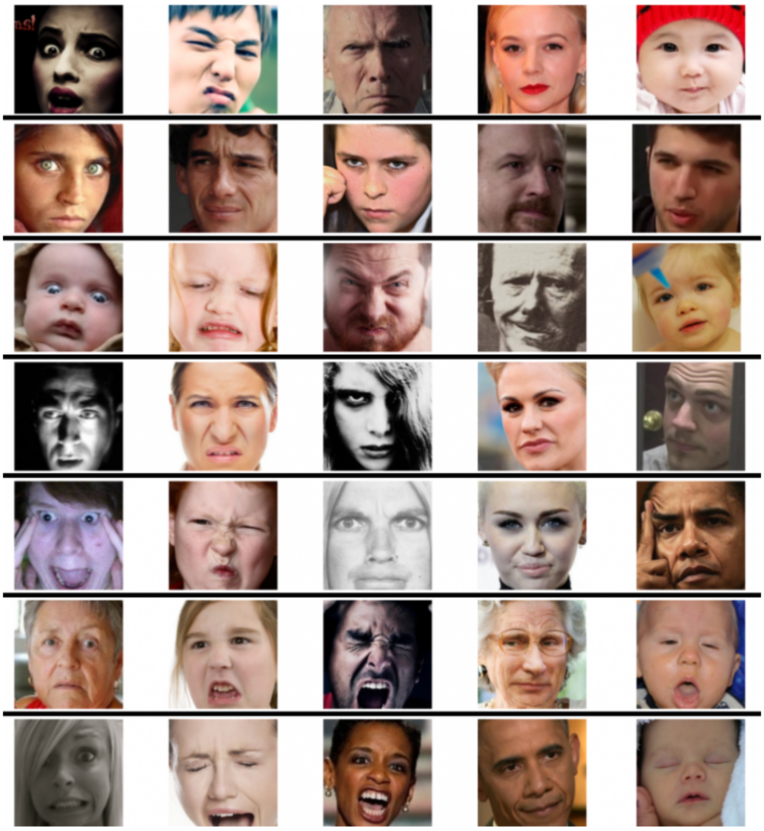
Until now, we conducted our experiments in the EmoReact database, which contains a limited amount of samples. As a result, it lacks diversity in the subjects' behaviour and appearance. On top of that, the subjects are within a restricted range of age (4-14). All of these issues would cause difficulties if we were to implement our emotion recognition system in a real-time setup, where we evaluate it on a real-world scenario. What we will do to address them, is train our model in other larger databases with a wider domain of samples.

AffectNet

AffectNet [132] is by far the largest database of facial expressions that provides both categorical and valence, arousal annotations. It contains more than one million images with faces and extracted facial landmark points (Fig. 6.1). Twelve human experts manually annotated 450,000 of these images in both categorical and dimensional (valence and arousal) models and tagged the images that have any occlusion on the face. It is a very challenging database as it contains images of people from different races and ethnic groups as well as high variety in the background, lighting, pose, point of view, etc.

BoLD

As for the bodily expression database, we will train our model on BoLD, as presented in section 2.4. It contains a larger amount and wider range of bodily expressed features, that will help towards model generalization.



(credit: [132])

Figure 6.1: AffectNet Samples

6.2 Lightweight Backbone Architecture

Real-time recognition demands lightweight networks, in order for inference time to be low. The ResNet-50 backbone architecture that we currently use, despite its great advantages, is computiationally heavy and cannot operate time-efficiently. For this reason, we choose MobileNetV2 [126], an architecture specifically tailored for mobile and resource constrained environments. This network pushed the state of the art, by significantly decreasing the number of operations and memory needed, while retaining the same accuracy. Its basic building block is a bottleneck depth-separable convolution with residuals, shown in Table 6.1, which transforms the input from k to k' channels, with stride s for height and width.

Table 6.1: Bottleneck Residual Block

Input	**Operator**	**Output**
$h \times w \times k$	1x1 conv2d, ReLU6	$h \times w \times (tk)$
$h \times w \times tk$	3x3 dwise s=s, ReLU6	$\frac{h}{s} \times \frac{w}{s} \times (tk)$
$\frac{h}{s} \times \frac{w}{s} \times tk$	linear 1x1 conv2d	$\frac{h}{s} \times \frac{w}{s} \times k'$

The first key idea is depthwise separable convolutions [133], which replace a full convolutional operator with a factorized version that splits convolution into two separate layers. The first layer is called a depthwise convolution, it performs lightweight filtering by applying a single convolutional filter per input channel. The second layer is a 1×1 convolution, called a pointwise convolution, which is responsible for building new features through computing linear combinations of the input channels. This way, computation is reduced by almost a factor of k^2, where k denotes the convolution kernel width/height dimension. MobileNetV2 uses $k = 3$ (3×3 depthwise separable convolutions) so the computational cost is 8 to 9 times smaller than that of standard convolutions at only a small reduction in accuracy [134].

Another technique that is exploited is linear bottlenecks. For an input set (batch) of real images, we say that the set of layer activations, for each layer, forms a "manifold of interest". Assuming that these manifolds of interest could be embedded in low-dimensional subspaces, we can optimize existing neural architectures, by inserting linear bottleneck layers into the convolutional blocks. Linear bottlenecks simply reduce the dimensionality of a layer, thus reduce the dimensionality of the operating space [134].

Lastly, inspired by the intuition that the bottlenecks actually contain all the necessary information, shortcuts directly between the bottlenecks are used, instead of classical residuals connections between layers with higher number of channels. This implementation is called inverted residuals [126], and is considerably more memory efficient, while also improving the ability of a gradient to propagate across multiplier layers.

The entire architecture is displayed in Table 6.2. Each line describes a sequence of 1 or more identical (modulo stride) layers, repeated n times. All layers in the same sequence have the same number c of output channels. The first layer of each sequence has a stride s and all others use stride 1. All spatial convolutions use 3×3 kernels. The expasions factor t is always applied to the input size. The non-linearity used is ReLU6 because of its robustness when used with low-precision computation [134].

Table 6.2: MobileNetV2 Architecture

Input	**Operator**	t	c	n	s
$224^2 \times 3$	conv2d	-	32	1	2
$112^2 \times 32$	bottleneck	1	16	1	1
$112^2 \times 16$	bottleneck	6	24	2	2
$56^2 \times 24$	bottleneck	6	32	3	2
$28^2 \times 32$	bottleneck	6	64	4	2
$14^2 \times 64$	bottleneck	6	96	3	1
$14^2 \times 96$	bottleneck	6	160	3	2
$7^2 \times 160$	bottleneck	6	320	1	1
$7^2 \times 320$	conv2d 1x1	-	1280	1	1
$7^2 \times 1280$	avgpool 7x7	-	-	1	-
$1^2 \times 1 \times 1280$	conv2d 1x1	k	-	1	

Table 6.3: Backbone Comparison for Real-Time Emotion Recognition

Backbone	**Parameters**	**Inference Time**	**FPS**	**Accuracy (%)**	
				ImageNet	AffectNet
ResNet-50	25.0M	170ms	∼6	77.6	57.9
MobileNetV2	3.5M	60ms	∼17	71.9	57.2

In Table 6.3, we compare the two backbone architectures that we experimented with, regarding the number of model trainable parameters, inference time, Frames Per Second (FPS) and accuracy (performance) achieved on the ImageNet and AffectNet datasets.

$$\text{FPS} = \frac{1}{(P + I)}$$

where $P \approx 0.015$ s, denotes preprocessing time, which includes face detection, mask application and frame rescaling, and I denotes inference time. The numbers reported stand for an 1,8 GHz Intel Dual-Core i5 CPU.

6.3 Online Demo

We have now adjusted our setup to operate effectively in real-time and move forward to the evaluation process. We note that throughout this whole thesis, as well as in this section, we utilize a single image modality, that is the RGB tensor. This not only enjoys higher training speed, but also real-time applicability, compared to multi-stream architectures that process a variety of different modalities, including RGB, flow, skeleton sequence, context, audio etc.

In Fig. 6.2, we present the pipeline of our online demo setup. At first, the webcam input frame is forwarded to the MediaPipe Face Mesh tool, to extract the facial landmarks and crop the face region. Then the medical face mask is optionally applied, and then the frame is rescaled to fit the model's input size. Our model is now a plain MobileNetV2 architecture, which produces the expressed emotion prediction, after passing the outputs through a sigmoid function layer and choosing the index with the maximum probability, which we call "confidence". In case of the body version, we skip the facial landmarks tracking step, and instead, simply place the camera and the user at the desired distance, to provide body information.

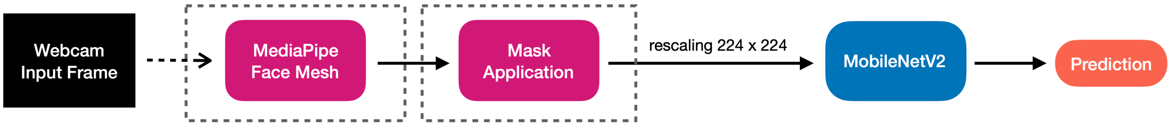


Figure 6.2: Online Demo Pipeline

Face

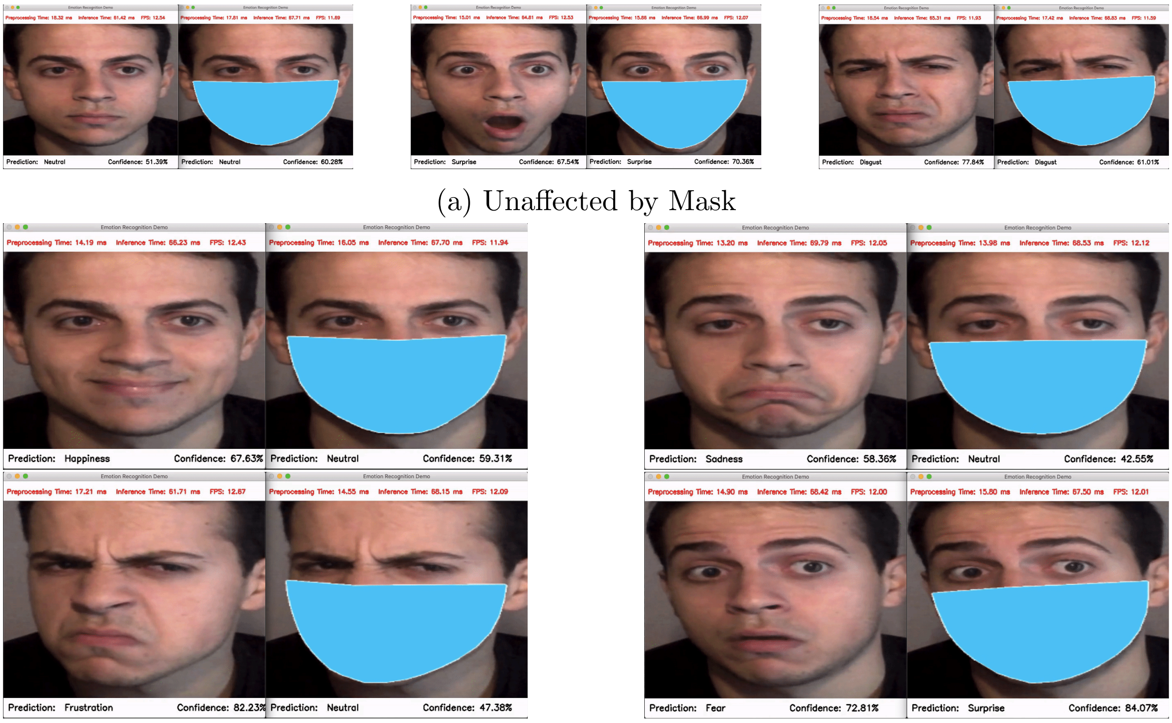
The face version of the demo classifies between the 8 AffectNet emotions, which are: Neutral, Happiness, Sadness, Surprise, Fear, Frustration, Disgust, Uncertainty. In Fig. 6.3, we present some real-time in-the-wild recognition demo examples, where we compare predictions between default and masked input.

Body

The body version of the demo classifies between the 26 BoLD emotions, which involve very complex emotions, like Disconnect, Yearning, Sensitivity, Anticipation and others. In this case, it would be very hard to handle such a big number of classes, so instead we map them into 3 general classes of: Positive, Neutral and Negative (Table 6.4). The prediction is the general class with the maximum confidence, computed as the sum of the individual emotion probabilities, produced by a softmax layer. In Fig. 6.4, we present some examples for the body demo.

Demo Results

Starting from the face, while the model can correctly predict Neutral, Suprise, and Disgust emotions (Fig. 6.3a), we observe that it is prone to mistakes when masked (Fig. 6.3b), as it misclassifies Happiness, Sadness, Frustration, and Fear. These emotions are all, more or less, depedent on facial features that get occluded by the mask. Regarding the body, we observe that the model remains unaffected by the medical face mask in recognizing emotions. One effect we can note is on the top-left



Preprocessing Time: 18.32 ms Inference Time: 61.42 ms FPS: 12.54
Prediction: Neutral Confidence: 51.39%

Preprocessing Time: 17.81 ms Inference Time: 67.71 ms FPS: 11.69
Prediction: Neutral Confidence: 60.28%

Preprocessing Time: 15.01 ms Inference Time: 64.81 ms FPS: 12.53
Prediction: Surprise Confidence: 67.54%

Preprocessing Time: 15.86 ms Inference Time: 66.99 ms FPS: 12.07
Prediction: Surprise Confidence: 70.36%

Preprocessing Time: 18.54 ms Inference Time: 65.31 ms FPS: 11.93
Prediction: Disgust Confidence: 77.84%

Preprocessing Time: 17.42 ms Inference Time: 68.83 ms FPS: 11.59
Prediction: Disgust Confidence: 61.01%

(a) Unaffected by Mask

Preprocessing Time: 14.19 ms Inference Time: 66.23 ms FPS: 12.43
Prediction: Happiness Confidence: 67.63%

Preprocessing Time: 16.05 ms Inference Time: 67.70 ms FPS: 11.94
Prediction: Neutral Confidence: 59.31%

Preprocessing Time: 13.20 ms Inference Time: 69.79 ms FPS: 12.05
Prediction: Sadness Confidence: 58.36%

Preprocessing Time: 13.98 ms Inference Time: 68.53 ms FPS: 12.12
Prediction: Neutral Confidence: 42.55%

Preprocessing Time: 17.21 ms Inference Time: 61.71 ms FPS: 12.67
Prediction: Frustration Confidence: 82.23%

Preprocessing Time: 14.55 ms Inference Time: 68.15 ms FPS: 12.09
Prediction: Neutral Confidence: 47.38%

Preprocessing Time: 14.90 ms Inference Time: 68.42 ms FPS: 12.00
Prediction: Fear Confidence: 72.81%

Preprocessing Time: 15.80 ms Inference Time: 67.50 ms FPS: 12.01
Prediction: Surprise Confidence: 84.07%

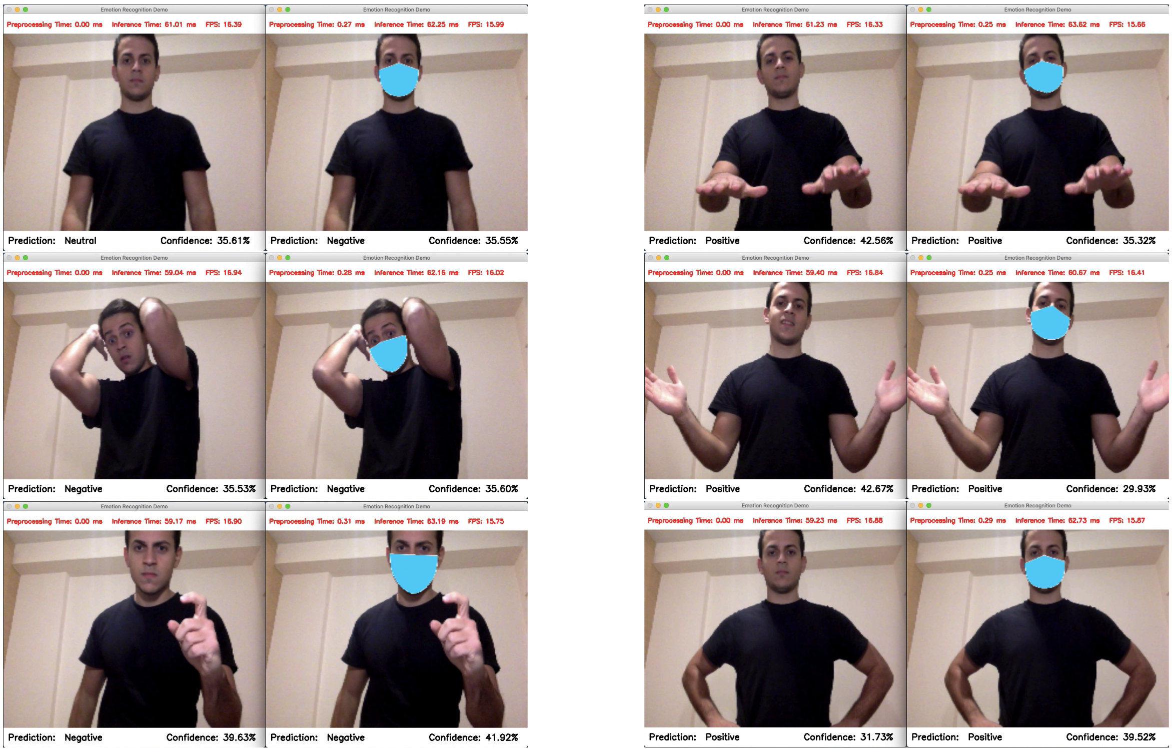
(b) Affected by Mask

Figure 6.3: Real-Time In-the-Wild Face Emotion Recognition Examples

Table 6.4: BoLD Emotion Mapping

General Class	BoLD Emotion
Positive	Happiness, Affection, Esteem, Pleasure, Excitement, Sympathy, Peace, Engagement
Neutral	Surprise, Sensitivity, Confusion Yearning, Anticipation, Confidence
Negative	Sadness, Pain, Anger, Fear, Fatigue, Disconnect, Annoyance, Embarrassment, Suffering, Disquietment, Disapproval, Aversion

comparison example, where the predictions are different. It seems, that the model classifies this body expression as neutrality when the face is unmasked and neutral itself, whereas it predicts negativity from the body when the face is occluded. [1]



```
Preprocessing Time: 0.00 ms   Inference Time: 61.01 ms   FPS: 16.39
Prediction:  Neutral            Confidence:  35.61%

Preprocessing Time: 0.27 ms   Inference Time: 62.25 ms   FPS: 15.99
Prediction:  Negative           Confidence:  35.55%

Preprocessing Time: 0.00 ms   Inference Time: 59.04 ms   FPS: 16.94
Prediction:  Negative           Confidence:  35.53%

Preprocessing Time: 0.28 ms   Inference Time: 62.16 ms   FPS: 16.02
Prediction:  Negative           Confidence:  35.60%

Preprocessing Time: 0.00 ms   Inference Time: 59.17 ms   FPS: 16.90
Prediction:  Negative           Confidence:  39.63%

Preprocessing Time: 0.31 ms   Inference Time: 63.19 ms   FPS: 15.75
Prediction:  Negative           Confidence:  41.92%

Preprocessing Time: 0.00 ms   Inference Time: 61.23 ms   FPS: 16.33
Prediction:  Positive           Confidence:  42.56%

Preprocessing Time: 0.25 ms   Inference Time: 63.62 ms   FPS: 15.66
Prediction:  Positive           Confidence:  35.32%

Preprocessing Time: 0.00 ms   Inference Time: 59.40 ms   FPS: 16.84
Prediction:  Positive           Confidence:  42.67%

Preprocessing Time: 0.25 ms   Inference Time: 60.67 ms   FPS: 16.41
Prediction:  Positive           Confidence:  29.93%

Preprocessing Time: 0.00 ms   Inference Time: 59.23 ms   FPS: 16.88
Prediction:  Positive           Confidence:  31.73%

Preprocessing Time: 0.29 ms   Inference Time: 62.73 ms   FPS: 15.87
Prediction:  Positive           Confidence:  39.52%
```

Figure 6.4: Real-Time In-the-Wild Body Emotion Recognition Examples

[1] The code for the emotion recognition demo is publicly available at: https://github.com/nkegke/emotion-demo

Chapter 7

Conclusion & Future Work

In this chapter, we summarize the conclusions that came out of this thesis:

- The COVID-19 pandemic has forced people to extensively wear medical face masks, in order to prevent transmission. This face occlusion results in considerable emotion recognition performance drop by models that exploit facial expressions. Therefore, we are urged to incorporate the whole body in the input, as it has to play a more major role in the task of recognition, despite its complementary nature.

- Model training heavily relies on the data it is fed with. Emotion recognition in children is not understudied by chance, as Child-Robot Interaction problems in general, are more challenging compared to adults. The dataset we choose, featuring children reacting to different topics, neither contains a very large amount of samples, nor is purely about bodily expression. This adds an extra degree of difficulty, demands fine-tuned configurations, and general facts do not necessarily apply. This also makes model generalization harder, and regarding in-the-wild recognition, we are in need of expanding the domain to larger databases.

- For affect perception from the body, both form (spatial) and motion (temporal) information can be analyzed. Form information can be instrumental in emotion recognition and yield great results. TSN-based models can exploit spatial information from multiple frames by sparsely sampling a video. On top of that, complex actions like emotional expressions, consist of multiple stages spanning over a period of time, which indicates we should also capture temporal emotional expression features. A TSN can operate on a sequence of frames, instead of a single frame, and utilize temporal information. Combining TSN with a residual backbone architecture, that performs an identity mapping, it comprises a powerful emotion recognition model, that can be used to overcome face occlusion.

- Visual explanation techniques give insights on model understanding and modality feature learning. Single stream RGB input models, like our TSN, seem to be able to learn both facial and bodily expressive features. The model's dependence on modality features to make predictions vary, and the recognition results per emotion can be intuitively validated.

- The TSN model that processes the full body masked image can achieve great performance, but there is room to pursue improvements, in order to to fully overcome the consequences of the face mask. In other words, we can enhance our model to perform as good as it does with the unmasked input. Its architecture can naturally support temporal modeling, by mingling information among neighboring snippet frames with the TSM module. Experimental results suggest that spatial structure plays a more important role for each emotional expression, while temporal structure is complementary.

- Although TSN can learn both face and body features in a single stream, this may lead to irrelevant information confusion. By processing those features separately and fusing their preliminary prediction scores with a late fusion scheme, we are more effectively taking advantage of both modalities. In combination with TSM, these techniques enhance the model and help achieve unmasked input performance.

- Real-world emotion recognition applications demand real-time performance. To operate efficiently, major changes in our setup are needed. Except for database domain expansion that we mentioned earlier, the backbone network architecture has to be lightweight, which means have a relatively low amount of trainable parameters, and therefore fast inference.

Part of this thesis was submitted as a paper [136] to the 16th PErvasive Technologies Related to Assistive Environments conference (PETRA 2023).

However, there is always room for improvement, as some aspects of the subject were not covered. For future work, we would suggest:

- The medical face mask application tool that we created, generates the 3D keypoint coordinates of the MediaPipe face mesh model, but only visualizes in 2D. By taking advantage of the depth dimension, one can create a visually more realistic mask surface. This kind of filters have been all over social media, but are implemented only on mobile-supported programming languages, not Python.

- Most datasets, like EmoReact, provide annotations not only for categorical but also for dimensional emotions (VAD). By defining the appropriate loss function and evaluation metrics, one can explore the face occlusion's effect on VAD.

- While we only processed the RGB image, one can take advantage of additional body information, e.g. from the body skeleton joints, by feeding them in a sequence or graph model, and possibly improve performance even more.

- Real-time recognition can be optimized with deep learning compilers [135], by taking models described in different frameworks and generating optimized code for diverse hardware.

Bibliography

[1] A. R. Damasio, *Descartes' Error: Emotion, Reason, and the Human Brain.* New York :G.P. Putnam, 1994.

[2] W. James, "II.—What Is An Emotion ?", *Mind*, vol. 9, pp. 188–205, Apr. 1884.

[3] C. Darwin, *The Expression of the Emotions in Man and Animals*, Cambridge University Press, 2013.

[4] J. Tao and T. Tan, "Affective Computing: A Review," in *Affective Computing and Intelligent Interaction*, vol. 3784, pp. 981–995, Oct. 2005.

[5] R. W. Picard, "Affective Computing," MIT Media Laboratory; Perceptual Computing Section, USA, Tech. Rep. 321, 1997.

[6] K. R. Scherer, "Psychological models of emotion," *The neuropsychology of emotion,* vol. 3, pp. 137–162, May 2000.

[7] D. Grandjean, D. Sander, and K. R. Scherer, "Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization", *Consciousness and Cognition,* vol. 17, pp. 484–495, June 2008.

[8] Tomkins, S. S., *Affect, imagery, consciousness: Vol. 1. The positive affects,* New York: Springer, 1962.

[9] Tomkins, S. S., *Affect, imagery, consciousness: Vol. 2. The positive affects,* New York: Springer, 1963.

[10] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *Journal of personality and social psychology,* vol. 17, pp. 124-129, Apr. 1971.

[11] P. Ekman et al. "Universals and cultural differences in the judgments of facial expressions of emotion," *Journal of Personality and Social Psychology* vol. 53 pp. 712-717, Oct. 1987.

[12] P. Ekman, "An argument for basic emotions," *Cognition and Emotion,* vol. 6, pp. 169–200, May 1992.

[13] P. Ekman, "Basic emotions", *Handbook of cognition and emotion,* T. Dalgleish & M. Power (Eds.), Chichester, England: Wiley, 1999, pp. 45–60.

[14] C. Yu, P. M. Aoki, and A. Woodruff, "Detecting User Engagement in Everyday Conversations", in *Proc. 8th Int'l Conf. on Spoken Language Processing*, ICSLP 2004, pp. 1329-1332.

[15] J. A. Russell, "Is There Universal Recognition of Emotion From Facial Expression? A Review of the Cross-Cultural Studies," *Psychological Bulletin,* vol. 115, pp. 102-141, Jan. 1994.

[16] J. A. Russell, "A circumplex model of affect", *Journal of Personality and Social Psychology,* vol. 39, pp. 1161–1178, Dec. 1980.

[17] A. Mehrabian and J. A. Russell, *An approach to environmental psychology,* The MIT Press, 1974.

[18] H. Gunes and M. Pantic, "Automatic, Dimensional and Continuous Emotion Recognition," *International Journal of Synthetic Emotions,* vol. 1, pp. 68–99, Jan. 2010.

[19] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen, "Music Emotion Classification: A Regression Approach", in *IEEE International Conference on Multimedia and Expo, ICME 2007,* Beijing, China, 02-05 Jul. 2007, pp. 208–211

[20] A. Tursunov, S. Kwon, and H.-S. Pang, "Discriminating Emotions in the Valence Dimension from Speech Using Timbre Features", *Applied Sciences,* vol. 9, 470, June 2019.

[21] O. Mitruț, G. Moise, Gabriela L. Petrescu, A. Moldoveanu, M. Leordeanu, and F. Moldoveanu, "Emotion Classification Based on Biophysical Signals and Machine Learning Techniques", *Symmetry,* vol. 12, 21, Dec. 2019.

[22] K. Scherer, A. Schorr, and T. Johnstone, *Appraisal Processes in Emotion: Theory, Methods, Research,* Oxford University Press, 2001.

[23] J. Kumari, R. Rajesh, and K. M. Pooja, "Facial Expression Recognition: A Survey," *Procedia Computer Science,* vol. 58, pp. 486–491, Aug. 2015.

[24] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," *IEEE Transactions of Affective Computing,* vol. 13, pp. 1195–1215, Mar. 2020.

[25] B. de Gelder, "Towards the neurobiology of emotional body language," *Nature Reviews Neuroscience,* vol. 7, pp. 242–249, Mar. 2006.

[26] B. de Gelder, "Why bodies? twelve reasons for including bodily expressions in affective neuroscience," *Philosophical Transactions of the Royal Society of London B: Biological Sciences,* vol. 364, pp. 3475–3484, Dec. 2009.

[27] H. Aviezer, Y. Trope, and A. Todorov, "Body cues, not facial expressions, discriminate between intense positive and negative emotions," *Science,* vol. 338, pp. 1225-1229, Nov. 2012.

[28] C.-C. Carbon, "Wearing Face Masks Strongly Confuses Counterparts in Reading Emotions," *Frontiers in Psychology,* vol. 11, 566886, Sep. 2020.

[29] M. Shiffrar, M. D. Kaiser, and A. Chouchourelou, "Seeing Human Movement as Inherently Social," *The Science of Social Vision*, in Reginald B. Adams and others (eds), pp. 248–263, Nov. 2010.

[30] P. P. Filntisis, N. Efthymiou, P. Koutras, G. Potamianos, and P. Maragos, "Fusing Body Posture with Facial Expressions for Joint Recognition of Affect in Child-Robot Interaction," *IEEE Robotics and Automation Letters,* vol. 4, pp. 4011–4018, Oct. 2019.

[31] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition,* vol. 44, pp. 572–587, Mar. 2011.

[32] B. Reeves and C. Nass, *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Pla.,* Bibliovault OAI Repository, the University of Chicago Press, 1996.

[33] C. Breazeal, "Emotion and sociable humanoid robots," *International Journal of Human-Computer Studies,* vol. 59, pp. 119–155, Jul. 2003.

[34] A. Tsiami, P. Koutras, N. Efthymiou, P. P. Filntisis, G. Potamianos, and P. Maragos, "Multi3: Multi-Sensory Perception System for Multi-Modal Child Interaction with Multiple Robots," in *IEEE International Conference on Robotics and Automation, ICRA 2018,* QLD, pp. 4585–4592, 2018.

[35] T. Belpaeme et al., "Child-Robot Interaction: Perspectives and Challenges," in *International Conference on Social Robotics, ICSR 2013,*, Springer, vol. 8239, pp. 452–459, 2013.

[36] B. W. Kort, *Personal Communication,* 1995.

[37] R. Pekrun, T. Götz, W. Titz, and R. P. Perry, "Academic emotions in students' self-regulated learning and achievement : a program of qualitative and quantitative research," *Educational Psychologist,* vol. 37, pp. 91-105, June 2002.

[38] G. Gordon et al., "Affective Personalization of a Social Robot Tutor for Children's Second Language Skills," in *Proceedings of the Thirtieth Conference on Artificial Intelligence, AAAI'16,* pp. 1343-1349, 2016.

[39] I. Leite, G. Castellano, A. Pereira, C. Martinho, and A. Paiva, "Modelling empathic behaviour in a robotic game companion for children: an ethnographic study in real-world settings," in *Proceedings of the seventh IEEE International Conference on Human-Robot Interaction, IEEE HRI '12,* Boston, Massachusetts, USA, 2012.

[40] Z. Yang, A. Kay, Y. Li, W. Cross, and J. Luo, "Pose-based Body Language Recognition for Emotion and Psychiatric Symptom Interpretation," in *International Conference on Pattern Recognition, ICPR 2020,* Milan, Italy, 2020.

[41] Z. Lian, Y. Guo, X. Cao, and W. Li, "An Ear Wearable Device System for Facial Emotion Recognition Disorders," *Frontiers in Bioengineering and Biotechnology* vol. 9, 703048, 2021.

[42] N. Savva, A. Scarinzi, and N. Bianchi-Berthouze, "Continuous Recognition of Player's Affective Body Expression as Dynamic Quality of Aesthetic Experience", in *IEEE Transactions on Computational Intelligence and AI in Games,* vol. 4, pp. 199-212, Sep. 2012.

[43] M. A. Giese and T. Poggio, "Neural Mechanisms for the Recognition of Biological Movements," *Neuroscience,* vol. 4, pp. 179-191, Mar. 2003.

[44] L. M. Vania, M. Lemay, D.C. Bienfang, A.Y. Choi, and K. Nakayama, "Intact Biological Motion and Structure from Motion Perception in a Patient with Impaired Motion Mechanisms: A Case Study," *Visual Neuroscience,* vol. 5, pp. 353-369, Oct. 1990.

[45] J. Lange and M. Lappe, "The Role of Spatial and Temporal Information in Biological Motion Perception," *Advances in Cognitive Psychology,* vol. 3, pp. 419-428, July 2007.

[46] M. Hirai and K. Hiraki, "The Relative Importance of Spatial versus Temporal Structure in the Perception of Biological Motion: An Event-Related Potential Study," *Cognition,* vol. 99, pp. B15-29, Feb. 2006.

[47] T. Bänziger, M. Mortillaro, and K. R. Scherer, "Introducing the Geneva Multimodal Expression Corpus for Experimental Research on Emotion Perception," *Emotion,* vol. 12, pp. 1161-1179, 2012.

[48] A. P. Atkinson, W.H. Dittrich, A.J. Gemmell, and A.W. Young, "Evidence for Distinct Contributions of Form and Motion Information to the Recognition of Emotions from Body Gestures," *Cognition,* vol. 104, pp. 59-72, 2007.

[49] P. Ross and T. Flack, "Removing Hand Form Information Specifically Impairs Emotion Recognition for Fearful and Angry Body Stimuli," *Perception,* vol. 49, pp. 98–112, 2020.

[50] P. Ekman and W.V. Friesen, *Facial Action Coding System: A technique for the measurement of facial movement,* Consulting Psychologists Press, Stanford University, Palo Alto, 1977.

[51] N. Dael, M. Mortillaro, and K. R. Scherer, "The Body Action and Posture Coding System (BAP): Development and Reliability," *Journal of Nonverbal Behavior,* vol. 36, pp. 97–121, June 2012.

[52] A. Kleinsmith and N. Bianchi-Berthouze, "Affective Body Expression Perception and Recognition: A Survey", in *IEEE Transactions on Affective Computing,* vol. 4, pp. 15–33, Jan. 2013.

[53] J. Ye, J. Li, M. G. Newman, R. B. Adams Jr., and J. Z. Wang, "Probabilistic Multigraph Modeling for Improving the Quality of Crowdsourced Affective Data", in *IEEE Transactions on Affective Computing,* vol. 10, pp. 115–128, Jan. 2019.

[54] A. Kleinsmith, P. R. De Silva, and N. Bianchi-Berthouze, "Cross-cultural differences in recognizing affect from body posture", *Interacting with Computers,* vol. 18, pp. 1371–1389, Dec. 2006.

[55] T. M. Mitchell, *Machine learning,* vol. 1. No. 9. New York: McGraw-hill, 1997.

[56] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," *Nature,* vol. 521, pp. 436-444, 2015.

[57] D. Rumelhart, G. Hinton, and R. Williams, "Learning representations by back-propagating errors," *Nature,* vol. 323, pp. 533–536, 1986.

[58] S. Ruder, "An overview of gradient descent optimization algorithms," arXiv:1609.04747 [cs.LG], 2016.

[59] F. Girosi, M. Jones, and T. Poggio, "Regularization Theory and Neural Networks Architectures," *Neural Computation,* vol. 7, pp. 219-269, 1995.

[60] D. T. Nguyen, W. Li, and P. O. Ogunbona, "Human detection from images and videos: A survey," *Pattern Recognition,* vol. 51, pp. 148–175, Mar. 2016.

[61] N. Dalal, "Finding people in images and videos," Ph.D. dissertation, Institut National Polytechnique de Grenoble - INPG, 2006.

[62] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," in *The 9th IEEE International Conference on Computer Vision, ICCV 2003,* Nice, France, vol. 1, pp. 734–741, 2003.

[63] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001,* Kauai, IH, USA, vol. 1, pp. 511-518, 2001.

[64] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting", In *European Conference on Computational Learning Theory, EUROCOLT 95,* pp. 23–37, Springer, 1995.

[65] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'05,* San Diego, CA, USA, vol. 1, pp. 886-893, 2005.

[66] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian Detection: An Evaluation of the State of the Art," in *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 34, no. 4, pp. 743–761, Apr. 2012.

[67] N. Dalal, B. Triggs, and C. Schmid, "Human Detection Using Oriented Histograms of Flow and Appearance," in *The 12th IEEE European Conference on Computer Vision, ECCV 2006,* vol. 3952, pp. 428–441, 2006.

[68] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008,* Anchorage, AK, USA, pp. 1–8, 2008.

[69] K. Mikolajczyk, C. Schmid, and A. Zisserman, "Human Detection Based on a Probabilistic Assembly of Robust Part Detectors," in *The 10th IEEE European Conference on Computer Vision, ECCV 2004,* vol. 3021, Prague, Czech Republic, pp. 69–82, 2004.

[70] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning,* vol. 20, pp. 273–297, Sep. 1995.

[71] A. Satpathy, X. Jiang, and H.L. Eng, "Human detection by quadratic classification on subspace of extended histogram of gradients," in *IEEE Transactions on Image Processing,* vol. 23, pp. 287–297, 2014.

[72] W. Ouyang and X. Wang, "A discriminative deep model for pedestrian detection with occlusion handling," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2012,* Providence, RI, pp. 3258–3265, 2012.

[73] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A Fast Learning Algorithm for Deep Belief Nets," *Neural Computation,* vol. 18, pp. 1527–1554, Jul. 2006.

[74] W. Ouyang and X. Wang, "Joint Deep Learning for Pedestrian Detection," in *IEEE International Conference on Computer Vision, ICCV 2013,* Sydney, Australia, pp. 2056–2063, 2013.

[75] P. Luo, Y. Tian, X. Wang, and X. Tang, "Switchable Deep Network for Pedestrian Detection," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014,* Columbus, OH, USA, pp. 899–906, 2014.

[76] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," arXiv:1611.08050 [cs.CV], Apr. 2017.

[77] H.-B. Zhang, Q. Lei, Bi.-N. Zhong, J.-X. Du, and J. Peng, "A Survey on Human Pose Estimation," *Intelligent Automation & Soft Computing,* vol. 22, pp. 483-489, 2016.

[78] M. A. Fischler and R. A. Elschlager, "The representation and matching of pictorial structures," *IEEE Transactions on Computers,*, vol. 22, pp. 67–92, 1973

[79] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient matching of pictorial structures," in *IEE Conference on Computer Vision and Pattern Recognition, CVPR 2000,* Hilton Head Island, SC, USA, 2000.

[80] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial Structures for Object Recognition," *International Journal of Computer Vision,* vol. 61, pp. 55–79, Jan. 2005.

[81] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in *IEEE CVPR Workshops 2009,* Miami, FL, USA, 2009.

[82] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 35, pg. 2878–2890, 2013.

[83] Q. Dang, J. Yin, B. Wang, and W. Zheng, "Deep learning based 2D human pose estimation: A survey", *Tinshhua Science and Technology,* vol. 24, pp. 663–676, Dec. 2019.

[84] C. Zheng et al., "Deep Learning-Based Human Pose Estimation: A Survey," arXiv:2012.13392 [cs.CV], Jan. 2021.

[85] A. Toshev and C. Szegedy, "DeepPose: Human Pose Estimation via Deep Neural Networks," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014,* pp. 1653–1660, Jun. 2014.

[86] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, "Human Pose Estimation with Iterative Error Feedback," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016,* Las Vegas, NV, USA, pp. 4733–4742, Jun. 2016.

[87] X. Sun, J. Shang, S. Liang, and Y. Wei, "Compositional human pose regression," in *the IEEE International Conference on Computer Vision, ICCV 2017,* Venice, Italy, 2017

[88] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional Pose Machines," arXiv:1602.00134 [cs.CV], Apr. 2016

[89] A. Newell, K. Yang, and J. Deng, "Stacked Hourglass Networks for Human Pose Estimation," arXiv:1603.06937 [cs.CV], Jul. 2016.

[90] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27, NeurIPS 2014,* Montreal, Canada, 2014.

[91] Y. Chen, C. Shen, X.-S. Wei, L. Liu, and J. Yang, "Adversarial posenet: A structure-aware convolutional network for human pose estimation," arXiv:1705.00389 [cs.CV], Apr. 2017.

[92] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, “BlazePose: On-device Real-time Body Pose tracking,” arXiv:2006.10204 [cs.CV], Jun. 2020.

[93] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, “Hand keypoint detection in single images using multiview bootstrapping,” arXiv:1704.07809 [cs.CV], Apr. 2017.

[94] H. Gunes and M. Piccardi, “Affect Recognition from Face and Body: Early Fusion vs. Late Fusion,” in *IEEE International Conference on Systems, Man and Cybernetics,* Waikoloa, HI, USA, vol. 4, pp. 3437–3443, 2005.

[95] M. Kipp and J.-C. Martin, “Gesture and emotion: Can basic gestural form features discriminate emotions?”, in *The third International Conference on Affective Computing and Intelligent Interaction and Workshops,* Amsterdam, Netherlands, Sep. 2009, pp. 1–8.

[96] D. Glowinski, A. Camurri, G. Volpe, N. Dael, and K. Scherer, “Technique for automatic emotion recognition by body gesture analysis,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops,* Anchorage, AK, USA, Jun. 2008, pp. 1–6.

[97] D. Glowinski, N. Dael, A. Camurri, G. Volpe, M. Mortillaro, and K. Scherer, “Toward a Minimal Representation of Affective Gestures,” IEEE Transactions on Affective Computing, vol. 2, pp. 106–118, Apr. 2011.

[98] T. Sapiński, D. Kamińska, A. Pelikant, and G. Anbarjafari, “Emotion Recognition from Skeletal Movements,” Entropy, vol. 21, p. 646, Jun. 2019.

[99] G. Castellano, S. D. Villalba, and A. Camurri, “Recognising Human Emotions from Body Movement and Gesture Dynamics,” in *Affective Computing and Intelligent Interaction,* vol. 4738, pp. 71–82, 2007.

[100] Y. Luo, J. Ye, J. Adams, J. Li, M. G. Newman, and J. Z. Wang, “ARBEE: Towards Automated Recognition of Bodily Expression of Emotion In the Wild,” *International Journal of Computer Vision,* vol. 128, pp. 1–25, Jan. 2020.

[101] R. Laban and L. Ullmann, “The Mastery of Movement,” ERIC. 1971.

[102] S. Herath, M. Harandi, and F. Porikli, “Going deeper into action recognition: A survey,” *Image and Vision Computing,* vol. 60, pp. 4–21, Apr. 2017.

[103] Y. Zhu et al., “A Comprehensive Study of Deep Video Action Recognition,” arXiv:2012.06567 [cs.CV], Dec. 2020.

[104] K. Simonyan and A. Zisserman, “Two-Stream Convolutional Networks for Action Recognition in Videos,” arXiv:1406.2199 [cs.CV], Nov. 2014.

[105] S. Yan, Y. Xiong, and D. Lin, “Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition,” arXiv:1801.07455 [cs.CV], Jan. 2018.

[106] I. Pikoulis, P. P. Filntisis, and P. Maragos, “Leveraging semantic scene characteristics and multi-stream convolutional architectures in a contextual approach for video-based visual emotion recognition in the wild,” arXiv:2105.07484, May 2021.

[107] J. Shi, C. Liu, C.T. Ishi, and H. Ishiguro, “Skeleton-Based Emotion Recognition Based on Two-Stream Self-Attention Enhanced Spatial-Temporal Graph Convolutional Network,” *Sensors,* vol. 21, 205, 2021.

[108] A. Vaswani et al. “Attention is all you need,” In *Proceedings of the Advances in Neural Information Processing Systems, NIPS 2017,* Long Beach, CA, USA, 2017, pp. 5998–6008.

[109] L. Shi, Y. Zhang, J. Cheng, and H. Lu, “Two-stream adaptive graph convolutional networks for skeleton-based action recognition,” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019,* Long Beach, CA, USA, 2019, pp. 12026–12035.

[110] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, “Emotion Recognition in Context,” in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017,* Honolulu, HI, 2017, pp. 1960–1968.

[111] H. Gunes and M. Piccardi, “A Bimodal Face and Body Gesture Database for Automatic Analysis of Human Nonverbal Affective Behavior”, in *The 18th International Conference on Pattern Recognition, ICPR’06,* Hong Kong, China, 2006, pp. 1148–1153.

[112] B. Nojavanasghari, T. Baltrušaitis, C. E. Hughes, and L.-P. Morency, “EmoReact: a multimodal approach and dataset for recognizing emotional responses in children,” in *Proceedings of the International Conference on Multimodal Interaction, ICMI 2016,* Tokyo Japan, 2016, pp. 137–144.

[113] P. P. Filntisis, N. Efthymiou, G. Potamianos, and P. Maragos, “Emotion Understanding in Videos Through Body, Context, and Visual-Semantic Embedding Loss,” In ECCV 2020 Workshops, vol. 12535.

[114] P. P. Filntisis, N. Efthymiou, G. Potamianos, and P. Maragos, "An Audiovisual Child Emotion Recognition System for Child-Robot Interaction Applications," in *European Signal Processing Conference, EUSIPCO 2021,* Dublin, Ireland, 2021.

[115] N. Efthymiou, P. P. Filntisis, G. Potamianos, and P. Maragos, "Visual Robotic Perception System with Incremental Learning for Child–Robot Interaction Scenarios," *Technologies,* vol. 9, 86, 2021.

[116] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the International Conference on Computer Vision and Pattern Recognition, CVPR 2016,* Las Vegas, Nevada, USA, 2016.

[117] L. Wang, Y.Xiong, Z.Wang, Y.Qiao, D.Lin, X.Tang, and L.VanGool, "Temporal segment networks: Towards good practices for deep action recognition," in *Proceedings of the 14th European Conference on Computer Vision, ECCV 2016,* Amsterdam, The Netherlands, vol. 5, 2016.

[118] A. P. Bradley, "The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms," *Pattern Recognition,* vol. 30, pp. 1145-1159, 1997.

[119] S. Theodoridis and K. Koutroumbas, *Pattern Recognition,* Fourth Edition (4th. ed.), Academic Press, Inc., USA, 2008.

[120] T. Adali, X. Liu, and K. Sonmez, "Conditional distribution learning with neural networks and its application to channel equalization," *IEEE Transactions on Signal Processing,* vol. 45, pp. 1051–1064, 1997.

[121] N. Qian, "On the momentum term in gradient descent learning algorithms," *Neural Networks,* vol. 12, pp. 145–151, 1999.

[122] F. Grundmann, K. Epstude, and S. Scheibe, "Face masks reduce emotion-recognition accuracy and perceived closeness," *PLoS ONE,* vol. 16: e0249792, Apr. 2021.

[123] M. Tsantani, V. Podgajecka, KLH Gray, and R. Cook, "How does the presence of a surgical face mask impair the perceived intensity of facial emotions?," *PLoS ONE,* vol. 17: e0262344, Jan 2022.

[124] M. Marini, A. Ansani, F. Paglieri et al. , "The impact of facemasks on emotion recognition, trust attribution and re-identification," *Scientific Reports* vol. 11, 5577, Mar. 2021.

[125] Y. Kartynnik, A. Ablavatski, I. Grishchenko, and M. Grundmann, "Real-time Facial Surface Geometry from Monocular Video on Mobile GPUs," *CVPR Workshop on Computer Vision for Augmented and Virtual Reality 2019,* IEEE, Long Beach, CA, USA.

[126] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," arXiv:1801.04381v4 [cs.CV], Mar 2019.

[127] V. Bazarevksy, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "BlazePose: On-device Real-time Body Pose tracking," *CVPR Workshop on Computer Vision for Augmented and Virtual Reality 2020,* IEEE, WA, USA.

[128] F. Zhang, V. Bazarevksy, A. Vakunov, G. Sung, C.-L. Chang, and M. Grundmann, "MediaPipe Hands: On-device Real-time Hand Tracking," *CVPR Workshop on Computer Vision for Augmented and Virtual Reality 2020,* IEEE, WA, USA.

[129] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *Proceedings of the 13th IEEE International Conference on Face and Gesture Recognition, FG 2018,* pp. 59–66, 2018.

[130] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," arXiv:1610.02391 [cs.CV], Oct. 2016.

[131] J. Lin, C. Gan, and S. Han, "TSM: Temporal Shift Module for Efficient Video Understanding," arXiv:1811.08383 [cs.CV], Nov. 2018.

[132] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing,* vol. 10, pp. 18–31, Jan. 2017.

[133] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," arXiv:1610.02357v3 [cs.CV], Apr. 2017.

[134] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv:1704.04861 [cs.CV], Apr. 2017.

[135] T. Chen, T. Moreau, Z. Jiang, L. Zheng, E. Yan, M. Cowan, H. Shen, L. Wang, Y. Hu, L. Ceze, C. Guestrin, and A. Krishnamurthy, "TVM: An Automated End-to-End Optimizing Compiler

for Deep Learning," in *Proceedings of the 13th USENIX conference on Operating Systems Design and Implementation,* Oct. 2018, pp. 579–594.

[136] N. Kegkeroglou, P. P. Filntisis, and P. Maragos, "Medical Face Masks and Emotion Recognition from the Body: Insights from a Deep Learning Perspective", arXiv:2302.10021 [cs.CV], Feb. 2023.