Politecnico di Torino

Towards Automated Facial Mimicry Assessment Using RGB-D Data and a commercial tracking software: preliminary results on healthy and parkinsonian subjects

Giacomo Saracino

Master's Degree in Biomedical Engineering
Biomedical Instrumentation
Academic Year 2024/2025

Candidate: **Giacomo Saracino**
Supervisors: **Prof. Andrea Cereatti**
Dr. Diletta Balta

Abstract

Hypomimia, the reduction of spontaneous facial movements, is an early and disabling symptom of Parkinson's disease (PD). Its clinical evaluation is currently subjective and based on qualitative rating scales. Gold standard (GS) methods including manual visual expert inspection and surface electromyography (EMG) provide objective assessments but require expensive equipment, trained personnel, and may interfere with natural facial expressivity. Recently explored markerless (ML) methods using RGB and RGB-D cameras, combined with deep learning-based facial landmark detection techniques, represent a non-invasive alternative. However, their clinical validation remains debated.

This thesis aimed to (i) propose a low-cost ML method using a single RGB-D camera to quantify facial muscle activity, (ii) validate it against GS measures (manual measurements and EMG signals), (iii) assess the impact of the depth sensor and RGB image resolution on its performance and (iv) evaluate its applicability in discriminating between young healthy (YH), elderly healthy (EH), and PD subjects during emotions.

Participants included 17 YH (25.5±3.7 y.o.), 13 EH (69.7±4.2 y.o.), and 11 PD patients (70.7±8.7 y.o.). Data were acquired using an Azure Kinect RGB-D camera (1280×720, 30 Hz), and EMG signals were recorded with a D360 amplifier (5 kHz). Each subject was recorded at rest, during maximum voluntary contraction (MVC) of the depressor anguli oris (DAO) muscle, and while expressing spontaneous happiness and sadness. The MediaPipe Face Mesh algorithm was used to extract 2D DAO landmarks, with depth data extracted from the depth image. To identify any statistical differences between methods (manual vs automatic) and patients' groups, a Mann–Whitney test ($\alpha = 0.05$) was applied.

For validation, the ML-derived DAO length variation (DAO-LV) was compared with manual measures during MVC, resulting in a Mean Absolute Error (MAE) of 1.5±1.6 mm (Mean Absolute Percentage Error = 27.9±27.2) across 41 subjects. No significant differences were found between automatic and manual measures confirming ML protocol validity. Both EMG RMS values and DAO-LV showed no significant differences between YH and EH, while a significant difference was observed between PD and EH, highlighting the ML protocol's sensitivity in distinguishing between groups during MVC.

To evaluate the adding value of the depth sensor with respect to the use of RGB image only and the influence of image resolution on the method's performance, only 2D DAO landmarks were considered and a static calibration phase was introduced to obtain DAO-LV in millimeters. Those values were compared with manual measurements. Removing the depth sensor and reducing image resolution to 640×480 resulted in an increase in MAE of 52% and 79%, respectively, confirming that both factors are critical for accurate facial expression analysis.

Emotional expression analysis showed no significant differences between YH and EH for either emotion. Happiness expression revealed significant differences between EH and PD for both DAO-LV and DAO contraction velocity, as happiness elicited greater muscle activation. On the other hand, during sadness, no significant difference between EH and PD were found likely due to high inter-subject variability and low DAO-LV in PD patients.

In conclusion, the proposed RGB-D ML method is a valid tool for the objective assessment of hypomimia in PD since it effectively differentiates between subject groups, particularly during MVC and happiness expression.

Contents

List of Figures

Acronyms

CV Contraction Velocity

DAO Depressor Anguli Oris

DAO-LV DAO Length Variation

EH Elderly Healthy

EMG Electromyography

GS Gold Standard

MAE Mean Absolute Error

ML Markerless

MAPE Mean Absolute Percentage Error

MVC Maximum Voluntary Contraction

PD Parkinson's Desease

ROI Region of Interest

YH Young Healthy

Chapter 1

Facial Mimicry and Hypomimia in Parkinson's Disease

1.1 Introduction

Facial expressions represent one of the most important and complex communicative tools of human beings, involving a sophisticated network of facial muscles whose activity can reveal emotions, intentions, and even pathological conditions. In this context, **quantitative analysis of facial movements** has gained increasing importance in various fields, from clinical diagnostics to emotion understanding, through human-machine interaction and virtual reality.

Among the various pathological conditions affecting facial expressiveness, **hypomimia** stands out as one of the most prevalent and socially impactful motor symptoms in Parkinson's disease (PD) [16]. This condition is characterized by reduced facial expressiveness and diminished spontaneous facial movements, significantly affecting patients' ability to convey emotions through facial expressions and leading to what is often described as a "masked face" appearance. Beyond the visible reduction in expression amplitude, Parkinsonian patients exhibit markedly impaired response readiness when attempting to express emotions, with muscle contraction velocities substantially slower than those observed in healthy subjects.

This facial muscle **bradykinesia** creates a complex cascade of motor impairments: it not only delays the onset of emotional expressions but also reduces their intensity and duration, ultimately resulting in profound communicative difficulties that can severely impact social

interactions and quality of life [2]. The underlying pathophysiology involves the degeneration of dopaminergic neurons in the substantia nigra, which disrupts the basal ganglia circuits responsible for the initiation and modulation of voluntary and involuntary facial movements. Consequently, the temporal dynamics of facial muscle activation are severely compromised, with patients requiring significantly longer latencies to initiate facial expressions and exhibiting reduced peak contraction velocities compared to age-matched healthy controls. This motor impairment extends beyond simple expression production to affect the entire spectrum of facial communication, including emotional responsiveness and social signaling.

Despite the clinical importance of accurately assessing hypomimia, traditional methodologies for the quantitative study of facial muscle activity present significant limitations. Current approaches have been based predominantly on invasive or semi-invasive techniques, ranging from *surface electromyography (sEMG)*, which requires the application of electrodes on the skin, to optoelectronic systems (*stereophotogrammetry*) that necessitate reflective markers positioned on the face. While these approaches offer undisputed precision, they present intrinsic drawbacks that limit their clinical applicability: they alter the spontaneity of expressions, prove uncomfortable for examined subjects, and require laborious setup procedures that hinder large-scale clinical implementation.

To address these methodological challenges, the present thesis work aims to overcome such limitations through the **development and validation of an innovative Python algorithm** that enables the identification and tracking of facial muscles in a completely markerless (ML) manner. The proposed algorithm integrates depth information acquired through an RGB-D camera (Azure Kinect camera) with facial landmark data provided by the Mediapipe library, thus enabling accurate **three-dimensional reconstruction** of facial landmarks and muscle movement analysis. This **non-invasive** approach represents a significant advancement in the study of hypomimia in Parkinson's disease, as it preserves the naturalness of facial expressions while providing precise quantitative measurements of muscle contraction dynamics and response timing, potentially opening new avenues for both clinical assessment and therapeutic monitoring.

1.2 Thesis Objectives

Given the limitations of current assessment methods and the clinical need for objective hypomimia evaluation, this thesis work was designed with four specific and interconnected objectives that collectively aim to establish a comprehensive framework for ML facial expression analysis in Parkinson's disease.

The **first objective** focused on developing a low-cost markerless ML method using a single RGB-D camera to quantify facial muscle activity. This involved creating an innovative algorithm that integrates MediaPipe Face Mesh facial landmark detection with depth information captured by an Azure Kinect camera. The method successfully detected eleven facial muscles. Although, the methodological approach centered on the right depressor anguli oris (DAO) muscle, given its critical role in facial expression. The algorithm was designed to extract 2D DAO landmarks with MediaPipe and combine them with corresponding depth data from Azure Kinect camera to enable three-dimensional analysis of muscle contraction dynamics, offering a non-invasive alternative to traditional assessment methods.

The **second objective** aimed to validate the proposed ML approach against established gold standard (GS) measures. This validation process involved two complementary comparisons: first, the algorithm-derived DAO length variations were compared with manual DAO length variations measurements performed by trained operators on the same facial recordings (*Maximum DAO Voluntary Contraction*); second, the consistency of the markerless method was assessed against surface electromyography (EMG) signals recorded simultaneously from the DAO muscle. This dual validation strategy was designed to ensure both geometric accuracy and physiological relevance of the proposed measurements, establishing the scientific credibility of the ML approach.

The **third objective** sought to systematically assess the impact of depth sensor specifications and resolution settings on method performance. Recognizing that different hardware configurations and imaging parameters could significantly influence measurement accuracy, this analysis involved comparing the performance of the RGB-D

approach against purely 2D methods using pixel-to-millimeter calibration. Multiple resolution conditions were tested, including RGB 1280×720 and RGB 640×480 configurations, to determine the minimum technical requirements necessary for reliable facial expression analysis and to provide practical guidelines for clinical implementation.

The **fourth objective** focused on evaluating the clinical applicability of the developed method to distinguish between different subject populations during emotional expression tasks. This involved recruiting three distinct groups: young healthy YH subjects, elderly healthy EH subjects, and Parkinson's disease PD patients. Each participant was recorded during spontaneous expression of happiness and sadness emotions. The clinical evaluation was designed to assess the method's sensitivity in detecting the subtle differences in facial expressiveness that characterize Parkinsonian hypomimia, while also investigating the differential response patterns across various emotional expressions.

These four objectives were pursued through a comprehensive experimental design that combines technological innovation with rigorous clinical validation, ultimately aiming to establish **a new standard for objective, non-invasive assessment of facial expression disorders in neurological conditions.**

Chapter 2

Literature Background

2.1 Anatomy and physiology of the facial musculature

A thorough understanding of the anatomy and physiology of facial muscles is the foundation for any quantitative study of facial expressions. The human face presents a complex network of more than 40 muscles, classified as muscles of facial expression, whose peculiarity lies in the fact that they do not connect skeletal elements to each other, but rather insert directly into the skin dermis. This anatomical feature allows for the great variety and finesse of facial movements that characterise human non-verbal communication.

Facial muscles have unique functional and anatomical specificities: unlike skeletal musculature, they are not enveloped by well-defined muscle bands and often intertwine with each other, creating an integrated system of forces that act simultaneously to produce complex expressions. Particularly relevant to the study of emotions are the **zygomaticus** muscles (involved in smiling), the **levator labii superioris** the **frontalis muscles** (crucial in expressions of anger and concentration), and the **depressor anguli oris muscles**, the specific object of the study presented in this thesis, whose role is crucial in the expression of sadness and disappointment.

The activity of these muscles is mainly regulated by the facial nerve (7th cranial nerve), the impairment of which, as in the case of peripheral paralysis or neurodegenerative diseases, can significantly alter the expressive capacity. Understanding these neuroanatomical mechanisms is essential to correctly interpret alterations of facial expressions in pathological contexts [6].

2.2 Facial expression analysis systems

2.2.1 Traditional approaches

The quantitative analysis of facial expressions has a relatively recent history in scientific research. The first standardised system for the objective coding of facial movements was the **Facial Action Coding System (FACS)**, developed by Ekman and Friesen in the 1970s [8]. FACS identifies Action Units (AU) corresponding to the activation of specific muscles or muscle groups, allowing any facial expression to be decomposed into its elemental components. Despite its undoubted validity, FACS requires expert coders and considerable analysis time, limiting its applicability in clinical or large-scale research contexts.

Other traditional approaches to facial motion analysis include:

- **Surface Electromyography (sEMG)**: considered the gold standard for measuring muscle activity, sEMG uses electrodes applied to the skin to detect electrical potentials generated by muscle contraction. Despite its high temporal accuracy ($\leq$1 ms) and the ability to detect even the smallest contractions, this technique has intrinsic limitations related to invasiveness, cross-talk between signals from adjacent muscles, and the difficulty of positioning the electrodes in particularly small or closely spaced facial areas [7].

- **Marker optoelectronic systems**: these systems use multiple cameras to track the three-dimensional position of reflective markers applied to the face. Commercial systems such as Vicon, Qualisys or OptiTrack offer excellent spatial (in the order of 0.1 mm) and temporal accuracy, but require controlled environments, elaborate setups and inevitably alter the naturalness of expressions [9].

- **Stereo-photogrammetry**: an advanced method that, like optoelectronic systems, uses multiple cameras to reconstruct facial movements three-dimensionally. This technique is based on simultaneous imaging of the face from different angles, allowing the precise measurement of deformations and muscle movements. Unlike traditional optoelectronic marker systems, facial

stereo-photogrammetry offers a more natural and minimally invasive reconstruction, with spatial accuracy that can reach fractions of a millimeter (≤0.1 mm). However, it presents challenges related to the need for controlled lighting conditions, computational complexity in image processing and high equipment costs.

Figure 2.1: Traditional approaches: i) sEMG system acquisition, ii) stereo-photogrammetry system

Figure 2.1 shows the complexity of the sEMG acquisition system and the optoelectronic system acquisition protocol.

2.2.2 Recent markerless technologies

In recent years, technological evolution has led to the development of ML systems that overcome many of the limitations of traditional approaches:

- **Depth Sensors**: RGB-D cameras such as Microsoft Azure Kinect, Intel RealSense or devices based on Time-of-Flight (ToF) technology have revolutionised the acquisition of three-dimensional data, allowing the reconstruction of facial morphology without the need for markers. These devices project infrared patterns or light pulses onto the scene, measuring the return time to calculate the distance to each point. Their accuracy, although lower than optoelectronic systems (errors in the order of 2-10 mm depending on distance), is sufficient for many facial motion analysis

applications [28].

- **Stereo-photogrammetry**: uses two or more calibrated cameras to reconstruct 3D coordinates by triangulating corresponding points identified in different views. This technology has benefited greatly from recent advances in computer vision algorithms, enabling accurate reconstructions without marker.

2.3 Libraries and frameworks for facial landmark detection

The field of automatic identification of facial landmarks has seen tremendous progress in the last decade, mainly due to the adoption of deep learning techniques. The main libraries available include:

- **MediaPipe**: developed by Google, this open-source library represents the state-of-the-art for real-time tracking of facial landmarks. MediaPipe Face Mesh identifies 468 three-dimensional points on the face with excellent robustness to variations in illumination, pose and partial occlusions. Its lightweight architecture allows it to run on mobile devices while maintaining high frame rates (>30 FPS). The latest version integrates optimised convolutional neural networks and accelerated inference techniques that guarantee sub-millimetre accuracy in the localisation of landmarks [14].

- **Dlib**: this Python-binding C++ library includes implementations of algorithms for detecting 68 face landmarks based on Ensemble of Regression Trees (ERT). Although less point-rich than MediaPipe, Dlib has been a reference in the scientific community for years for its robustness and accuracy, and continues to be widely used in applications that do not require the point density offered by more recent solutions [12].

- **OpenCV**: although not specialised in facial landmark detection, this library provides implementations of fundamental algorithms for computer vision and easily integrates with other solutions. Recent versions include DNN (Deep Neural Networks) modules

that allow the use of pre-trained models for facial detection and landmark extraction [3].

- **Face Alignment Network (FAN)**: based on convolutional neural network architectures with hourglass networks, this academic solution offers high accuracy in landmark identification even under difficult conditions, but requires more computational resources than mobile-optimised alternatives [4].

- **OpenFace**: comprehensive facial analysis framework integrating face detection, landmark tracking, head pose estimation, action unit recognition and gaze analysis. Particularly useful in applications requiring multimodal analysis of facial expressions [1].

2.3.1 Evaluation of MediaPipe and dlib for Facial Landmark Detection

For the main purposes of this thesis project, an evaluation of the main open-source libraries for facial landmarks identification was required. Of the libraries presented in Section 2.3, only two have been deepen for this thesis purposes: **MediaPipe** and **dlib**. Both were tested for facial landmark detection in video acquisitions, with the objective of measuring parameters related to facial muscle mobility.

Here is a brief comparison between the two libraries:

- **Number of landmarks**: Since MediaPipe library offers 400 more landmarks than Dlib (468 vs. 68), the first one is more suitable for detailed analysis of expressions and micro-movements, such as DAO contraction during expressions.

- **Real-time performance**: MediaPipe is optimized for video and real-time applications, while maintaining a high frame rate. Although accurate, Dlib may slow down with real-time video acquisitions.

- **Stability**: MediaPipe better handles variations in facial position and lighting, ensuring more stable detection in videos. Dlib tends to be more sensitive to these variations, reducing reliability under suboptimal conditions.

- **Ease of use**: MediaPipe provides preconfigured pipelines and intuitive documentation, making integration into complex projects easier. Dlib requires more code customization and familiarity with its API.

- **Accuracy and applicability**: For applications requiring precise measurement of facial muscles, MediaPipe offers higher granularity thanks to its greater number of landmarks and ability to track dynamic movements in videos. Dlib is more suitable for basic detections or static contexts.

2.3.2 MediaPipe Face Mesh Architecture

MediaPipe's Face Mesh architecture implements a two-stage approach for precise facial landmark detection and localization [11]. The first stage employs **BlazeFace**, a lightweight convolutional neural network based on Single Shot MultiBox Detector (SSD) with MobileNetV1 backbone, optimized for real-time face detection. The BlazeFace network operates on input images resized to 128×128 pixels and generates face bounding boxes using a composite loss function:

$$L_{total} = L_{classification} + \alpha \cdot L_{regression} \tag{2.1}$$

where $L_{classification}$ employs focal loss to handle class imbalance and $L_{regression}$ uses smooth L1 loss for coordinate regression. The second stage consists of a landmark regression network that processes normalized facial regions of 192×192 pixels, extracted using BlazeFace predictions. This network directly predicts 3D coordinates (x, y, z) of the 468 characteristic points through a linear activation function in the final layer:

$$\mathbf{p}_i = \mathbf{W}_{out} \cdot \mathbf{h}_{final} + \mathbf{b}_{out} \tag{2.2}$$

where $\mathbf{p}_i \in \mathbb{R}^3$ represents the coordinates of the i-th landmark, $\mathbf{h}_{final}$ is the feature vector from the final hidden layer, and $\mathbf{W}_{out} \in \mathbb{R}^{1404 \times d}$ is the output weight matrix (1404 = 468 landmarks × 3 coordinates). The architecture incorporates residual connections and utilizes separable convolutions to reduce computational complexity

while maintaining sub-pixel accuracy with mean error below 1.5 pixels on benchmark datasets. The complete pipeline operates at over 30 FPS on mobile devices through INT8 quantization and neural accelerator-specific optimizations, enabling real-time augmented reality applications and facial analysis.

2.4 Clinical applications of facial expression analysis

Neurodegenerative diseases are often associated with significant alterations in the ability to produce and control facial expressions, which may manifest early in the course of the disease:

- **Parkinson's disease**: A typical feature of Parkinsonian patients is hypomimia, up to complete amimia. Quantitative studies have shown asymmetries in muscle contractions, reduction in the amplitude and speed of movements, and alterations in the synchrony between agonist and antagonist muscles. Of particular interest is the observation that some of these alterations may precede clinical diagnosis by several years, suggesting potential for **early disease identification**. The evaluation of therapeutic efficacy in PD through the analysis of facial expressions has revealed important indicators of progression and response to treatment. Quantitative analysis of facial movements has proven to be a sensitive method for monitoring the effects of drug therapy, particularly levodopa and dopaminergic agonists. Expression analysis methods make it possible to assess not only the reduction of motor symptoms, but also the impact of treatments on the patient's emotional and communicative component. Some studies have shown that improved facial expressiveness may be an early indicator of therapeutic efficacy, even before clear motor improvements appear.

- **Alzheimer's disease**: patients show a progressive impoverishment of emotional expressiveness, with particular impairment in the production of spontaneous expressions in response to emotional stimuli, although the ability to produce expressions on

demand in the early stages is relatively preserved [5]. In the evaluation of therapeutic efficacy in Alzheimer's disease, the analysis of facial expressions has assumed a crucial role as a tool for monitoring disease progression and the impact of therapeutic interventions. Quantitative analysis techniques make it possible to accurately document the decline in emotional and cognitive expressiveness, offering an objective biomarker of disease progression. In particular, studies have shown that the reduction in the ability to produce spontaneous expressions can be a more sensitive indicator of cognitive decline than traditional neuropsychological tests. Pharmacological and rehabilitation treatments are now increasingly evaluated also through the ability to slow down or partially recover facial expressiveness, considering it a key element of the patient's quality of life and emotional communication.

2.5 Technical challenges in markerless facial motion analysis

2.5.1 Calibration and standardisation of measurements

One of the most critical aspects in markerless quantitative analysis concerns the calibration and standardisation of measurements across different subjects and acquisition sessions:

- **Spatial calibration**: conversion from pixel coordinates to metric units is a key challenge, particularly in 2D systems where the relationship between camera distance and apparent size introduces variables that are difficult to control. Recent approaches include the use of reference objects with known dimensions, as proposed in this thesis work, and self-calibration techniques based on anthropometric features [27].
- **Morphological normalisation**: significant anatomical differences between individuals require normalisation strategies to make measurements comparable. Techniques based on deformable templates, statistical shape models and non-rigid registration methods represent the state of the art in this field.

- **Compensation of head movements**: even minimal head movements can introduce significant artefacts in facial movement measurements. Robust systems must implement compensation algorithms based on 3D head pose estimation and registration techniques that isolate intrinsic facial movements from those resulting from global head displacement.

2.5.2 Robustness to changing environmental conditions

The performance of markerless systems is strongly influenced by environmental conditions:

- **Lighting variations**: represent a major challenge, as they can significantly alter facial appearance and compromise accurate landmark identification. Modern approaches include data augmentation techniques during model training, adaptive illumination normalisation and the integration of illumination invariant sensors such as infrared cameras.

- **Partial occlusions**: masks, glasses, hair or hands partially covering the face can compromise tracking. More advanced solutions use neural networks specifically trained to handle occlusions and inference algorithms that reconstruct missing parts based on anatomical models and temporal coherence constraints [10].

- **Image resolution and quality**: degradation of image quality in low-light conditions or the use of low-resolution cameras significantly impacts tracking accuracy. Super-resolution techniques based on deep learning are increasingly being integrated into processing pipelines to mitigate these effects.

2.6 Future perspectives and emerging research directions

2.6.1 Multimodal integration

The integration of different sensory modalities is one of the most promising directions:

- **Fusion of visual and thermal data**: thermal cameras can provide complementary information on muscle activation through the detection of subtle changes in skin temperature associated with muscle contraction, particularly useful in unfavourable lighting conditions.
- **Synchronised audio-visual analysis**: integration of acoustic speech analysis with lip movement tracking improves understanding of communication dynamics, with applications in the early diagnosis of neuromotor disorders affecting both verbal articulation and facial expressiveness [23].
- **Integration with physiological data**: the correlation between facial expressions and physiological parameters such as heart rate variability, skin conductance or electroencephalographic activity opens up new perspectives in the holistic understanding of emotional states and their manifestation [21].

2.6.2 Generative models and biomechanical simulation

The development of detailed biomechanical models of the face is a rapidly developing field of research:

- **Musculoskeletal models**: physics-based simulations that explicitly model muscle anatomy, viscoelastic tissue properties and mechanical interactions between facial structures allow for a better understanding of muscle activation patterns and their pathological alterations [25].
- **Facial digital twins**: the creation of customised digital twins of a patient's face allows predictive simulations of the effect of surgical or therapeutic interventions, optimising treatment planning [17].
- **Realistic synthesis of expressions**: advanced generative models allow the synthesis of photorealistic facial expressions from muscle activation parameters, with applications in the creation of synthetic datasets for the training of more robust algorithms [24].

Chapter 3

Materials & Methods

This chapter presents the materials and methodological framework employed in the development of this thesis project. The discussion covers the dataset proprieties, the characteristics of RGB-D video acquisition systems, the machine learning algorithms implemented, and the specific processes employed for data evaluation and analysis. Each component of the experimental methodology is examined to provide a comprehensive understanding of the research approach and to ensure reproducibility of the results.

3.1 Dataset Description

The dataset comprised a total of 41 subjects, categorized into three distinct groups to enable comparative analysis of facial muscle contraction patterns across different populations, *Table 3.1*.

Group	**n**	**Years Old**
Young Healthy **YH**	17	23.53 ± 3.71
Elderly Healthy **EH**	13	69.69 ± 4.21
Parkinsonian **PD**	11	70.72 ± 8.67

Table 3.1: Demographic characteristics of study participants

The gender distribution across the entire dataset consisted of 21 female subjects (F) and 20 male subjects (M).

The dataset provides comprehensive multimodal data for each participant. Visual data consists of 123 total **video acquisitions**, with each of the 41 subjects performing three distinct facial expressions:

maximum voluntary contraction (**MVC**) of the DAO muscle, expression of **happiness** and expression of **sadness**. These specific expressions were selected to capture different patterns of DAO muscle activation, ranging from isolated muscle contraction to complex emotional expressions involving coordinated activation of multiple facial muscle groups.

Electromyographic recordings of the DAO muscle were obtained exclusively during the MVC expression, resulting in 41 EMG recordings across all subjects.

Manual measurements of the DAO muscle length were performed for all 41 subjects under two different conditions, yielding a total of 82 measurements. These measurements were conducted during both **rest** and **DAO MVC conditions** by trained operators using standardized anatomical landmarks to ensure consistency and accuracy across all subjects.

The manually obtained measurements and the EMG signals served as the gold standard GS reference for the subsequent validation process of the ML facial muscle tracking algorithm, providing a comprehensive dataset for validation and comparative analysis across various demographic and pathological conditions.

3.2 Acquisition Protocol

Visual data acquisition was performed using an **Azure Kinect RGB-D camera** positioned at approximately *50 cm* from each subject, as in Figure 3.1, maintaining consistent spatial configuration throughout all recording sessions [26, 19]. The camera was consistently placed at the subject's face height and kept at a fixed distance to ensure measurement consistency and accuracy. All acquisitions were performed in the same room under **identical lighting conditions** to avoid any bias due to surrounding factors.

The Azure Kinect camera is capable of producing RGB images (**color maps**), **depth maps**, and three-dimensional **point clouds** simultaneously for each frame. The RGB image provides visual information about the surface of the face, the depth map indicates the distances of points from the camera, and the point cloud reconstructs

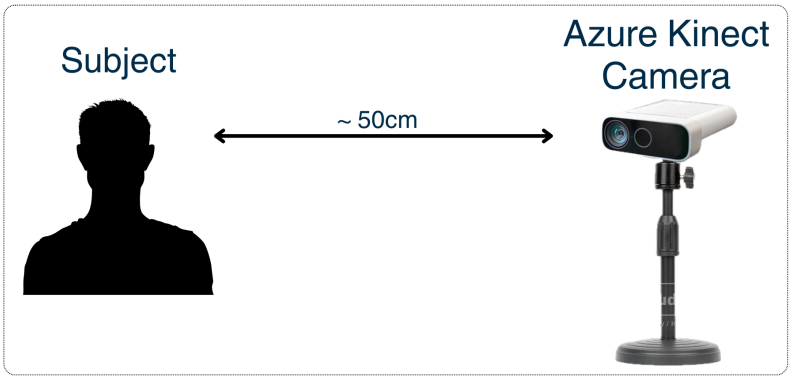

Figure 3.1: Acquisition Protocol Scheme

the 3D geometry of the scene. This combination of data is essential for obtaining precise measurements of facial muscle lengths and conducting detailed motion analysis, with particular focus on key parameters for studying patients affected by PD.

Each participant was instructed to perform the three facial expressions in a controlled sequence, allowing for systematic data collection across all modalities. The standardized protocol ensured consistent data quality and comparability across different subject groups.

Figure 3.2: EMG Acquisition Setup

Electromyographic recordings were obtained using a D360 amplifier (Digitimer Ltd, Welwyn Garden City, UK) and an EMG electrode, Figure 3.2, **exclusively** during the **MVC** expression. The EMG signals were amplified (×1000), filtered with a bandpass of 3-3000 Hz, and sampled at 5 kHz using a 1401 power analog-to-digital converter and

Signal 6 software (Cambridge Electronic Design, Cambridge, UK). This configuration provided high-quality EMG data for quantitative analysis of muscle activation patterns during MVC.

The **manual measurement protocol** involved trained operators conducting measurements using standardized anatomical landmarks during both **rest** and **MVC conditions**. This approach ensured consistent and accurate reference measurements across all subjects, establishing reliable ground truth data for algorithm validation purposes.

3.3 RGB-D Markerless Method

This section provides a detailed analysis of the code developed for this thesis project, the core of this thesis project. The main purposes and functionalities of the code will be presented, clarifying the objectives of its execution. Subsequently, the crucial steps and the reasoning behind achieving these goals will be examined, analyzing the functions and techniques employed. The motivations behind the programming choices will then be discussed, highlighting why certain solutions were preferred over others. Finally, examples of the expected outputs will be shown to provide a clear idea of the code's capabilities.

Figure 3.3 represents the RGB-D ML algorithm pipeline, used for the detection of the following muscle: the depressor anguli oris DAO (left and right), levator labii superioris (left and right), orbicularis oculi (left and right), zygomaticus major (left and right), orbicularis oris, and the left and right frontalis muscles (measured as distance between eyebrows and the center point). For the main objectives of this thesis project, the **DAO muscle is the most important**.

The system requires as **input** a subject folder, which includes three emotion acquisitions (DAO MVC, happiness, sadness). As **output**, an image is generated showing the mentioned muscle lengths for each frame of the acquisition, along with two Excel files: one contains the spatial coordinates (xyz) of each landmark, while the other provides the distance values for each frame. The Excel files were useful for visualizing the landmarks in 3D via MATLAB, *Figure 3.6*, which offers a dynamic and more intuitive representation of the patient's face.

A step-by-step explanation is now provided on how the algorithm uses the data acquired from the Azure Kinect acquisitions to achieve these results.

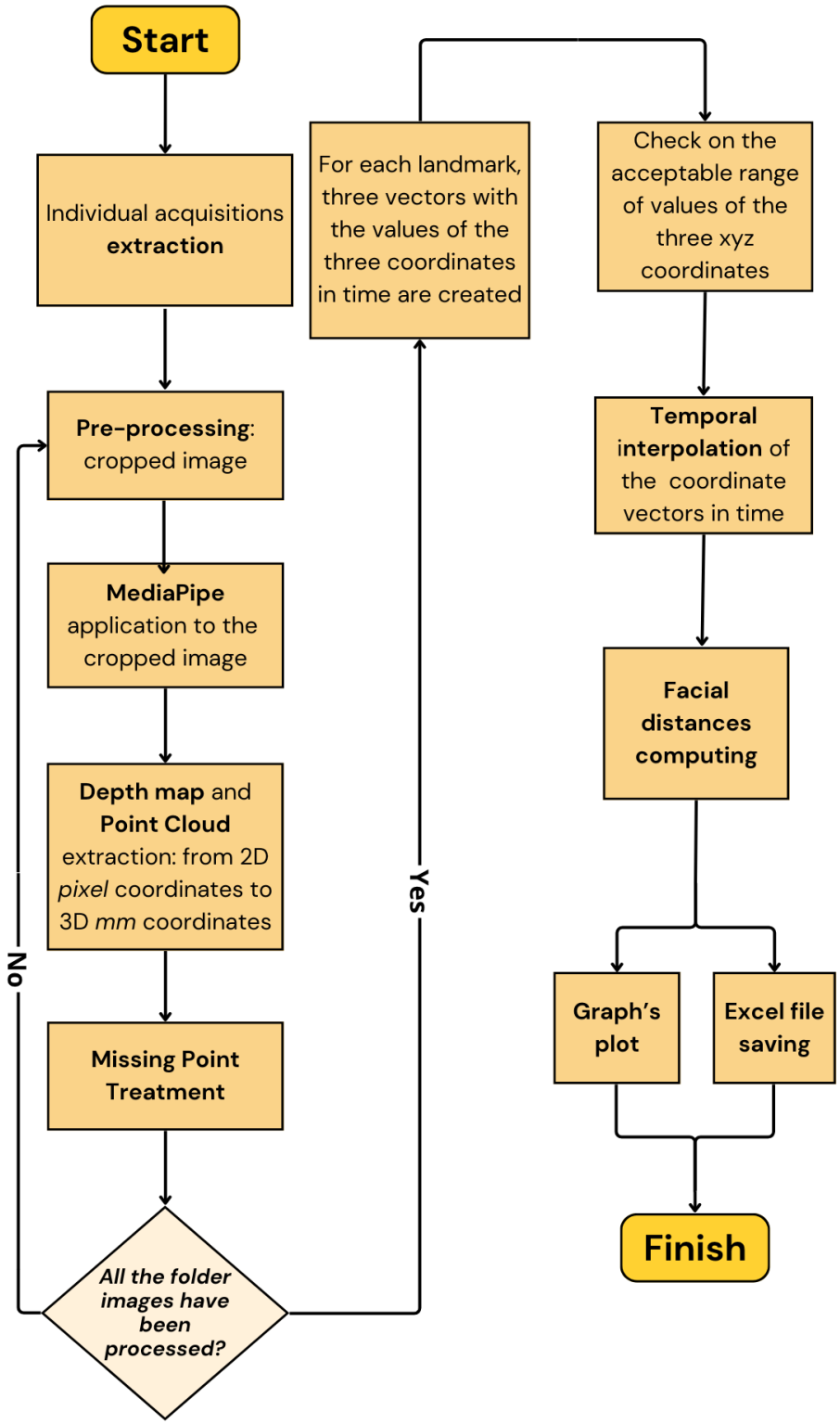

Figure 3.3: Algorithm's Pipeline

3.3.1 Image Pre-processing

The process begins with a subject folder, which contain the respective acquisitions. Within each acquisition folder, for each frame, color map, depth map, and point cloud files from the Azure Kinect camera are available, as described in Section 3.2.

The algorithm processes each folder by identifying image files that begin with the prefix *"color"* and end with the *".png"* extension. These files are sorted numerically to ensure the correct temporal sequence of the frames.

The code processes the RGB images contained in the patient folders, focusing on a specific region of interest (ROI) defined within each frame. As illustrated in *Figure 3.4*, the ROI is identified through specific coordinates, defined to ensure that the analysis algorithm uniquely detects the patient's face. This guarantees that the face is contained within a *500x600 pixel* box, roughly positioned at the center of the image. The RGB image is then cropped accordingly.

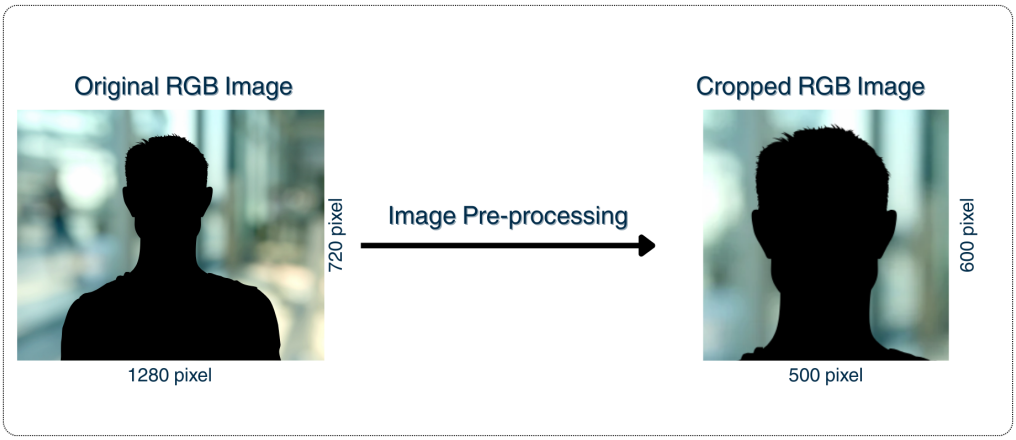

Figure 3.4: RGB Image Pre-processing

For each RGB image of the acquisition, the code checks the existence of the file and reads the corresponding image, crops the ROI from the original frame, and finally processes the cropped image with a chosen landmark detection library.

The pre-processing workflow shown in *Figure 3.4* represents the systematic approach where the algorithm focuses library's evaluation

on the area of interest. This operation **reduces noise** from external areas and **optimizes computational performance** by focusing only on the relevant region for the study and subsequently increasing landmark detection accuracy.

3.3.2 Commercial Tracking Software

In addiction to the MediaPipe and dlib comparison, in Section 2.3.1, here are two examples of how these two libraries work in cases similar to the case in this thesis. *Figure 3.5 i* shows an application of the Dlib library on a random female subject picture. While, *Figure 3.5 ii* shows an application of the Mediapipe library on a picture of random men.

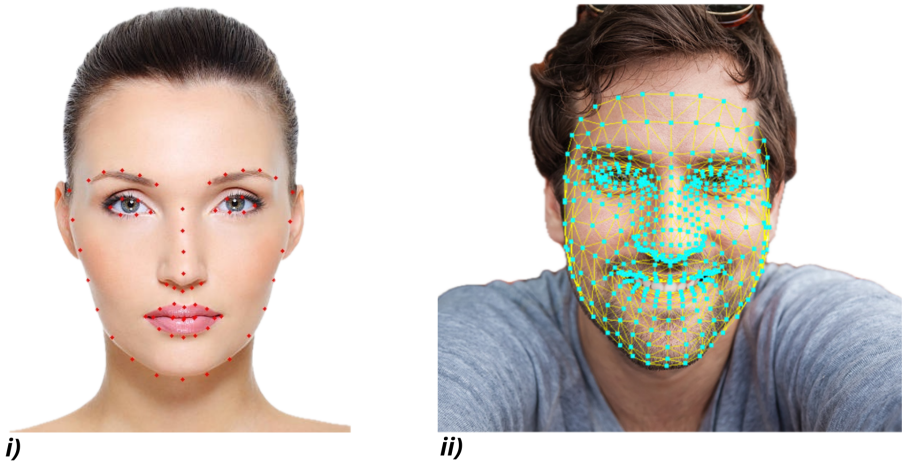

Figure 3.5: Application example of *i)* dlib library and *ii)* Mediapipe library

In light of this comparison, **MediaPipe** was selected as the most suitable choice for this study. Its ability to detect 468 facial landmarks, combined with optimization for real-time video processing and stability in detection, enables a more accurate and detailed analysis of facial movements. These elements are crucial for studying facial muscle mobility in patients with PD, where precision and the ability to capture small movements are essential.

3.3.3 Landmarks' Three-Dimensional Representation

After applying MediaPipe to the cropped image and obtaining the 2D pixel coordinates for 468 landmarks, the information related to

the **depth map** and **point cloud** associated with each frame is integrated. This process represents the main innovation of this thesis project. The integration of depth information (Azure Kinect) with the 2D landmark coordinates detected on the subject's face (MediaPipe) enables the **3D reconstruction** of all 468 landmarks.

For each facial landmark detected by MediaPipe (468 in total), the code proceeds as follows:

- **Point Cloud Extraction:** The corresponding *point cloud* file, for the *color* file being analyzed, is located and read using a specific Python function. The function returns an array containing the three-dimensional coordinates (x, y, z) in millimeters [mm] for each pixel of the original frame.

- **Depth Map Extraction:** Similarly, the associated *depth map* file is identified and read using another dedicated Python function. This returns the distance of each pixel from the camera (in millimeters).

- **Landmark-Depth Association:** For each detected landmark, the corresponding (x, y) coordinates are converted into linear indices to directly access the corresponding data in the *point cloud* and *depth map*. From the *point cloud*, the x_{mm} and y_{mm} values are extracted in millimeters. From the *depth map*, the associated depth values are obtained, thus completing the three-dimensional information set (x_{mm}, y_{mm}, z_{mm}) for each landmark.

This approach allows for a **precise three-dimensional representation of facial features**, Figure 3.6, combining the geometric information provided by the Azure Kinect camera (*point cloud* and *depth map*) with the topological data obtained from Mediapipe (landmark detection).

By studying the landmark-muscle correspondence, each facial muscle can be uniquely identified and its length can be automatically computed in every frame, using the RGB-D method. In this thesis project, the landmark-muscle correspondence has been evaluated only for the right DAO muscle.

To further develop this study, it would be necessary to establish

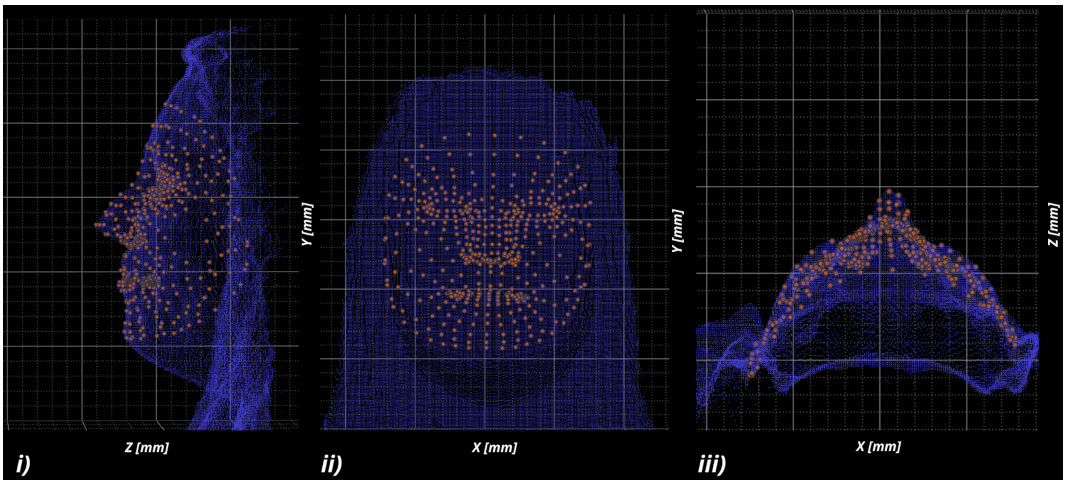

Figure 3.6: MATLAB 3D Landmark Reconstruction: *i)* lateral, *ii)* frontal, *iii)* upper view

the landmark correspondence for all relevant facial muscles identified by the algorithm. This would allow the creation of a more robust dataset, allowing a comprehensive assessment of **hypomimia** in subjects affected by PD.

3.3.4 Landmark-Muscle Anatomical Correspondence

Through an anatomical landmark-muscle correspondence approach, specific facial muscles were approximately identified on the subjects' faces using the 3D reconstruction of landmarks detected by MediaPipe. This methodology enabled the precise localization and analysis of key facial muscles responsible for emotional expression.

Figure 3.7 shows the identified muscles:

- **Depressor anguli oris** - left and right
- **Orbicularis oris**
- **Levator labii superioris** - left and right
- **Orbicularis oculi** - left and right
- **Zygomaticus major** - left and right
- **Frontalis muscles** - left and right

The 3D landmark reconstruction provided by MediaPipe facilitated the accurate mapping of these anatomical structures, allowing for comprehensive analysis of facial muscle activation patterns during

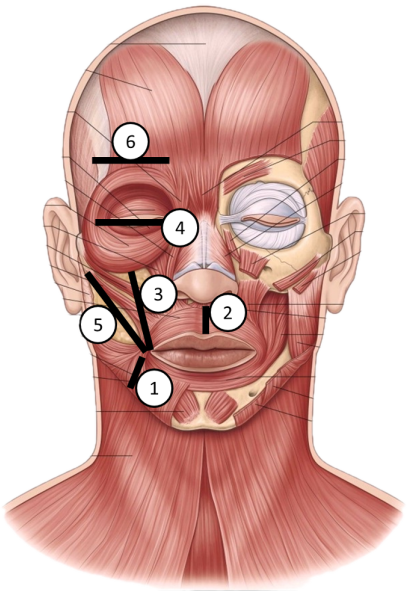

Figure 3.7: Muscles detected by the algorithm

emotional expressions. This landmark-based approach ensured consistent identification of muscle regions across all subjects in the study, providing a standardized framework for subsequent biomechanical and kinematic analyses.

3.3.5 Missing Point Detection and Temporal Interpolation

In the MediaPipe identification of facial landmarks and the Azure Kinect camera acquisition, it is inevitable that some data may be **missing** or **invalid**. Such conditions can occur due to depth sensor limitations, detection errors, or boundary effects inherent in depth sensing technology. To effectively handle these issues, the method implements an articulated strategy that combines the detection of invalid and missing values with temporal interpolation techniques.

Depth sensors, including the Azure Kinect, commonly experience challenges in accurately measuring depth values near object boundaries, a phenomenon known as the **"invalid black area"** problem, green circle in Figure 3.8. This issue arises because depth sensors use structured light or time-of-flight principles that can struggle to provide reliable depth measurements at the edges of objects, where the transition between foreground and background creates ambiguous depth information. Additionally, areas with insufficient reflected light or surfaces that absorb infrared light can result in missing depth

data.

In such cases, some coordinates may be represented as `missing point` to indicate invalid or missing values. Therefore, the code includes a section for handling `missing point` values that follows specific criteria to identify invalid or missing data:

- If all coordinates of a point (x, y, z) are equal to 0, they are replaced with `missing point` to indicate that the point is completely invalid. This condition typically corresponds to the invalid black area problem, where the depth sensor fails to provide any reliable measurement.
- If only the z-coordinate is equal to 0, it is set to `missing point`, while x and y remain unchanged. This occurs when the 2D position can be detected but depth information is unavailable due to sensor limitations.

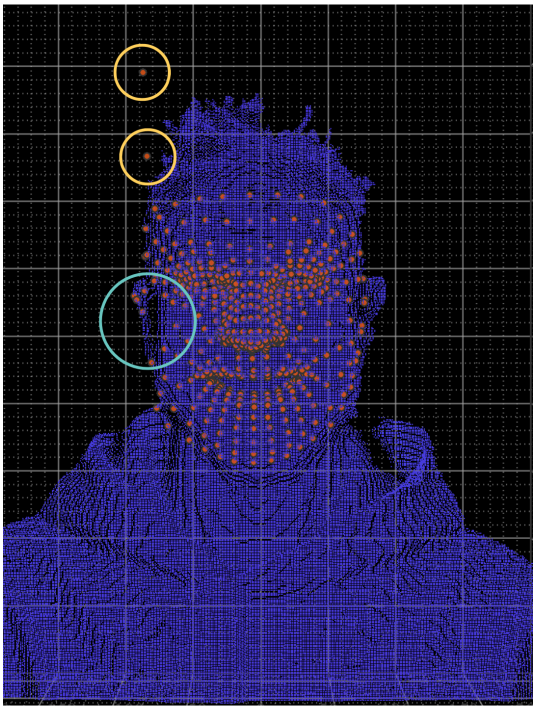

Figure 3.8: Invalid black area and noisy point examples

Another possible cause for the presence of `missing point` in the

data is related to defining acceptable ranges for the facial landmark coordinates, yellow circles in Figure 3.8 . The variables med (acceptable mean value) and var (variable tolerance) define a range within which the coordinates are considered valid. If a coordinate (x, y, z) falls outside of this acceptable range ([med−var, med+var]), it is set to `missing point`. This approach is useful for filtering out anomalous or erroneous values that may be caused by sensor errors or detection issues.

The processing of each single frame allows to obtain a matrix of dimensions 468 landmarks × (frame_number × 3), where the rows represent the 468 detected landmarks and the columns represent the three spatial coordinates (x, y, z) for each frame. In this way, for each landmark, its spatial position in xyz, expressed in millimeters (mm), is provided throughout the entire acquisition. This data structure allows for a detailed analysis of the trajectory and spatial variation of each point detected over time and facilitates the resolution of `missing point` values.

These precautions are helpful as they address the issue of missing data. Following this data representation, the treatment of `missing point` values, such as those caused by occlusions or detection errors, can be addressed through a **temporal interpolation method** that ensures the consistency of the data over time. The code performs the interpolation of the `missing point` values and also applies a **Savitzky-Golay filter** [18] to ensure that the variations in the data are temporally coherent.

In summary, a function is used to interpolate the `missing point` values in the dataset using linear interpolation, allowing to fill the `missing point` gaps caused by small malfunctions of the camera. To refine the temporal consistency, a Savitzky-Golay filter is then applied, which reduces noise and smooths the data. This integrated approach ensures that the trajectory and spatial variations of the landmarks remain consistent and do not present temporal anomalies, guaranteeing temporal stability in the position of the landmarks and enabling reliable analysis of facial movements throughout the acquisition.

3.3.6 Algorithm's Outputs

At this point, the algorithm proceeds to compute all the previously defined facial distances using the landmark numbering system provided by MediaPipe. The processing generates a comprehensive visualization consisting of a single figure with 11 subplots each dedicated to one of the considered muscles, representing the temporal variation of muscular length throughout the analyzed sequence. The main output example of the method is displayed in Figure 3.9. In addition to the graphical output, the same dataset is exported to an Excel file, ensuring immediate compatibility and reusability with other software platforms, such as MATLAB, for further analysis and post-processing applications.

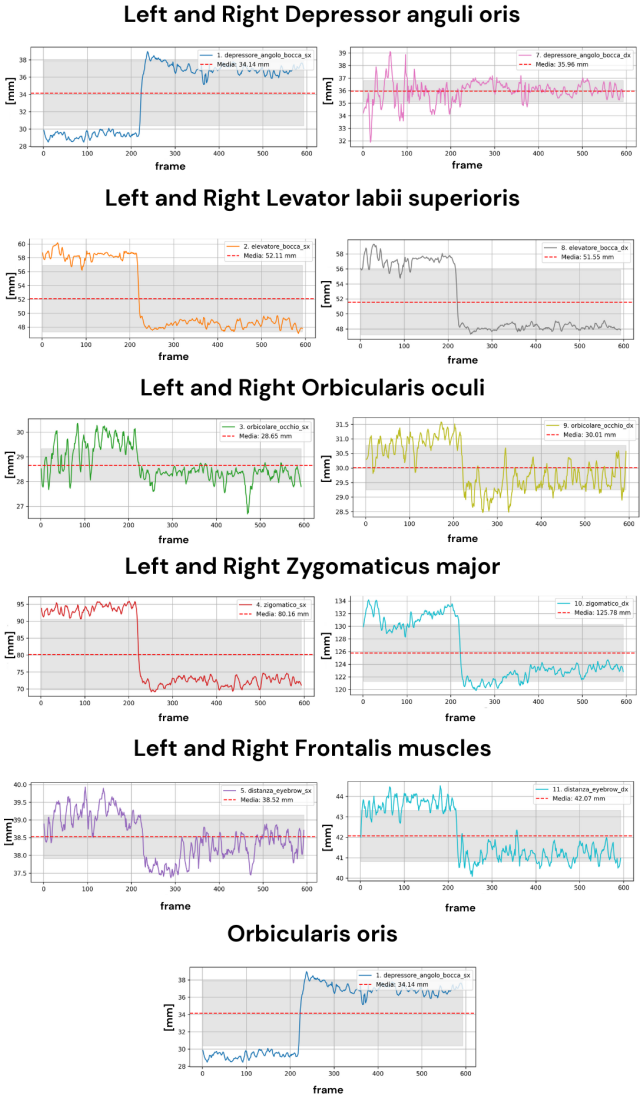

Figure 3.9: Output Muscular Distances during Happiness Expression

3.4 MAE and MAPE Computing

To validate the RGB-D ML method and to evaluate any 2D-method against manual DAO measurements (considered GS), the variation in DAO length between rest and contraction was selected as the primary measure rather than the absolute DAO length. This approach was adopted to eliminate any potential bias arising from individual differences in muscle contraction patterns and general muscle expressiveness across subjects. Additionally, this methodology helps reduce potential errors in landmark identification required for muscle detection, as it focuses on the relative change rather than absolute positioning accuracy.

Two main metrics were selected for evaluation: **Mean Absolute Error (MAE)** and **Mean Absolute Percentage Error (MAPE)**. The mathematical formulations and computation procedures for both metrics are presented below:

- **DAO Length Variation:** for both manual and automatic measurements, the DAO contraction amplitude is calculated using:

$$\Delta DAO = \text{abs}(DAO_{\text{rest}} - DAO_{\text{contr}}) \quad [\text{mm}] \tag{3.1}$$

  where DAO_{rest} represents the DAO muscle dimension at rest and $\text{DAO}_{\text{contract}}$ represents the dimension during MVC.

- **Mean Absolute Error MAE Assessment -** The algorithm's accuracy is quantified by computing MAE between manual and automatic measurements:

$$MAE = \frac{\sum_{i=1}^{n} \text{abs}(\Delta DAO_{i,\text{manual}} - \Delta DAO_{i,\text{automatic}})}{n} \quad [\text{mm}] \tag{3.2}$$

  where n is the number of subjects, $\Delta DAO_{i,\text{manual}}$ is the i-th DAO length variation derived from manual measurements and $\Delta DAO_{i,\text{automatic}}$ is the i-th DAO length variation derived from automatic measurements.

  This metric provides a **quantitative measure** of the discrepancy between the DAO detection of the automatic method and the

DAO GS manual measurements.

- **Mean Absolute Percentage Error (MAPE) Assessment -** In addition to MAE, the MAPE is calculated as:

$$MAPE = \frac{\sum_{i=1}^{n} \frac{\text{abs}(\Delta DAO_{i,\text{manual}} - \Delta DAO_{i,\text{automatic}})}{\Delta DAO_{i,\text{manual}}}}{n} \times 100 \quad [\%] \qquad (3.3)$$

where n is the number of subjects, $\Delta DAO_{i,\text{manual}}$ is the i-th DAO length variation derived from manual measurements and $\Delta DAO_{i,\text{automatic}}$ is the i-th DAO length variation derived from automatic measurements.

Also this metric provides a quantitative measure of the accuracy of the automatic method, by normalizing the MAE with respect to the reference measurement magnitude. Unlike MAE, which expresses the error in absolute units, MAPE enables comparison across different contraction amplitudes and provides insight into the proportional accuracy of the detection algorithm. So, these two metrics give complementary information.

3.5 Dynamic Contraction Parameters

To facilitate comparative dynamical analysis between different subject groups, two key contraction parameters were computed from the DAO muscle activity data. Using MATLAB visualization of DAO contraction profiles, obtained from the RGB-D method application, two critical time points were **manually** identified for each recording acquisition:

- **Contraction onset point t_0:** The temporal instant marking the initiation of the requested emotional expression (DAO MVC, happiness or sadness emotions)
- **Peak contraction point t_1:** The moment of maximum muscular activation during emotional expression

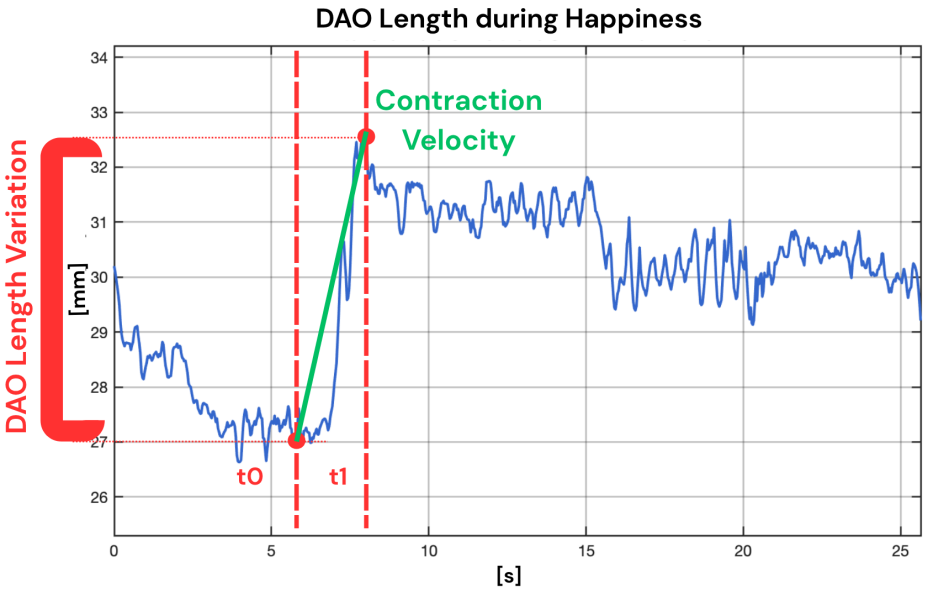

Figure 3.10: Time point identification and Parameters computing

Based on these manually identified time points, t_0 and t_1, the analysis software automatically computed two primary dynamic parameters:

- **DAO Length Variation:** Representing the total contraction range of the muscle during expression.

$$DAO - LV = \text{abs}(DAO_{t1} - DAO_{t0}) \quad [mm] \tag{3.4}$$

where DAO is the length of the DAO muscle at the reference time.

- **Contraction velocity:** Quantifying the rate of muscle activation, thereby providing insight into the subject's neuromuscular control capacity.

$$CV = \frac{DAO - LV}{t_1 - t_0} \quad [mm/s] \tag{3.5}$$

The **DAO-LV** serves as an indicator of an individual's capacity for emotional expression, with reduced values potentially reflecting hypomimia commonly observed in subjects with PD.

The **DAO CV** measures the rate of change in facial muscle movement over time, reflecting an individual's intentionality and speed in displaying emotions, with decreased velocity indicative of **bradykinesia** characteristic of motor impairment in PD patients.

The experimental hypothesis predicts that PD subjects will demonstrate significantly reduced DAO length variation and diminished contraction velocity compared to healthy control groups, including both YH and EH participants.

3.6 Statistic Test between two groups

To compare two generic data groups, a structured methodological pipeline was implemented in two sequential phases, designed to ensure the application of the most appropriate statistical test based on the distributional characteristics of the analyzed samples.

Phase 1: Assessment of Distribution Normality

Before proceeding with the comparative analysis, it was necessary to determine whether the data from both groups followed a normal distribution. This evaluation was carried out using a statistical testing approach. Two complementary statistical tests were applied to verify normality:

- **Shapiro-Wilk Test:** This was used as the primary test for normality assessment, particularly effective for small-to-medium sized samples. The test verifies the null hypothesis that the data come from a normal distribution by calculating the W statistic, comparing the observed distribution with the theoretical normal distribution [20].
- **Lilliefors Test:** This was employed as a supplementary test, representing a variant of the Kolmogorov-Smirnov test specifically adapted to test normality when the parameters of the normal distribution are not known a priori. This test provides additional verification of the normality assumption, increasing the robustness of the preliminary analysis [13].

Phase 2: Application of Statistical Test

Based on the results obtained from the normality assessment, the appropriate statistical test was selected and applied for the comparison between the two groups.

- **Student's t-test for normal distribution**: When both groups showed a normal distribution, the Student's t-test for independent samples was applied. This parametric test compares the means of the two groups assuming that the data follow a normal distribution and that the variances are homogeneous [22].
- **Mann-Whitney Test**: When one or both groups violated the normality assumption, the non-parametric Mann-Whitney U test

was used. This test compares the distributions of the two groups based on the ranks of the observations, proving robust with respect to the shape of the distribution and not requiring the normality assumption [15].

Figure 3.11 graphically shows the statistical test process.

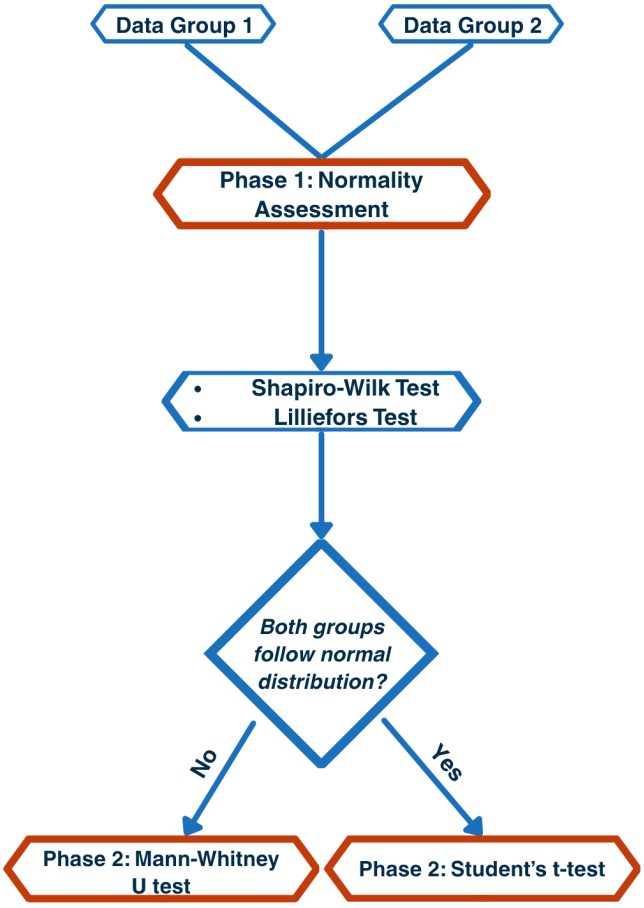

Figure 3.11: Statistical test Pipeline

Chapter 4

Results

This chapter evaluates the proposed method across three key dimensions. First, the method is validated against GS measurements to establish its accuracy and reliability. Second, method performance is assessed under reduced technical specifications to assess its robustness under suboptimal conditions. Finally, method clinical utility is examined through group discrimination during emotional expression tasks.

4.1 Method Validation Against GS Measurements

This section evaluates the performance of the proposed method under optimal video specifications (RGB-D, 30Hz, 1280x720 resolution). The method's accuracy is assessed by comparison with GS measurements provided in the dataset, Section 3.1, for the right DAO muscle. The GS measurements include manual annotations of DAO muscle length at rest and during contraction, as well as corresponding electromyographic (EMG) signals, both recorded during MVC expressions.

4.1.1 Validation Against Manual Right DAO Measurement

This subsection addresses the research question: 'What is the measurement error of the RGB-D ML method relative to manual measurements for right DAO Length Variation?'.

As described in the methodology section, Section 3.4, two comparison metrics were used to evaluate the method performance against the gold standard: Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE). The comparison between automatic

method measurements and manual GS measurements for right DAO length variation produced a MAE of 1.5 ± 1.6 mm and a MAPE of 27.9 ± 27.2%.

Methods	**MAE [mm]**	**MAPE [%]**
Automatic vs. Manual	1.5 ± 1.6	27.9 ± 27.2

Table 4.1: Automatic vs. Manual Method

Statistical Differences between Manual and Automatic Measurements

A statistical comparison was performed between manual and automatic right DAO length variation measurements. As described in the statistical analysis section, Section 3.6, a Mann-Whitney test was applied to assess potential differences between the two measurement groups. As shown in Figure 4.1, the analysis revealed no significant

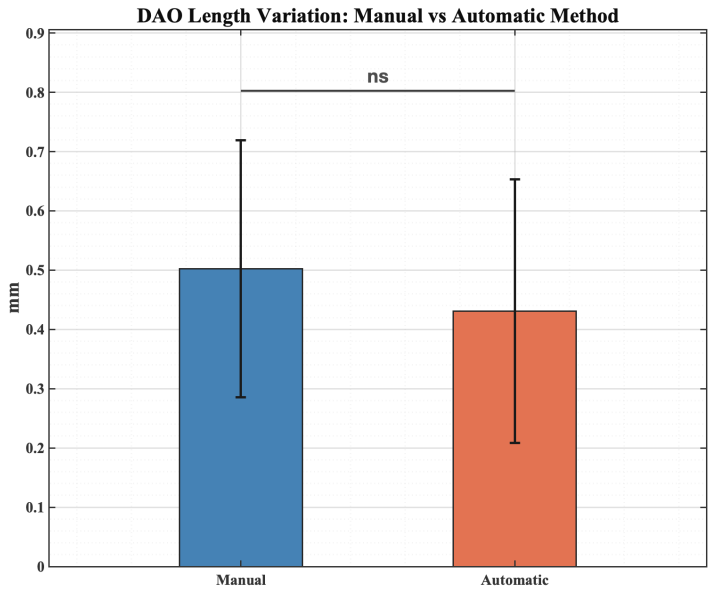

Figure 4.1: Statistic Test between Manual and Automatic DAO Measurements

differences between automatic and manual measures ($p > 0.05$).

4.1.2 EMG Signals Analysis

As mentioned in Section 3.1, the available dataset includes EMG signals from subjects, specifically related to the DAO muscle in the execution of DAO MVC expression. This subsection presents a comprehensive analysis of the EMG data using three key parameters: **EMG Mean, EMG RMS, and EMG Amplitude**. The complete dataset was systematically divided into three distinct subgroups (YH, EH, PD), then each EMG parameter analyzed through consistent statistical methodology by calculating the mean and standard deviation for each group. Figure 4.2 presents the obtained results. The analysis

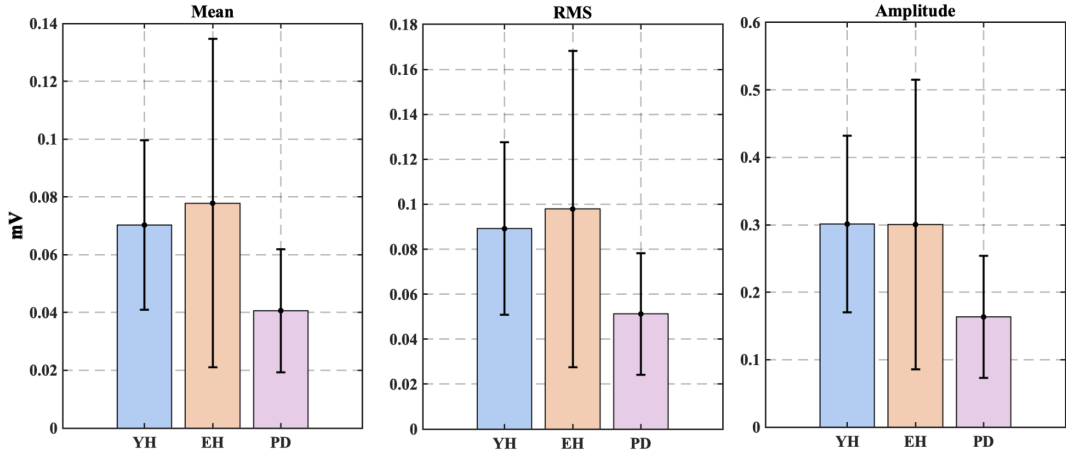

Figure 4.2: EMG Mean, EMG RMS, EMG Amplitude

showed that all three EMG parameters exhibit similar measurement patterns across groups. The healthy control groups (YH and EH) demonstrated consistent EMG RMS values. The pathological group (PD) exhibited reduced EMG RMS values compared to the healthy groups.

4.1.3 Method Validation Against EMG measurements

For additional validation of RGB-D ML method reliability, a comparative analysis was performed between DAO Length Variation measurements, method index, and EMG RMS signal trends during MVC expression. The RMS was obtained as described in Section 4.1.2. The analysis examined the correlation between the proposed method's

DAO Length Variation output and the average EMG RMS signals across the three participant groups (YH, EH, and PD).

Statistical Correspondence

Figures 4.3 and 4.4 show the results of the statistical analysis between the two measurement approaches. The statistical analysis re-

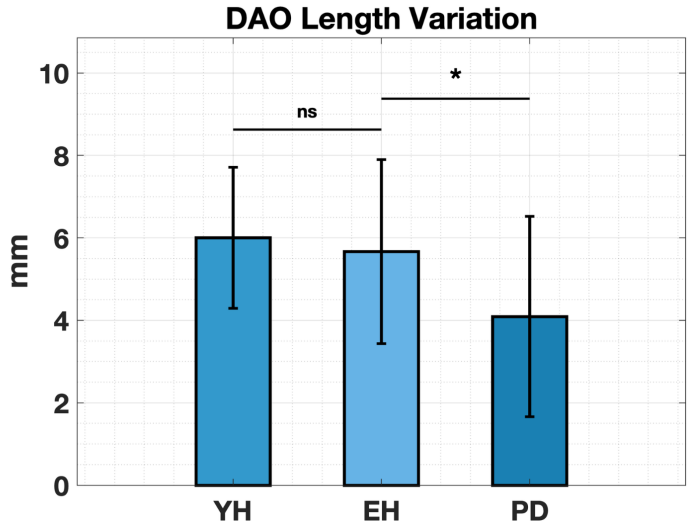

Figure 4.3: DAO Length Variation during MVC expression

vealed:

Both DAO Length Variation and EMG RMS showed no significant differences between Young Healthy (YH) and Elderly Healthy (EH) groups ($p > 0.05$) Both measures demonstrated significant differences when comparing PD patients with EH participants (DAO-LV $p = 0.035$, EMG RMS $p = 0.0418$)

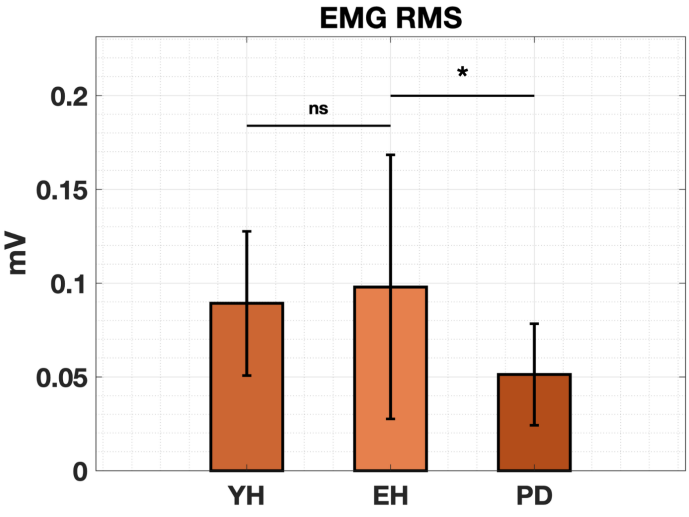

Figure 4.4: EMG RMS during MVC expression

Both measurement modalities demonstrated consistent group discrimination patterns.

4.2 Method Performance Under Reduced Technical Specifications

This section presents the algorithm's performance under various experimental conditions, including different dimensionality configurations (2D vs 3D), frame rates (15 vs 30 Hz), and image resolutions (1280x720 vs 640x480). The analysis aims to quantify how these parameters affect the algorithm's accuracy in detecting DAO muscle contractions.

The results are expected to demonstrate measurable performance degradation in DAO muscle contraction identification as technical constraints are introduced, providing insights into the algorithm's robustness and optimal operating conditions.

4.2.1 Calibration Method for 2D Analysis

The RGB-D ML method was not the only approach evaluated in this study. To assess the robustness and practical applicability of the proposed method, it was essential to evaluate its performance under suboptimal conditions that deviate from the ideal experimental setup. These challenging scenarios included the removal of depth sensor information and the reduction of image resolution. Such evaluations necessitated the development of alternative analysis pipelines and, critically, the calculation of a **pixel-to-millimeter conversion factor** to ensure dimensional consistency across different implementation approaches.

The calibration approach differs significantly between the 2D and 3D implementations. For 3D analysis, no additional calibration was necessary, since the Azure Kinect camera directly provides all required spatial coordinates in millimeters. However, 2D analysis relied solely on RGB images without utilizing Kinect depth information. Since Python's MediaPipe library returns 2D coordinate values in pixels, a **pixel-to-millimeter conversion factor** was required to maintain dimensional consistency.

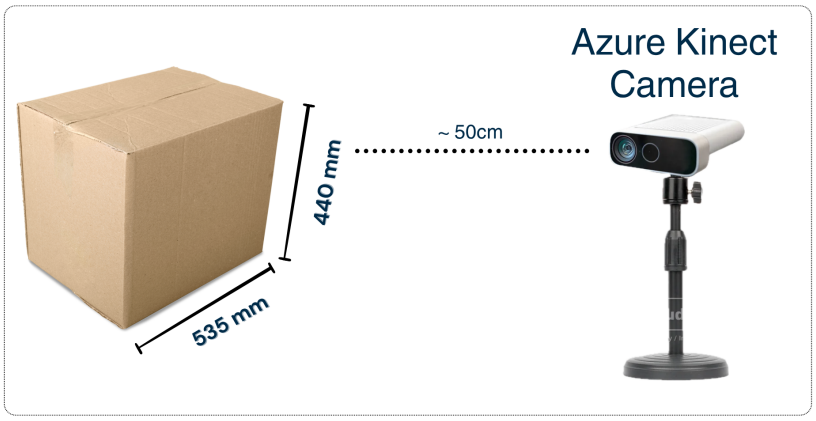

Figure 4.5: Calibration object

The conversion factor was calculated using a rigorous calibration procedure with a reference object of known dimensions (*535 x 440 mm*) positioned at *50 cm* from the Azure Kinect camera, maintaining the same spatial configuration used for all facial acquisitions. Using a dedicated MATLAB script, two reference points were manually selected on the calibration object within one acquired RGB image (*Figure 4.5*), and the Euclidean distance between these points was computed in pixel coordinates. The conversion factor was determined by dividing the known real-world distance (*535 mm*) by the measured pixel distance, yielding a calibration factor of **0.833 mm/pixel**. This factor was subsequently applied to all 2D measurements to transform pixel-based distances into metric units [mm], ensuring accurate quantitative analysis and enabling proper validation against established measurement standards.

After exploiting the 2D configuration specifics, the algorithm's performance was exploited under different experimental conditions in 4.2.

Configuration	**Dim**	**Frame Rate (Hz)**	**Resolution**
#1	3D	30	1280x720
#2	2D	30	1280x720
#3	2D	30	640x480

Table 4.2: Camera configuration parameters

4.2.2 MAE and MAPE Across Video Configurations

The video configuration analysis focused exclusively on **maximum DAO contraction acquisitions**, as these represent the only conditions for which certified manual measurements are available. The manual measurements of the DAO muscle at rest and during maximum contraction serve as the gold standard (GS) for validation, enabling objective assessment of the algorithm's accuracy in detecting DAO muscle contractions. MAE and MAPE were calculated for each configuration described in Table 4.2. The comparative MAE analysis across all tested configurations is presented in Figure 4.6, while MAPE analysis in Figure 4.7. **Configuration #1** (3D, 30Hz, 1280x720)

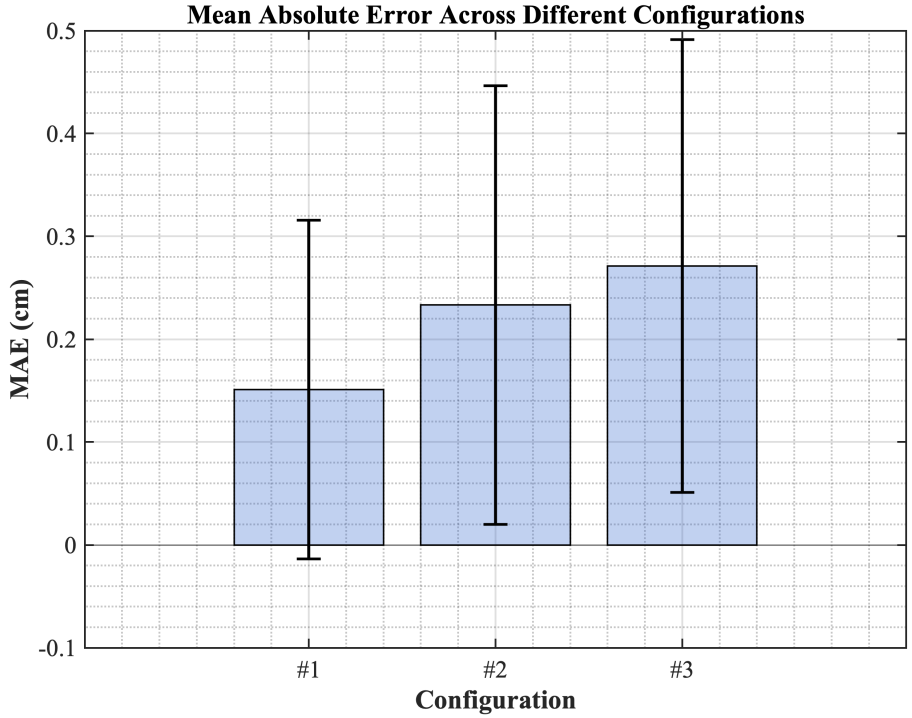

Figure 4.6: Mean Absolute Error MAE per configuration

achieved the lowest MAE value (0.15 cm) and MAPE value (27%). **Configurations #2** (2D, 30Hz, 1280x720) exhibited MAE values of (0.23-0.25 cm) and MAPE values of (45-47%), while **Configurations #3** (2D, 30Hz, 640x480) demonstrated the highest MAE values (0.27-0.26 cm) and MAPE values (48-50%). The transition from 3D to 2D acquisition (Configuration #1 vs #2) resulted in approximately 60% increase in both parameters. The reduction from 1280x720 to 640x480

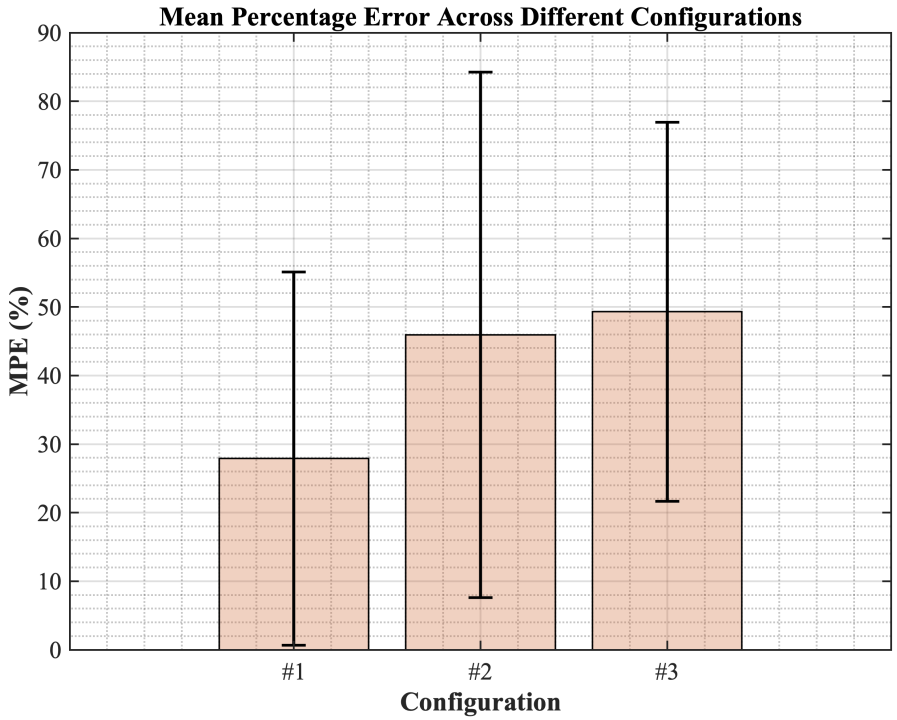

Figure 4.7: Mean Absolute Percentage Error MAPE per configuration

(comparing #2 vs #3) produced MAE and MAPE increases of approximately 15-20%. A brief analysis on the impact of the **frame rate** on the evaluation of DAO measurements indicated that temporal sampling frequency has minimal influence on contraction detection accuracy.

Configuration	MAE (cm)	MAPE (%)
3D, 30Hz, 1280x720	0.15 ± 0.16	27.9 ± 27.22
2D, 30Hz, 1280x720	0.23 ± 0.21	45.91 ± 38.3
2D, 30Hz, 640x480	0.27 ± 0.22	49.31 ± 27.65

Table 4.3: Experimental results on different video configurations

4.2.3 DAO Measurements Statistical Analysis

To further validate the observed performance differences between configurations, a statistical analysis was conducted using the Mann-Whitney test. The analysis specifically compared the 3D configuration (Configuration #1) against each 2D configuration to assess whether the performance differences are statistically significant. The statistical comparison results are presented in Table 4.4. Both p-values are below

Configuration	p-value
2D, 30Hz, 1280x720	0.037
2D, 30Hz, 640x480	0.002

Table 4.4: 2D configurations vs. RGB-D Method

the conventional significance threshold of $\alpha = 0.05$, confirming that the observed performance improvements of the 3D configuration are statistically significant.

4.3 Clinical Application: Group Discrimination During Emotional Expression

This section presents a comprehensive statistical emotion analysis examining dynamic contraction parameters, calculated as in Section 3.5, with particular emphasis on left and right DAO muscle contraction differences across the study groups described in *Section 3.1*. To enhance the statistical power of the analysis, data from both left and right DAO muscles were combined into a single expanded dataset, effectively doubling the available sample size. Prior to conducting the expressive emotion analysis, **outlier detection** and **removal** were performed using the boxplot method, which identifies anomalous values based on quartile calculations and removes data points falling outside the range defined by 1.5 times the interquartile range from the first and third quartiles.

In Section 4.1.3 the method capacity to discriminate between different study groups has been tested, comparing the method's output with the GS EMG signals. That assessment is now used to discriminate between study groups during happiness and sadness expression analysis. The final objective of the analysis is to evaluate the clinical application of the method, in assessing hypomimia and bradykinesia in PD subject with respect to healthy subjects, YH and EH.

4.3.1 Happiness expression

Happiness expression represents a highly dynamic facial movement that requires an elevated range of DAO muscle contraction compared to other emotional expressions, such as sadness. The analysis of happiness expression focuses on two key parameters: DAO Length Variation (DAO-LV) and Contraction Velocity (CV), as defined in equations 3.4 and 3.5.

Statistical analysis revealed distinct patterns across the study groups. No significant differences were observed between Young Healthy (YH) and Elderly Healthy (EH) subjects for both DAO-LV and CV parameters ($p > 0.05$). However, significant differences emerged when comparing EH subjects with PD patients. Both DAO LV and CV showed statistically significant reductions in the PD group compared

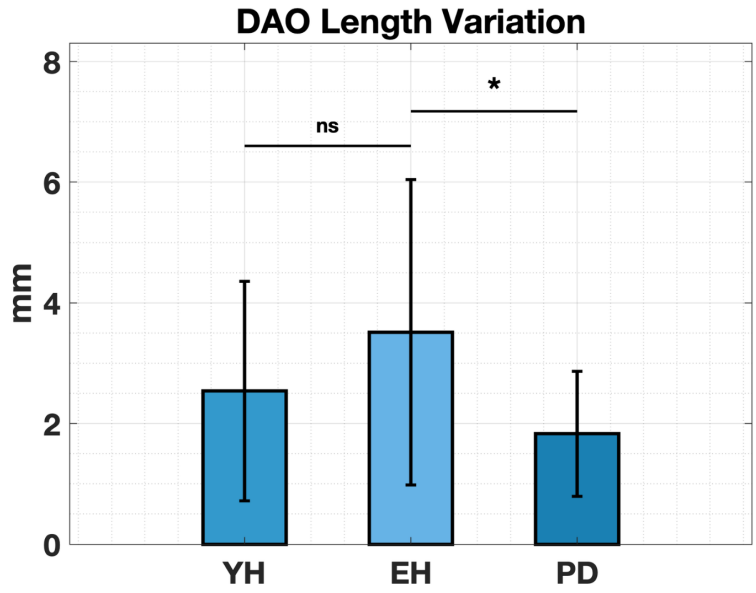

Figure 4.8: DAO Length Variation - Happiness Dynamic Analysis

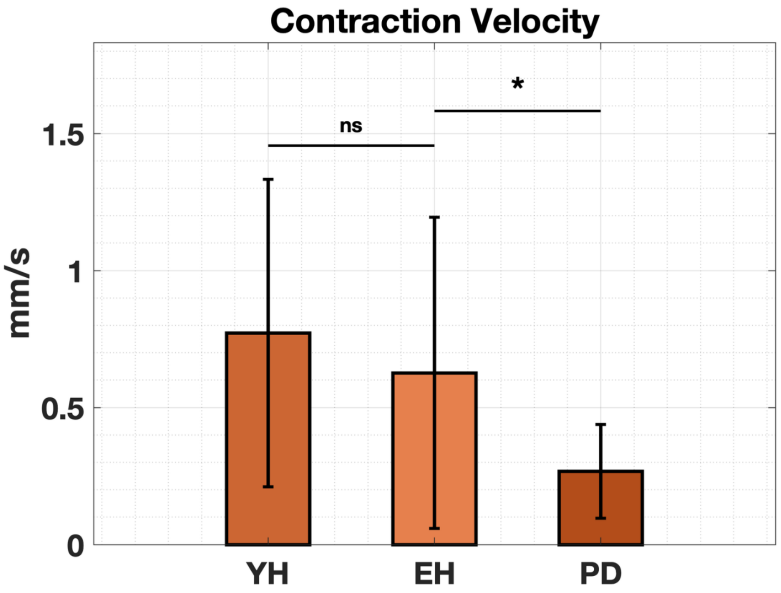

Figure 4.9: DAO Contraction Velocity - Happiness Dynamic Analysis

to the EH group ($p = 0.03$).

4.3.2 Sadness expression

Sadness expression involves minimal DAO muscle displacement that closely resembles the resting facial condition. Statistical analysis revealed no significant differences between Young Healthy (YH) and Elderly Healthy (EH) subjects for both DAO-LV and CV parameters ($p > 0.05$), consistent with the findings observed in happiness expression analysis. No significant differences were detected between EH

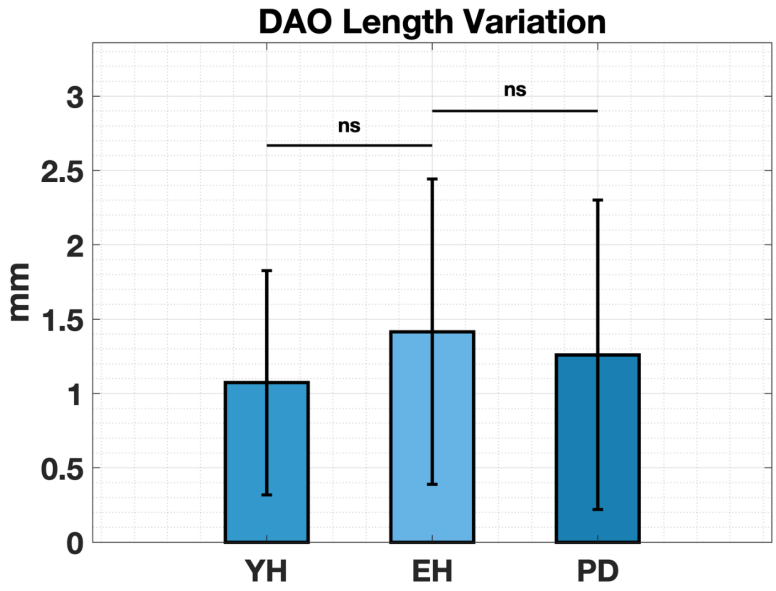

Figure 4.10: DAO Length Variation - Sadness Dynamic Analysis

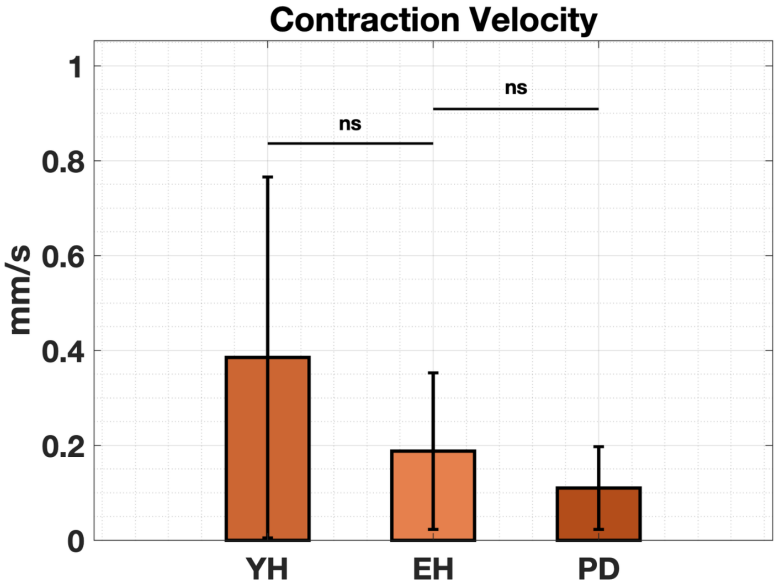

Figure 4.11: DAO Contraction Velocity - Sadness Dynamic Analysis

subjects and PD patients for either parameter ($p > 0.05$).

Chapter 5

Discussions

5.1 Interpretation of Method Validation Results

MAE Performance Analysis. Considering that data acquisition was performed at a *50 cm* distance between subjects' faces and the Azure Kinect camera, a MAE of **1.5 mm** demonstrates excellent performance, as reported in Table 4.1. The RGB-D method shows **high accuracy** in identifying DAO contraction compared to manual measurements, with sub-millimeter precision that is remarkable for automated facial muscle tracking at this distance.

MAPE Performance Analysis. The average DAO contraction range during MVC expression is approximately *6-7 mm*. Within this limited range, as reported in Table 4.1, an MAPE of **27.9%** indicates that while the absolute error is small, even minimal deviations can result in disproportionately high percentage errors. This highlights the critical importance of precise muscle identification and the inherent challenge of measuring such small physiological movements. The relatively high percentage error reflects the sensitivity required for accurate DAO contraction assessment rather than a fundamental limitation of the method.

Statistical Validation Implications. As shown in Figure 4.1, the statistical validation demonstrates that the automatic method produces measurements that are statistically equivalent to manual GS annotations, supporting the method's validity for clinical and research applications. This confirms **the reliability of the muscle identification method** and establishes that this validation establishes the fundamental accuracy of the method when operating under ideal

technical conditions.

Interpretation of EMG Signals Analysis. The results in Figure 4.2 demonstrate that all three EMG parameters exhibit remarkably similar measurement patterns across groups, thereby making the selection of any single parameter for subsequent analysis essentially equivalent in terms of discriminative power.

Focusing specifically on the **EMG RMS** parameter, several clinically significant observations emerge from the comparative analysis between groups. The healthy control groups (YH and EH) demonstrate remarkably consistent EMG RMS values, indicating no substantial age-related differences in muscle activation patterns between younger and elderly healthy subjects during standardized facial expressions. This finding suggests that **normal aging does not significantly compromise the electrical activity of the DAO muscle** during facial movement tasks.

In marked contrast, the pathological group (PD) exhibits significantly diminished EMG RMS values when compared to their healthy counterparts. This reduction in muscle electrical activity is particularly pronounced in PD patients, providing **quantitative confirmation** of the expected impairment **in muscle contraction capacity** that characterizes this neurodegenerative condition.

These findings demonstrate strong alignment with established clinical expectations, as the observed reduction in EMG activity within pathological groups provides objective **quantitative evidence for the hypomimia** documented in PD patients. The EMG data thus serves as a valuable biomarker, offering measurable support for clinical observations of reduced facial expressiveness commonly associated with neurodegenerative conditions.

Interpretation of EMG Validation Results

The cross-validation approach provides independent physiological confirmation of the method's accuracy by comparing mechanical muscle movement detection, Figure 4.3, with electrical muscle activity measurement, Figure 4.4. The statistical analysis revealed a remarkable concordance between the two measurement approaches. This parallel statistical behavior confirms that the RGB-D ML protocol captures physiologically meaningful muscle activity patterns that correspond directly with EMG-detected electrical muscle activation.

The method's ability to distinguish between clinical populations with the same sensitivity as EMG measurements validates its potential as a non-invasive alternative for DAO muscle assessment and demonstrates the **RGB-D ML protocol's sensitivity in discriminating between groups** with different neuromuscular characteristics.

The consistent group discrimination patterns between RGB-D and EMG measurements support the validity of the RGB-D approach, providing evidence that the proposed method can reliably detect physiological differences that correlate with established electrophysiological measures.

5.2 Interpretation of Method Performance Under Reduced Technical Specifications

The analysis of MAE, Figure 4.6, and MAPE results, Figure 4.7, across different video configurations reveals significant insights into the factors influencing the algorithm's performance in detecting DAO muscle contractions. Configuration #1, being the only one with the depth sensor information, represented optimal performance conditions. The results establish a clear hierarchical relationship between technical parameters and measurement accuracy: **dimensionality emerges as the primary influence factor**, with the transition from 3D to 2D acquisition resulting in approximately 60% increase in both parameters. This substantial impact demonstrates that three-dimensional data provides essential depth information for accurate characterization of muscle tissue deformation during contraction, which cannot be fully captured through planar imaging alone.

Image resolution represents a secondary influence factor. While the reduction from 1280x720 to 640x480 represents measurable performance degradation, the impact remains substantially lower than the dimensionality effect, suggesting that the algorithm demonstrates reasonable robustness to resolution variations within clinically relevant ranges. The findings indicate that **temporal sampling frequency has minimal influence** on contraction detection accuracy, suggesting that lower frame rates can be employed without compromising measurement quality. These findings have important implications

for clinical implementation. The predominant influence of dimensionality indicates that 3D acquisition should be prioritized when maximum accuracy is required, while the moderate impact of spatial resolution provides flexibility for different clinical scenarios when 3D acquisition is not feasible. The minimal frame rate dependency offers practical advantages for system implementation, allowing for reduced computational overhead and storage requirements without significant accuracy compromise, which is particularly valuable in clinical environments where real-time processing capabilities may be limited.

DAO Measurements Statistical Analysis. The Table 4.4 results provide strong statistical evidence supporting the superiority of depth sensor-based acquisition over conventional 2D approaches. The highly significant p-value for the Configuration #1 vs. #3 comparison ($p = 0.002$) particularly emphasizes that the performance gap becomes more pronounced as both dimensionality and resolution factors compound their effects. This statistical confirmation reinforces the clinical relevance of the observed differences, demonstrating that the enhanced accuracy achieved through 3D acquisition represents a genuine improvement rather than random variation, and establishes that 2D configurations cannot reliably match the measurement precision achievable with depth sensor information.

5.3 Interpretation of Emotional Expression

Happiness Expression.

The finding that **age alone does not significantly affect the dynamic characteristics of happiness expression in healthy individuals** suggests that the facial muscle coordination required for expressing happiness remains relatively preserved during normal aging. The significant differences between EH subjects and PD patients provide quantitative evidence of **hypomimia** and **bradykinesia** in PD subjects during happiness expression. The reduced DAO-LV, Figure 4.8, reflects the diminished capacity for emotional expression characteristic of hypomimia, while the decreased CV, Figure 4.9,demonstrates the motor slowness (bradykinesia) that affects facial muscle movement

in PD patients. These results confirm the method's sensitivity in detecting the subtle motor deficits that characterize PD-related facial expression impairments. This characteristic makes happiness a particularly suitable emotion for identifying inter-group differences, as the more expressive nature of the movement amplifies the detection of subtle motor impairments.

Sadness Expression.

Sadness expression presents a fundamentally different challenge compared to happiness analysis due to its inherently subtle nature and reduced movement amplitude. This similarity to the neutral state introduces **significant inter-subject variability** in expression, as individual baseline facial configurations and personal expressive tendencies become more influential factors in the measurement outcomes. The reduced movement range characteristic of sadness expression limits **the method's discriminative power** between study groups. The absence of significant differences, Figure 4.10 and Figure 4.11, between EH and PD groups in sadness expression can be attributed to several factors. First, the **minimal movement amplitude** required for sadness expression approaches the detection threshold of the method, making subtle motor impairments less distinguishable from normal inter-subject variability. Second, the proximity of sadness expression to the resting facial state **reduces the contrast** needed to identify bradykinesia and hypomimia effects. These findings highlight the importance of selecting appropriate emotional stimuli when assessing facial motor function in clinical populations, emphasizing that more expressive emotions provide superior diagnostic sensitivity for detecting PD-related motor deficits.

Chapter 6

Conclusions

In conclusion, returning to the main objectives of this thesis, the developed RGB-D method has proven to be more than reliable and clinically valid. One of the main future development directions consists in establishing a systematic association between facial landmarks and muscles, to encompass all facial muscles that can be analyzed by the method and not limit the analysis to the DAO muscle alone. In this way, it would be possible to obtain real-time measurements of eleven muscles during the execution of specific expressions or facial movements. Having multiple muscular parameters would mean obtaining more comprehensive information regarding the same expressive movement, thus making the concept of a comprehensive evaluation parameter a concrete prospect. By considering all muscles and integrating them within an index that evaluates their contraction (both in terms of range of movement and contraction velocity), it would be possible to provide an indicator of subjects' global contractile capacity, potentially reaching the definition of a Parkinson's disease progression parameter based exclusively on hypomimia and bradykinesia symptoms. Naturally, for each identified muscle, it would be necessary to perform specific validation of muscle identification, similar to that carried out for the DAO (MAE = 1.51 mm and MAPE = 27.9%). This would require the acquisition of gold standard measurements for all muscles to be included in the analysis. The method has also demonstrated good discriminative capacity between different groups (YH-EH-PD), as confirmed by comparison with DAO EMG data, provided that the movement performed is sufficiently broad and expressive (such as in happiness expression, rather than sadness

expression). Regarding the technical specifications of the camera, the depth sensor has proven to be the critical element in this type of analysis, since the method's performance without it decreases drastically, resulting in an error increase of approximately 52% for both MAE and MAPE. Concerning the image resolution used, the impact is significant but of lesser magnitude. Indeed, resolution reduction leads to a percentage error increase of approximately 79%, therefore less influential than the absence of the depth sensor, but still relevant for the overall system accuracy. The frame rate, as expected, is a parameter that has limited influence in this type of application, dealing with controlled and relatively slow movements, where a 30 Hz frame rate does not substantially differ from a 15 Hz one in terms of detection efficiency. Once all the proposed methodological improvements have been implemented, it would be interesting to conduct a more in-depth study of emotional expressions, including additional emotions beyond happiness and sadness, such as anger, surprise, or disgust. The objective would remain the quantitative evaluation of how much the expressive movement of PD subjects is reduced compared to healthy subjects, defining auxiliary parameters to understand disease progression and allowing the developed method to be considered a full-fledged clinical evaluation tool.

Bibliography

[1] T. Baltrušaitis, P. Robinson, and L. P. Morency. "OpenFace: An open source facial behavior analysis toolkit". In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2016, pp. 1–10.

[2] M. Bologna et al. "Facial bradykinesia". In: *Journal of Neurology, Neurosurgery & Psychiatry* 84.6 (2013), pp. 681–685.

[3] G. Bradski. "The OpenCV Library". In: *Dr. Dobb's Journal of Software Tools* (2000).

[4] A. Bulat and G. Tzimiropoulos. "How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks)". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 1021–1030.

[5] K. W. Burton and A. W. Kaszniak. "Emotional experience and facial expression in Alzheimer's disease". In: *Aging, Neuropsychology, and Cognition* 13.3-4 (2006), pp. 636–651.

[6] L. Cattaneo and G. Pavesi. "The facial motor system". In: *Neuroscience & Biobehavioral Reviews* 38 (2014), pp. 135–159.

[7] C. J. De Luca. "The use of surface electromyography in biomechanics". In: *Journal of Applied Biomechanics* 13.2 (1997), pp. 135–163.

[8] P. Ekman and W. V. Friesen. *Facial action coding system: A technique for the measurement of facial movement*. Consulting Psychologists Press, 1978.

[9] D. A. Furtado et al. "A specialized motion capture system for real-time analysis of mandibular movements using infrared cameras". In: *Biomedical Engineering Online* 12.1 (2013), pp. 1–16.

[10] G. Ghiasi and C. C. Fowlkes. "Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 2385–2392.

[11] Yury Kartynnik et al. "Real-time facial surface geometry from monocular video on mobile GPUs". In: *arXiv preprint arXiv:1907.06724* (2019). URL: `https://arxiv.org/abs/1907.06724`.

[12] D. E. King. "Dlib-ml: A machine learning toolkit". In: *The Journal of Machine Learning Research* 10 (2009), pp. 1755–1758.

[13] H.W. Lilliefors. "On the Kolmogorov-Smirnov test for normality with mean and variance unknown". In: *Journal of the American Statistical Association* 62.318 (1967), pp. 399–402. DOI: 10.1080/01621459.1967.10482916. URL: https://www.jstor.org/stable/2283970.

[14] C. Lugaresi et al. "MediaPipe: A framework for building perception pipelines". In: *arXiv preprint arXiv:1906.08172* (2019).

[15] H.B. Mann and D.R. Whitney. "On a test of whether one of two random variables is stochastically larger than the other". In: *Annals of Mathematical Statistics* 18.1 (1947), pp. 50–60. DOI: 10.1214/aoms/1177730491. URL: https://projecteuclid.org/euclid.aoms/1177730491.

[16] T. Maycas-Cepeda et al. "Hypomimia in Parkinson's Disease: What Is It Telling Us?" In: *Frontiers in Neurology* 11 (2021), p. 603582. DOI: 10.3389/fneur.2020.603582. URL: https://www.frontiersin.org/articles/10.3389/fneur.2020.603582/full.

[17] S. Raith et al. "Computational modeling in maxillofacial surgery". In: *Perspectives in Medicine* 12.1 (2021), p. 100112.

[18] Abraham Savitzky and Marcel JE Golay. "Smoothing and differentiation of data by simplified least squares procedures". In: *Analytical Chemistry* 36.8 (1964), pp. 1627–1639. DOI: 10.1021/ac60214a047.

[19] F. Schlagenhauf, S. Sreeram, and W. Singhose. "Accuracy and repeatability of the Microsoft Azure Kinect for clinical measurement of motor function". In: *PLOS ONE* 17.12 (2022), e0279697. DOI: 10.1371/journal.pone.0279697. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0279697.

[20] S.S. Shapiro and M.B. Wilk. "An analysis of variance test for normality (complete samples)". In: *Biometrika* 52.3-4 (1965), pp. 591–611. DOI: 10.1093/biomet/52.3-4.591. URL: https://academic.oup.com/biomet/article/52/3-4/591/336553.

[21] L. Shu et al. "A review of emotion recognition using physiological signals". In: *Sensors* 18.7 (2018), p. 2074.

[22] Student. "The probable error of a mean". In: *Biometrika* 6.1 (1908), pp. 1–25. DOI: 10.2307/2331554. URL: https://www.jstor.org/stable/2331554.

[23] P. Tzirakis et al. "End-to-end multimodal emotion recognition using deep neural networks". In: *IEEE Journal of Selected Topics in Signal Processing* 11.8 (2017), pp. 1301–1309.

[24] T. C. Wang, A. Mallya, and M. Y. Liu. "One-shot free-view neural talking-head synthesis for video conferencing". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 10039–10049.

[25] C. Wu et al. "An anatomically-constrained local deformation model for monocular face capture". In: *ACM Transactions on Graphics* 35.4 (2016), pp. 1–12.

[26] H. Zhu et al. “Evaluating the Accuracy of the Azure Kinect and Kinect v2”. In: *Sensors* 22.7 (2022), p. 2588. DOI: 10.3390/s22072588. URL: https://www.mdpi.com/1424-8220/22/7/2588.

[27] X. Zhu and D. Ramanan. “Face detection, pose estimation, and landmark localization in the wild”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2012, pp. 2879–2886.

[28] M. Zollhöfer et al. “State of the art on monocular 3D face reconstruction, tracking, and applications”. In: *Computer Graphics Forum* 37.2 (2018), pp. 523–550.