ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ ΣΤΗΝ ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΤΕΧΝΟΛΟΓΙΕΣ ΠΛΗΡΟΦΟΡΙΑΣ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αντιστοίχιση των αρθρώσεων του σκελετού σε κινήσεις του Avatar στον χώρο νοηματισμού

Δημήτριος Ε. Καραμανίδης

Επιβλέπων: **Χαρίλαος Παπαγεωργίου,** Διευθυντής Έρευνας

ΑΘΗΝΑ

ΣΕΠΤΕΜΒΡΙΟΣ 2023

NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCES
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS

INTERDEPARTMENTAL PROGRAM OF POSTGRADUATE STUDIES IN DATA SCIENCE AND INFORMATION TECHNOLOGIES

MSc THESIS

Mapping of skeleton keypoints to avatar motions in signing space

Dimitios E. Karamanidis

Supervisor: **Charilaos Papageorgiou,** Research Director

ATHENS

SEPTEMBER 2023

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αντιστοίχιση των αρθρώσεων του σκελετού σε κινήσεις του Avatar στον χώρο νοηματισμού

Δημήτριος Ε. Καραμανίδης
Α.Μ.: DS1200004

ΕΠΙΒΛΕΠΩΝ : **Χαρίλαος Παπαγεωργίου,** Διευθυντής Έρευνας

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ:

Χαρίλαος Παπαγεωργίου, Διευθυντής Έρευνας

Βασίλειος Κατσούρος, Διευθυντής Έρευνας

Ελένη Ευθυμίου, Διευθύντρια Έρευνας

ΣΕΠΤΕΜΒΡΙΟΣ 2023

MSc THESIS

Mapping of skeleton keypoints to avatar motions in signing space

Dimitios E. Karamanidis
S.N.: DS1200004

SUPERVISOR: **Charilaos Papageorgiou,** Research Director

EXAMINATION COMMITTEE:

Charilaos Papageorgiou, Research Director

Vasileios Katsouros, Research Director

Eleni Efthimiou, Research Director

SEPTEMBER 2023

ΠΕΡΙΛΗΨΗ

Η Νοηματική γλώσσα αποτελεί τον κύριο τρόπο επικοινωνίας για άτομα που είναι κωφά ή αντιμετωπίζουν προβλήματα στην ακοή. Η αναπαράσταση της Νοηματικής γλώσσας αποτελεί μια πολύπλοκη διαδικασία, η οποία εμπλέκει ανθρώπινες δραστηριότητες που απαιτούν πολύ χρόνο. Για να αντιμετωπίσουμε αυτήν την πρόκληση, προτείνουμε μια αυτοματοποιημένη μέθοδο που να αντιστοιχεί τις αρθρώσεις του σκελετού σε κινήσεις του Avatar στον χώρο νοηματισμού, χρησιμοποιώντας προηγμένες τεχνικές βαθιάς μάθησης. Αυτή η αντιστοίχιση επιτυγχάνεται με την ακριβή εξαγωγή συντεταγμένων 3Δ αρθρώσεων του σώματος από βίντεο, χρησιμοποιώντας τελευταίας τεχνολογίας αλγόριθμους για την εκτίμηση της ανθρώπινης πόζας. Στη μελέτη μας, εξετάζουμε συγκεκριμένες προσεγγίσεις που εντοπίζουν τα 2Δ σημεία του σκελετού από βίντεο και στην συνέχεια τα μετατρέπουν στο 3Δ χώρο, τις οποίες αξιολογούμε σε ένα μικρό συνθετικό σύνολο δεδομένων που περιλαμβάνει πέντε βίντεο με το avatar Paula. Η έρευνα μας επικεντρώνεται στις κινήσεις των χεριών, δίνοντας έμφαση στους ώμους, τους αγκώνες και τους καρπούς, αναγνωρίζοντας τη σημασία τους στην κατανόηση της νοηματικής γλώσσας. Λόγω της εκπαίδευσης των αξιολογημένων μεθόδων σε γενικά σύνολα δεδομένων και όχι σε συγκεκριμένα για τη νοηματική γλώσσα, κάναμε ορισμένες προσαρμογές προκειμένου να επιτύχουμε την αντιστοίχιση των σημείων του σκελετού. Επίσης, παρέχουμε μια ολοκληρωμένη ανάλυση των πλεονεκτημάτων και των αδυναμιών για κάθε μέθοδο και αναφέρουμε συγκεκριμένα μοτίβα της απόδοση τους που παρατηρήθηκαν σε κάθε άξονα. Σημαντικό είναι ότι η προσέγγιση που χρησιμοποιεί το μοντέλο BlazePose του Mediapipe για την εκτίμηση της 2Δ πόζας και το VideoPose3D για την 3Δ ανακατασκευή, υπερτερεί των υπολοίπων, επιτυγχάνοντας ένα μέσο σφάλμα αρθρώσεων (MPJPE) ίσο με 72.2 χιλιοστά.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Όραση Υπολογιστών

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Εκτίμηση ανθρώπινης πόζας, 3Δ ανακατασκευή, αναπαράσταση νοηματικής γλώσσας, κίνηση Avatar

ABSTRACT

Sign Language constitutes the primary means of communication for the deaf and hard-of-hearing individuals. Sign Language Representation is a complex task, which involves human labor-intensive processes. To address this challenge, we propose an automated method that maps skeleton keypoints to avatar motions by leveraging advanced deep learning approaches. This mapping can be achieved by extracting accurate 3d body joints coordinates from monocular videos using state-of-the-art human pose estimation algorithms. In our study, we investigate certain approaches which detect the 2D body joints in videos and subsequently convert them into 3D space, evaluated on a small synthetic dataset of five videos, featuring the Paula avatar. Our work focuses on arm motions, emphasizing on keypoints related to shoulders, elbows, and wrists, acknowledging the significance of their movements in sign language understanding. Due to the training of evaluated methods on generic dataset rather than those specific to sign language, we had to make certain adjustments to ensure the accordance of skeleton keypoints. We provide a comprehensive analysis of the benefits and drawbacks of each method and report special patterns of performance on different axes. Notably, the approach, which uses the BlazePose of Mediapipe as the 2D detector and the VideoPose3D for 3D reconstruction, outperforms its competitors, achieving an average Mean Per Joint Position Error (MPJPE) of 72.2 mm.

SUBJECT AREA: Computer Vision

KEYWORDS: human pose estimation, 3D reconstruction, sign language representation, avatar motion

ΕΥΧΑΡΙΣΤΙΕΣ

Καταρχάς, θα ήθελα να ευχαριστήσω τα μέλη της τριμελούς μου επιτροπής και ιδιαίτερα την διευθύντρια ερευνών στο Ινστιτούτο Επεξεργασίας του Λόγου του Ε.Κ. Αθηνά, Ελένη Ευθυμίου. Η καθοδήγησή και η υποστήριξή της σε όλα τα στάδια της έρευνάς μου ήταν ανεκτίμητη. Επίσης, είμαι ευγνώμων για τις εισηγήσεις και τις συμβουλές που μου έδωσε η διευθύντρια ερευνών στο αντικείμενο «Τεχνολογίες Φωνής και Νοηματικής Γλώσσας για υποστήριξη ατόμων με αναπηρία», Ευίτα-Σταυρούλα Φωτεινέα. Επιπλέον, θέλω να εκφράσω τις ευχαριστίες μου προς την R. Wolfe, και τον J. McDonald για την εξαιρετική μας επικοινωνία, την ανταλλαγή ιδεών και την παροχή των δεδομένων πάνω στα οποία έγιναν τα πειράματα.

AKNOLEDGMENTS

First of all, I would like to thank the members of my three-member committee and especially the research director at the Institute for Language and Speech Processing (ILSP) of the ATHENA RC, Eleni Efthimiou. Her guidance and support throughout all stages of my research have been invaluable. I am also appreciative of the insights and advice provided by the research director on "Speech Technologies with an emphasis on prosodic rules for text-to-speech synthesis," Stavroula-Evita Fotinea. Additionally, I would like to express my thanks to R. Wolfe and J. McDonald for our outstanding collaboration, exchange of ideas and for providing the data on which the experiments were conducted.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 Motivation

According to the World Federation of the Deaf, there are over 70 million deaf people worldwide [1], [2]. In addition about 430 million people need rehabilitation because of hearing loss [3]. This includes those who primarily use sign language to communicate.

Sign languages are visual and manual systems of communication used by individuals who are deaf and hard of hearing. They are natural languages that incorporate hand movements, facial expressions, body postures, and other non verbal elements to convey meaning [4]. Sign languages are also distinct from spoken languages and are independent and unique to each country or region. Similar to spoken languages across the globe there exist sign languages, such as Greek Sign Language (GSL), British Sign Language (BSL), American Sign Language (ASL), among others. Each sign language has its unique grammar rules, vocabulary, and cultural idiosyncrasies [5].

However, despite being a field of research with tremendous potential for significant impact Sign Language Processing (SLP) has not progressed at the same pace, as its spoken language counterpart.

Consequently individuals who use sign language often face communication obstacles. These obstacles can pose challenges in accessing information and opportunities to those individuals, fact that emphasizes the significance of research and development in the field of SLP. Advancements in this domain can significantly contribute to eliminating communication barriers and improving the quality of life for members of the sign language community.

To delve deeper into this subject, SLP refers to the field of research and technology that focuses on creating methods and systems for analyzing, understanding and generating sign language content [6]. It involves utilizing computational techniques such as computer vision, machine learning, avatar signing and natural language processing to capture interpret and generate sign language data.

In essence SLP is crucial since it facilitates effective communication while also enhancing educational resources and opportunities. Additionally it promotes employment practices while ensuring accessibility for all. Moreover it plays a role in preserving cultural identity. By leveraging the advancement of technology to bridge communication gaps SLP contributes towards building a society that is more inclusive, equitable and accessible for individuals with hearing loss.

1.2 Problem definition

Sign language, being a gestural form of communication heavily utilizes 3D space to convey information. Unlike spoken languages that heavily depend on auditory cues, sign languages rely on hand movements, body positions and facial expressions in relation to the surrounding space. Within the field of SLP various tasks such as sign language recognition, translation, representation and resource creation and maintenance play a significant role. These tasks involve understanding and processing the cues present in sign language through video or image data.

As mentioned earlier sign language representation is an essential task within the field of SLP. It requires employing approaches to capture and represent sign language in a format that allows effective analysis, processing and communication [7]. Video recordings are commonly used as a method for representing sign language. Additionally other

approaches like glosses, written notation systems or avatars are frequently utilized to represent sign language data. Each method contributes to achieving the goal of facilitating effective communication and understanding of sign language.

However it's worth noting that the aforementioned methods, for representing sign language share a common limitation; they heavily rely on human effort. These methods require the involvement of humans (annotators or linguists) who manually transcribe, annotate or analyze sign language data. This manual work is time consuming and requires expertise. Therefore there is a need for an automated approach that can address these challenges effectively.

In the context of avatar based representation techniques a successful approach would involve extracting 3D skeleton joints from videos and using them to animate the avatar replicating the corresponding movements. This is where Computer Vision (CV) comes into play. It’s a field dedicated to extracting valuable insights from visual data, including both still images and dynamic videos.

Computer vision algorithms specifically designed for pose estimation can be utilized to track and recognize hand movements identify facial expressions analyze body postures and extract spatial and temporal information from sign language videos. By leveraging these computer vision techniques, SLP systems can automatically recognize signs while interpreting the grammatical and semantic aspects of sign language. Consequently the task of sign language representation can be approached as both a computer vision problem and a human pose estimation problem effectively.

1.3 Contribution

Our thesis makes a versatile contribution in several areas. Firstly recognizing the importance and complexity of sign language representation, which typically involves labor-intensive manual processes, we propose an automated method for mapping signer skeleton keypoints to avatar motions. Specifically, we explore existing human pose estimation algorithms that can accurately extract 3D joint information from RGB videos captured with a single camera. This sets our research apart from previous works, which often relied on multiple cameras and depth data to infer 3D poses.

Secondly, incorporating such an accurate 3D estimator in sign language video processing can greatly improve recognition and translation tasks. This advancement enables efficient and accurate analysis of sign language videos. Moreover our work extends to creating sign language databases which are valuable resources for research and training in this field.

We conducted experiments using continuous sign language videos performed by the Paula avatar – an advanced system capable of dynamically generating new signed phrases, representing the current state of the art in this field. We present the results and findings obtained from these experiments. Utilizing Paula as an avatar allows for consistent evaluation and comparison across different scenarios and approaches, within sign language processing.

In general our thesis seeks to enhance sign language representation, recognition, translation and corpora creation through automated methods and thorough experimentation using Paula as the reference avatar.

1.4 Structure

In Chapter 2 we introduce the background of sign language representation methods and human pose estimation algorithms and we investigate representative, previously published works in these fields.

In Chapter 3 we present the experimental setup used for conducting the research, including details about the models utilized and how Paulas videos were selected and prepared for analysis.

In Chapter 4 we study the qualitative and quantitative results we obtain and discuss the findings in relation to our research objectives.

In Chapter 5 we summarize our findings and present our conclusions based on the analysis. We also discuss any limitations that we encountered during our study.

In Chapter 6 we suggest areas of improvement and future research directions and Identify opportunities for further development and exploration within the field.

2 BACKGROUND

This chapter starts with a brief introduction to Sign Language Processing (SLP). We explore the key tasks within SLP and place particular emphasis on sign language representation methods thoroughly analyzing their benefits and drawbacks. After that we move on to pose estimation field giving important background information and conducting a thorough review of state of the art models used in this field.

2.1 Sign language processing

SLP is a field that combines disciplines to develop methods and systems for analyzing, understanding and generating sign language. This involves utilizing techniques such as computer vision, machine learning and natural language processing. These approaches aim to capture, interpret, and generate sign language data, making communication more feasible and accessible for the deaf and hard of hearing communities.

To gain a clearer understanding of sign language processing, it's helpful to explore the tasks associated with it. SLP comprises several important tasks [22]:

Sign Language Recognition focuses on creating algorithms and models that can automatically recognize and interpret signs, from video or image data [8]. It typically involves detecting and tracking hand and body movements analyzing handshapes and facial expressions and mapping them to signs.

Sign Language Translation aims to convert sign language into spoken or written language [9] [10]. It entails comprehending the meaning of signs and generating textual or spoken translations. To accomplish this task we need to bridge the gap between sign language and spoken language by considering their different structures, meanings and cultural subtleties.

Sign language production involves converting text or spoken language into sign language [11]. This allows individuals who primarily communicate through sign language to express themselves to non-signers or create sign language content.

Sign language detection focuses on recognizing of signing activity in visual contexts [12]. Its main objective is to determine if a person is using sign language in a video by identifying specific hand and body movements or facial expressions associated with sign language communication. Sign language detection serves as an initial step in developing systems that can interpret and respond to sign language automatically.

Sign language identification goes beyond detection and aims to recognize and differentiate between different types of sign languages or variations [13]. This involves determining whether the signer is using American Sign Language (ASL) British Sign Language (BSL) or another regional sign language. The accurate identification of the sign language being used is crucial for providing precise translation and interpretation services.

Sign language segmentation refers to the process of breaking down sequences of signs into meaningful units [14]. This task requires identifying the boundaries, between signs, gestures or expressions within a signing sequence. Accurate segmentation plays a vital role in various tasks such as sign language recognition, translation, and production since it enables the system to process and analyze signs effectively.

Sign language corpora creation involves gathering, annotating and organizing datasets comprising sign language videos, annotations and linguistic information [15]. These

corpora are worthwhile resources for training and evaluating SLP systems as well as conducting linguistic research. Developing comprehensive and diverse sign language corpora is vital, for advancing the accuracy and reliability of sign language recognition and translation models. Additionally it contributes to the exploration of sign languages and their linguistic characteristics.

Sign language representation refers to the methods and systems used for capturing, conveying and interpreting sign language. It involves converting sign language expressions into a format that can be analyzed, processed and communicated through different means. This can be accomplished with the following approaches:

Sentence: "I LIKE TO DANCE"

Gloss: "I DANCE LIKE"

Figure 1: Illustration of Sentence-to-Gloss conversion.

Gloss is a technique that represents sign language using written or printed words from a language (Figure 1). It employs a system of symbols or abbreviations to indicate signs, fingerspelling and grammatical elements in sign languages. Previous works like [16] and [17] have provided guidelines for gloss annotation; however there is currently no established protocol for gloss annotation.

Videos offer a representation of sign language by capturing the movements and gestures of signers. Video recordings of sign language performances, conversations or instructional content enable comprehension of sign language communication. Videos are extensively employed in sign language learning, interpreter training and multimedia applications.

Written Notation Systems aim to capture the spatial and temporal aspects of sign languages using written symbols, diagrams or annotations. These systems provide a standardized way to transcribe and represent signs along, with their movements. Examples include Stokoe notation, HamNoSys [19], and SignWriting [20] (Figure 2).

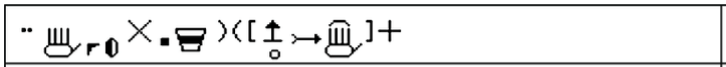
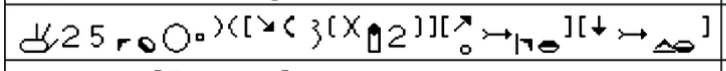
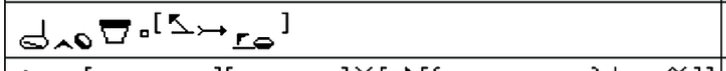
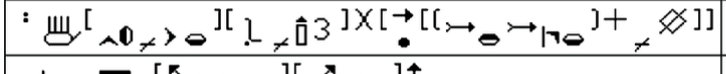
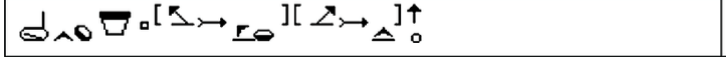
	bears
	Goldilocks
	somewhere wandering
	deep forest
	somewhere wandering

Figure 2: Example of HamNoSys notation system

3D Motion Capture is used to track the movements of signers using equipment like multiple cameras or depth sensors. This method captures positions and orientations of body parts and gestures in three dimensional space. Its applications include research on sign language recognition systems, animation and virtual reality.

Pose Estimation algorithms analyze video or image data to estimate the positions and orientations of body parts such as hands, arms and face. These algorithms use computer vision techniques like deep learning models for extracting and tracking relevant body

keypoints [23] [24] [25]. Pose estimation finds its application in a plethora of interactive sign language technologies.

Avatar or animation based representations: Another approach involves creating characters or avatars that mimic sign language movements and gestures. These representations can be generated using motion capture data, animation techniques or manually created by artists. Avatars and animations are commonly used for purposes, like sign language interpretation services and educational resources [7] [26].

Different methods of representation have their advantages and the choice of method depends on the specific goals, requirements and limitations of the task. Selecting the appropriate representation method is crucial for the accurate analysis and interpretation of sign language.

Undoubtedly SLP involves a range of technologies that underline its inherent complexity. This complexity emphasizes the importance of adopting an interdisciplinary approach to address the various challenges within this field [27]. Therefore by bringing together experts from disciplines such as deaf culture, linguistics, computer vision, NLP, machine translation, computer graphics and human computer interaction, we can gain comprehensive insights and develop practical solutions that tackle the complex challenges in SLP. Through this effort across disciplines we can consider the needs of the sign language community and result in effective technologies and services that promote inclusivity and accessibility. Ultimately embracing a such approach allows us to make meaningful advancements with tangible real world impact. This benefits sign language users, by fostering accessibility and promoting communication equality.

2.2 Human pose estimation

Human pose estimation (HPE) is defined as the positioning of human joints, like shoulders, elbows, wrists, etc, in images or videos. HPE is used in a plethora of real-world applications across different areas, such as healthcare, sports analytics, robotics and gaming. It comprises techniques used to detect and track humans and recognize their actions [28]. Although it has received much attention in the computer vision community for decades, it still remains a challenging task. Intricate body postures, occlusions, changes in lighting conditions and clothing are some of the issues that reveal the complexity of HPE task .

Moreover, HPE can be categorized into two main categories based on the output space:

- 2D HPE: Involves estimation of the two-dimensional (2D) coordinates (x, y) of keypoints in an image or video.
- 3D HPE: Aims to estimate the three-dimensional (3D) coordinates (x, y, z) of joints or landmarks in an image or video, i.e. it also provides depth information.

2.2.1 Human body modeling, datasets and evaluation metrics

Before presenting the most important works in this area, we provide information on the modeling of the human body. In general, due to the complex structure of the human body, different models were adopted by the HPE methods [29][DK1]. However, the two most frequently employed models are the skeleton and shape models. Additionally, in a recent study [30][DK2], a surface-based representation, called DensePose, was proposed.

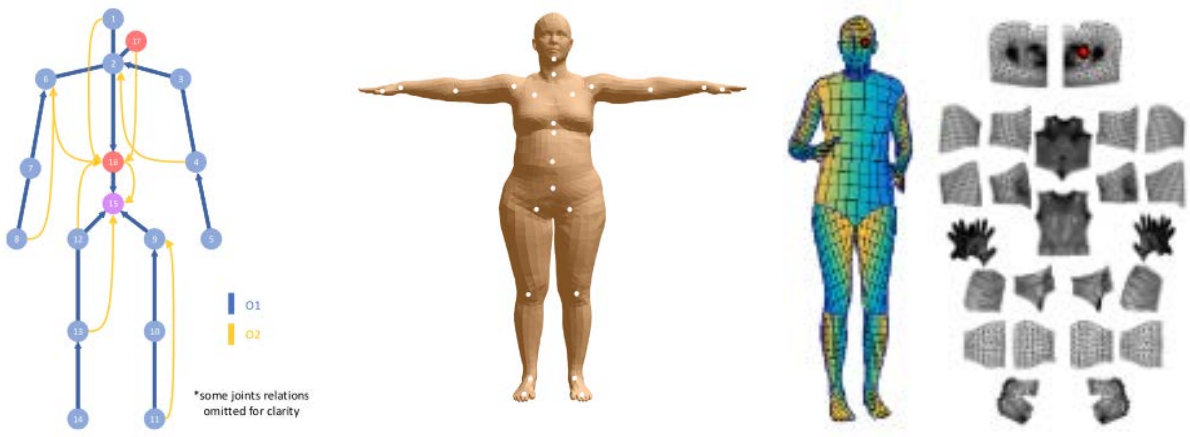


Figure 3: Human Body modeling - Skeleton (left), Shape (center) and Surface (right) models.

In skeleton-based model, the human body is treated as a tree structure [31][DK3] which consists of many keypoints and edges that connect the natural adjacent joints as illustrated in Figure 3-left. Regarding the shape model, researchers have adopted the skinned multi-person linear (SMPL) model [32][DK4], as depicted in Figure 3-center. In this model, the human skin is represented as a triangulated mesh containing 6890 vertices which is parameterized by both shape and pose parameters. On the other hand, DensePose was created to represent the human body in a denser structure since the sparse correspondence of the image and keypoints might not suffice to accurately capture the configuration of the human body (see Figure 3-right).

Consequently, a new dataset called DensePose-COCO has been developed, which demonstrates the dense correspondences between image pixels and a surface-based representation of the human body. This dataset consists of 50K properly annotated images of COCO (Common Objects in Context) dataset [33], a widely used dataset for various tasks including Keypoint Detection. COCO comprises more than 200,000 images and 250,000 person instances labeled with keypoints. It employes the skeleton-based model with 17 keypoints. In the 3d HPE field, Human3.6M [34][DK5] is one of the most extensive motion capture datasets which consists of 3.6 million human poses, each accompanied by corresponding image. Similar to COCO, it adopts the skeleton model comprising 17 keypoints. However, it's worth noting that the configuration of these keypoints differ slightly from those in the COCO dataset as illustrated in Figure 4.

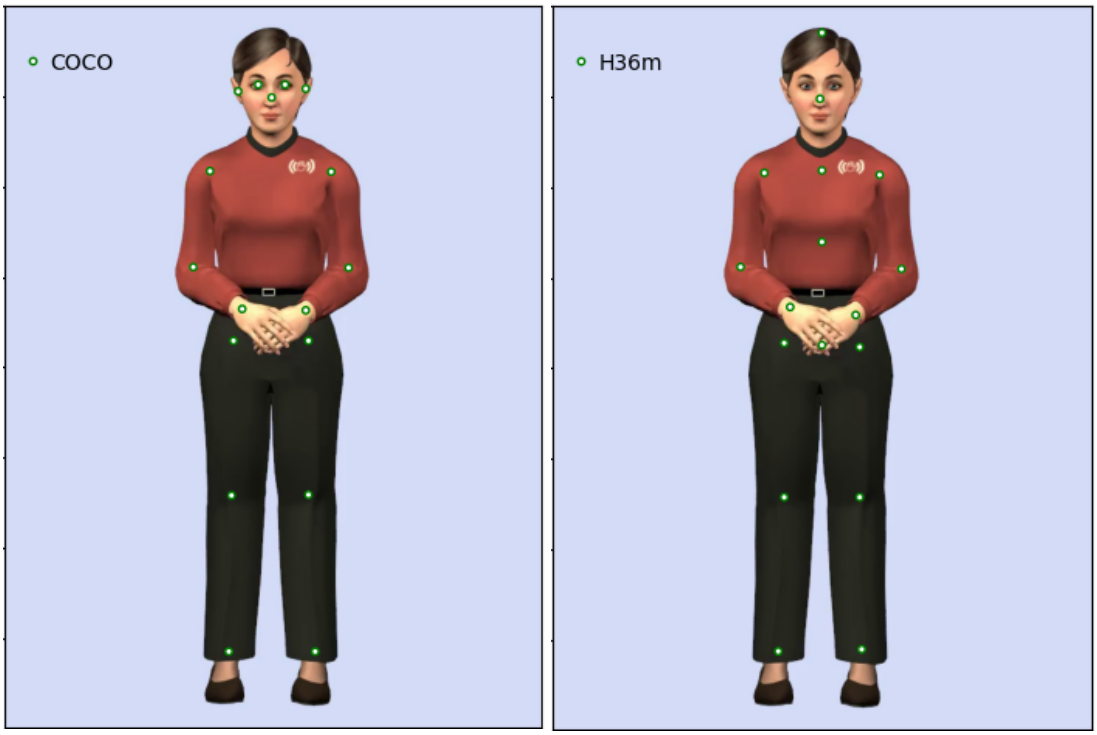


Figure 4: Keypoints in Coco (left) and H36m format (right).

Furthermore, a frequently used metric in 2D HPE is the Percentage of Correct Keypoints (PCK) [35][DK6] which measures the percentage of correctly detected keypoints compared to the ground truth. An estimated keypoint is regarded correct if its distance from the corresponding ground truth falls below a predefined threshold.

$$PCK = \frac{1}{N}\sum_{i=1}^{N}\begin{cases}1, & if\ \dfrac{||p_i - G_i||}{head\ or\ torso\ length} \leq T \\ 0, & otherwise\end{cases}$$

Where:

- N is the total number of keypoints.
- p_i is the predicted 2D position of the i-th keypoint.
- G_i is the ground truth 2D position of the i-th keypoint.
- T is the defined distance threshold.

Another evaluation metric which is commonly used in 3D HPE is the Mean Per Joint Position Error (MPJPE). It measures the average Euclidean distance between corresponding predicted joints and the ground truth.

$$MPJPE = \frac{1}{N}\sum_{i=1}^{N} ||p_i - G_i||$$

Where:

- N is the total number of joints.
- p_i is the predicted 3D position of the i-th joint.
- G_i is the ground truth 3D position of the i-th joint.

2.2.2 Human Pose Estimation Methods

After the rise of deep learning in recent years, the field of HPE has also undergone earth-shaking changes. However, before deep learning, other traditional approaches were being used to face that problem. Specifically, in the 2D HPE concept, a classical approach is the Pictorial Structures [DK7](PS) model [36]. In this framework, the basic idea is to represent an object like human body as a collection of its parts. The parts are not considered rigidly fixed in place and they can move or deform relative to each other. Therefore, a body transformation is treated as a set of local part deformations. Human structure is represented as a graph and each node corresponds to a part. In Figure 5, springs show the spatial relations between limbs and an appearance model is used for each part. The model tries to find the arrangement of parts and connections that best matches the human body in the image.

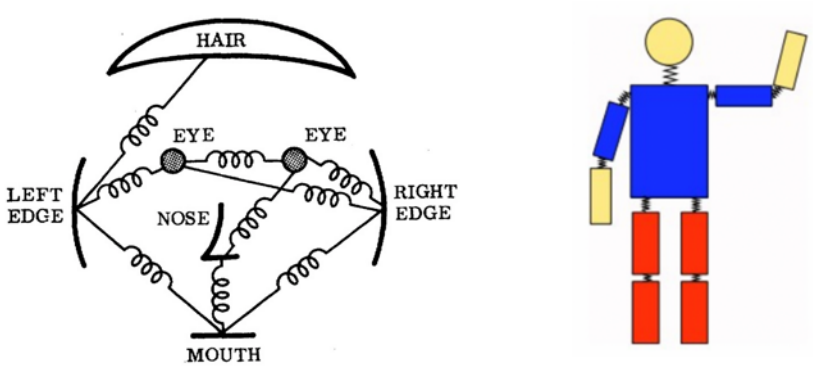


Figure 5: Pictorial Structures Model connects rigid body parts together through the use of springs to create a tree-like structure of the entire body

However, the optimization method used, depends on initial solutions and doesn't guarantee finding the global optimal solution. Felzenszwalb et al. utilized the probability statistical model to overcome this limitation [37]. Nevertheless, this approach was struggling to capture the connections between occluded parts. To solve these issues, Andriluka [38] proposed a generic approach based on the PS model.

Another well-known classical approach is the Deformation Part model (DPM) which was introduced by Felzenszwalb [39]. In this framework, the human structure is represented as a star one which involves a root filter, part detectors and a part deformation model. [40][DK8](see Figure 6). The limitations of this model are due to the fact that it focuses on modeling the spatial relationships between body parts without explicitly taking into account changes caused by rotation, scale or size variations. To address these problems, Yang and Ramanan use a mixture model of parts to represent complex object structures, known as the Flexible Mixtures-of-Parts model (FMP) [41]. [DK9] FMP builds upon the

ideas of DPM incorporating features which increase the versatility of how parts can be arranged handle variations in rotation and scale.

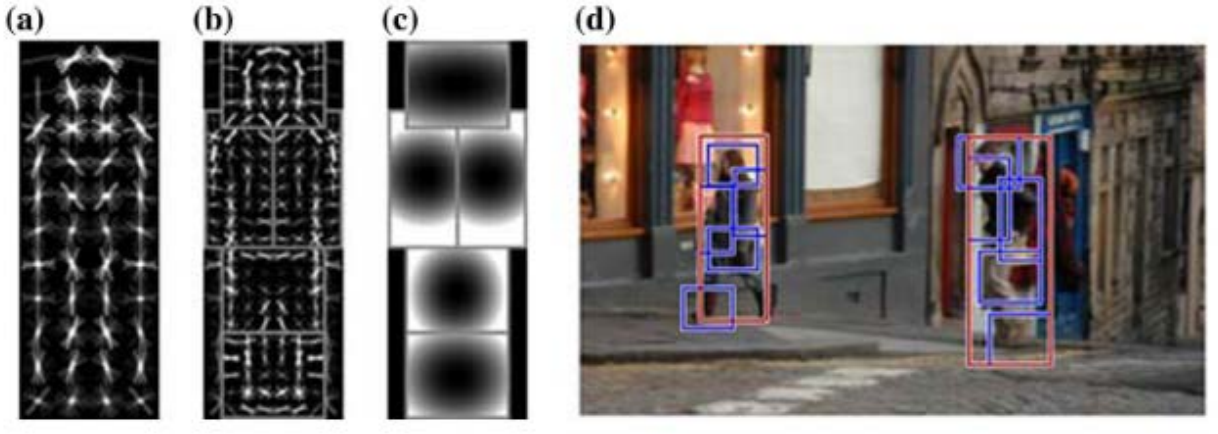


Figure 6: The template of the DPM and the Detection Result - (a) the Root filter; (b) the part Detection; (c) the Deformation model; (d) the Sample of the Detection and pose estimation.

In 2014, Toshev proposed DeepPose, the first major method for HPE based on Deep Neural Networks (DNNs) [42][DK10]. In this work, pose estimation was considered as a DNN-based regression problem. Specifically, DeepPose leverages the power of convolutional neural networks (CNNs) to achieve accurate posture estimation even if several joints are not directly visible in the images since CNNs inherently have the ability to reason about poses in a holistic manner. Moreover, a significant innovation of DeepPose is its progressive refinement of pose estimation, as illustrated in Figure 7.

Figure 7: Left: General view of the DeepPose architecture. Right: The refinement stage where a regressor is applied on a cropped image to refine the prediction from the previous stage.

A different approach which implements heatmap regression, is introduced by Tompson [43][DK11]. This method, instead of indicating directly the body joints as in the previous work, estimates the probability of a keypoint occurring in each pixel of the image and its output is a heatmap showing these probabilities. As shown in Figure 8, the framework [DK12]comprises the coarse heat-map model for coarse localization, the component for extracting and cropping convolutional features at defined (x, y) positions for each joint and an extra convolutional model for fine-tuning.

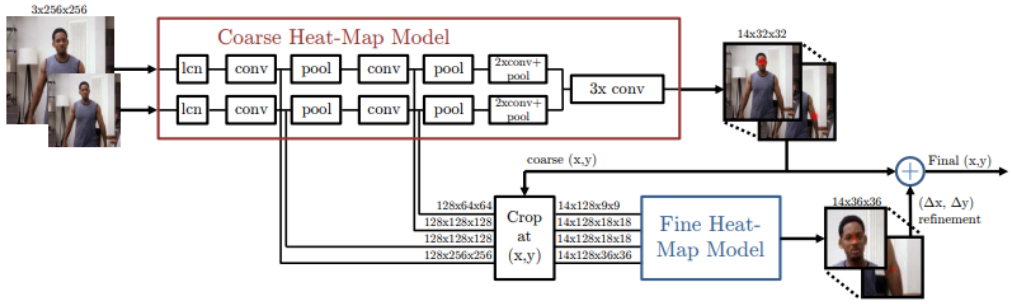


Figure 8: Overview of cascaded architecture

In 2016, Wei et al. harnessing the power of Pose Machines, introduced a novel framework called Convolutional Pose Machines (CPMs) [44][DK13]. A Pose machine comprise the image feature computation stage followed by a sequence of prediction stages (Figure 9a and 7b). CPMs is end-to-end framework which integrates convolutional networks into the Pose Machine model, capturing long-range dependencies between image and multi-part cues. The stacked convolutional networks as depicted (Figure 9), operate on belief maps from previous stages, progressively improving the precision of part location predictions.

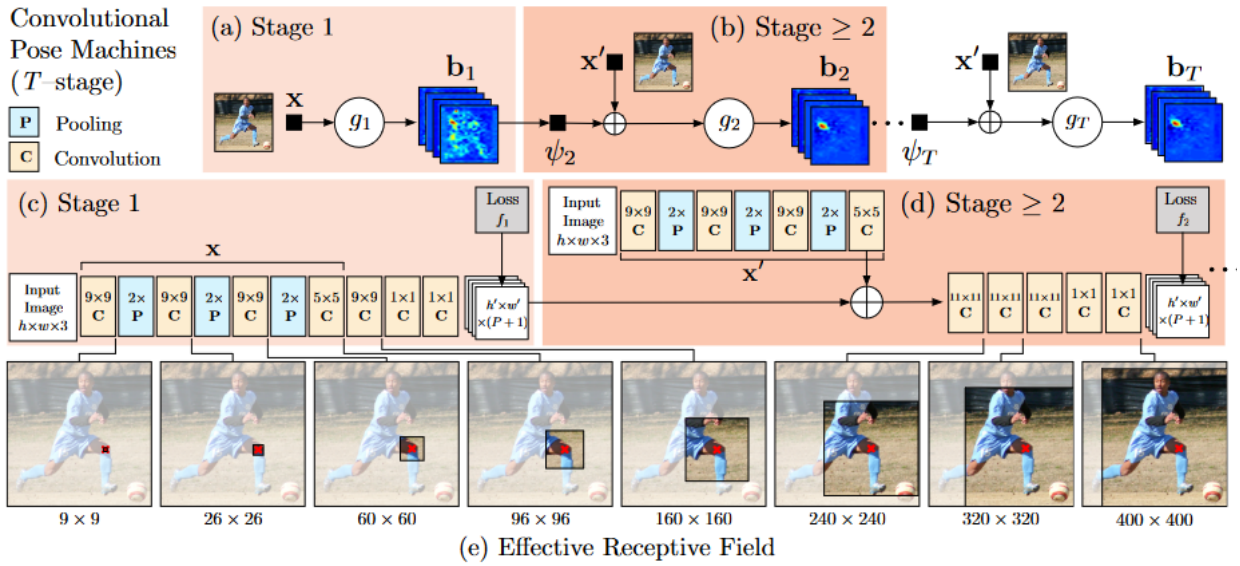


Figure 9: Overview of CPMs architecture

In contrast with the aforementioned works which handle the single person pose estimation, OpenPose is a real time approach which achieves pose estimation of multiple people in an image [31]. It utilizes a nonparametric representation, known as Part Affinity Fields (PAFs), to learn to connect body parts with individuals in the image. PAFs are a set of 2d vectors that model the relationships between different body limbs, indicating the orientation and the strength of this affinity.

As illustrated in Figure 10, OpenPose starts by taking the entire image as input. This image, which can contain more than one person, is processed using a CNN to create confidence maps for body parts detection. In addition to confidence maps, CNNs estimate a set of PAFs that denotes the level of association between parts. In the next step, bipartite graphs are performed between the associated parts of the body. Depending on the PAF values, weaker connections in bipartite graphs are removed and finally, by assembling them into whole-body poses, the skeleton of each person in the image is constructed.

OpenPose is the first bottom-up real-time multi-person framework to simultaneously detect human body, hand, foot, and facial landmarks.

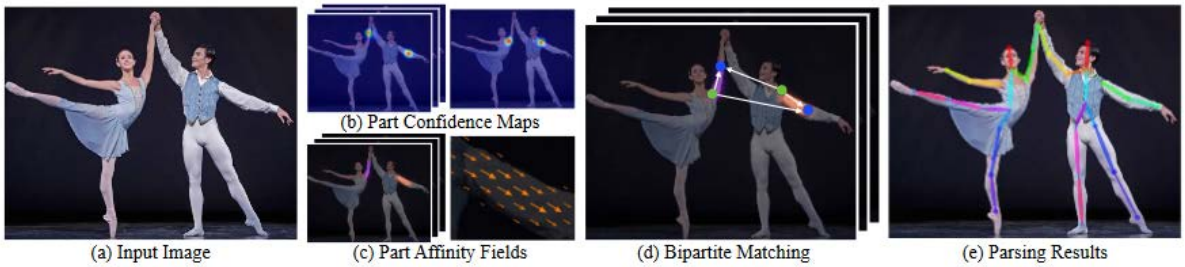


Figure 10: OpenPose pipeline.

In 2017, He et al introduced a conceptually simple and flexible framework for object instance segmentation, called Mask R-CNN [45]. It extends the well-known Faster R-CNN, designed primarily for object detection (Figure 11-left), by adding a third distinct branch that outputs a binary mask for each Region of Interest (RoI). These pixel-wise segmentation masks indicate which pixels belong to the object and which do not and encode the semantic information about its spatial structure.

Moreover, Mask R-CNN can be extended to HPE. The main differences of the Mask R-CNN are the output size and the way of encoding keypoints within the keypoint mask. As illustrated in Figure 11 (right), Mask R-CNN predicts 17 (one for each keypoint) one-hot 56 x 56 binary masks where only one pixel is labeled as foreground. This extension of Mask R-CNN is known as Keypoint R-CNN and is included in Facebook's AI library, Detectron2.

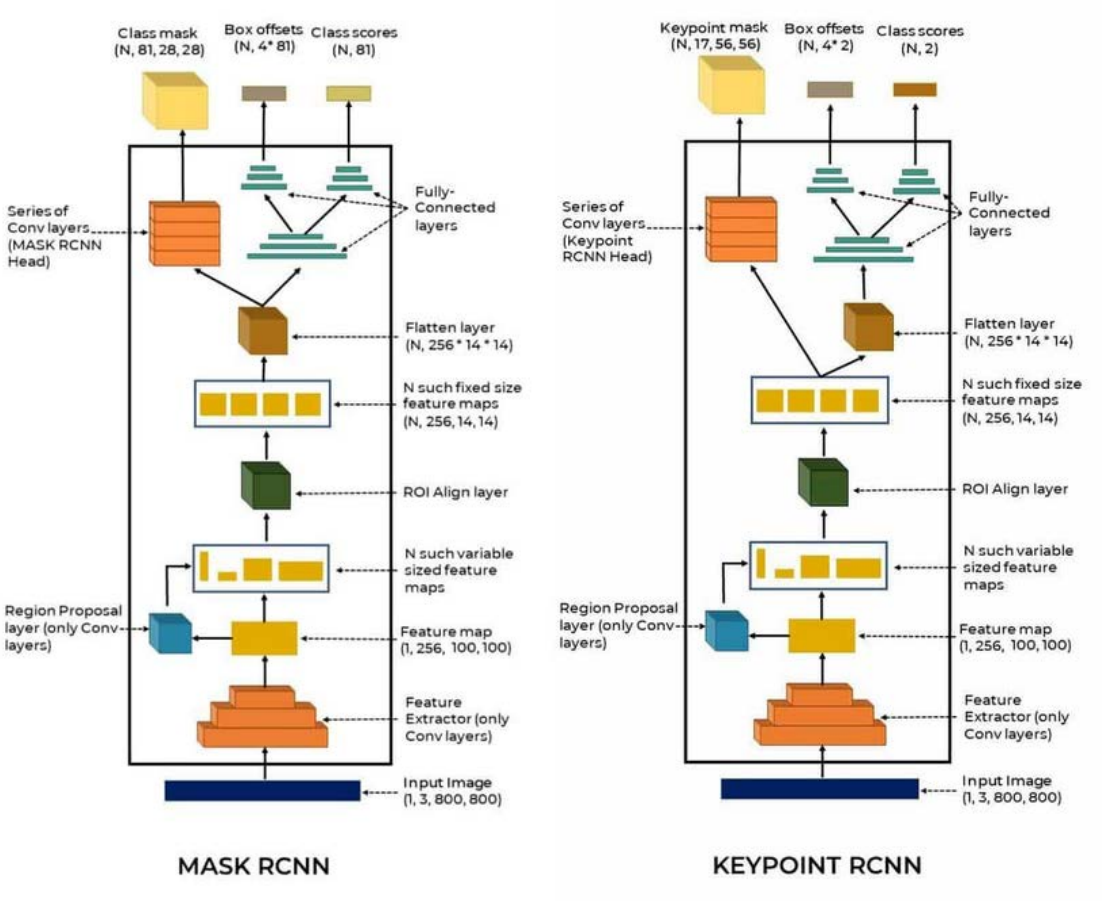


Figure 11: Architectures of Mask R-CNN (left) and Keypoint R-CNN (right).

In contrast with 2D HPE, reconstructing a 3D human pose from a monocular image have to overcome a fundamental challenge that different 3D poses can correspond to the same 2D image. Motivated by the rapid advancement of 2D HPE algorithms, many studies have tried to leverage the promising 2d HPE results for 3D HPE. Indeed, Martinez et al. [46] introduced a simple yet highly effective baseline for lifting 2d poses into 3d space. As illustrated in Figure 12, the model takes as input the 2d coordinates of keypoints, parses them through a deep, multilayer neural network and outputs the corresponding 3d coordinates.

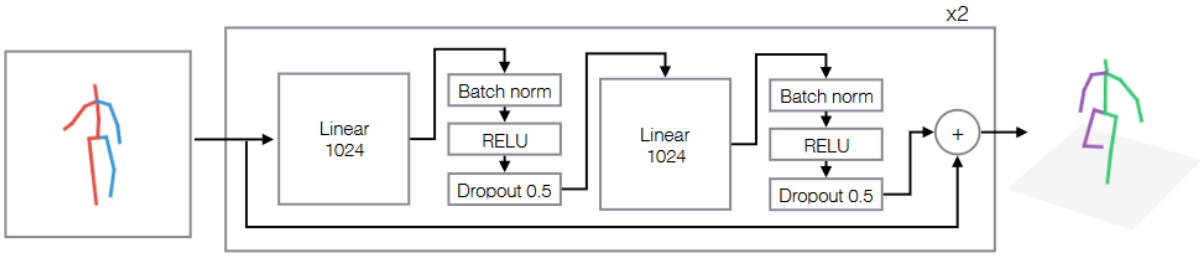


Figure 12: Overview of network architecture.

On the other hand, in the context of 3D HPE from a sequence of monocular images, the exploitation of temporal information is an efficient technique for reducing the inherent depth ambiguity. Therefore, Pavllo et al. [47] proposed a fully convolutional model to learn long-term information. Particularly, VideoPose3D, as it's called, employs dilated temporal convolutions over 2d keypoints to predict 3d positions Figure 13.

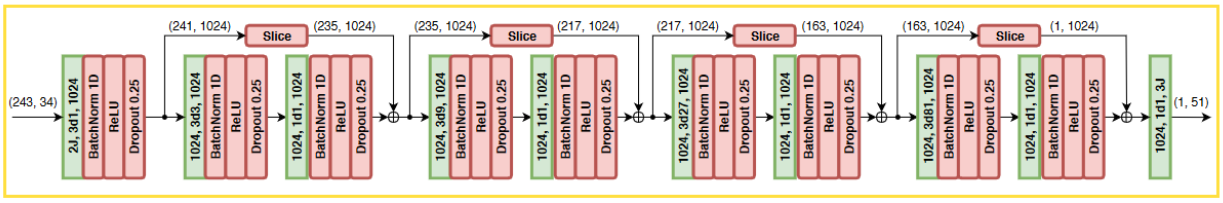


Figure 13: VideoPose3D architecture

Another interesting work which is designed to estimate both 2D and 3D human pose from a single monocular (2D) image or video, is the BlazePose [48]. As depicted in Figure 14, it comprises two models: the pose detector and the pose tracker. Specifically, the detector is used to identify the ROI where the human is located and afterwards the tracker predicts the coordinates of 33 keypoints. In video cases, the detector is applied only to the first frame since for subsequent frames, the ROI were derived from the previous frames. BlazePose employs heatmaps and regression techniques, as shown in Figure 15, to predict the 2D keypoints and then extends this information to estimate the 3D pose. To estimate the full 3D body pose in images or videos, BlazePose uses GHUM, a 3D human shape modeling pipeline [49].

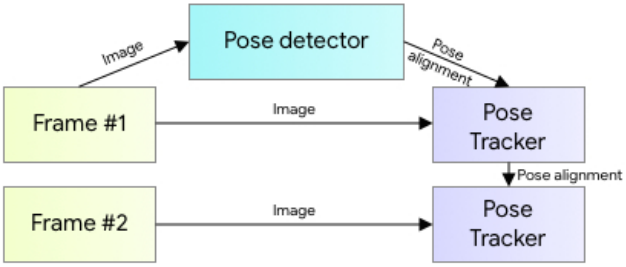


Figure 14: Inference pipeline of BlazePose.

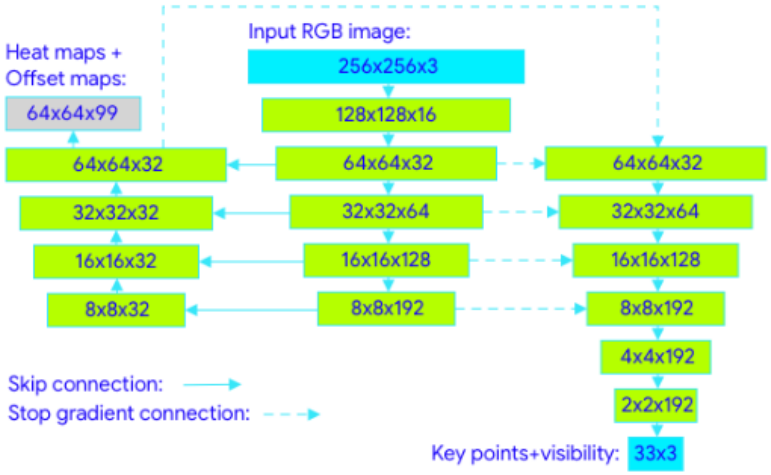


Figure 15: BlazePose architecture

3 Experiments

In this chapter, we discuss our experiments conducted on Continuous Sign Language (CSL). Due to limited computational resources and the unavailability of considerable datasets it is not practical to train models from scratch. Instead we have chosen to focus on evaluating existing state of the art models. By utilizing pre trained models we can assess their performance and capabilities on specific sign language videos. This approach enables us to make comparisons and gain valuable insights. While training models from scratch may be preferable evaluating existing models provides meaningful information without requiring extensive training efforts.

3.1 Data

This approach allows us to maximize the use of resources while still gaining valuable insights into the effectiveness of sign language processing models. As a result we conducted experiments using a synthetic dataset consisting of five videos featuring an avatar named Paula specifically selected for evaluation purposes.

Paula developed at DePaul University[1] [50] and enhanced within the EASIER project by ATHENA/ILSP, is a computer-based sign language avatar initially designed for teaching sign language to hearing adults. This avatar takes string of glosses of sign languages including ASL (American), LSF (French), DGS (German), DSGS (Swiss-German), and GSL (Greek) sign languages, then applies morphological adjustments determines appropriate phonemes and timing and combines these elements to create a 3D animation featuring the avatar. Over time significant efforts have been made to improve the realism and expressiveness of Paula.

Several notable advancements have been made in this regard, including refining eyebrow movements in order to achieve a more natural appearance [51] enhancing animation smoothness while avoiding robotic motions [52] and enabling simultaneity [53]. Additionally adaptations have been made to make Paula compatible, with sign language notation systems like Azee [54] further enhancing mouth animations [55] [56]. The latest advancements include incorporating layered facial textures and makeup [57]. These ongoing developments constantly enhance the authenticity and adaptability of the Paula avatar in sign language communication.

In our research videos Paula demonstrates GSL by showcasing a range of complex signs, including phonological signs with occlusions. Although the videos capture the whole body of Paula, only her upper body is in motion while her lower body remains still. We chose to use full body videos because the evaluated models were trained on data that encompasses the full body.

These videos are accompanied by ground truth data which provides precise world coordinates (x, y, z) for 14 primary skeleton keypoints in each frame. The arrangement of these keypoints can be seen in Figure 16. The ground truth data used in our experiments was preciously collected by a team at DePaul University ensuring accuracy and reliability of the keypoints position. Additionally detailed information about camera setup and characteristics such as intrinsic camera parameters has been provided.

Although the dataset is not extensive enough it serves as a controlled environment for evaluating the effectiveness and abilities of the tested models. The use of this dataset

[1] http://asl.cs.depaul.edu

allows us to make meaningful comparisons and draw introductory conclusions. These experiments conducted on Paulas videos provide valuable insights as a starting point, which can pave the way for further advancements, in sign language processing.

By using this synthetic dataset, we aim to make meaningful comparisons and draw initial conclusions that can guide further research in sign language representation and recognition. As a starting point, these experiments on Paula's videos offer valuable insights that can lay the foundation for future work and potential advancements in sign language processing.

Table 1: Overview of Evaluated Approaches

Approach	2D Detector	2D keypoints format (input)	3D Reconstruction	3D keypoints format (output)
OpenPose	OpenPose	H36m	Videopose3D (pretrained_h36m_cpn.bin)	H36m
Detectron	Keypoint-RCNN	COCO	Videopose3D (pretrained_h36m_Detectron_coco.bin)	H36m
MpCoco	BlazePose	COCO	Videopose3D (pretrained_h36m_Detectron_coco.bin)	H36m
Mp3D	BlazePose	COCO	BlazePose (GNUM)	H3.6m

3.2 Evaluation: Design and Methods

As detailed in Table 1, we constructed and evaluated four approaches by combining state-of-the-art pretrained models in HPE. All approaches in our experiments involve a two-step process; initially estimate the 2D skeleton in image space taking as input an rgb video and then reconstruct it in 3D space utilizing the predicted 2D keypoints from the previous step.

To ensure a proper configuration and smooth integration into a unified pipeline we need to apply certain adjustments since each model has its own input and output setup. These modifications were necessary in order to facilitate the evaluation and comparison of four approaches in our experiments.

Specifically the first approach employs the OpenPose framework to estimate the 2D coordinates of keypoints in the H3.6m format. Unfortunately OpenPose does not provide predictions for the Midhip and Spine keypoints. To address this incompatibility issue with the h36m format we artificially generated these keypoints by leveraging information from adjacent ones. We define the Midhip as the midpoint between the right and left hip while the Spine as the midpoint between the Thorax and Midhip keypoints. In 3D reconstruction step, a pretrained model from Videopose3D framework was applied to predict the coordinates of the keypoints in 3D space in the same format (H3.6m).

The next two approaches are quite similar. Both approaches share the input and output format (COCO) at both stages and utilize the same pretrained model for 3D reconstruction. The main difference between these two approaches lies in their 2D detectors. Specifically the "Detectron" approach uses the Keypoint R-CNN model from the Detectron2 library while the "MpCoco" approach utilize the BlazePose from Mediapipe framework.

The last approach, called “Mp3D” relies exclusively on the mediapipe framework in both the first and second stages. In this approach, for consistent configuration of 3D output (H3.6m), apart from Midhip and Spine, Thorax and Headtop are artificially generated by using information from neighboring keypoints. It is important to note that our analysis primarily focuses on arm trajectory and therefore these extra keypoints do not affect our results.

Indeed, studying the trajectory of arm joints in sign language is crucial for advancing sign language technologies. Analyzing movements of arm joints like shoulders, elbows and wrists plays an important role in improving sign language recognition systems and developing natural and expressive representation since they convey meaningful information. Given the importance of arm trajectory analysis, we have decided to narrow down our research scope to analyze primarily the movements of the arms. This allows us to delve deeper into this aspect. However, we acknowledge that a thorough understanding of sign language demands the comprehensive study of facial expressions and finger movements as well and we opt to leave it for future work, since it deserve a separate dedicated research.

In addition we have opted to employ a right hand coordinate system where the X axis represents the width, pointing towards the left, the Y axis represents the depth pointing towards the camera while the Z represents the height pointing upwards (Figure 16).

Figure 16: The upper body keypoints are indicated by the pink dots and serve as the reference points for the Ground Truth data. Additionally, the white segments emphasize the keypoints associated with the arms.

4 RESULTS

In this chapter, we delve into our comprehensive analysis. As previously indicated, our analysis is centered around arm movements, with particular emphasis on the six keypoints: R/L Shoulder, R/L Elbow, and R/L Wrist. For the sake of brevity, we provide commentary and visually insightful diagrams for the most notable cases where error peaks are observed while the remaining results can be located in Appendix.

Figure 17: Screenshot captured from the videos of the H36M dataset. The arrow points to the Thorax keypoint.

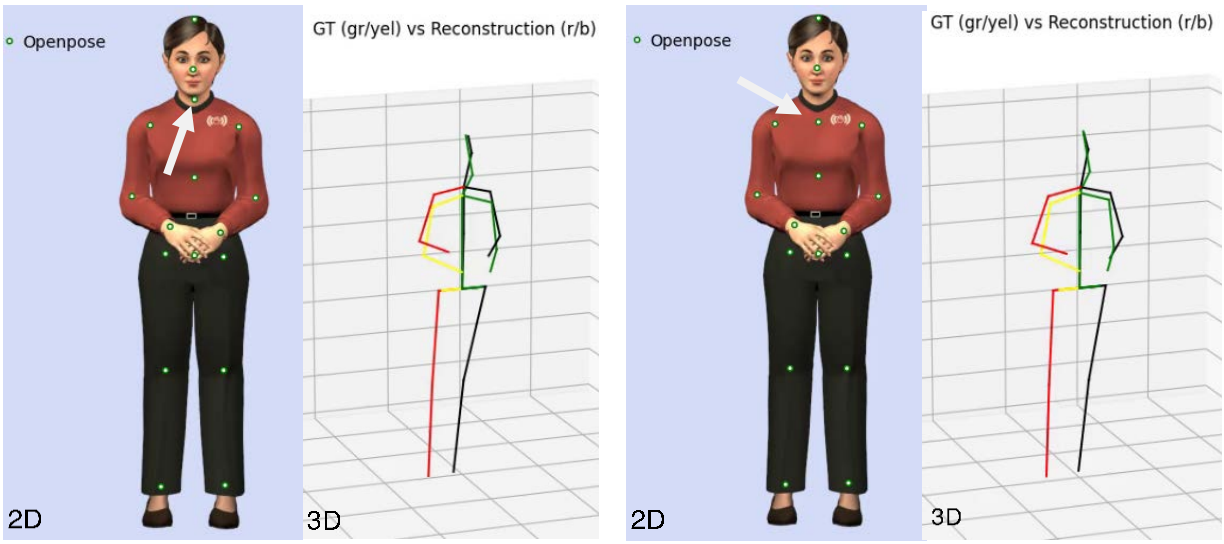


Figure 18: OpenPose Approach - Before (left) and after (right) 2D Thorax keypoint correction. In the 3D plots, the real skeleton is illustrated in green/yellow, while the predicted one is shown in black/red.

Prior to presenting the results, it is essential to acknowledge some adjustments in order to ensure the accordance of skeleton keypoints [2]. In terms of 2D detections, particularly within the OpenPose approach, there exists a notable divergence in the predicted Thorax keypoint. It is positioned above the corresponding keypoint used during training (Figure 17). As demonstrated in Figure 18, this disparity leads to the visual outcome of the predicted 3D skeleton appearing elevated. Consequently, rectifying this by lowering the

[2] Quantitative results without applying arrangements can be found in Appendix.

Thorax keypoint to the level of the shoulders, akin to the training data, yields a substantial reduction in error of approximately 8-10 mm within 3D space.

In the realm of 3D reconstruction, especially within the approaches that employ the COCO skeleton in their 2D detection stage (such as Detectron, MpCoco, Mp3D), a consistent error pattern is observable across all analyzed videos. To be specific, the predicted 3D skeleton consistently appears lower than the actual skeleton.

To address this systematic error, a potential solution involves raising the mid hip joint by approximately 6cm. This is due to the fact that the predicted 3D skeleton is dependent on the position of the mid hip joint. This heuristic adjustment would effectively elevate the entire skeleton, resulting in a reduction of the MPJPE by approximately 35-40 mm (Figure 19. By aligning the predictions more accurately with the real-world skeleton, this corrective action can significantly enhance the overall accuracy of the model's 3D reconstructions.

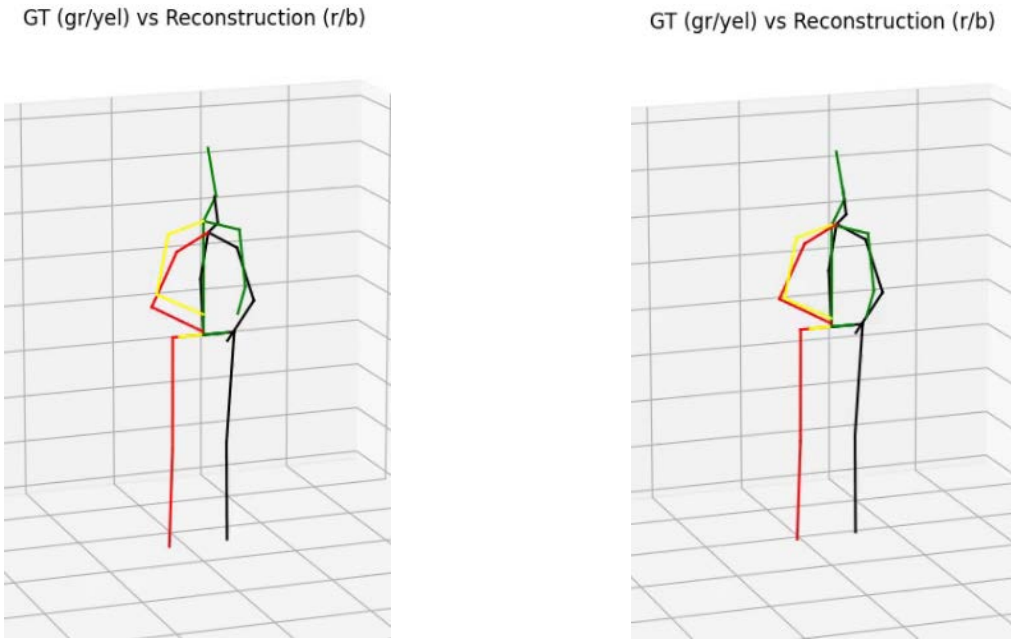


Figure 19: MpCoco approach - Before (left) and after (right) 3D midhip joint elevation.

Considering the insights gained from the above observations and implementing the necessary corrections, we proceed with our analysis.

4.1 Video1

Starting with the error analysis of the video1, it is evident that the OpenPose, Detectron, and MpCoco approaches follow a similar pattern throughout the entire video. The Figure 20 clearly illustrates two prominent peaks in errors, occurring specifically on the 66th and 104th frames. Notably, the Mp3D approach deviates from this trend, as it additionally displays relatively elevated errors on intermediate frames.

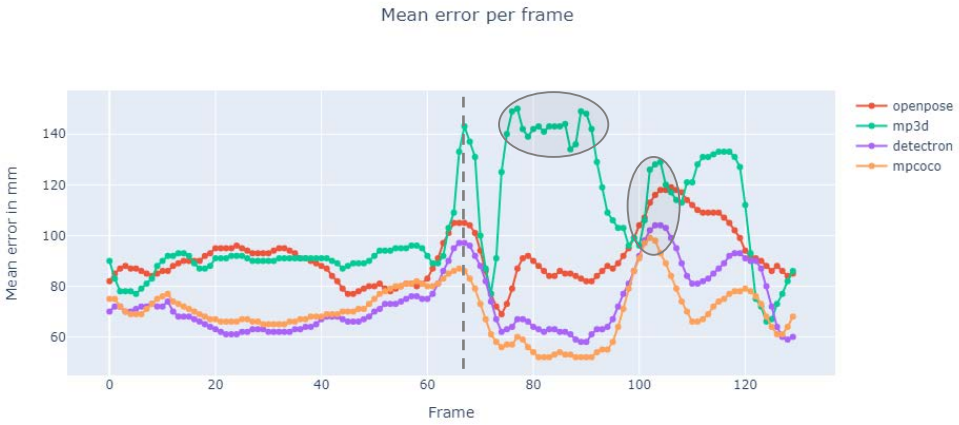


Figure 20: The average error of the six examined keypoints for different approaches across the sequence of frames.

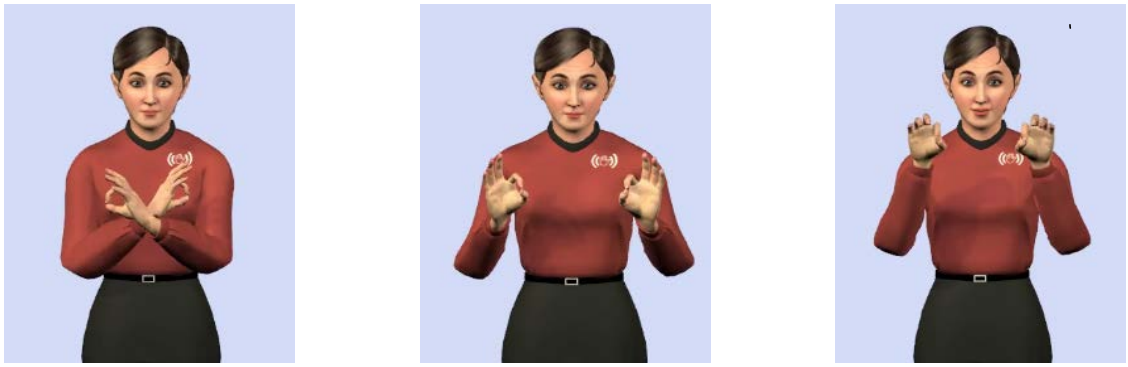


Figure 21: The 66th (left), 78th (center) and 104th frame (right) of the Video1.

In particular, on the 66th frame when Paula is crossing her wrists (Figure 21-left), the notably elevated average error can be attributed primarily to the inaccurate prediction of the left wrist and, to a lesser extent, the left elbow along the y-axis (Figure 22). The disparity between the predicted and actual skeleton is clearly evident in Figure 24.

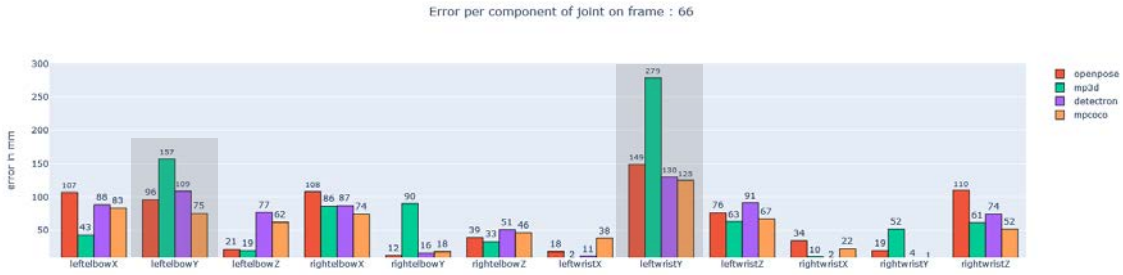


Figure 22: Joint Component Errors on 66th frame for various approaches.

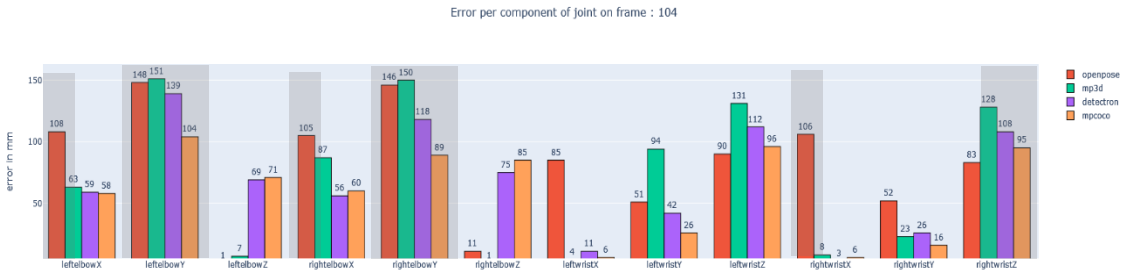


Figure 23: Joint Component Errors on 104th frame for various approaches

Furthermore, on the 104th frame, when Paula raises her arms to shoulder level (Figure 21-right), all approaches struggle to accurately predict the left and right elbow positions along the y-axis (Figure 23). Additionally, a minor error is observable in the z-component of her wrists across most approaches, except for OpenPose. However, OpenPose's predictions for the x-direction of both elbows and wrist are notably inaccurate. Figure 25 visually demonstrates that the predicted skeleton's arms appear to be spread wider apart than in reality.

Regarding observed errors of Mp3D approach on 75th-90th frames corresponding to Paula's movement as illustrated in figure (Figure 21-center), it is struggling to estimate the y coordinate of both elbows and wrists (Figure 26 & Figure 27).

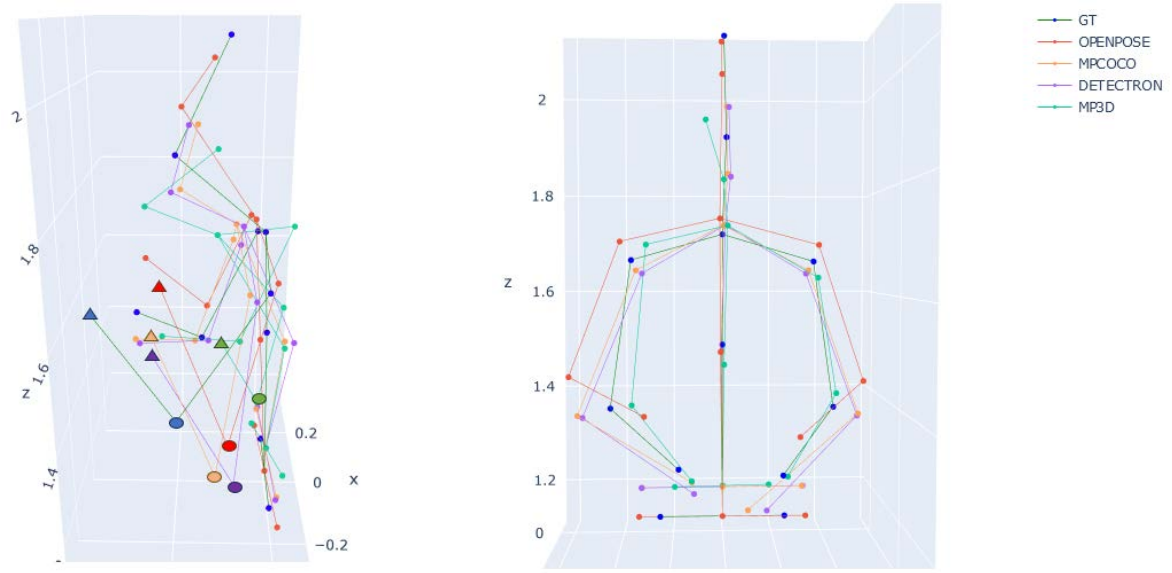


Figure 24: Comparison of approaches with Ground Truth on 66th frame. Left View (left): The left elbow is illustrated with circle and the left wrist with triangle - Front View (right).

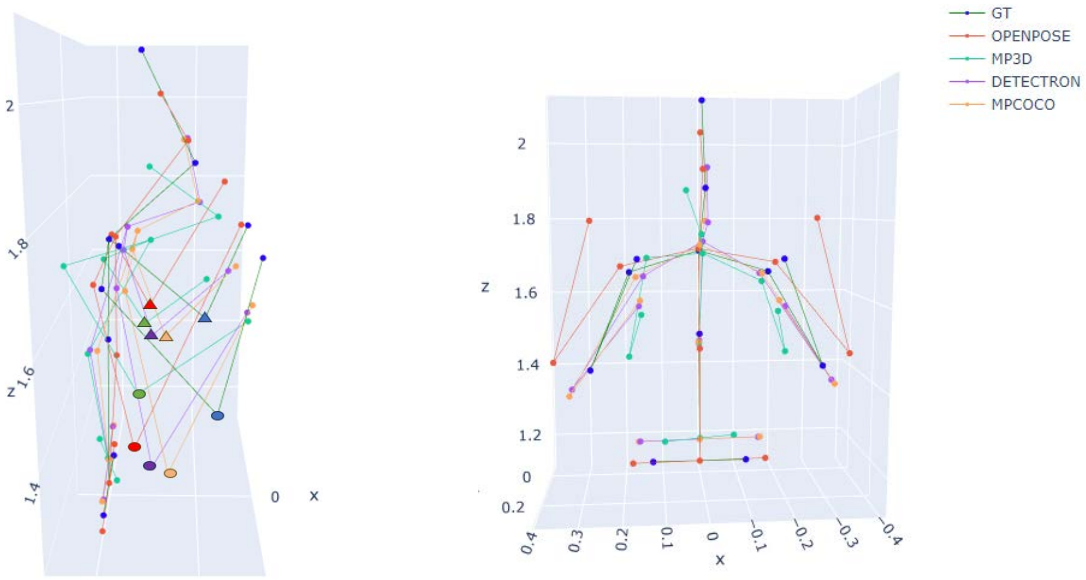


Figure 25: Comparison of approaches with Ground Truth on 104th frame. Right View (left): Right Elbow (circle) and Left Elbow (triangle) - Front View (right).

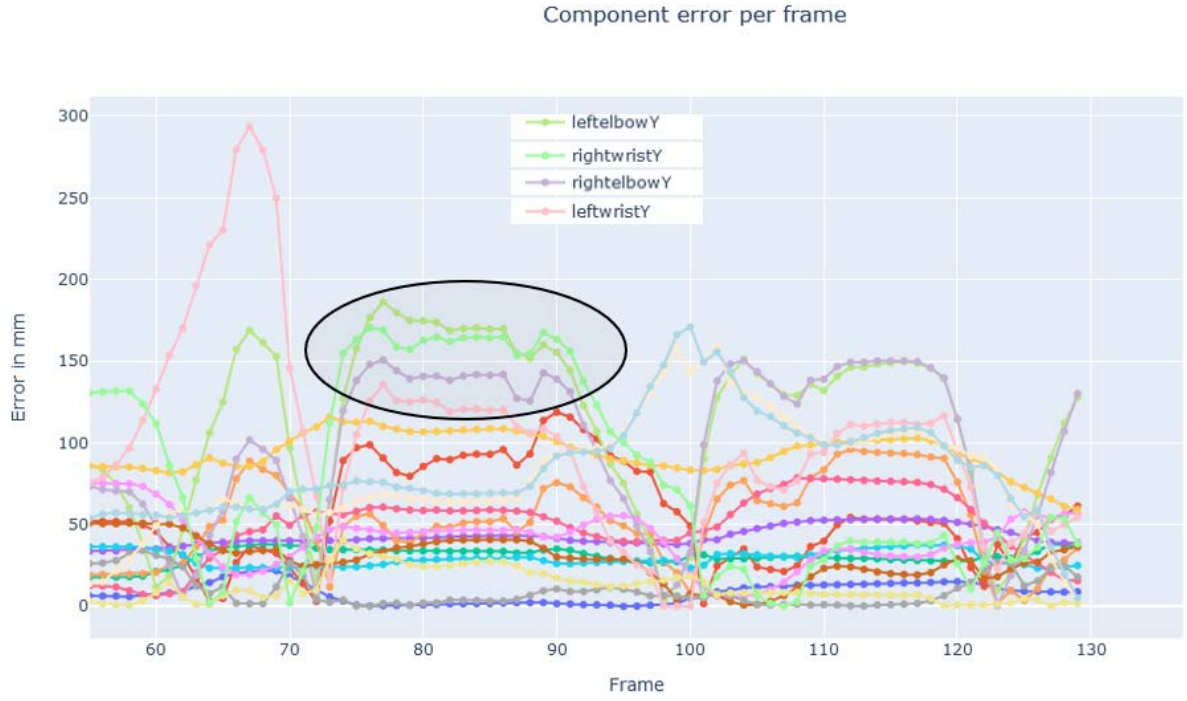


Figure 26: Joint component error of Mp3D approach across the sequence of frames.

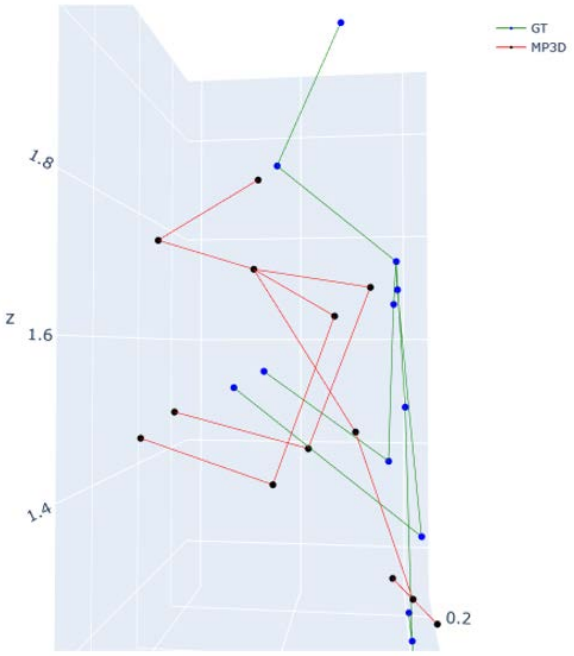


Figure 27: Comparison of Mp3D approach with Ground Truth on 78th frame.

4.2 Video2

In the context of video2, upon observing the Figure 28, it becomes evident that there are two distinct segments within the videos where most of our approaches exhibit a substantial error of around 100 mm. To be precise, from 10th to the 30th frame when Paula points towards the camera with her right index finger (see Figure 29-left), a pronounced error surfaces in predicting the position of the right wrist. With regards to the approaches, except for OpenPose, this error manifests mainly in the z-coordinate (Figure 30). However, for OpenPose, the error is observed in the y-coordinate of the right wrist (Figure 31). Moreover, in the case of OpenPose, the error in the y-component of the right elbow also contributes to the overall mean error.

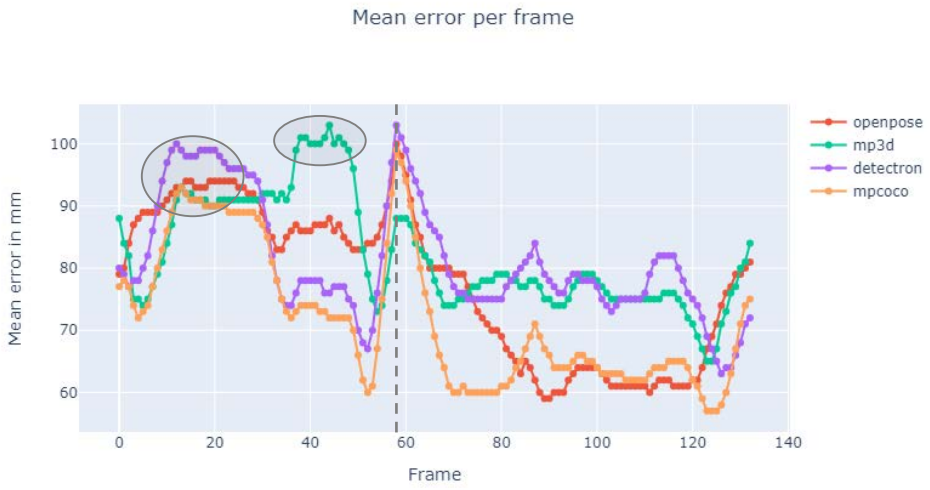


Figure 28: The average error of the six examined keypoints for different approaches across the sequence of frames.

Figure 29: The 12th (left), 44th (center) and 58th frame (right) of the Video2.

During the 58th frame, as Paula moves both her hands, only the Mp3D approach achieves a low error score (Figure 28). In contrast, the other approaches exhibit a high error rate, primarily attributed to the inaccurate prediction of the right wrist in the y dimension (Figure 33). Furthermore, OpenPose encounters challenges in accurately reconstructing the right elbow, as evidenced by elevated errors in both the x and y dimensions.

It is worth noting the elevated error observed in the Mp3D approach during the interval of the 38th to 48th frame when Paula lowers her right hand (Figure 29-center). During this sequence, the Mp3D approach struggles to accurately predict the position of the right wrist in both the y and z dimensions (Figure 32).

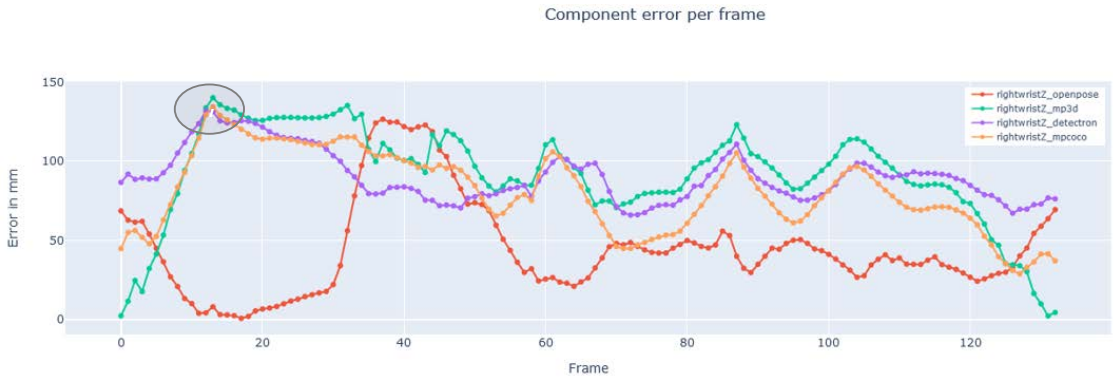


Figure 30: Right wrist z-component error for different approaches across the sequence of frames.

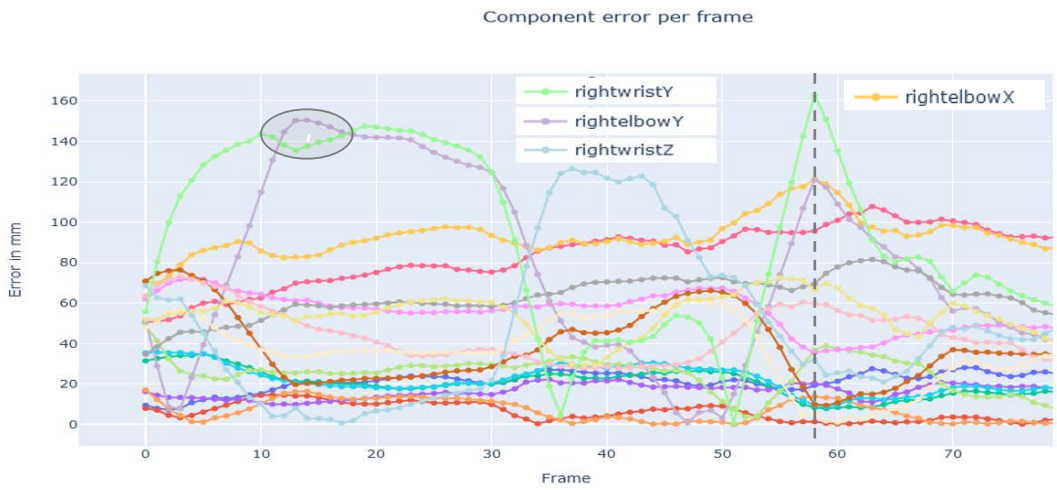


Figure 31: Joint components error for OpenPose across the sequence of frames.

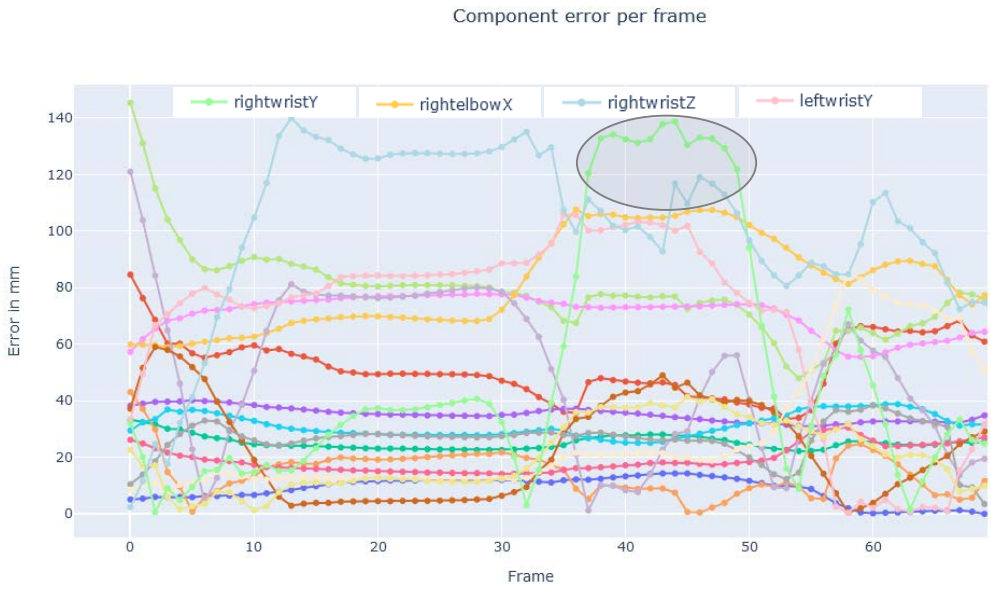


Figure 32: Joint components error for Mp3D across the sequence of frames.

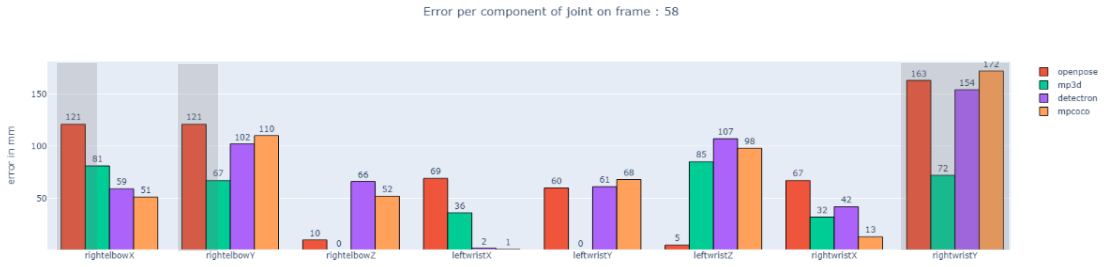


Figure 33: Joint Component Errors on 58th frame for various approaches

4.3 Video3

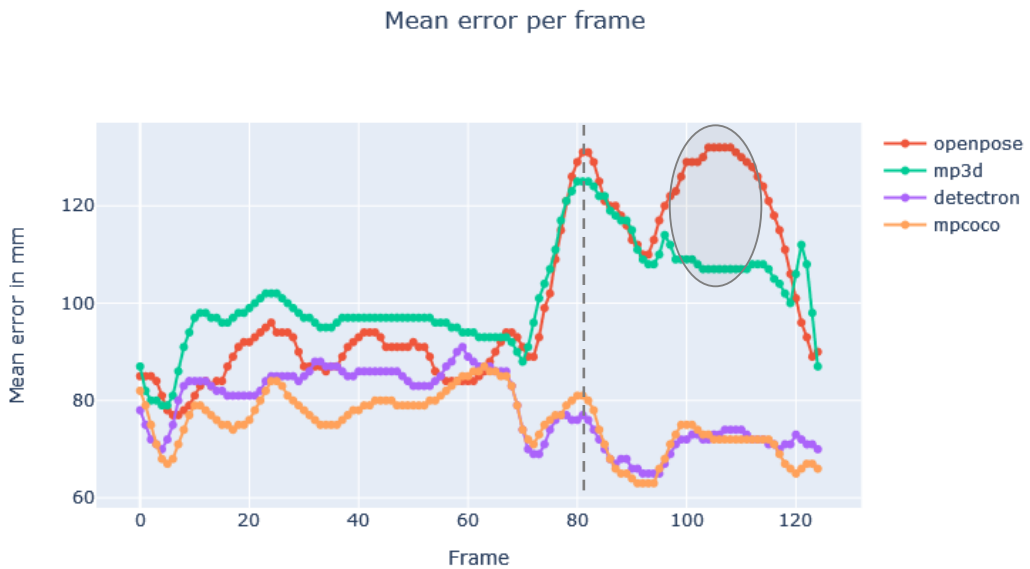


Figure 34: The average error of the six examined keypoints for different approaches across the sequence of frames.

The case of video3 presents an intriguing pattern. As illustrated in Figure 34, beyond the 70th frame, there is a noticeable increase in error for the OpenPose and Mp3D approaches, while the Detectron and MPCOCO approaches display a decrease in error. Particularly, Figure 35-left illustrates that within this timeframe, Paula raises her right hand to the level of her eyes. In the case of OpenPose, the error originates from the inaccurate prediction of the z-coordinate of the right wrist (Figure 37). This significant error could be attributed to the fact that the nose is reconstructed much higher than its actual position in reality, as shown in Figure 38.

Conversely, in the Mp3D approach, the elevated error is mainly due to the incorrect prediction of the y-coordinate of the right elbow (Figure 36, Figure 38). Furthermore, Figure 36 reveals a notable increase in error in the y-dimension of the left wrist, particularly on the last frames. This is intriguing considering the left wrist remains relatively stationary throughout the entire video.

Figure 35: The 80th (left) and 145th frame (right) of the Video4.

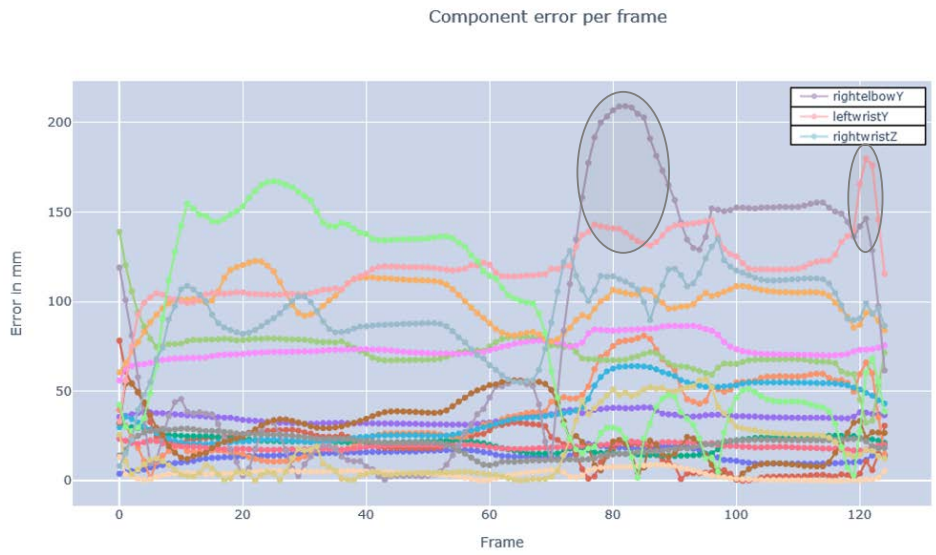


Figure 36: Joint components error for Mp3D across the sequence of frames.

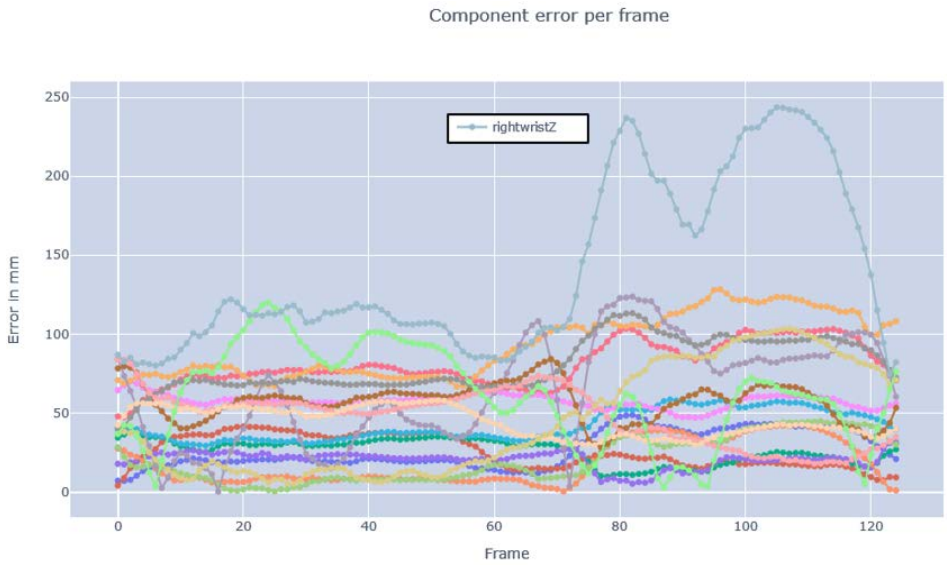


Figure 37: Joint components error for OpenPose across the sequence of frames.

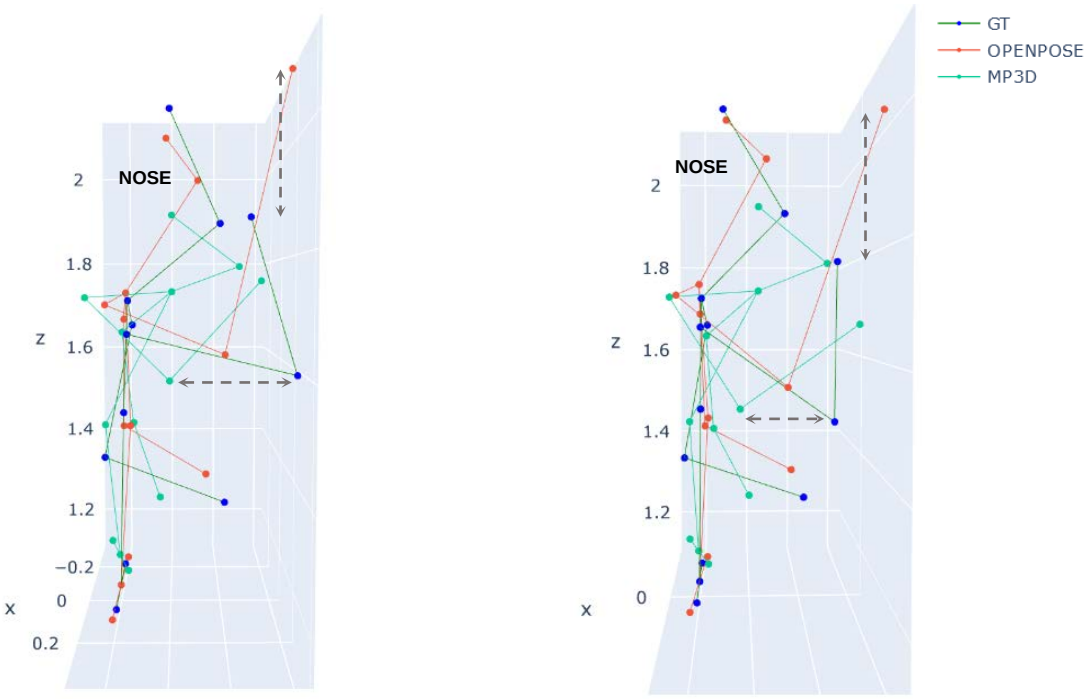


Figure 38: Comparison of Mp3D and OpenPose approaches with Ground Truth on 82nd (left) and 106th frame (right) .

4.4 Video4

In this video, a peak in the total error is evident on the 80th frame (Figure 39). During this frame, Paula positions her right wrist in front of her thorax (Figure 40-left) and notably, none of the four approaches accurately predicted the z-component of the right wrist (Figure 41). Additionally, the Mp3D approach encountered even more significant difficulties in estimating the y-dimension location of the right elbow and left wrist. Similarly, both the MpCoco and Detectron approaches displayed suboptimal predictions in the z-direction for the left wrist and right elbow joints as well. Furthermore, OpenPose reconstructed the elbows wider than their actual position, leading to errors in the x-dimension (Figure 43).

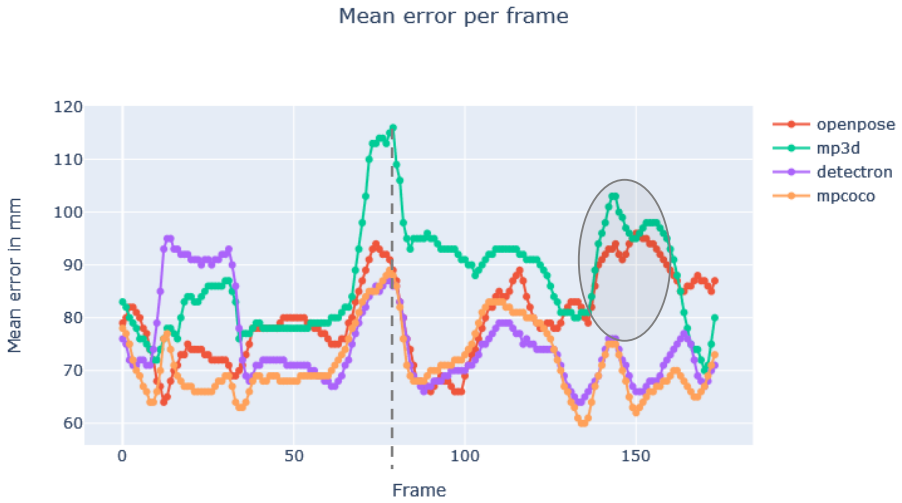


Figure 39: The average error of the six examined keypoints for different approaches across the sequence of frames.

Figure 40 : The 80th (left) and 145th frame (right) of the Video4.

For OpenPose, this pattern of errors is observed on the 145th frame as well (Figure 44). Moreover, the Mp3D approach exhibits a substantial error on this frame. As evident in Figure 44, the predicted positions of the right elbow and left wrist are noticeably closer to the body. For the former, the error is apparent in the x-dimension, while for the latter, it relates to the y-dimension (Figure 42).

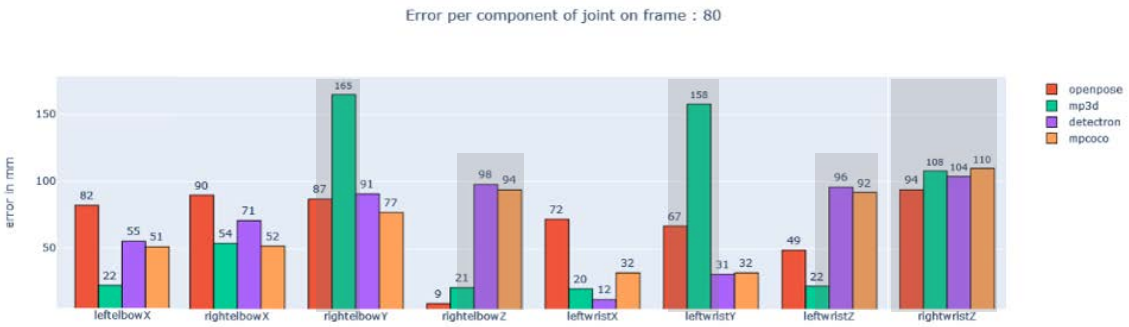


Figure 41: Joint Component Errors on 80th frame for various approaches

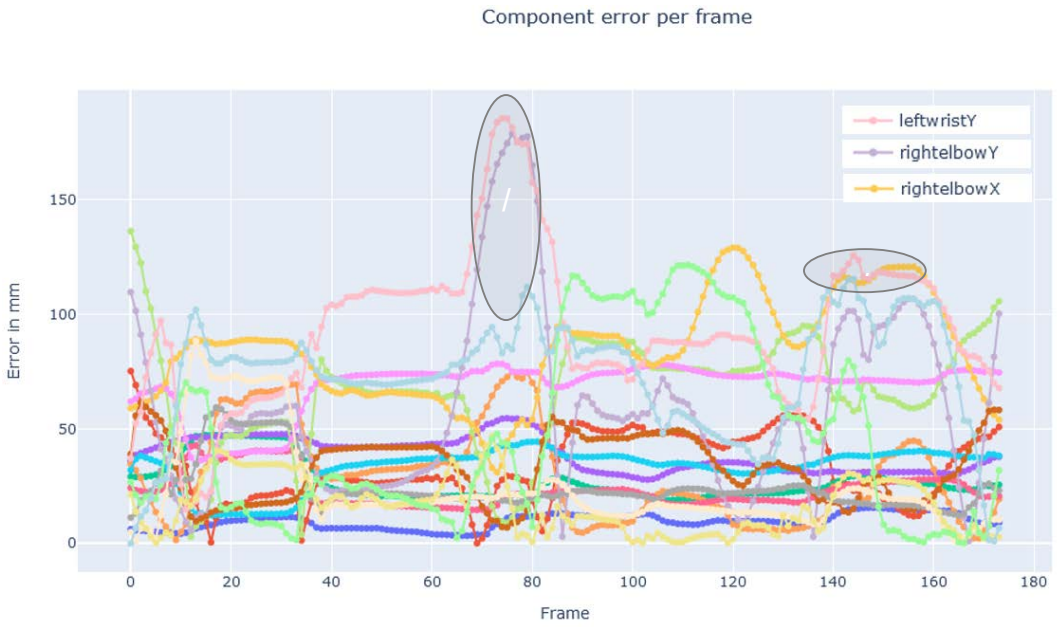


Figure 42: Joint components error for Mp3D across the sequence of frames.

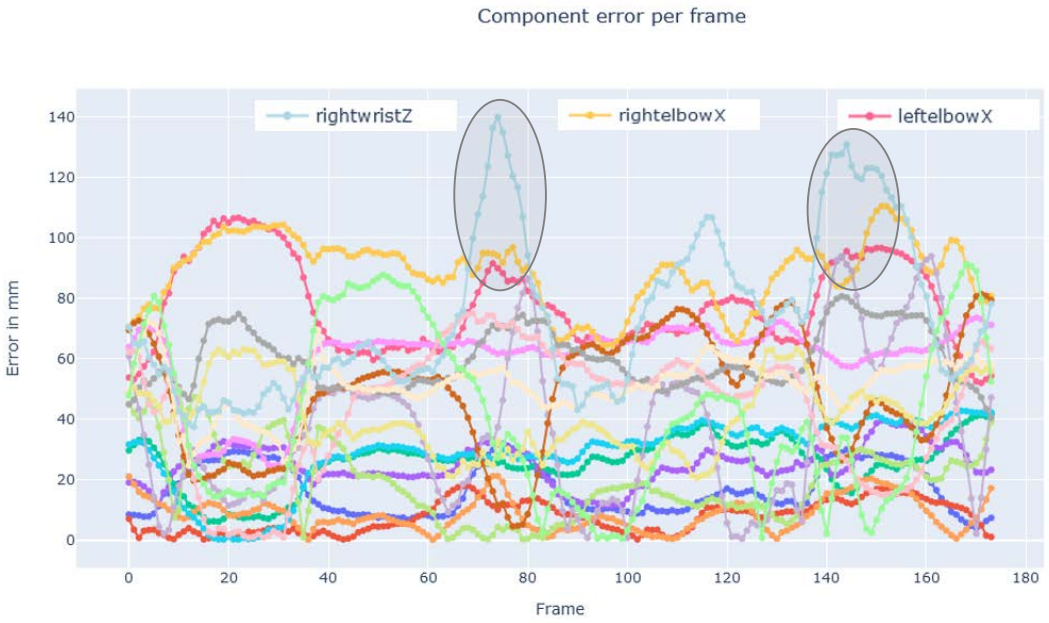


Figure 43: Joint components error for OpenPose across the sequence of frames

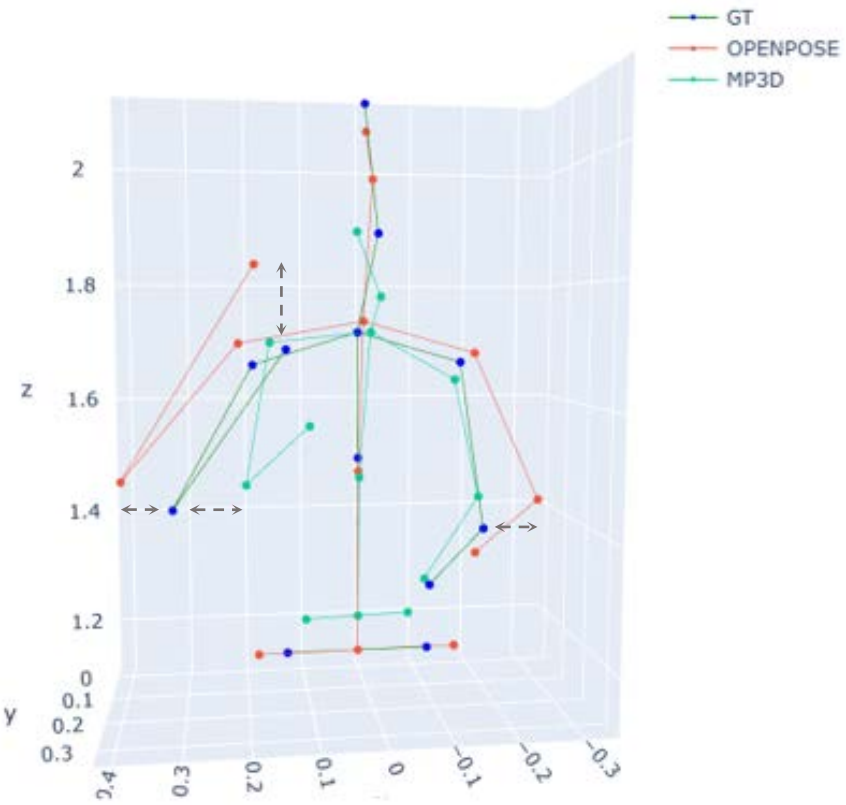


Figure 44:Comparison of Mp3D and OpenPose approaches with Ground Truth on 145th frame.

4.5 Video5

Figure 45: The average error of the six examined keypoints for different approaches across the sequence of frames.

In this video, it's important to highlight two specific intervals during which both OpenPose and Mp3D approaches display prominent peaks in error analysis (Figure 45). Firstly, the initial interval between the 15th and 30th frames, during which Paula raises her right arm to the level of her thorax (Figure 46-left), results in a significant error in the prediction of the right wrist's y-component for both approaches (Figure 47 & Figure 48).

Figure 46: The 24th (left), 40th (center) and 184th frame (right) of the Video5.

Moving further into the video, when Paula elevates her right hand above her right shoulder, as illustrated in the Figure 45-right, both approaches encounter difficulties in accurately estimating the y-coordinate of the right elbow. Notably, OpenPose exhibits suboptimal predictions for the z-coordinate of the right wrist, while Mp3D struggles with the y-coordinate of the left wrist. These discrepancies for both approaches are visually represented in Figure 50.

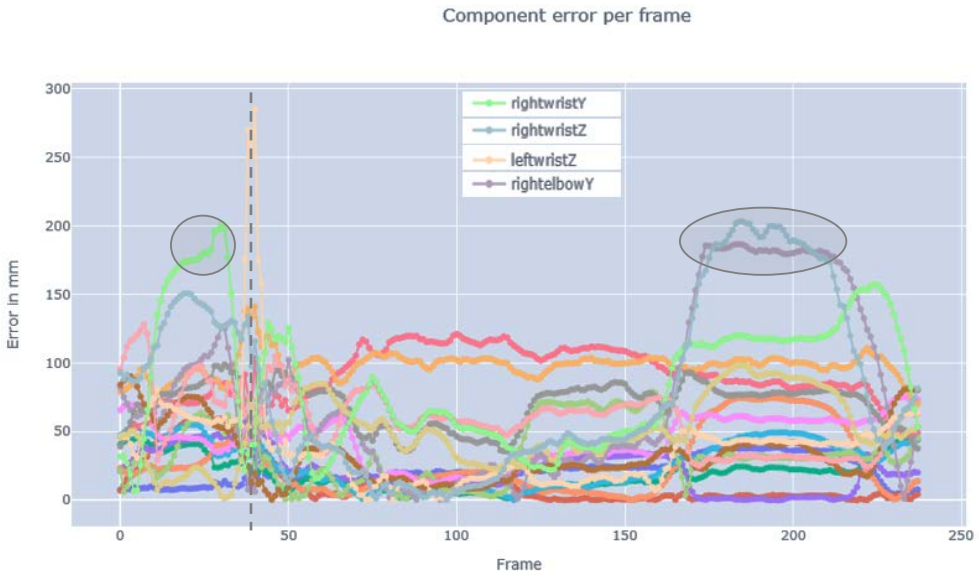


Figure 47: Joint components error for OpenPose across the sequence of frames.

Figure 48: Joint components error for Mp3D across the sequence of frames.

Additionally, there is another notable peak in error for the OpenPose approach on the 40th frame when Paula points to her left side Figure 46-center). This error is primarily attributed to inaccuracies in the z-coordinate prediction of her left wrist (Figure 47). A closer examination reveals that this error originates from inaccurate estimations during the 2D detection stage. As depicted in Figure 49, OpenPose exhibits difficulties in accurately detecting the left wrist for several frames.

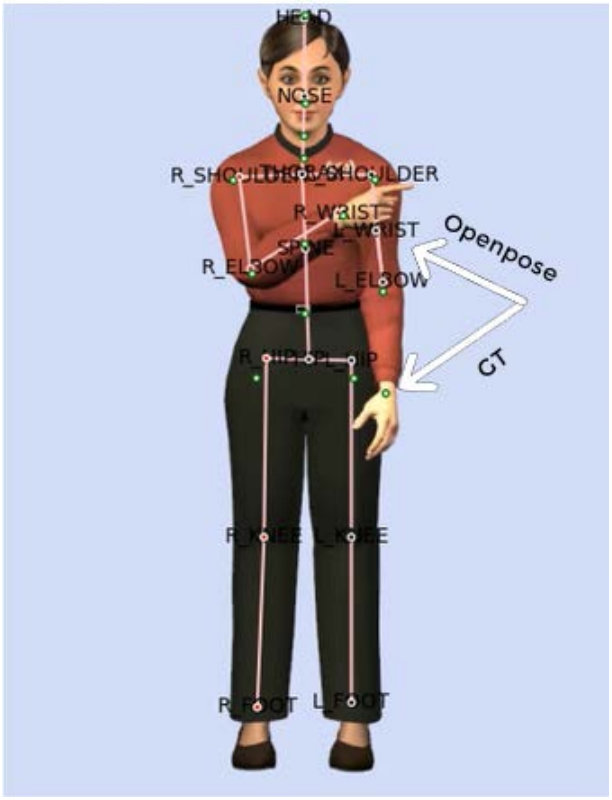


Figure 49: Comparison of OpenPose 2D Detections with Ground Truth. Notably, a significant error concerning the left wrist is highlighted.

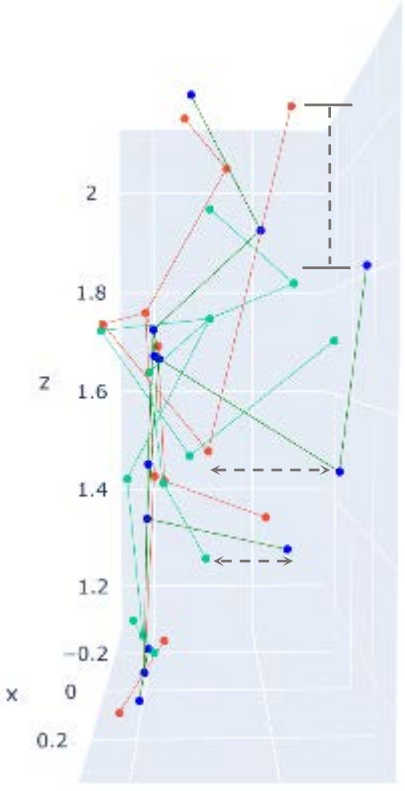


Figure 50: Comparison of Mp3D and OpenPose approaches with Ground Truth on 184th frame.

4.6 Aggregated analysis

After providing a detailed qualitative analysis for each video individually and discussing the relevant diagrams, we proceed to present the comprehensive quantitative results in both Table 2 and Table 3.

Table 2: MPJPE for each approach across all videos. The columns labeled x, y, and z represent the average error observed along each respective axis. All values are in millimeters (mm).

Videos	EUDs1				EUDs2				EUDs3				EUDs4				EUDs5			
Approaches	MPJPE	X	Y	Z	MPJPE	X	Y	Z	MPJPE	X	Y	Z	MPJPE	X	Y	Z	MPJPE	X	Y	Z
OpenPose	91	59	**37**	45	78	53	**33**	**33**	101	57	**39**	61	80	51	**25**	47	96	58	48	40
Detectron	73	30	38	44	81	33	35	56	78	32	43	43	75	33	32	50	83	33	47	49
MpCoco	**70**	**29**	42	**35**	**72**	**31**	38	41	**75**	**31**	46	**35**	**72**	33	38	**38**	**72**	32	**45**	**35**
Mp3D	104	34	73	48	82	**31**	42	47	101	34	69	42	87	**32**	54	43	95	**30**	62	45

Table 3: Aggregate errors averaged across all videos.

Approach	MPJPE	X	Y	Z
OpenPose	89.2	55.6	**36.4**	45.2
Detectron	78	32.2	39	48.4
MpCoco	**72.2**	**31.2**	41.8	**36.8**
Mp3D	93.8	32.2	60	45

We initiate the discussion with the MpCoco approach, which demonstrates superior performance compared to its counterparts, achieving the lowest MPJPE on every video. The aggregated MPJPE across all videos is noted to be 72.2 mm. Analyzing the dimensions separately, it is evident that MpCoco consistently yields the lowest error on the x and z dimensions as well, with aggregated errors of 31.2 mm and 36.8 mm, respectively.

Next in line is the Detectron approach, which closely follows MpCoco's performance with an aggregated error of 78 mm. While it maintains a comparable level of excellence, even surpassing MpCoco in certain cases, along the x and y dimensions, it exhibits a relatively higher error in the z dimension. Specifically, it records the highest error in comparison to all approaches for the z dimension, with an aggregated error of 48.4 mm.

Securing the third position is the OpenPose approach, with an aggregated error of 93.8 mm. An intriguing aspect of OpenPose's performance is its remarkable accuracy in the y-dimension across most videos, outperforming the second-best approach (Detectron) by a margin of 2.6 mm (with an error of 36.4 mm). However, in the x-dimension, OpenPose displays the highest error among all approaches, which ultimately contributes to its elevated total error.

The final position is occupied by the Mp3D approach, registering a total error of 93.8 mm. Evidently, this substantial error can be attributed to its notable error in the y-dimension, which reaches 60 mm, the highest among all approaches. Nevertheless, the Mp3D approach demonstrates commendable performance in the x and z dimensions, particularly excelling in the x-dimension with a marginal difference of only 1.0 mm from the best approach (with an error of 32.2 mm).

5 CONCLUSIONS

In this thesis, we extensively studied the problem of sign language representation. Acknowledging its importance and complexity, which demands labor-intensive manual processes, we propose an automated method for mapping skeleton keypoints to avatar motions. Our rationale relies on the fact that an accurate 3D HPE technique from a video can be utilized to animate the avatar, reproducing the corresponding sign. In particular, we evaluated four approaches which involve state-of-the-art HPE algorithms to “lift” 2D body joint locations to the 3D plane.

We conducted our experiments on a small synthetic dataset consisting of five videos featuring the Paula avatar. Our research is focused on studying the trajectory of arm joints i.e., R/L Shoulder, R/L Elbow, and R/L Wrist, since their movements convey essential information for sign language understanding. Due to the fact that the evaluated algorithms have been trained on generic dataset and have specific skeleton configurations, we had to make certain adjustments for achieving accordance of skeleton keypoints.

Among the evaluated methods, MpCoco emerges as the frontrunner in terms of performance. Demonstrating consistent superiority across all videos, it showcases an impressive ability to minimize errors across different axes. This reliability positions MpCoco as a formidable contender for accurate pose estimation.

The analysis reveals that Detectron delivers a competitive performance, although slightly trailing behind MpCoco. This disparity can be attributed to the notable edge that Mediapipe holds in 2D pose estimation, which indirectly influences Detectron's performance. This observation underscores the interconnectedness of different stages in pose estimation.

OpenPose's performance unfolds as a story of axis-specific competency. It excels in depth estimation along the y-axis, showcasing commendable proficiency. However, its performance falters on the x-axis, where the predicted skeleton tends to diverge from the actual ground truth, hinting at potential challenges in width estimation.

Mp3D's performance varies significantly across different dimensions. While it attains notable accuracy on the x-axis, indicating precision in width estimation, its performance suffers on the y-axis (depth). This is because y-coordinate is derived from synthetic data using the GHUM model, fitted via an algorithm to the 2D key point projection. Therefore, the y-coordinate doesn't represent exact distance but rather it provides relative depth information within an image.

In summation, our analysis provides a comprehensive view of the strengths and limitations of each method. The distinct patterns of performance on different axes underscore the complexity of accurate pose estimation and offer a roadmap for further advancements in the field. The outcomes of this analysis serve as a foundation for refining methodologies and steering the evolution of sign language representation technology.

6 FUTURE WORK

First and foremost, the scope of analysis can be extended to capture the entirety of the upper body, including facial and finger landmarks. By incorporating these extra keypoints into our analysis we can achieve a more comprehensive understanding of sign language gestures and expressions. This broader perspective could illuminate the interplay between different components of the upper body in signing space.

Moreover, we observed that several failures in our findings are caused by disparities in motion capture data. There was no absolute correspondence of keypoints between the training and the test dataset. To address this, one promising direction is to train or fine tune models using avatar data. This strategy has the potential to improve prediction accuracy and leveraging the flexibility of avatar data allows us to tailor our models to better align with sign language motions intricacies in real world scenarios.

In addition to error analysis, as presented in this research, exploring the aspect of sign perception is vital in drawing any conclusion in SLP. Conducting studies to investigate how human signers perceive and interpret signs predicted by the developed approaches can provide valuable information which cannot be extracted from an error analysis. For instance, a subtle discrepancy in movement (low error) may lead to a different meaning (and vice versa). Therefore, taking into account the aspect of sign perception, we can refine our models and their applications resulting in more accurate and meaningful representations of sign language gestures.

In conclusion these future paths have the potential to improve the precision and real world applicability of the proposed approaches. By exploring the whole range of upper body movements, refining models using data driven techniques and receiving feedback from human signers we can obtain a more comprehensive understanding of sign language technologies.

ABBREVIATIONS – ACRONYMS

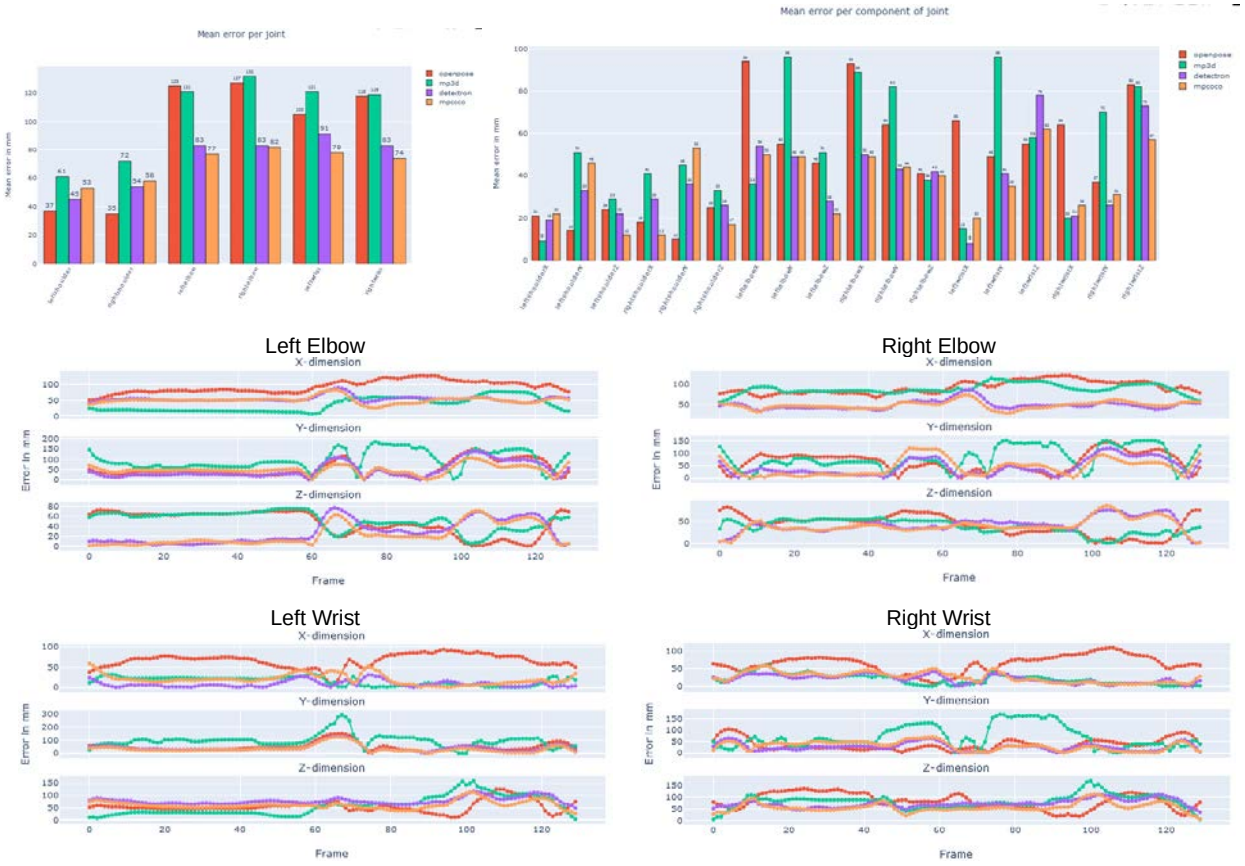
SLP	Sign Language Processing
HPE	Human Pose Estimation
GSL	Greek Sign Language
ASL	American Sign Language
LSF	French Sign Language
DGS	German Sign Language
DSGS	Swiss-German Sign Language
FMP	Flexible Mixtures-of-Parts model
DPM	Deformation Part Model
DNN	Deep Neural Networks
PAF	Part Affinity Fields
R-CNN	Region-based Convolutional Neural Networks
CPM	Convolutional Pose Machines
GHUM	Generative 3D Human Shape and Articulated Pose Models
MPJPE	Mean Per Joint Position Error
GT	Ground Truth
COCO	Common Objects in Context
H36m	Human 3.6 million

APPENDIX

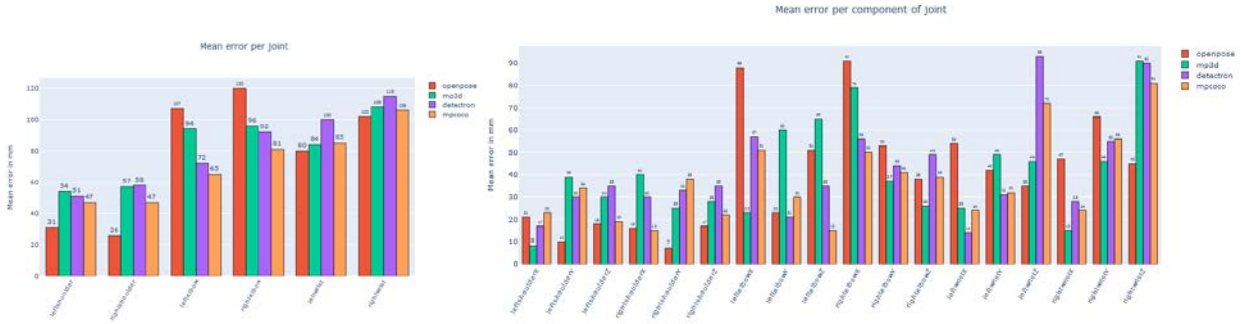
In this section, we provide the remaining diagrams that were not included in chapter 4. We exclude those that related to shoulders, as the observed error was consistently low in the majority of cases.

Code, data, models and supplementary materials associated with our research are available on my GitHub repository: https://github.com/JKaraman93/2dTo3d_Paula .

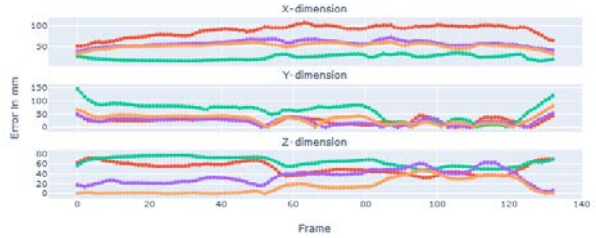
Video1:



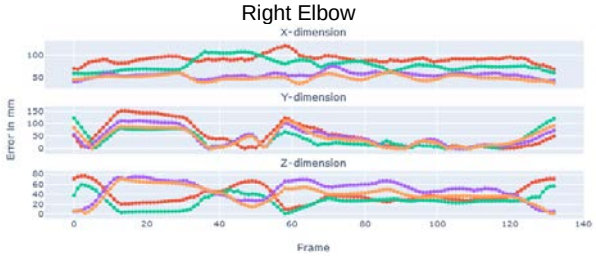
Video2:



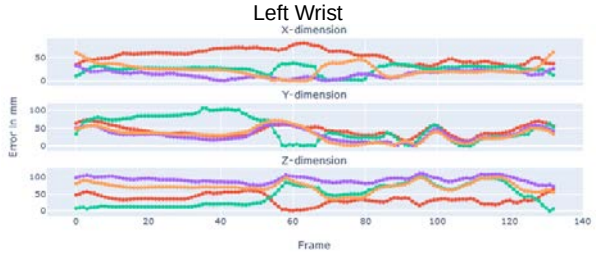
Left Elbow



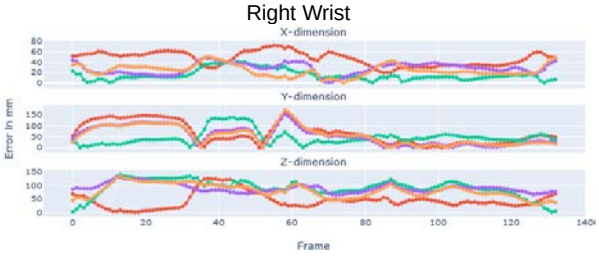
Right Elbow



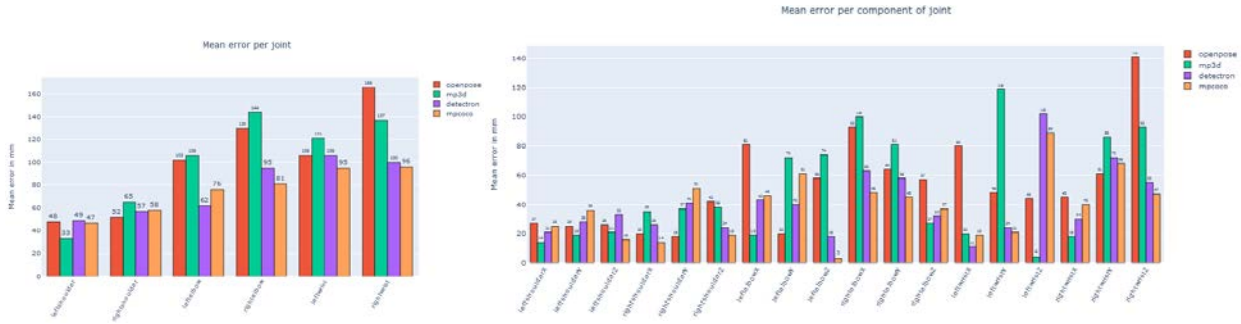
Left Wrist



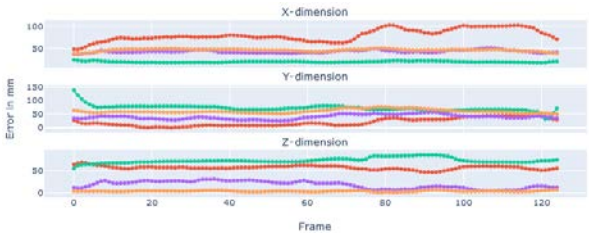
Right Wrist



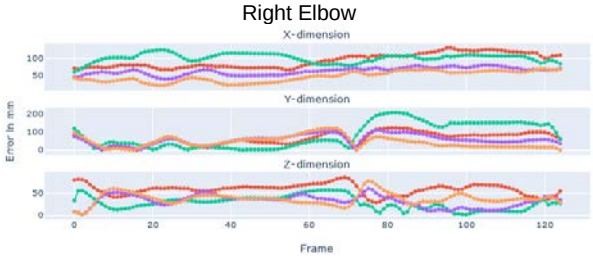
Video3:

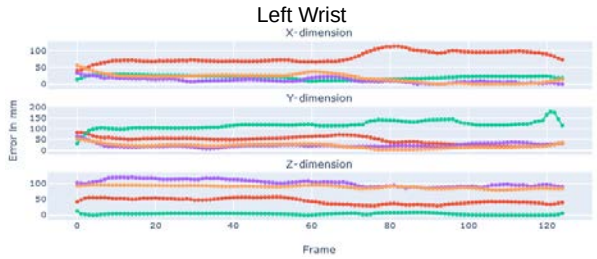


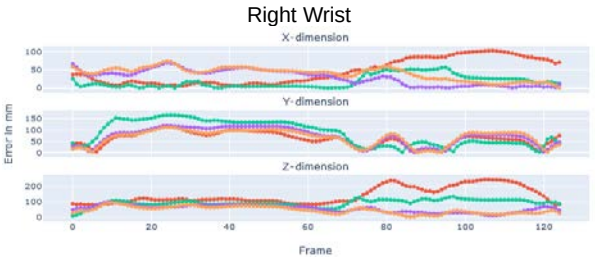
Left Elbow



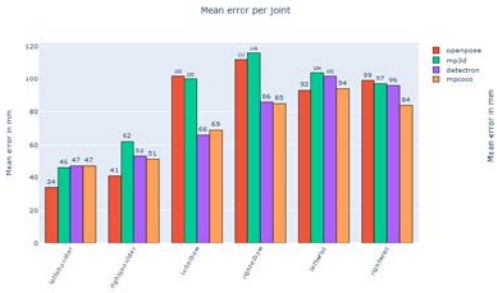
Right Elbow

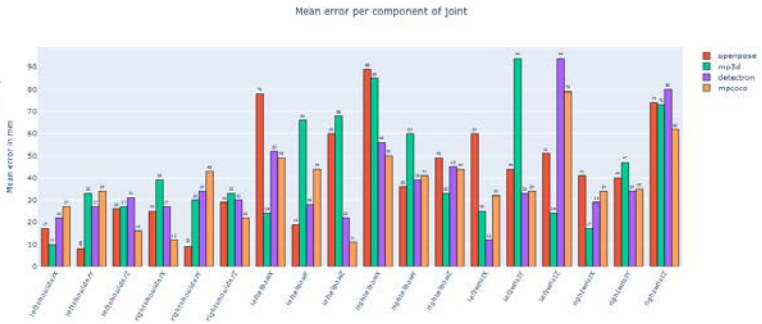


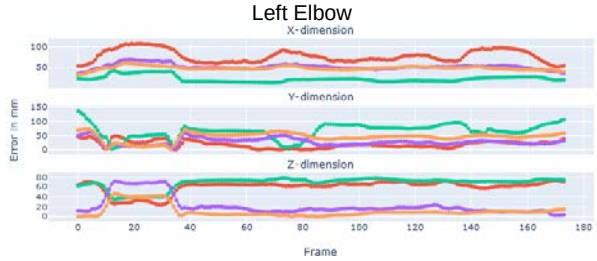


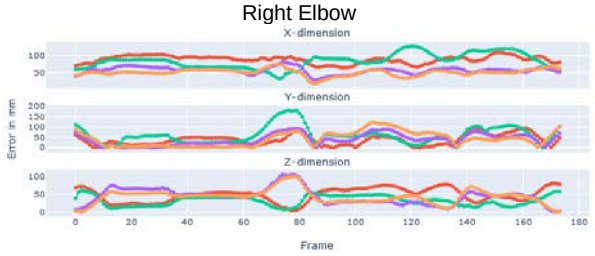


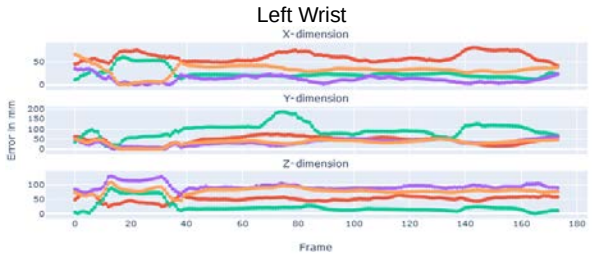
Video4:

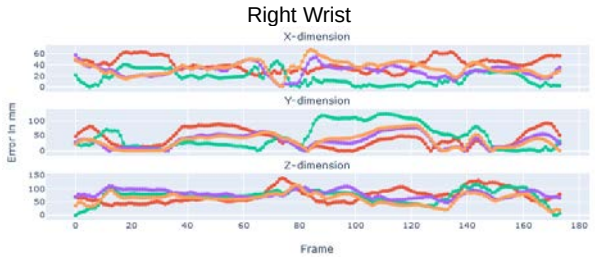




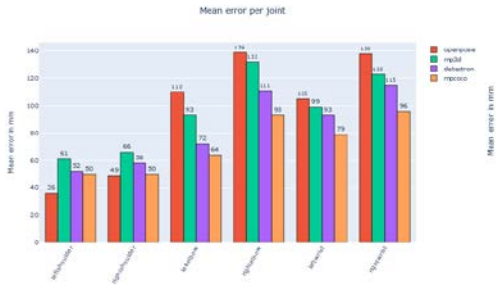


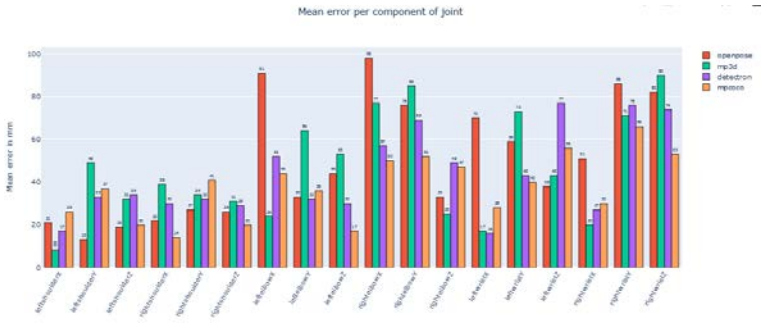


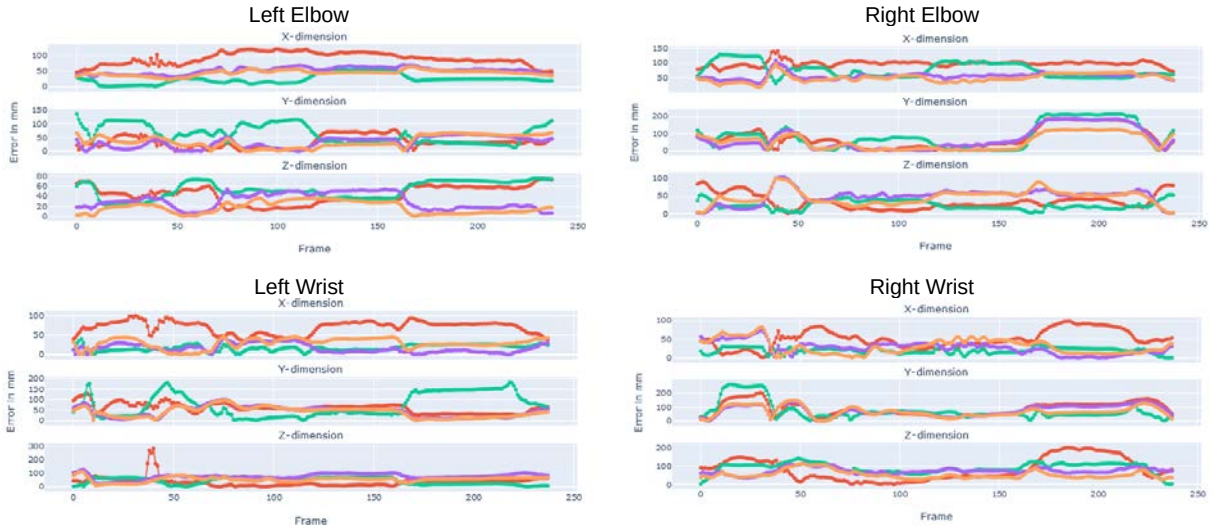




Video5:



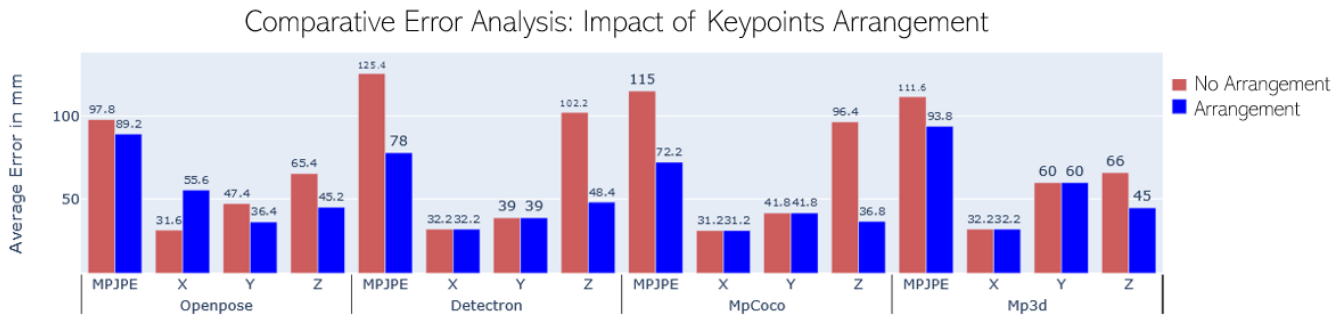




Results (without adjustments)

Videos	EUDs1				EUDs2				EUDs3				EUDs4				EUDs5			
Approaches	MPJPE	X	Y	Z	MPJPE	X	Y	Z	MPJPE	X	Y	Z	MPJPE	X	Y	Z	MPJPE	X	Y	Z
OpenPose	**96**	34	49	**62**	**89**	**29**	47	**53**	**106**	33	**40**	**82**	**94**	**27**	41	70	**104**	35	60	**60**
Detectron	120	30	**38**	69	131	33	**35**	117	122	32	43	104	124	33	**32**	111	130	33	47	110
MpCoco	113	**29**	42	94	117	31	38	102	115	**31**	46	93	115	33	38	98	115	32	**45**	95
Mp3D	123	34	73	69	103	31	42	71	115	34	69	60	104	32	54	**62**	113	**30**	62	68

Approach	**MPJPE**	**X**	**Y**	**Z**
OpenPose	**97.8**	31.6	47.4	**65.4**
Detectron	125.4	32.2	**39**	102.2
MpCoco	115	**31.2**	41.8	96.4
Mp3D	111.6	32.2	60	66



REFERENCES

[1] “World Health Organization. 2021. ‘Deafness and Hearing Loss.’”, [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss

[2] “World Federation of the Deaf. 2022. ‘World Federation of the Deaf - Our Work.’”, [Online]. Available: https://wfdeaf.org/our-work/

[3] “United Nations. 2022. ‘International Day of Sign Languages.’”, [Online]. Available: https://www.un.org/en/observances/sign-languages-day

[4] D. Brentari, "Sign Language Phonology," in The Handbook of Phonological Theory, J. Goldsmith, J. Riggle, and A. C. L. Yu, Eds. [Online]. Available: https://doi.org/10.1002/9781444343069.ch21, 2011.

[5] W. Sandler and D. Lillo-Martin, Sign Language and Linguistic Universals. Cambridge: Cambridge University Press, 2006.

[6] K. Yin, A. Moryossef, J. Hochgesang, Y. Goldberg, and M. Alikhani, "Including Signed Languages in Natural Language Processing," in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Aug. 2021, pp. 7347-7360, Online. DOI: 10.18653/v1/2021.acl-long.570. Available: https://aclanthology.org/2021.acl-long.570

[7] R. J. Wolfe, J. C. McDonald, T. Hanke, S. Ebling, D. Van Landuyt, F. Picron, V. Krausneker, E. Efthimiou, E. F. Fotinea, and A. Braffort, "Sign Language Avatars: A Question of Representation," Inf., vol. 13, p. 206, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:248281127.

[8] N. Adaloglou, T. Chatzis, I. Papastratis, A. Stergioulas, G. Th. Papadopoulos, V. Zacharopoulou, G. J. Xydopoulos, K. Atzakas, D. Papazachariou, and P. Daras, "A Comprehensive Study on Deep Learning-Based Methods for Sign Language Recognition," IEEE Transactions on Multimedia, vol. 24, pp. 1750-1762, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:234354860.

[9] M. De Coster, D. Shterionov, M. Van Herreweghe, and J. Dambre, "Machine Translation from Signed to Spoken Languages: State of the Art and Challenges," ArXiv, vol. abs/2202.03086, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:246634045.

[10] N. C. Camgöz, O. Koller, S. Hadfield, and R. Bowden, "Sign Language Transformers: Joint End-to-End Sign Language Recognition and Translation," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10020-10030, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:214728269.

[11] R. Rastgoo, K. Kiani, S. Escalera, and M. Sabokrou, "Sign Language Production: A Review," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 3446-3456, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:232417318.

[12] Moryossef, I. Tsochantaridis, R. Aharoni, S. Ebling, and S. Narayanan, "Real-Time Sign Language Detection using Human Pose Estimation," ArXiv, vol. abs/2008.04637, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:221095609.

[13] D. D. Monteiro, C. M. Mathew, R. Gutierrez-Osuna, and F. M. Shipman, "Detecting and Identifying Sign Languages through Visual Features," 2016 IEEE International

Symposium on Multimedia (ISM), pp. 287-290, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:13552976.

[14] M. De Sisto, D. Shterionov, I. Murtagh, M. Vermeerbergen, and L. Leeson, "Defining meaningful units. Challenges in sign segmentation and segment-meaning mapping (short paper)," in Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL), Aug. 2021, pp. 98-103, Virtual. DOI: 10.18653/v1/2021.mtsummit-at4ssl.11. Available: https://aclanthology.org/2021.mtsummit-at4ssl.11.

[15] T. A. Johnston, "From archive to corpus: transcription and annotation in the creation of signed language corpora," in Proceedings of the Pacific Asia Conference on Language, Information and Computation, 2008. [Online]. Available: https://api.semanticscholar.org/CorpusID:1266398.

[16] J. Mesch and L. Wallin, "Gloss annotations in the Swedish Sign Language Corpus," International Journal of Corpus Linguistics, vol. 20, pp. 102-120, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:62241141.

[17] T. E. Johnston, "Auslan Corpus Annotation Guidelines," 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:201605087.

[18] W. C. Stokoe Jr, "Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf," Journal of Deaf Studies and Deaf Education, vol. 10, no. 1, pp. 3-37, 2005.

[19] S. Prillwitz and H. Zienert, “Hamburg Notation System for Sign Language: Development of a Sign Writing with Computer Application.,” in Proceedings of the 3rd European Congress on Sign Language Research, 1990, pp. 355–79.

[20] V. Sutton, "Lessons in Sign Writing," SignWriting, 1990.

[21] B. Garcia and Marie-Anne Sallandre, “Transcription systems for sign languages: a sketch of the different graphical representations of sign language and their characteristics,” 2013, doi: 10.13140/RG.2.1.4760.2404.

[22] Moryossef and Y. Goldberg, "Sign Language Processing," [Online]. Available: https://sign-language-processing.github.io/, 2021.

[23] M. Parelli, K. Papadimitriou, G. Potamianos, G. Pavlakos, and P. Maragos, "Exploiting 3D Hand Pose Estimation in Deep Learning-Based Sign Language Recognition from RGB Videos," in ECCV Workshops, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:230795205.

[24] M. Parelli, K. Papadimitriou, G. Potamianos, G. Pavlakos, and P. Maragos, "Spatio-Temporal Graph Convolutional Networks for Continuous Sign Language Recognition," in ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8457-8461, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:249437248.

[25] M. Ivashechkin, O. A. Mendez Maldonado, and R. Bowden, "Improving 3D Pose Estimation For Sign Language," in 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW), pp. 1-5, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:260387723.

[26] R. Wolfe, J. McDonald, E. Efthimiou, E. Fotinea, F. Picron, D. Van Landuyt, T. Sioen, A. Braffort, M. Filhol, S. Ebling, T. Hanke, and V. Krausneker, "The myth of signing avatars," Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL), Aug. 2021, pp. 33-42.

[27] D. Bragg, O. Koller, M. Bellard, L. Berke, P. Boudreault, A. Braffort, N. K. Caselli, M. Huenerfauth, H. Kacorri, T. Verhoef, C. Vogler, and M. R. Morris, "Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective," Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:201645446.

[28] Aggarwal, J. K., & Ryoo, M. S. (2011). Human activity analysis: A review. ACM Computing Surveys, 43: 194–218. doi: 10.1145/1922649.1922653.

[29] J. Wang, S. Tan, X. Zhen, S. Xu, F. Zheng, Z. He, and L. Shao, "Deep 3D human pose estimation: A review," Comput. Vis. Image Underst., vol. 210, p. 103225, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:236241840.

[30] R. A. Güler, N. Neverova, and I. Kokkinos, "DensePose: Dense Human Pose Estimation in the Wild," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7297-7306, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:13637778.

[31] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," arXiv preprint arXiv:1812.08008, 2018.

[32] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A Skinned Multi-Person Linear Model," ACM Trans. Graph. (TOG), vol. 34, no. 6, p. 248, 2015.

[33] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in European Conference on Computer Vision, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:14113767.

[34] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014.

[35] Y. Yang and D. Ramanan, "Articulated Human Detection with Flexible Mixtures of Parts," IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 12, pp. 2878-2890, 2012.

[36] Fischler, M. A., & Elschlager, R. A. (1973). The representation and matching of pictorial structures. IEEE Transactions on Computers, C22, 67–92.

[37] Felzenszwalb, P. F., & Huttenlocher, D. P. (2005). Pictorial structures for object recognition. International Journal of Computer Vision, 61, 55–79. doi: 10.1023/B:VISI.0000042934.15159.49.

[38] Andriluka, M., Roth, S., & Schiele, B. (2009). Pictorial structures revisited: People detection and articulated pose estimation. Paper presented at the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009, June 20, 2009 - June 25, 2009, Miami, FL, United states

[39] Felzenszwalb, P., McAllester, D., & Ramanan, D. (2008). A discriminatively trained, multiscale, deformable part model. Paper presented at the 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR, June 23, 2008 - June 28, 2008, Anchorage, AK, United states.

[40] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part-Based Models," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, pp. 1627-1645, 2010. doi: 10.1109/TPAMI.2009.167.

[41] C. Wang, Y. Wang, and A. L. Yuille, "An Approach to Pose-Based Action Recognition," in Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2013, June 23, 2013 - June 28, 2013, Portland, OR, United States.

[42] A. Toshev and C. Szegedy, "DeepPose: Human Pose Estimation via Deep Neural Networks," in Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, June 23, 2014 - June 28, 2014, Columbus, OH, United States.

[43] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient Object Localization Using Convolutional Networks," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 648-656, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:206592615.

[44] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional Pose Machines," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4724-4732, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:163946.

[45] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:54465873.

[46] J. Martinez, R. Hossain, J. Romero, and J. Little, "A Simple Yet Effective Baseline for 3D Human Pose Estimation," 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2659-2668, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:206771080.

[47] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3D Human Pose Estimation in Video With Temporal Convolutions and Semi-Supervised Training," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7745-7754, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:53806352.

[48] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. L. Zhu, F. Zhang, and M. Grundmann, "BlazePose: On-device Real-time Body Pose Tracking," arXiv preprint arXiv:2006.10204, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:219793039.

[49] H. Xu, E. G. Bazavan, A. Zanfir, W. T. Freeman, R. Sukthankar, and C. Sminchisescu, "GHUM & GHUML: Generative 3D Human Shape and Articulated Pose Models," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6183-6192, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:219964093.

[50] M. J. Davidson, “PAULA: A Computer-Based Sign Language Tutor for Hearing Adults”.

[51] R. Wolfe, P. Cook, J. C. McDonald, and J. Schnepp, “Linguistics as structure in computer animation: Toward a more effective synthesis of brow motion in American Sign Language”.

[52] J. McDonald et al., “An automated technique for real-time production of lifelike animations of American Sign Language,” Univers. Access Inf. Soc., vol. 15, no. 4, pp. 551–566, Nov. 2016, doi: 10.1007/s10209-015-0407-2.

[53] J. McDonald, R. Wolfe, S. Johnson, S. Baowidan, R. Moncrief, and N. Guo, “An Improved Framework for Layering Linguistic Processes in Sign Language Generation: Why There Should Never Be a ‘Brows’ Tier,” in Universal Access in Human–Computer Interaction. Designing Novel Interactions, M. Antona and C. Stephanidis, Eds., in Lecture

Notes in Computer Science, vol. 10278. Cham: Springer International Publishing, 2017, pp. 41–54. doi: 10.1007/978-3-319-58703-5_4.

[54] M. Filhol, J. McDonald, and R. Wolfe, “Synthesizing Sign Language by Connecting Linguistically Structured Descriptions to a Multi-track Animation System,” in Universal Access in Human–Computer Interaction. Designing Novel Interactions, M. Antona and C. Stephanidis, Eds., in Lecture Notes in Computer Science, vol. 10278. Cham: Springer International Publishing, 2017, pp. 27–40. doi: 10.1007/978-3-319-58703-5_3.

[55] R. Johnson, M. Brumm, and R. Wolfe, “An Improved Avatar for Automatic Mouth Gesture Recognition”.

[56] R. Wolfe et al., “Supporting Mouthing in Signed Languages: New innovations and a proposal for future corpus building”.

[57] R. Wolfe, E. Jahn, R. Johnson, and J. C. McDonald, “The case for avatar makeup”.