Deep Neural Labeling: Hybrid Hand Pose Estimation Using Unlabeled Motion Capture Data With Color Gloves in Context of German Sign Language

1st Kristoffer Waldow
TH Köln, Computer Graphics Group
Cologne, Germany
kristoffer.waldow@th-koeln.de
FAU Erlangen-Nürnberg
Nürnberg, Germany

2nd Arnulph Fuhrmann
TH Köln, Computer Graphics Group
Cologne, Germany
arnulph.fuhrmann@th-koeln.de

3rd Daniel Roth
Technical University of Munich
Klinikum rechts der Isar
Orthopedics and Sports Orthopedics
Munich, Germany
daniel.roth@tum.de

***Abstract*—Hands are fundamental to conveying emotions and ideas, especially in sign language. In the context of virtual reality, motion capture is becoming essential for mapping real human movements to avatars in immersive environments. While current hand motion capture methods feature partly great usability, accuracy, and real-time performance, they have limitations. Industry-standard motion capture methods with sensor gloves lead to acceptable results, but still produce occasional errors due to proximity of the fingers and sensor drifts. This, in turn, requires time-consuming correction and manual labeling of optical markers during post-processing for offline use cases and prohibits the use in real-time scenarios as VR communication. To overcome these limitations, we introduce a novel hybrid hand pose estimation method that leverages both an optical motion capture system and a color-coded fabric glove. This approach merges the strengths of both techniques, enabling the automated labeling of 3D marker positions through a data-driven machine-learning approach. Using a spherical capture rig and a deep learning algorithm, we improve efficiency and accuracy. The labeled markers then drive a robust optimization procedure for solving hand posture, accounting for limitations in finger movements and validation checks. We evaluate our system in the context of German sign language where we achieve an accuracy of 97% correct marker assignments. Our approach aims to enhance the accuracy and immersion of sign language communication in VR, making it more inclusive for both deaf and hearing people.**

***Index Terms*—Hand Pose Estimation, Optical Motion Capture, Virtual Reality, Sign Language, Accessibility**

I. Introduction

Hands play an essential role in communication and expression across cultures, serving as powerful tools for conveying emotions, ideas, and intentions. This universal significance is particularly pronounced in sign language, where hands act as the primary mode of communication, enabling the deaf and hard-of-hearing communities to express complex thoughts and feelings. Slight variations in hand placement, orientation, and movement can drastically alter the message. Thereby, precise hand positioning and finger posture are essential to accurately capture these details, ensuring that the intended meanings are correctly communicated.

Integrating the deaf and hearing impaired community for avatar-based communication in virtual reality, another challenge emerges. Often, the lack of accuracy in tracking does not allow to faithfully reproduce all communication channels. As a consequence, speech becomes the dominant information channel and accurate finger postures play a minor role. For mixed avatar-mediated interaction in VR (hearing and deaf) there is the possibility to have a one-sided communication via speech bubbles to integrate the deaf and hearing impaired community [1] or haptic feedback for guidance [2]. However, adding gesture enhances immersion and engage in interactions that mimic real-life experiences, thereby deepening the sense of presence and engagement within an embodied virtual environment [3]–[6].

Currently, there are three distinct methods to capture hand movements. The first relies on optical motion capture which uses cameras to track markers in 3D space. The second method is using external gloves with sensors to estimate the current hand pose. The third method involves data-driven approaches utilizing machine learning, where hand poses are estimated from images through training data.

While classical optical motion capture has high accuracy but also the need for cost-intensive systems, data-driven approaches often use only a single camera to estimate the posture but have limited accuracy in space. Sensor-based gloves tend to overcome both of these drawbacks, however, they are often bulky, deal with drift problems over time, and are hard to calibrate.

For data-driven approaches, assumptions are made and depth ambiguities arise when the camera cannot accurately distinguish the distance of certain hand parts. Just relying on visual data leads to inaccuracies resulting in distorted or incorrect hand pose tracking, which is particularly problematic in scenarios like sign language communication where preci-

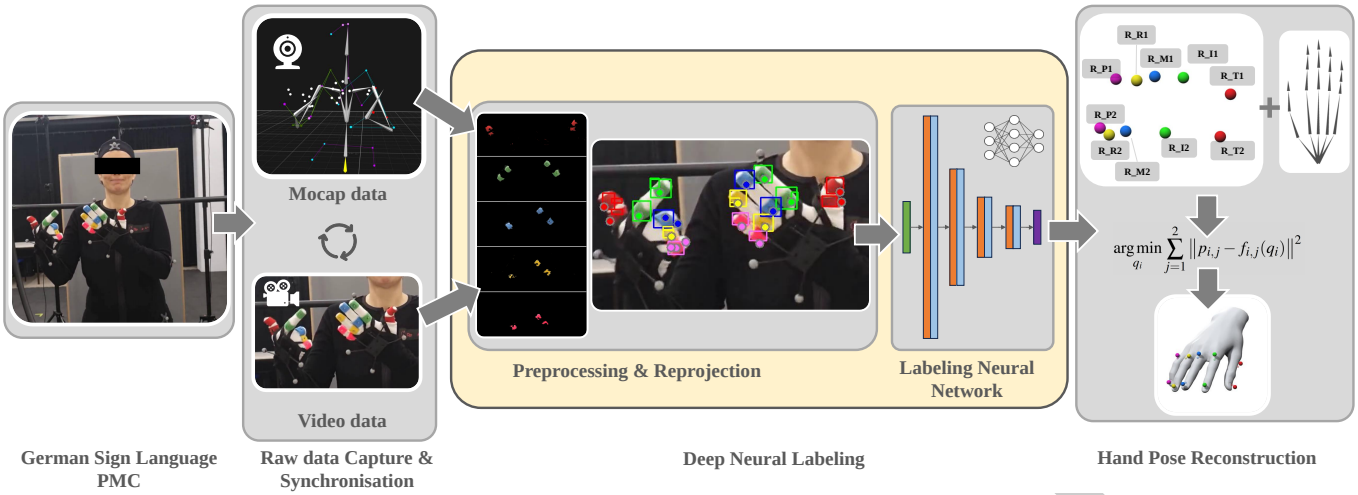


Fig. 1: Overview of our **Deep Neural Labeling** method with hand pose reconstruction based on German sign language motion capture performances and a color glove. Motion capture data is captured in addition to video information. The unlabeled point cloud of finger joint markers is assigned by an image reprojection and a subsequent neural network that has learned the point cloud shape for a regression classification task. Finally, the assigned points are used to animate the hand by solving an optimization problem to find the best joint angles q_i by minimizing the positional error of the markers $p_{i,j}$ and $f_{i,j}(q_i)$.

sion is crucial. That is also why, most available sign language datasets primarily consist of static data like hand poses and lack animation sequences [7], [8]. Yet, this transition of signs is necessary in order to reproduce natural-looking sequences since blending between static poses would lead to animations that would look robotic.

However, solely using an optical motion capture system does not result in perfect hand tracking. Estimating the hand poses from unlabeled markers is still a demanding task due to the proximity of the potential finger markers where occlusions and marker swaps occur frequently. The high number of degrees of freedom for fingers, self-similarity, and limited space make it difficult to accurately capture poses such as those needed for sign language.

By strategically positioning markers in critical areas by reducing or optimizing the marker layout, occlusion-related problems can be minimized [9]–[11]. Yet, the problem of manually correcting the labeling of markers in a subsequent process remains.

In this paper, we tackle the labeling problem by providing a hybrid hand pose estimation based on a passive marker-based optical motion capture system and a color-coded fabric glove. By combining the strengths of both approaches and coupling them with a data-driven approach we achieve a high accuracy without the need for manual labeling.

In section III we introduce a hardware setup and a method with two essential steps to animate hands for sign language for a wide variety of applications such as avatar-based communication in immersive environments. In the first step, we overcome the problem of the manual labeling process in section IV by introducing a novel color glove with markers that are captured with a unique spherical recording rig. In combination with synchronously taken video and computer vision, the 3D marker positions are preliminary labeled using color information. Since this information is not sufficient, our novel machine learning approach (**Deep Neural Labeling**) refines the labels by uniquely assigning markers for every finger joint based on pre-trained data. In section V, these labeled markers drive a custom skeleton hand model with our optimization approach to animate the hand. Our hand skeleton solver considers finger limitations and validation checks including a collision test.

II. Related Work

A. Motion Capture of Human Hands

Capturing complex 3D hand poses accurately from human hand movements is a challenging task, considering the complex dynamics of the flexibility and rapid angular movements of the hands. However, extensive research has led to several effective approaches, including optical, non-optical, and hybrid methods, to address this challenge [12]. It can be categorized into three groups [13], [14]:

1) Computer Vision: The fusion of camera technology and computer vision for hand pose recognition provides precise results, as depth imaging cameras or RGB cameras capture the three-dimensional movements and gestures of the hands [15]. Using multiple cameras in distinct layouts, 3D points can be reconstructed via optical motion capture, and a full skeleton pose is estimated [16].

2) Data gloves: Data gloves or wearable devices provide an alternative way to record hand movements during sign language. This technology does not require special cameras and allows free movement of the hands, which is particularly convenient for users. The recorded hand data is analyzed to identify the hand poses being displayed. Chen et al. [13]

categorize data gloves into four groups: Electromagnetic transmitters [17], bending and stretch sensors [18], [19], inertial measuring unity (IMU) [20] and exoskeletons [21]. These data gloves and the respective sensors are suitable for mobile applications, making the interaction with digital devices in different environments easier [22], [23].

3) Machine Learning: By using machine learning techniques such as Convolutional Neural Networks (CNNs), complex hand postures can be recognized and classified in real-time. This approach is particularly suitable for applications that do not require specialized hardware, such as depth imaging cameras, as conventional RGB cameras are sufficient to detect hand poses. The use of machine learning enables cost-effective and widespread implementation in various systems and applications [24], [25]. However, due to their spatial proximity to the body, the hands prove to be a critical aspect that requires more detailed evaluation. Although the initial impressions may seem plausible, they may not be sufficient for the requirements of a performance motion capture application [26]. Often, the first step is to extract depth information from the images to generate a more precise pose based on this information. However, the results obtained do not always show the predicted accuracy. The evaluation is mostly performed on pixel level and requires, in addition to a precise pose reconstruction, the creation of an individual reconstructed mesh [25]. The approach of animating hands in sign language using machine learning and artificial intelligence offers several advantages. By training models with large datasets of hand poses from motion capture systems, highly accurate and realistic animations can be generated [24], [27], [28].

B. Labeling & Hand Pose Models in Motion Capture

There are several approaches to label markers in motion capture. A promising approach is to group markers based on pairwise distances and propagate this information into future frames [29]. Another approach is to follow a starting position, such as the T-pose, using the coordinate axes as an orientation to define an initial layout [30]. Alternatively, large databases of predefined poses can be used as presets to improve marker assignment [31], [32]. Determining hand pose requires a certain number of markers per finger [33]. However, such dense marker layouts on the hand cause problems, especially in large spaces and in combination with full-body tracking [34]. For this reason, the study of specific marker layouts for the hands is a widely studied area of research. Wheatland et al. have investigated various layouts, including a maximally reduced layout (one marker per finger joint) and a minimal layout [10], [11], [35]. Despite the small number of markers, all degrees of freedom of the hand can be adequately captured and realized in the context of the application. It is possible to further reduce the marker layout with inverse kinematics techniques [36], [37]. The process of assigning the markers can therefore also be seen as sorting the marker data. Therefore, different markers are arranged into a certain structure, which can be seen by e.g. an inverse of a permutation matrix [38].

1) Color-coded gloves: Previous work demonstrated how colors can be used to achieve hand pose information only using a color-coded glove. Already in the 90s Dorner and Hagen [39] used a glove with colored rings to recognize short sequences of American sign language. These rings correspond to individual finger joints. Data-driven approaches extend this idea by using a fabric glove with color patterns resulting in a non-restrictive and inexpensive hand tracking, but with lack of accuracy [40].

C. Requirements of Sign Language

For sign language, only the upper body plays an important role in communication, which can be observed in most publications [41]. However, the context and position of a sign within a sentence become very crucial, as the same sign can translate to different meanings [42]. Wu and Huang [43] categorize hand gestures into different types, which include conversational, controlling, manipulative, and communicative gestures. This classification provides a comprehensive understanding of the various functions and applications of hand gestures in sign language. The structured nature of sign language makes it an interesting domain for testing computer vision algorithms [44]. Its well-defined grammar and syntax provide an excellent opportunity to develop and refine computer-based systems for accurate recognition and translation of sign language, which benefits both the deaf community and the field of human-computer interaction [1], [45], [46].

III. Hardware Setup

To combine the advantages of optical motion capture and gloves, we started by investigating different camera configurations for our *Optitrack* system with 11 Flex13 cameras (1.3 MP resolution, ± 0.2 mm accuracy, and 120 FPS). Standard camera layouts in a box shape are optimized for a large capture area so that the actors can walk around and perform actions. In the context of sign language, the performer stays in one place. Therefore, we could ensure optimal performance by designing a special camera layout that optimizes marker visibility while minimizing occlusion.

By using a spherical camera rig, we can capture a person from all angles, with more cameras focused on the forward space of the person (see Fig. 2). This gives us a capture area of about 3 m³. Though a spherical rig is not necessary for recording, it increases the visibility of markers for our subsequent learning processes. Additionally, we use a *Samsung S21* to record a video of the person in front of a neutral background and an LED bar to make it easier to distinguish the subject from the background in the videos. Additionally, the LED bar provides consistent lighting, which helps to identify the color features of our gloves in the following labeling steps.

A. Color Glove

Although color-coded gloves enable nonrestrictive and inexpensive hand tracking previous approaches lack in terms of accuracy [39], [40]. To tackle this we combine a custom color glove with the advantages of optical motion capture. Our glove consists of a white fabric and 10 elastic colored

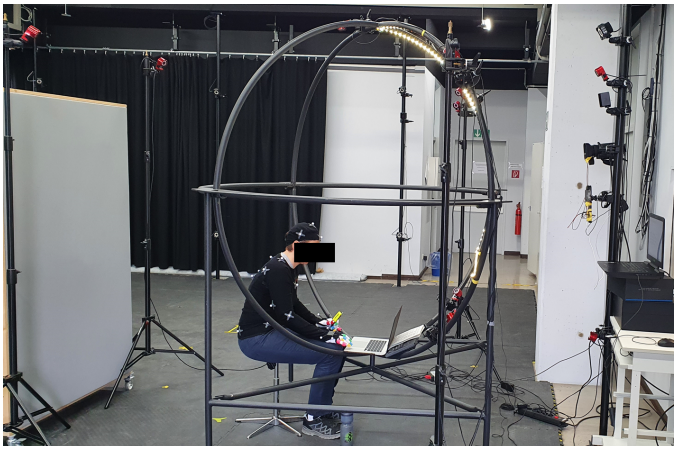


Fig. 2: The spherical arrangement of the cameras for our capturing process to get the best visibility of the markers with a neutral background and an LED Bar.

bands in 5 colors (red, green, blue, yellow, and pink) for each finger (see Fig. 3). In addition to colored bands, our glove has small attached reflective markers on the upper side of each band resulting in a total of 10 reflective markers per hand. To ensure that markers can be easily identified, the band colors are chosen to be distinguishable and easily isolated by image processing afterward.

B. Synchronization

In dealing with multimodal data, it is crucial to maintain coherence and synchronization among the data. Therefore, a microcontroller was designed to trigger optical pulses via LEDs in both the visible and infrared light ranges when an external command-based procedure is fired by the recording manager. These pulses serve as synchronization in and out points of the videos and animation data marking the beginning and end of a recording take. Consequently, the motion capture system recognizes the infrared LED pulse as a marker, while the white visible LED is detected in the video. To get the best results, the videos and the animation data are recorded at 120 fps.

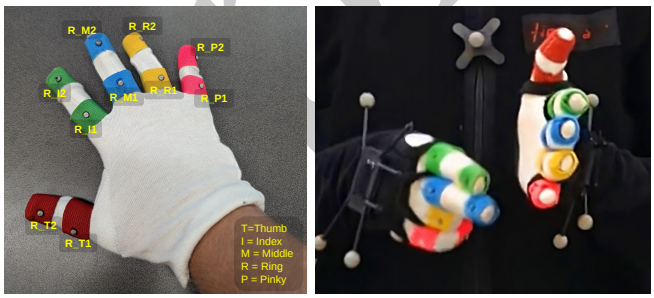


Fig. 3: Our glove for labeling finger data consists of 10 elastic colored bands in 5 colors (red, green, blue, yellow, and pink) with a small reflective marker on the top of each band. *Left:* Static hand pose with label naming convention. *Right:* Color gloves worn together with the motion capture suit and rigid bodies to track the palm of the hand.

IV. Deep Neural Labeling

The core of our approach is the labeling process, which allows us to use the motion capture data to maintain high positional accuracy and overcome the limitations of dense marker layouts using our custom color glove. This allows us to automatically classify and label finger markers when occlusions and marker swaps occur. Our method, called Deep Neural Labeling (**DNL**), consists of two essential steps (see Fig. 1): First, the motion capture data needs to be transformed to reproject the marker information in the image plane of the appropriate video frame. Together with the colored bands of our glove, *preliminary labels* are assigned with a simple search for the nearest neighbor. However, this information is incomplete and contains occasionally errors, as the video for assigning is fixed in space, and self-occlusions of the performer's body and hands can occur. Therefore, a subsequent machine learning algorithm is necessary that takes the preliminary marker information to refine each marker with a unique label. This results in an accurate labeling of the finger markers.

A. Data Preprocessing & Reprojection

In the first step, the unlabeled optical motion capture markers of the fingers are provided with a unique ID by the motion capture system as long as they are continuously tracked. However, in cases where markers vanish due to occlusion or experience occasional positional jumps, they reemerge with new IDs. Relying on this information is therefore not applicable and needs a custom solution. We use a model-based tracking approach to reconstruct missing markers by extrapolating the movement. This only works for a certain number of frames, as finger movement can be quite fast. If the maximum frames for extrapolation are exceeded, the hand marker set stays reduced, waiting for those missing markers to reappear, triggering our labeling method. We also trigger our labeling method at the start of every take.

For preliminary labeling, synchronized video is used to detect and identify relevant color glove regions. To optimize color detection, the method of *Contrast Limited Adaptive Histogram Equalization* (CLAHE) [47] was applied to the color values, resulting in an improved and normalized coherent image. To find the relevant colors of the glove, the image is processed further with color isolation and Canny edge detection [48] to find the contours of each segment of the finger.

The calibration of the virtual camera involves obtaining both intrinsic and extrinsic camera parameters to initially label the finger markers. This calibration process utilizes a combination of Zhang's method for intrinsic calibration [49] and 3D-2D point correspondences for extrinsic calibration [50] to get the transformation matrix T_{cam} using the RANSAC scheme [51]. Specifically, the method involves reprojection of the position $p_{world} \in R^3$ of motion capture markers into image

No.	Set	Count	Accumulated Frames
1	Signs	343	35568
2	Sentences	89	76816
Total		**432**	**112384**

TABLE I: Overview of the used dataset. The accumulated frames are treated separately as individual frames in the training process.

space $p_{screen} \in R^2$. Finally, each finger marker's position is assigned with the closest detected color contour through a nearest-neighbor search.

B. Labeling Neural Network

The preliminary labels can still contain faulty assignments and additional errors in joints and hands. Hence, a subsequent final refinement step is necessary. To this end, we use a data-driven machine-learning approach.

1) Dataset: The dataset for our **DNL** model consists of normalized and labeled 3D marker data. Therefore, each point is assigned to a specific part of the hand and finger. To obtain a certain amount of samples, a combination of real recorded and artificial data is used. The real data consists of the marker data in hand space derived from the motion capture body data and of the color data from the reprojection step. To enlarge our dataset, artificial data was created by using fully animated skeletons with finger poses from German Sign language and extracting the virtually attached finger markers and colors. At this point, there were about 432 records of movements, of which about 80 contained more than 1000 usable frames. These individual 110k frames from the motion data were used in a supervised learning algorithm to finally assign the finger markers (see Table I).

a) Normalized Hand Space: To create a more consistent dataset, the 3D marker positions of the fingers were transformed to be in the space of the respective hand so that they no longer contained body motion information. This has the advantage that the space is significantly reduced and becomes more predictable. Additionally, the data is scaled to a smaller space. To combine both hands in one designated space, each hand is shifted on one axis to combine both hands in one normalized space (see Fig. 4). It can be identified that the marker positions of each finger follow certain patterns and can be categorized due to distinct spatial separation.

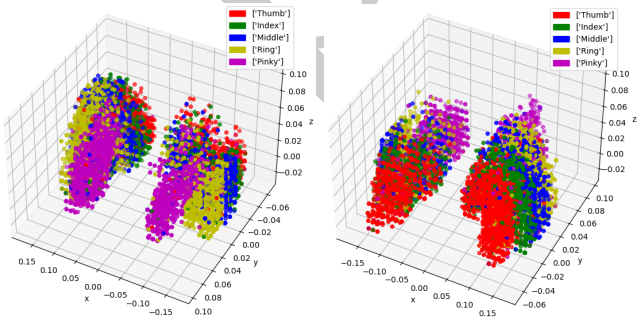


Fig. 4: An overview of a selection of the marker positions of each finger in a scaled and normalized hand space in different views.

2) Machine Learning Architecture: A fully-connected layer architecture with batch normalization was chosen to perform regression classification (see Fig. 5). This allows multiple classifications to be performed simultaneously for multiple input values. The problem can therefore be broken down into a sorting algorithm, as described by Ghorbani et al. [38]. Instead of a direct classification, they use an approach to determine the components of a permutation matrix.

In the initial stage, the color values are converted to integer values from -2 to 2, where each value represents a specific color (for example, red = -2, etc.). Combined with the normalized positional information, they are flattened to an input vector of 80 nodes. The positional components and color values are tightly coupled to form a 4D vector (T) for each of the 20 preliminary labeled markers. Each input node is then fed through five fully-connected layers with decreasing sizes starting with 2048 cells with a subsequent batch normalization step and ending at the 20-cell output layer. Each layer has a rectified linear unit (*ReLU*) activation function besides the output layer which has a linear activation function. Each of the 20 output cells represents a value between -1 and 1. To interpret the information of the model, a discretization step is necessary that rounds the values to distinct values that represent a unique label that describes each marker by the three categories of hand, finger, and joint. For example, a value of 0.3 represents the first joint on the left index.

To improve the accuracy of the **DNL** and to gain some robustness, several data augmentation methods were applied based on the observed data sets. These included adding noise to marker positions ($\pm 0.1cm$), adding a 10% of misinformation to the data to make it more capable of detecting and correcting such errors, and a 10% chance of completely removing data to become accustomed to dealing with incomplete or missing information. Additionally, the order of the frames and markers were changed to train the model with different temporal sequences and arrangements.

3) Performance Analysis: The machine learning algorithm achieved an accuracy of 97% after being trained for 500 epochs, with a batch size of 32, using the ADAM optimizer, and the MSQ loss function. The prediction runs on the GPU (Nvidia 4070Ti) and needs 35 ms for both hands combined. The preprocessing with color detection with the glove is CPU-based and takes an average of 20 ms resulting in a total of 55 ms for the **DNL** method.

To validate accuracy, the data set was split, with 30% of the data reserved solely for this validation task. To eliminate any temporal dependencies in the evaluation and validation, we ran the system with shuffled frames. This approach allows the algorithm to act independently in time and trains without considering sequence-related information in the data. A comparison with a smaller architecture or removing the color information reveals that the **DNL** model has the highest

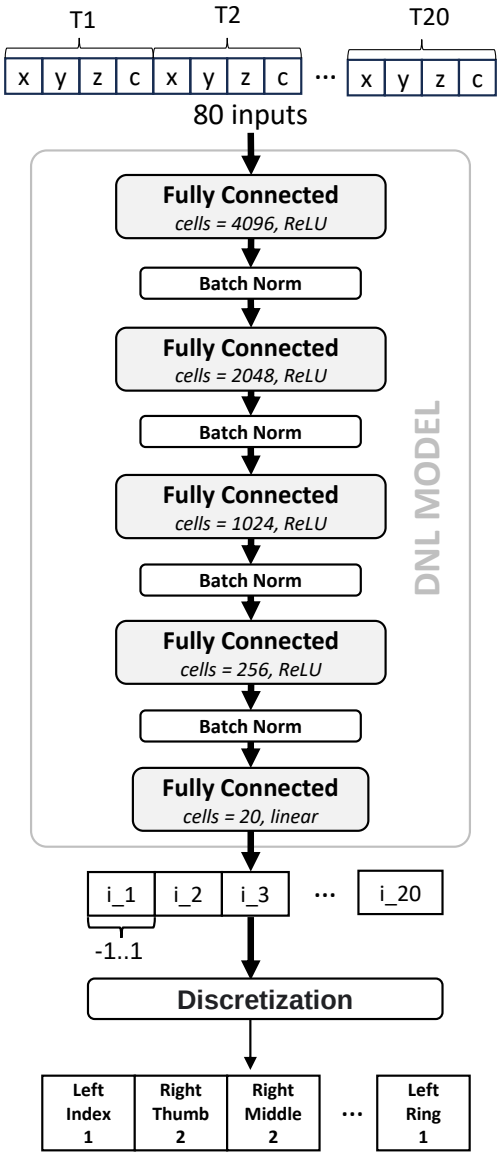


Fig. 5: The Deep Neural Labeling Algorithm flattens 20 4D vectors T containing positional and color information into 80 inputs values. Each input node is fed through five fully-connected layers with decreasing sizes. The model generates continuous values between -1 and 1, needing a subsequent discretization step for complete assignment.

validation accuracy (see Fig. 6). There is a clear difference between the curves with (*Glove*) and without (*No Glove*) color information. The model without color glove information shows a 10% worse performance compared to the model with color information. This leads to the assumption that using the color information from our gloves in the preliminary labeling step results in a significant increase in accuracy, even though these labels are sometimes false.

In addition to the function of the color information, an investigation of the hyperparameters was also carried out. Adjustments were made to use smaller dimensioned layers in the

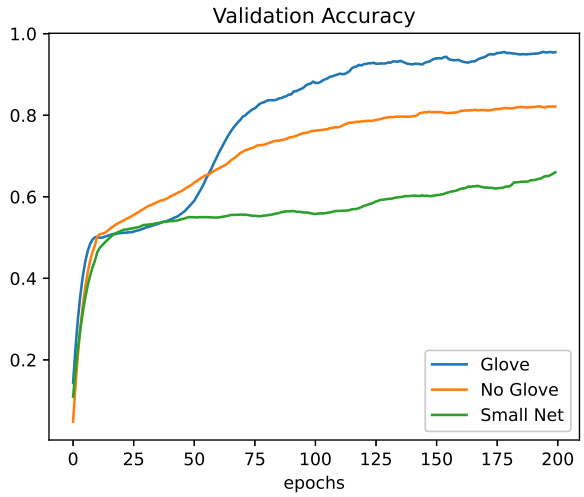


Fig. 6: Overview of the performances of the dataset and machine learning architecture while learning. The validation accuracy increases by adding color information and using a larger architecture.

architecture. The results obtained indicate a notable reduction in the accurate identification of markers, as illustrated in the graph (see Figs. 6 - *Small Net*). This underlines the sensitivity of the hyperparameters and shows that the right choice is essential for the performance of the model.

C. Limitations

Our labeling algorithm does not use temporal information from previous frames. The algorithm considers each frame as a separate individual frame and handles them independently. Information about the temporal context or previous frames is not incorporated. This approach can have both advantages and disadvantages. On the positive side, it allows efficient processing and labeling of the data without having to perform additional calculations for temporal context. However, in some scenarios, it may be necessary to take into account the temporal progression of the underlying data to achieve better results. This depends on the specific task and the context of the data on which it is based.

V. Pose Estimation

Given our labeled marker information, it is now possible to animate the skeleton model of the hands. Therefore, a mathematical hand model is created that precisely specifies the marker locations for each finger while using only two markers, since the two interphalangeal joints often move together, as described by Alexanderson et al. [9]. The kinematic rotation space of each finger is reduced to 5 degrees of freedom (DoF), due to natural axis restrictions of the human hand. An overview of the joints and limitations is listed in Table II and Fig. 7.

No.	Finger	Joints	DOF	Limits
1	Thumb	3	5	Joint 1: $x \in [\pm 80°]$,$y \in [-150°, 30°]$,$z \in [\pm 90°]$, Joint 2-3: $y \in [0°, 90°]$
2	Index, Middle, Ring, Pinky	4	5	Joint 1: $x \in [\pm 8°]$, Joint 2: $x \in [0°, 110°]$, $y \in [-10°, 25°]$, Joint 3-4: $x \in [0°, 130°]$

TABLE II: Overview of the finger configuration for each hand. The recorded limits are in degree and tightly coupled with our underlying skeleton structure for hand pose reconstruction. All joints are oriented forward on the z-axis.

A. Solving Algorithm

To determine the overall pose of the hand, five separate optimization problems need to be solved. We compute the forward kinematics of each finger $i \in \{1, ...5\}$ for the two markers $m_{i,j}$ for $j \in \{1, 2\}$ using the current joint angle configuration $q_i \in \mathbb{R}^5$. Therefore, $f_{i,j}(q_i)$ transforms the local marker $m_{i,j}$ by using the current configuration q_i with the corresponding joint chain transformation matrix $T_{i,j}(q_i)^{[Local \to World]}$.

We solve the minimization problem by calculating the distance as an error of the expected markers from $f_{i,j}(q_i)$ and the real corresponding point $p_{i,j}$ with the following optimization:

$$q_i^* = \arg\min_{q_i} \sum_{j=1}^{2} \|p_{i,j} - f_{i,j}(q_i)\|^2 \cdot w_j \tag{1}$$

with:

$$\text{with} \quad f_{i,j}(q_i) = T_{i,j}(q_i)^{[Local \to World]} \cdot m_{i,j} \tag{2}$$

and the limitations from Table II:

$$q_i^{min} \leq q_i \leq q_i^{max} \in \mathbb{R}^5 \tag{3}$$

The optimization error is multiplied by different weights w_j to ensure larger errors for the critical markers close to the fingertips and to maintain higher accuracy in positioning. Thereby, a 30% penalty produces the best results. We solve these optimization problems using the limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS-B) method with simple box constraints and numerically estimated gradients [52], [53], starting with the configuration of the joint angles of the previous frame.

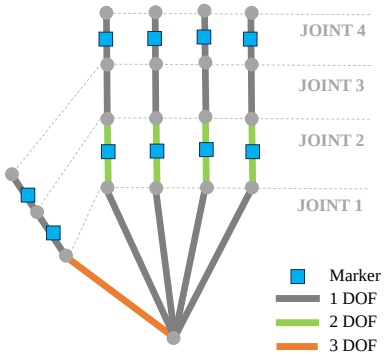


Fig. 7: Our underlying skeleton structure with DOF for our hand pose estimation.

B. Pose Evaluation

Combining the degrees of freedom of the thumb and finger joints results in a total number of 25 DoF. This number allows the model to cover all possible hand poses, and thus provide a comprehensive capture of hand movements. Consequently, this model allows for an optimal balance between flexibility and complexity in optimization performance. Figure 9 provides an overview of some possible hand poses based on actual German sign language poses captured with our system.

1) Pose Validation: An additional simple collision detection of capsules was implemented to validate the estimated pose. Therefore, capsules are attached to the finger joints. If an overlap is detected between these capsules, the pose is considered invalid and the previous pose is used. However, this collision test can be expensive. Hence, a more simple cross-check is performed before that compares the angles between the normalized finger joint vectors to determine if there is an unrealistic crossover. The scalar product is used to check the normalized vectors.

2) Pose Performance: The computational effort needed varies significantly due to the inconsistent number of iterations required to estimate the correct hand pose (see Fig. 8). Using a current system (Intel i9-13900k with 24 cores, 32 GB RAM, Nvidia GTX 4070Ti), on average the optimization method needs 125 ms for the dominant hand in motion. The computational effort increases when both hands are active and the sign language actor changes the hand poses. In this case, the fingers undergo a large movement from one frame to the next. These large finger movements lead to large positional changes of the markers, resulting in an increased time to solve the optimization problem.

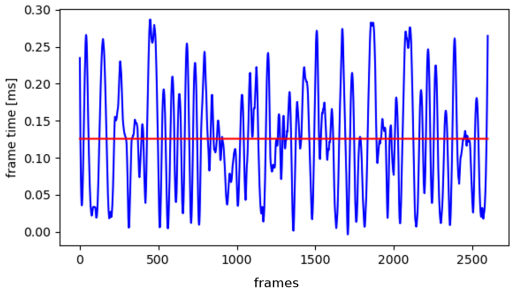


Fig. 8: Optimization times for the dominant hand in one example take. The graph shows a large variation in the calculation time but averages out at 125 ms.

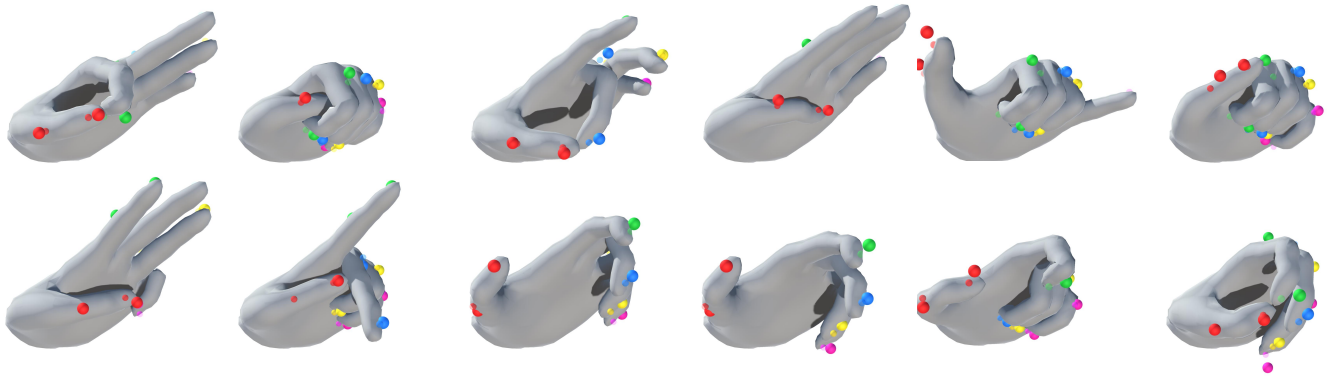


Fig. 9: Overview of the variety of possible hand poses of our system based on the German sign language. The finger postures of the right dominant hand are shown as an example.

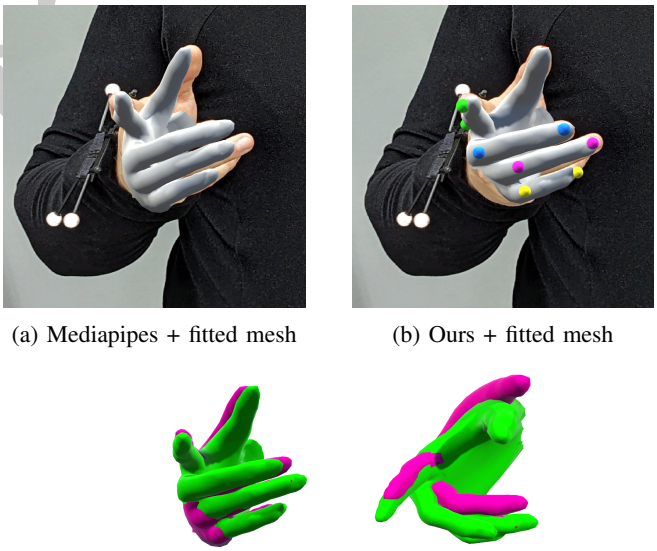
VI. Discussion

The system is specifically trained on sign language gestures and enables highly accurate capturing of hand poses. This design focus on accuracy makes the system particularly attractive for use in avatar-based immersive environments, where highly accurate representation is critical to ensure natural interpersonal communication. This is especially meaningful for deaf people communicating in VR or AR environments, as it allows them to express and communicate via sign language. This contributes substantially to the linguistic inclusion of deaf people and has both social and ethical implications. Thus, deaf users can express and communicate naturally in virtual worlds. Moreover, its accuracy provides the opportunity to develop hybrid systems in which signs could be translated in real-time [54]. This extends the scope of applications beyond sign language and opens up new accessibility possibilities for immersive communication in VR and AR applications.

Another interesting application area is outside of sign language, as the system can be used in other scenarios due to its precision and hand pose recognition capability. For example, using a colored glove for hand pose recognition in virtual reality can greatly improve interaction in VR environments. This allows users to move and gesture more naturally, which significantly increases immersion and the overall experience. At the same time, low latency is of great importance and must be taken into account to ensure a realistic experience.

The current system has some limitations that require further refinement. Most notably, real-time capability has not yet been achieved, but there is an opportunity to address this with further improvements. Two ways are promising to improve the performance of the system. First, optimizing the neural network with temporal information and including non-signing gestures in the dataset could significantly increase overall performance. This could lead to a reduction in the size of the neural network, improving the efficiency and accuracy of hand pose recognition in VR environments. In addition, we see great potential in simplifying the kinematic model for our solving algorithm to a less dimensional problem. Although our system was only evaluated on our spherical rig, it is important to note that this configuration helps to improve visibility but is not required. Further investigation of the camera layout is needed to evaluate different setups.

It is important to highlight that the visualization of the hand did not play a leading role in our scenario. Nevertheless, we are aware that it can make a significant contribution to realism. Especially in VR, the aspect of self-embodiment is a decisive factor for presence and must be taken into account. We are aware of the challenges of such representations, such as the deformation of the skin and the formation of wrinkles, but they are not included in this work.



(a) Mediapipes + fitted mesh

(b) Ours + fitted mesh

(c) Mediapipe + Ours overlay

Fig. 10: Comparison of our hand pose compared to state-of-the-art image based method from *Mediapipe* [55]. In the overlay 10c the estimated poses are compared from different angles (pink presents the hand of *Mediapipe* and green ours), which visualize distinct deviations.

A. Comparison to image-only methods

For image-only methods, there are two main problems that our method eliminates: First, the depth information calculated is only an estimate and cannot produce a precise depth representation. On the other hand, the estimated skeleton is based on detected 2D feature points and estimated depth, resulting in inconsistent bone length in motion. Hence, it is necessary to handle a consistent skeleton for animation in an additional step.

A direct comparison between a complex hand pose using Zhang et al.'s method [55] and our developed method reveals a mean square error of 3.2% normalized by diagonal palm size to a synthetic reference pose compared to Zhang et al.'s method with 16.2%. By overlaying these two poses, noticeable differences become visible. A direct joint comparison of only the skeleton reveals the differences by using synthetically generated poses (see Fig. 10).

VII. Conclusion & Future Work

In this paper, we presented a novel hybrid approach for estimating hand poses for applications such as avatar-based communication in immersive environments. The method combines a passive marker-based motion capture system, a glove, and a data-driven approach. To this end, we used a specially designed colored glove with optical markers that classify unlabeled finger markers with computer vision and by a subsequent machine learning algorithm. In the second step, these labeled markers drive a custom hand skeleton model, considering factors like finger limitations and collision detection. This approach offers an effective solution for animating hands, overcoming manual labeling challenges, and providing accurate hand pose estimation.

For future projects, we want to continuously improve the machine learning algorithm to potentially eliminate the requirement of the glove in the end. A potential idea is to use a 3D convolutional layer and recurrent nodes to increase the accuracy and efficiency of the hand pose estimation and reduce the amount of physical aids needed. Our database of reference data provides a good foundation to develop and evaluate new models. This comprehensive database allows us to cover a wide range of scenarios and to apply the performance of the algorithms to a wide variety of application fields for sign language. Despite the challenge in terms of real-time capability, there is potential to adapt the method so that hand pose settings can be determined directly. This would allow the optimization problem to be bypassed, which could ultimately lead to an increase in overall performance.

Regarding real-time communication, we see great potential, especially in the area of avatar-based communication in the context of sign language, to further enhance accessibility. Our developed method helps to support more inclusiveness for deaf people and to make avatar-based communication in sign language more efficient and accurate as shown in previous publications [1], [56].

Acknowledgment

This work was funded by the German Federal Ministry of Education and Research (BMBF) under grant number 16SV8492 as part of the project *AVASAG*.

References

[1] K. Waldow and A. Fuhrmann, "Addressing deaf or hard-of-hearing people in avatar-based mixed reality collaboration systems," in *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 2020, pp. 594–595.

[2] M. Mirzaei, P. Kan, and H. Kaufmann, "Earvr: Using ear haptics in virtual reality for deaf and hard-of-hearing people," *IEEE transactions on visualization and computer graphics*, vol. 26, no. 5, pp. 2084–2093, 2020.

[3] S. W. Greenwald, Z. Wang, M. Funk, and P. Maes, "Investigating social presence and communication with embodied avatars in room-scale virtual reality," in *Immersive Learning Research Network: Third International Conference, iLRN 2017, Coimbra, Portugal, June 26–29, 2017. Proceedings 3*. Springer, 2017, pp. 75–90.

[4] B. Yoon, H.-i. Kim, G. A. Lee, M. Billinghurst, and W. Woo, "The effect of avatar appearance on social presence in an augmented reality remote collaboration," in *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2019, pp. 547–556.

[5] D. Roth, K. Waldow, M. E. Latoschik, A. Fuhrmann, and G. Bente, "Socially immersive avatar-based communication," in *2017 IEEE Virtual Reality (VR)*. IEEE, 2017, pp. 259–260.

[6] K. Waldow, A. Fuhrmann, and S. M. Grünvogel, "Investigating the effect of embodied visualization in remote collaborative augmented reality," in *International Conference on Virtual Reality and Augmented Reality*. Springer, 2019, pp. 246–262.

[7] N. Pugeault and R. Bowden, "Spelling it out: Real-time asl fingerspelling recognition," in *2011 IEEE International conference on computer vision workshops (ICCV workshops)*. IEEE, 2011, pp. 1114–1119.

[8] A. Barczak, N. Reyes, M. Abastillas, A. Piccio, and T. Susnjak, "A new 2d static hand gesture colour image dataset for asl gestures," 2011.

[9] S. Alexanderson, C. O'Sullivan, and J. Beskow, "Real-time labeling of non-rigid motion capture marker sets," vol. 69, pp. 59–67, publisher: Elsevier.

[10] M. Schröder, J. Maycock, and M. Botsch, "Reduced marker layouts for optical motion capture of hands," in *Proceedings of the 8th ACM SIGGRAPH Conference on Motion in Games*, pp. 7–16.

[11] N. Wheatland, S. Jörg, and V. Zordan, "Automatic hand-over animation using principle component analysis," in *Proceedings of motion on games*, pp. 197–202.

[12] A. Fabisch, M. Uliano, D. Marschner, M. Laux, J. Brust, and M. Controzzi, "A modular approach to the embodiment of hand motions from human demonstrations," in *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids)*, pp. 801–808, ISSN: 2164-0580.

[13] W. Chen, C. Yu, C. Tu, Z. Lyu, J. Tang, S. Ou, Y. Fu, and Z. Xue, "A survey on hand pose estimation with wearable sensors and computer-vision-based methods," vol. 20, no. 4, p. 1074, publisher: MDPI.

[14] R. Li, Z. Liu, and J. Tan, "A survey on 3d hand pose estimation: Cameras, methods, and datasets," *Pattern Recognition*, vol. 93, pp. 251–272, 2019.

[15] C. V. Nguyen, S. Izadi, and D. Lovell, "Modeling kinect sensor noise for improved 3d reconstruction and tracking," in *2012 second international conference on 3D imaging, modeling, processing, visualization & transmission*. IEEE, 2012, pp. 524–530.

[16] M.-C. Silaghi, R. Plänkers, R. Boulic, P. Fua, and D. Thalmann, "Local and global skeleton fitting techniques for optical motion capture," in *Modelling and Motion Capture Techniques for Virtual Environments: International Workshop, CAPTECH'98 Geneva, Switzerland, November 26–27, 1998 Proceedings*. Springer, 1998, pp. 26–40.

[17] K.-Y. Chen, S. N. Patel, and S. Keller, "Finexus: Tracking precise motions of multiple fingertips using magnetic sensing," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016, pp. 1504–1514.

[18] G. Saggio, F. Riillo, L. Sbernini, and L. R. Quitadamo, "Resistive flex sensors: a survey," *Smart Materials and Structures*, vol. 25, no. 1, p. 013001, 2015.

[19] M. Bianchi, R. Haschke, G. Büscher, S. Ciotti, N. Carbonaro, and A. Tognetti, "A multi-modal sensing glove for human manual-interaction studies," *Electronics*, vol. 5, no. 3, p. 42, 2016.

[20] B. O'Flynn, J. T. Sanchez, J. Connolly, J. Condell, K. Curran, P. Gardiner, and B. Downes, "Integrated smart glove for hand motion monitoring," in *The Sixth International Conference on Sensor Device Technologies and Applications*. International Academy, Research, and Industry Association, 2015.

[21] I. Sarakoglou, A. Brygo, D. Mazzanti, N. G. Hernandez, D. G. Caldwell, and N. G. Tsagarakis, "Hexotrac: A highly under-actuated hand exoskeleton for finger tracking and force feedback," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 1033–1040.

[22] E. Richter, N. Hendrich, and J. Zhang, "Hand pose reconstruction using a low-cost three-camera stereo vision system," in *5th International Conference on Cognitive Systems, COGSYS (POSTER*.

[23] T. Magnenat, R. Laperrière, and D. Thalmann, "Joint-dependent local deformations for hand animation and object grasping," Canadian Inf. Process. Soc, Tech. Rep., 1988.

[24] S. Han, B. Liu, R. Wang, Y. Ye, C. D. Twigg, and K. Kin, "Online optical marker-based hand tracking with deep labels," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 1–10, 2018.

[25] J. Wang, F. Mueller, F. Bernard, S. Sorli, O. Sotnychenko, N. Qian, M. A. Otaduy, D. Casas, and C. Theobalt, "Rgb2hands: real-time tracking of 3d hand interactions from monocular rgb video," *ACM Transactions on Graphics (ToG)*, vol. 39, no. 6, pp. 1–16, 2020.

[26] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 975–10 985.

[27] J. L. J. Bascones, "Cloud point labelling in optical motion capture systems," Ph.D. dissertation, Ph. D. thesis, Universidad del País Vasco-Euskal Herriko Unibertsitatea, 2019.

[28] Y. Endo, M. Tada, and M. Mochimaru, "Reconstructing individual hand models from motion capture data," *Journal of Computational Design and Engineering*, vol. 1, no. 1, pp. 1–12, 2014.

[29] M. Ringer and J. Lasenby, "A procedure for automatically estimating model parameters in optical motion capture," *Image and Vision Computing*, vol. 22, no. 10, pp. 843–850, 2004.

[30] J. Meyer, M. Kuderer, J. Müller, and W. Burgard, "Online marker labeling for fully automatic skeleton tracking in optical motion capture," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2014, pp. 5652–5657.

[31] T. Schubert, A. Gkogkidis, T. Ball, and W. Burgard, "Automatic initialization for skeleton tracking in optical motion capture," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 734–739.

[32] T. Schubert, K. Eggensperger, A. Gkogkidis, F. Hutter, T. Ball, and W. Burgard, "Automatic bone parameter estimation for skeleton tracking in optical motion capture," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 5548–5554.

[33] M. Kitagawa and B. Windsor, *MoCap for artists: workflow and techniques for motion capture*. CRC Press, 2020.

[34] N. Wheatland, Y. Wang, H. Song, M. Neff, V. Zordan, and S. Jörg, "State of the art in hand and finger modeling and animation," in *Computer Graphics Forum*, vol. 34, no. 2. Wiley Online Library, 2015, pp. 735–760.

[35] S. Alexanderson, C. O'Sullivan, and J. Beskow, "Robust online motion capture labeling of finger markers," in *Proceedings of the 9th International Conference on Motion in Games*, 2016, pp. 7–13.

[36] J. Maycock, T. Rohlig, M. Schroder, M. Botsch, and H. Ritter, "Fully automatic optical motion tracking using an inverse kinematics approach," in *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2015, pp. 461–466.

[37] A. Aristidou and J. Lasenby, "Motion capture with constrained inverse kinematics for real-time hand tracking," in *2010 4th International Symposium on Communications, Control and Signal Processing (ISCCSP)*. IEEE, 2010, pp. 1–5.

[38] S. Ghorbani, A. Etemad, and N. F. Troje, "Auto-labelling of markers in optical motion capture by permutation learning," in *Advances in Computer Graphics: 36th Computer Graphics International Conference, CGI 2019, Calgary, AB, Canada, June 17–20, 2019, Proceedings 36*. Springer, pp. 167–178.

[39] B. Dorner and E. Hagen, "Towards an american sign language interface," *Integration of Natural Language and Vision Processing: Computational Models and Systems*, pp. 143–161, 1995.

[40] R. Y. Wang and J. Popović, "Real-time hand-tracking with a color glove," *ACM transactions on graphics (TOG)*, vol. 28, no. 3, pp. 1–8, 2009.

[41] U. Bellugi and S. Fischer, "A comparison of sign language and spoken language," *Cognition*, vol. 1, no. 2-3, pp. 173–200, 1972.

[42] R. Yang, S. Sarkar, and B. Loeding, "Handling movement epenthesis and hand segmentation ambiguities in continuous sign language recognition using nested dynamic programming," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 3, pp. 462–477, 2009.

[43] Y. Wu and T. S. Huang, "Human hand modeling, analysis and animation in the context of hci," in *Proceedings 1999 international conference on image processing (Cat. 99CH36348)*, vol. 3. IEEE, 1999, pp. 6–10.

[44] K. Bantupalli and Y. Xie, "American sign language recognition using deep learning and computer vision," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 4896–4899.

[45] F. Nunnari, J. Bauerdiek, L. Bernhard, C. España-Bonet, C. Jäger, A. Unger, K. Waldow, S. Wecker, E. André, S. Busemann, C. Dold, A. Fuhrmann, P. Gebhardt, Y. Hamidullah, M. Hauck, Y. Kossel, M. Misiak, D. Wallach, and A. Stricker, "AVASAG: A german sign language translation system for public services (short paper)," in *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pp. 43–48.

[46] L. Bernhard, F. Nunnari, A. Unger, J. Bauerdiek, C. Dold, M. Hauck, A. Stricker, T. Baur, A. Heimerl, E. André, M. Reinecker, C. España-Bonet, Y. Hamidullah, S. Busemann, P. Gebhardt, C. Jäger, S. Wecker, Y. Kossel, H. Müller, K. Waldow, A. Fuhrmann, M. Misiak, and D. Wallach, "Towards automated sign language production: A pipeline for creating inclusive virtual humans," in *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments*, pp. 260–268.

[47] K. Zuiderveld, "Contrast limited adaptive histogram equalization," *Graphics gems*, pp. 474–485, 1994.

[48] B. Green, "Canny edge detection tutorial," *Retrieved: March*, vol. 6, p. 2005, 2002.

[49] Q. Zhang and R. Pless, "Extrinsic calibration of a camera and laser range finder (improves camera calibration)," in *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)*, vol. 3. IEEE, pp. 2301–2306.

[50] E. Marchand, H. Uchiyama, and F. Spindler, "Pose estimation for augmented reality: a hands-on survey," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 12, pp. 2633–2651, 2015.

[51] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[52] C. G. Broyden, "The convergence of a class of double-rank minimization algorithms 1. general considerations," *IMA Journal of Applied Mathematics*, vol. 6, no. 1, pp. 76–90, 1970.

[53] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, "Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization," *ACM Transactions on mathematical software (TOMS)*, vol. 23, no. 4, pp. 550–560, 1997.

[54] S. Sen, S. Narang, and P. Gouthaman, "Real-time sign language recognition system," in *2023 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA)*. IEEE, 2023, pp. 1–6.

[55] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann, "Mediapipe hands: On-device real-time hand tracking," *arXiv preprint arXiv:2006.10214*, 2020.

[56] N. K. Zirzow, "Signing avatars: Using virtual reality to support students with hearing loss," *Rural Special Education Quarterly*, vol. 34, no. 3, pp. 33–36, 2015.