2024 International Conference on IT and Industrial Technologies (ICIT)

Evaluation of Head Pose Estimation Algorithms for Sign Language Analysis

Allah Bux Sargano
Department of Linguistic, Literary and Aesthetic Studies
University of Bergen
Bergen, Norway

Sébastien Vandenitte
Department of Language and Communication Studies
University of Jyväskylä
Jyväskylä, Finland

Tommi Jantunen
Department of Language and Communication Studies
University of Jyväskylä
Jyväskylä, Finland

Vadim Kimmelman
Department of Linguistic, Literary and Aesthetic Studies
University of Bergen
Bergen, Norway

***Abstract*—Head pose estimation (HPE) can be used in sign language linguistics and gesture studies, particularly for quantitative assessment of head movements in terms of yaw, pitch, and roll. These measurements can be used to quantify and compare different types of head movements within and across sign languages and co-speech gesture. While optoelectronic camera-based motion capture systems are considered the gold standard for this purpose, their practicality is limited due to high costs, accessibility issues, and the need for markers on the signer's face. Despite the popularity of HPE as a research area, there has been limited validation of HPE algorithms based on RGB videos, which is more suitable for sign language analysis, and almost no testing using sign-language specific data. This study addresses this gap by providing an overview of existing HPE algorithms and evaluating three state-of-the-art algorithms—MediaPipe, OpenFace, and 6DRepNet—using RGB sign language videos from a MoCap dataset of Finnish Sign Language. The accuracy of these algorithms is compared against an optoelectronic camera-based motion capture system recordings of the same data. The results indicate a good performance of all three algorithms for measuring yaw (with some advantages of MediaPipe), and worse performance for measuring pitch and roll.**

***Index Terms*—HPE, Motion Capture, Sign Language**

I. Introduction

Head pose estimation (HPE) is aimed at predicting the human head orientation from images and videos, and it is a crucial step in many computer vision-based applications. These applications include, but are not limited to, attention estimation [1], face recognition [2], human-robot interaction [3], and sign language analysis [4]. Head pose estimation (HPE) is a challenging task due to factors such as occlusion, varying illumination conditions, and facial expression variations, all of which can negatively impact the performance of computer vision algorithms. Additionally, the orientation of the head relative to the camera's field of view is crucial for accurate face recognition and movement estimation [5]. Head pose estimation typically involves three key steps: face detection, face landmarks localization, and head angle estimation. Face detection identifies the presence or absence of a face in an image or video. Face landmarks localization pinpoints the locations of key facial landmarks. Finally, head angle estimation determines the 3D orientation of the head. This orientation is expressed through Euler angles—pitch (rotation around the x-axis), yaw (rotation around the y-axis), and roll (rotation around the z-axis) [6].

Initially, handcrafted feature-based methods were predominantly used for face detection. These methods are computationally efficient, capable of running on CPUs, and require less storage than deep learning-based approaches. Among these, the Viola-Jones algorithm [7] has been widely applied in various computer vision tasks. This method is effective when the person directly faces the camera but struggles with partial occlusions, such as when the face is partially obscured. Generally, handcrafted methods have several limitations, such as their reliance on non-robust features, inability to handle unstructured data, and low performance when learning more complex patterns [8].

However, the literature indicates that deep neural networks, with their learned features, are robust to significant variations in facial appearances, consistently delivering state-of-the-art results [8]. For instance, Zhang et al. [8] introduced FaceBoxes, a face detector that demonstrated superior performance across several benchmark datasets, including Annotated Faces in the Wild (AFW) [9], Face Detection Dataset and Benchmark (FDDB) [10], and WIDER FACE [11]. Despite its effectiveness in face detection, FaceBoxes is limited to this task and does not address other critical steps in HPE.

Several algorithms have been developed that extend beyond basic face detection to include landmark estimation. For example, the DLIB toolkit [12] can estimate 68 key facial landmarks. However, DLIB primarily targets frontal faces

This project is funded by the European Union (ERC, Nonmanual, project number 101039378). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. This study has also been supported by the Research Council of Finland under grant 339268.

and encounters difficulties with profile views or faces that are heavily occluded. Other landmark detectors have been introduced to overcome challenges related to posture, angle, and distortion. The Multitask Cascaded Convolutional Neural Network (MTCNN) [13], designed for face alignment and pose variance assessment, identifies five key landmark points in real-time. This approach was later enhanced into EMTCNN [14], which adds layers to estimate 68 facial landmark points in real-time.

Subsequently, Lugaresi et al. [15] introduced the Google MediaPipe library, an open-source, cross-platform tool for face and body detection and pose estimation. MediaPipe supports real-time video and data stream processing and can detect up to 468 facial landmark points, providing their precise coordinates. However, MediaPipe does not provide a mechanism to computer the Euler angles directly. Fortunately, Pouw [16] provides a method to derive Euler angles from facial landmarks.

After identifying facial landmarks, the final step in head pose estimation is to determine the 3D orientation of the head. Cao et al. [17] developed the OpenPose toolkit for estimating both head and body poses. While OpenPose provides robust pose estimation, it is computationally demanding, limiting its effectiveness for real-time applications, and shows reduced performance on benchmark datasets like FDDB [10] due to challenges in handling pose variability.

In contrast, Baltrusaitis et al. [18] introduced OpenFace 2.0, a tool designed for comprehensive facial behavior analysis, which includes facial landmark localization, head pose estimation, eye gaze tracking, and facial expression recognition, leveraging the DLIB face detector. OpenFace 2.0 performs reliably under challenging conditions, such as partial occlusion, non-frontal views, and low illumination. However, when a significant portion of the face is occluded, certain landmarks cannot be accurately detected, resulting in less precise head orientation estimation.

Recently, Hempel et al. [19] proposed a 6DRepNet and 6DRepNet360 for estimating the full range of head orientations, especially the later one for calculating the rotation angles at Six degrees of freedom (6DoF). This model was trained on the enhanced pose variation dataset CMU Panoptic [20]; for this purpose, an automatic head pose labeling process was applied, and samples for the back of the head were generated. Then, these samples were combined with other benchmark datasets to achieve expanded head rotation variations, which were subsequently used to train the model.

Head pose estimation (HPE) tools offer significant potential for enhancing sign language analysis. Sign languages are natural languages used primarily in deaf communities around the world. Importantly, sign languages employ not just the hands, but also body and head movements and facial expressions to convey linguistic information [21]. Specifically, sign languages use head nods and tilts (pitch and roll movements) to mark questions, affirmation, and prosodic boundaries, and head shakes (yaw movement) to express negation. It is crucial to study these aspects of sign language production for a better understanding of the grammar and use of these languages, as well as for applied purposes such as sign language recognition and translation [22]. Some recent studies have used CV solutions, including HPE, for conducting linguistic analysis of sign language data. For example, Kimmelman et al. [23] used OpenFace to measure and study negative headshakes in the Sign Language of the Netherlands, see also [24] for a study of rhythm of head movements in Finnish Sign Language. However, while such methods are being applied, the measurements typically are not and cannot be validated against ground truth, as the majority of sign language data consists of simple RGB recordings. The dataset employed in this study is unique as it consists of motion capture (MoCap) data and synchronized video recordings. The MoCap system is used to track the three-dimensional locations of the movement with the help of reflective markers placed on the participants' bodies.The data obtained from these markers is considered ground truth for evaluating HPE algorithms on corresponding videos.

While numerous head pose estimation methods have been proposed in the literature, their application in sign language recognition and analysis remains underexplored and largely unvalidated. Earlier work in this direction was performed by Karppa et al. [25]. They used the Viola-Jones cascade face detector [7] for signer face detetcion. Then, skin colour regions were determined using multivariate Gaussian distribution; thus, interconnected skin regions were classified as hands or face using heuristic rules. For the motion tracking, they used Kanade-Lucas-Tomasi (KLT) [26]. Luzardo et al. [27] also developed and evaluated a HPE approach using a single sign language video with MoCap data, and achieved high results for yaw and roll estimation. Since then, there has been massive progress in computer vision tools with the emergence of deep learning strategies. However, little attention has been paid to comparing modern-day computer vision solutions against the MoCap gold standard for sign language analysis.

The proposed framework is not only crucial for sign language analysis but also has important applications in other fields, such as speech and gesture analysis. For instance, research has explored how speech and gesture interact within communication, highlighting the synchronization between verbal and non-verbal cues. Moreover, studies suggest that, like in sign languages, speaker head movement follows specific patterns and serves different semantic, communicative, and discourse functions [28]. Visual body signals such as facial expressions, gaze, and gestures are fundamental to human communication, both spoken and signed. Human languages have evolved as a multimodal flexible system that uses not only spoken words but also non-iconic gestures and body signals, adding depth to how we communicate in everyday life [29].

This study evaluates the effectiveness of the MediaPipe [15], OpenFace [18], and 6DRepNet [19] algorithms by benchmarking them against an optoelectronic gold standard for sign language applications using a real dataset from Finnish Sign Language (FinSL). The findings from this study provide valuable insights into the suitability of these approaches for sign

language analysis. This paper's contributions are threefold: (1) a comprehensive review of state-of-the-art algorithms for head pose estimation, highlighting the strengths and limitations of each; (2) a comparative assessment of three promising algorithms, namely OpenFace, MediaPipe, and 6DRepNet—against an optoelectronic gold standard baseline using a sign language dataset; and (3) an analysis of key findings and their implications for future research, offering guidance for both sign language research and broader applications in human pose estimation (HPE) and gesture analysis. The overall framework is illustrated in Figure 1, with further details provided in the following sections.

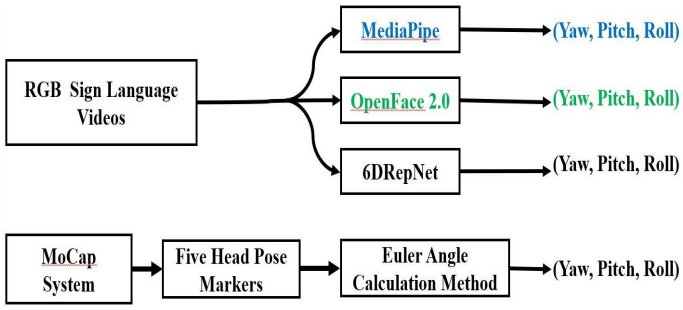

Fig. 1. Overall Proposed Framework for Experimentation and Analysis

II. Methods

This section describes the head pose estimation algorithms used to evaluate performance against the gold standard and explains the experimental setup employed for the gold standard and HPE algorithms in detail.

A. Dataset

The Finnish Sign Language MoCap Corpus consists of videos, numerical motion capture data, and text annotations of short narratives in FinSL collected in 2017 at the MoCap Laboratory of the University of Jyväskylä [30]. The participants were 6 signers (3 females and 3 males) aged 30-60 who were exposed to FinSL from birth by their primary caregivers or by their peers at an early age. Participants told at least 5 textless comic strips to an addressee, resulting in a total of 33 stories (21 minutes). Motion capture data was collected using an 8-camera Qualisys Oqus optical motion capture system operating at 120Hz. This system tracked the three-dimensional locations of 25 ball-shaped reflective markers placed on the participants' bodies. The video data was captured by a Full HD 30 fps camera synchronized with the Qualisys system. Participants also wore a head-mounted eye movement tracker that independently recorded the movement and gaze direction of the left eye. All data were imported into the ELAN annotation software [31], synchronized and annotated with signs and translations. In the present study, only the video and numerical motion capture data were used.

Finnish Sign Language MoCap Corpus provides 3D coordinates of each motion capture marker placed on the signer's body. For head pose estimation, it provides five markers: two on the front of the head (left and right), two on the back of the head (left and right) and one on the chin. However, the corpus does not provide Euler angles directly. Thus, these need to be computed from the motion markers. The following algorithm was developed to calculate Euler angles from these five markers (see the project repository linked below for the full algorithm).

The MoCap Algorithm:

1) Input: XYZ coordinates of five markers (HeadFL, HeadFR, HeadBL, HeadBR, Chin)
2) Calculate the midpoints (mid-front, mid-back, mid-left, and mid-right) and centroid of the four head markers.
3) Calculate front, side, and vertical orientation vectors
4) Normalize orientation vectors to unit vectors
5) Check the orthogonality and correct using the Gram-Schmidt process if required
6) Compute rotation matrix R from orthonormal orientation vectors
7) Convert rotation matrix R to Euler angles (yaw, pitch, and roll)

B. HPE Algorithms

The number of pose estimation algorithms designed for both constrained and unconstrained environments has significantly increased over the last decade [32]. These methods initially included appearance-based, geometric-based, multi-task, tracking-based, embedding-based, and regression-based approaches, which were prevalent before the deep learning era. However, the advent of deep learning models has substantially improved the performance of Human Pose Estimation (HPE) algorithms in diverse environments [33]. A recent analysis by Hammadi et al [6] highlighted that deep learning-based methods, particularly MediaPipe [15] and OpenFace [18], achieved superior performance on standard benchmark datasets like 300VW [34], as measured by Normalized Mean Error (NME). Based on this analysis, MediaPipe [15], OpenFace [18], and 6DRepNet [19] were selected for experimentation on the Finnish Sign Language dataset, with their results compared against the optoelectronics gold standard (ground truth). These algorithms, applicable to both images and videos, are available as open-source tools for the research community. Further details on the selected algorithms are provided in the subsequent sections.

1) MediaPipe: MediaPipe [15] is a deep learning-based framework developed by Google for estimating body and head poses from a continuous stream of images. This framework consists of multiple models, including MediaPipe Holistic (MPH), which is one of the multifaceted tools consisting of various sub-models. These sub-models include the pose sub-model to define the skeletal structure of the human body, the face sub-model to capture facial landmarks, and the hands sub-model to track hand gestures and movement.These sub-models together create a comprehensive landmark representation of the entire human body. The MediaPipe Holistic (MPH) model outputs a total of 543 landmarks in real time, including 33

pose landmarks, 468 facial landmarks, and 21 landmarks per hand.

As this study is focused on head pose estimation, 468 face landmarks are considered in three dimensions for head rotation estimation for sign language analysis. It is worth mentioning that MediaPipe does not provide a mechanism to automatically compute the Euler angles from these face landmarks. In this regard, we employed the mechanism provided by Thierfelder [35] and Pouw [16] to compute the yaw, pitch, and roll for the sign language videos.

2) OpenFace: OpenFace 2.0 [18] is a popular open-source toolkit for facial behavior analysis, supporting facial landmark detection, pose estimation, gaze tracking, and action unit recognition. It uses the Convolutional Experts Constrained Local Model (CE-CLM) [36], which localizes landmarks through pixel-level probability response maps. CE-CLM consists of two main components: the Point Distribution Model (PDM) for regulating landmark position and shape, and patch experts for handling local appearance variations. OpenFace accepts input from webcams, video files, image sequences, and individual images, outputting data such as gaze vectors, facial landmarks, shape parameters, head pose, and action units. It also automatically computes Euler angles for head pose, eliminating the need for additional algorithms.

3) 6DRepNet: 6DRepNet [19] is a recently proposed model for estimating the extended range of head orientations. This model has two variations, 6DRepNet and 6DRepNet360. It utilises ResNet50 as its backbone deep learning network for feature extraction. While 6DRepNet360 is designed to cover the full 6D rotation, it requires depth input to function effectively. The model was developed to estimate head poses across the full range of possible orientations and to provide a continuous 6D rotation matrix representation. This approach avoids the ambiguity and instability issues commonly associated with Euler angles and quaternions. Additionally, the 6D representation is paired with a geodesic loss function, which stabilises the learning process by more accurately reflecting the geometry of rotation spaces. In this study, we evaluated a pre-trained 6DRepNet model on a sign language dataset.

III. Results

The outputs of the three algorithms and the head rotation measurements calculated from the MoCap data were analyzed with R [37] in Rstuido [38]. The full data files and scripts can be accessed in the following repository https://osf.io/7scmj/?view_only=9027c7a144bf4cc78ca3059461933627.

A total of 33 videos from the The Finnish Sign Language MoCap Corpus were analyzed using OpenFace 2.0, MediaPipe, and 6DRepNet. It is important to note that the head-mounted eye movement tracker worn on the head of the signers partially occluded the head, leading to instances where these algorithms failed to detect the head, resulting in no Euler angle calculations. Among the three solutions, MediaPipe proved to be more robust, with a detection rate of 96% percent, compared to approximately 70% for the other two algorithms.

TABLE I
Correlations between the CV solutions and MoCaP

Solution	Yaw (range)	Pitch (range)	Roll (range)
MediaPipe	0.87 (0.54:0.96)	0.02 (-0.32:0.79)	0 (-0.5:0.4)
OpenFace	0.8 (0.26:0.95)	0.21 (-0.53:0.9)	0.54 (0.04:0.87)
6DRepNet	0.51 (-0.19:0.9)	0.1 (-0.28:0:58)	0.2 (-0.2:0.84)

Figure 2 illustrates the general trend of these three algorithms in comparison to the MoCap gold standard, for one video recording (sl_p6_201702130004_video2), where the head detection rate was highest for all the three algorithms, only for frames 200 to 500. Note that the CV solutions are scaled and centered approximately to the MoCaP outputs for the ease of comparison. It can be observed that, for yaw, the three solutions are in good agreement with the MoCap, although the 6DRepNet shows some outliers. For pitch again, MediaPipe and OpenFace seem to perform well, albeit with some outliers, and, for roll, MediaPipe shows the worst performance, and 6DRepNet again shows some outliers.

The performance of these algorithms was assessed by calculating the correlations between the gold standard (ground truth) and the angles computed by MediaPipe, OpenFace, and 6DRepNet, as shown in Table I. Note that, due to the low head detection rate for OpenFace and 6DRepNet, for these two algorithms we decided to only use intervals where the head was successfully detected in at least 10 consecutive frames, which leaves approximately half of the data out of consideration. The results indicate that MediaPipe achieved the best performance for yaw with a correlation of 0.87, followed by OpenFace 2.0 with 0.8, and 6DRepNet with 0.51. For pitch, the performance is on average bad for all the solutions; OpenFace seems to perform best, but some files show extremely low performance. For roll, the performance is extremely low for MediaPipe, and it is best for OpenFace, even though some files are wi low performance.

In Table II, we report Mean Absolute Errors to evaluate the performance of the three solutions. Because the solutions and the MoCap data are centered and also scaled differently (see the full code for details), to calculate the error, we build mixed effect linear regressions predicting MoCap measures from each model with random intecrept per video file, but without random slopes for scaling, and calculate the error based on the difference between the predicted and observed values of MoCaP. In order to interpret the magnitude of the error, which is reported in degrees, we also report it scaled by IQR of the corresponding MoCap measure. We use IQR and not SD for scaling due to the presence of some outlier errors in the MoCap measurements.

From this table we can see that, for yaw, MediaPipe performs considerably better than the other models, with much smaller errors (also relative to variation in the data), while 6DRepNet performs worse, with the error twice as large. For pitch, the performance of MediaPipe and 6DRepNet is very low, with high errors, and while the error of OpenFace is lower, it is still higher than for yaw. For roll, focusing on the error

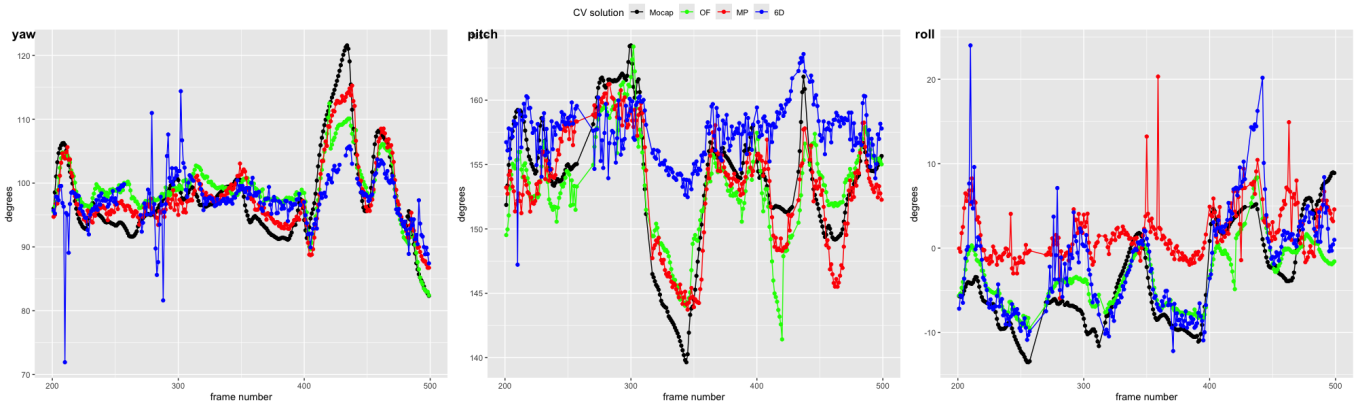

Fig. 2. Visualizing yaw, pitch and roll for MoCap (black), MediaPipe (red), OpenFace (green) and 6DRepNet (blue). The three CV solutions are shifted and scaled towards the MoCap measures.

TABLE II
MAE FOR THE CV SOLUTIONS, ALSO SCALED BY IQR

Solution	Yaw (scaled)	Pitch (scaled)	Roll (scaled)
MediaPipe	3.5 (0.26)	13.7 (0.94)	5.3 (0.67)
OpenFace	4.3 (0.37)	6.7 (0.52)	4.4 (0.63)
6DRepNet	7.3 (0.54)	16.9 (1.26)	5.38 (0.73)

scale by variation, all solutions perform worse than for yaw but, with the exception of OpenFace, better than for pitch.

IV. DISCUSSION

This study evaluated the effectiveness of MediaPipe, OpenFace, and 6DRepNet for estimating head pose from videos, using an optoelectronic camera system as the gold standard. The results indicate that MediaPipe is more robust for head rotation estimation in sign languages compared to OpenFace and 6DRepNet. First, even in the presence of a head-mounted eye movement tracker occluding the face, MediaPipe was able to detect the head in over 95% of cases, in comparison to 70% for the other two solutions.

Second, the estimations of yaw for MediaPipe were closer to the gold standard than of the other instruments. Specifically, MediaPipe achieved a Mean Absolute Error (MAE) of 3.5° for yaw, which is approximately 25% of the IQR of MoCap yaw in the dataset. In contrast, OpenFace exhibited a yaw MAE of 4.3° and 6DRepNet of 7.3°, which is also higher relative to variation. For pitch, only OpenFace shows relatively good performance, while the other two instruments perform very poorly. For roll, all the instruments perform worse than for yaw. The lower performance for roll, and especially for pitch, can be due to the occlusion of the head and also the camera position (the camera was located higher than the signer). More research with other datasets (also without occlusion of the face) is necessary to more realistically evaluate the performance of CV solutions for head rotation estimation.

However, some practical conclusions can also be drawn for the current study. In the context of sign language analysis, where measurement accuracy is critical and some degree of occlusion is often unavoidable (as the signers' hands often move in front of their faces), MediaPipe's ability to provide reliable face detection and head rotation measurements at least for yaw, even in challenging environments, offers a significant advantage. Furthermore, for yaw specifically, all the instruments have demonstrated relatively good performance, and especially so MediaPipe. We find the MAE for yaw with MediaPipe to be 3.5°, which makes it possible to use MediaPipe to measure this type of head rotation in sign languages and gesture data for quantitative analysis. Previous research [23] indicates that average maximal amplitude of yaw rotation used in negative headshake in sign languages is around 15°, so at least relatively large changes in head rotation can be detected with MediaPipe.

V. CONCLUSION

This study evaluated three advanced computer vision algorithms—MediaPipe, OpenFace, and 6DRepNet—for head pose estimation for sign languages, using RGB video data and benchmarking them against a MoCap-based optoelectronic gold standard. Results show that while all three algorithms align well with ground truth for yaw estimation, their accuracy for roll and especially pitch is less stable, indicating an area for future improvement.

MediaPipe demonstrated the highest robustness in head pose estimation, especially for yaw, maintaining accuracy even with large head movements and partial occlusions—common in dynamic sign language videos. This robustness suggests that MediaPipe is well-suited for real-world sign language analysis, where variable head orientations and occlusions are typical.

Conversely, while OpenFace and 6DRepNet provided satisfactory yaw estimates, they were less reliable under challenging conditions, struggling with occlusions and large head movements. This suggests their suitability for controlled environments with minimal occlusions but limited applicability in dynamic settings like sign language interpretation.

Overall, this evaluation highlights each algorithm's strengths and limitations, offering insights into their suitability for head pose estimation in sign languages. MediaPipe stands out as

a robust option for yaw estimation under varied conditions, though improvements in pitch and roll estimation remain essential to meet the specific demands of sign language and gesture analysis. Future research could explore adaptive or hybrid approaches to better handle occlusions and large movements.

REFERENCES

[1] Veronese, A., Racca, M., Pieters, R. S., & Kyrki, V. (2017). Probabilistic Mapping of human Visual attention from head Pose estimation. Frontiers in Robotics and AI, 4, 53.

[2] Chang, F. J., Tuan Tran, A., Hassner, T., Masi, I., Nevatia, R., & Medioni, G. (2017). Faceposenet: Making a case for landmark-free face alignment. In Proceedings of the IEEE International Conference on Computer Vision Workshops (pp. 1599-1608).

[3] Strazdas, D., Hintz, J., & Al-Hamadi, A. (2021). Robo-hud: Interaction concept for contactless operation of industrial cobotic systems. Applied Sciences, 11(12), 5366.

[4] Kuznetsova, A., & Kimmelman, V. (2024). Testing MediaPipe Holistic for Linguistic Analysis of Nonmanual Markers in Sign Languages. arXiv preprint arXiv:2403.10367.

[5] Murphy-Chutorian, E., & Trivedi, M. M. (2008). Head pose estimation in computer vision: A survey. IEEE transactions on pattern analysis and machine intelligence, 31(4), 607-626.

[6] Hammadi, Y., Grondin, F., Ferland, F., & Lebel, K. (2022). Evaluation of various state of the art head pose estimation algorithms for clinical scenarios. Sensors, 22(18), 6850.

[7] Viola, P., & Jones, M. (2001, December). Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001 (Vol. 1, pp. I-I). IEEE.

[8] Zhang, S., Wang, X., Lei, Z., & Li, S. Z. (2019). Faceboxes: A CPU real-time and accurate unconstrained face detector. Neurocomputing, 364, 297-309.

[9] Zhu, X., & Ramanan, D. (2012, June). Face detection, pose estimation, and landmark localization in the wild. In 2012 IEEE conference on computer vision and pattern recognition (pp. 2879-2886). IEEE.

[10] Jain, V., & Learned-Miller, E. F. A Benchmark for Face Detection in Unconstrained Settings. University of Massachusetts; Amherst, MA. USA: 2010. Technical Report, UMass Amherst Technical Report.

[11] Yang, S., Luo, P., Loy, C. C., & Tang, X. (2016). Wider face: A face detection benchmark. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5525-5533).

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

[13] Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. IEEE signal processing letters, 23(10), 1499-1503.

[14] Kim, H., Kim, H., & Hwang, E. (2019, February). Real-time facial feature extraction scheme using cascaded networks. In 2019 IEEE International Conference on Big Data and Smart Computing (BigComp) (pp. 1-7). IEEE.

[15] Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., ... & Grundmann, M. (2019). Mediapipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172.

[16] Pouw, W. (2024). Wim Pouw's EnvisionBOX modules for social signal processing (Version 1.0.0), Computer software. https://github.com/WimPouw/envisionBOX_modulesWP

[17] Cao, Z., Simon, T., Wei, S. E., & Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7291-7299).

[18] T. Baltrusaitis, A. Zadeh, Y. C. Lim and L. -P. Morency, "OpenFace 2.0: Facial Behavior Analysis Toolkit," 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 2018, pp. 59-66, doi: 10.1109/FG.2018.00019.

[19] Hempel, T., Abdelrahman, A. A., & Al-Hamadi, A. (2024). Toward Robust and Unconstrained Full Range of Rotation Head Pose Estimation. IEEE Transactions on Image Processing, 33, 2377-2387.

[20] Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., ... & Sheikh, Y. (2015). Panoptic studio: A massively multiview system for social motion capture. In Proceedings of the IEEE international conference on computer vision (pp. 3334-3342).

[21] R. Pfau and J. Quer, "Nonmanuals: their prosodic and grammatical roles," in Sign Languages, D. Brentari, Ed., Cambridge: Cambridge University Press, 2010, pp. 381–402.

[22] El-Din, S. A. E., & Abd El-Ghany, M. A. (2020, October). Sign Language Interpreter System: An alternative system for machine learning. In 2020 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES) (pp. 332-337). IEEE.

[23] Kimmelman, V., M. Oomen & R. Pfau (2024). "Headshakes in NGT: Relation between Phonetic Properties & Linguistic Functions". In Proceedings of LREC-SL 2024.

[24] Jantunen, T., Mesch, J., Puupponen, A., Laaksonen, J. (2016) On the rhythm of head movements in Finnish and Swedish Sign Language sentences. Proc. Speech Prosody 2016, 850-853, doi: 10.21437/SpeechProsody.2016-174

[25] Karppa, M., Jantunen, T., Viitaniemi, V., Laaksonen, J., Burger, B., & De Weerdt, D. (2012, May). Comparing computer vision analysis of signed language video with motion capture recordings. In LREC (pp. 2421-2425).

[26] Shi, J. (1994, June). Good features to track. In 1994 Proceedings of IEEE conference on computer vision and pattern recognition (pp. 593-600). IEEE.

[27] Luzardo, M., M. Karppa, J. Laaksonen & T. Jantunen (2013). Head pose estimation for Sign Language video. In J.-K. Kamarainen & M. Koskela (Eds.), Image Analysis [18th Scandinavian Conference, SCIA 2013, Espoo, Finland, June 17-20, 2013. Proceedings], pp. 349-360. Lecture Notes in Computer Science, Vol. 7944. Springer.

[28] P. Wagner, Z. Malisz, and S. Kopp, "Gesture and speech in interaction: An overview," *Speech Communication*, vol. 57, pp. 209-232, 2014.

[29] J. Holler, "Visual bodily signals as core devices for coordinating minds in interaction," *Philosophical Transactions of the Royal Society B*, vol. 377, no. 1859, pp. 20210094, 2022.

[30] Jantunen, T., Wainio, T., & Burger, B. (2022). Project data of ShowTell–Finnish Sign Language MoCap corpus. Jantunen, Tommi; Puupponen, Anna; Hernández Barros, Doris; Wainio, Tuija; Keränen, Jarkko; Salonen, Juhana; & Burger, Birgitta. Project data set of ShowTell–Showing and telling in Finnish Sign Language. University of Jyväskylä. https://doi.org/10.17011/jyx/dataset/89372. https://doi.org/10.17011/jyx/dataset/89371.

[31] Crasborn, O., & Sloetjes, H. (2008). Enhanced ELAN functionality for sign language corpora. In 6th International Conference on Language Resources and Evaluation (LREC 2008)/3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora (pp. 39-43).

[32] Khan, K., Khan, R. U., Leonardi, R., Migliorati, P., & Benini, S. (2021). Head pose estimation: A survey of the last ten years. Signal Processing: Image Communication, 99, 116479.

[33] Asperti, A., & Filippini, D. (2023). Deep learning for head pose estimation: A survey. SN Computer Science, 4(4), 349.

[34] Shen, J., Zafeiriou, S., Chrysos, G. G., Kossaifi, J., Tzimiropoulos, G., & Pantic, M. (2015). The first facial landmark tracking in-the-wild challenge: Benchmark and results. In Proceedings of the IEEE international conference on computer vision workshops (pp. 50-58).

[35] https://medium.com/@susanne.thierfelder/head-pose-estimation-with-mediapipe-and-opencv-in-javascript-c87980df3acb

[36] Zadeh, A., Chong Lim, Y., Baltrusaitis, T., & Morency, L. P. (2017). Convolutional experts constrained local model for 3d facial landmark detection. In Proceedings of the IEEE international conference on computer vision workshops (pp. 2519-2528).

[37] R Core Team, R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing, 2022. [Online]. Available: https://www.R-project.org/

[38] Posit team, "RStudio: Integrated development environment for R," Posit Software, PBC, Boston, MA, manual, 2024. [Online]. Available: http://www.posit.co/