A study on estimating the accurate head IMU motion from Video

Minyen LU[1] Chenhao CHEN[1] Shigemi ISHIDA[2] Yugo NAKAMURA[1]
Yutaka ARAKAWA[1]

Abstract: Inertial measurement unit (IMU) data have been utilized in human activity recognition (HAR). In recent studies, deep learning recognition for IMU data has caught researchers' attention for the capability of automatic feature extraction and accurate prediction. On the other hand, the challenge of data collection and labeling discourages researchers to step into it. IMUTube provides a solution by building up a pipeline to estimate virtual IMU data from YouTube videos for body motion. For head motion data, several methods, such as OpenFace 2.0 provide the function of predicting facial landmarks and calculating head facing angle from video. However, to our knowledge, there is no study focusing on estimating IMU data from human head motion. In our previous work DisCaaS, we created the M3B dataset which contains IMU and 360-degree video data from the meeting. We exploit head motion data extraction models to predict participants' nodding and speaking gestures. In order to further improve the performance of nodding recognition, in this paper, we are interested in understanding the quality of estimated gyro data calculated from these existing head motion models. We investigate the difference between the motion data estimated from video and those measured by a 9-axis sensor not only in the time domain but also in the frequency domain. Finally, we discuss the future direction of the result.

1. Introduction

Recent years, under influence of COVID-19, the number of online meetings has increased noticeably. According to the meetings statistics provided by Zippa [1], there are around 55 million meetings held each week in the United States. That means over a billion of meetings are held per year. However, although employees spent such amount of times and energy on meetings, if the meeting turns out to be unproductive, it becomes a huge waste of companies' resources. As their research shows, 24 billion is a considerable amount of productive work hours being lost. Without a doubt, the requirement to improve the productivity and efficiency of meetings is crucial.

After the publication survey, several characteristics have been found to be determinative of the behavior of participants during the meeting. These can be concluded to three categories: appearances [2], verbal information [3], and non-verbal information [4]. Our previous work DisCaaS [5] aimed to use not content-sensitive information to protect users' privacy, in other words, we applied a sensor device that does not collect the text or content of a meeting. Meanwhile, we also want to ensure the system's reproductivity, so a contactless device is selected.

In human activity recognition (HAR), researchers are attracted by the magnificent performance of deep learning solutions on sensor-based tasks. However, the challenge of data collection and labeling discourages researchers from exploring sensor-based applications based on Deep Learning methods. In order to solve the data sparsity problem in this field, several researchers start to propose solutions that simulate IMU data from video. OpenFace [6] estimates the facial and eye landmarks' location of the person in the video. Based on landmark information, they provide functions including head pose estimation, gaze estimation, and facial expression prediction. IMUtube [7] transforms the full body of each person in a video into a skeleton body in 3-dimensional space. With the full 3D motion information, they tracked the acceleration and orientation changes in the body's joints, and extract the virtual IMU information from the video. Compared to IMUtube, recent research, CROMOSim [8], emphasizes that instead of simulation based on skeleton structure, they simulated virtual IMU data based on 3D body extracted from 3D-skinned models. The sensor position can be located closer to reality, which enables the system to simulate virtual sensor data with higher fidelity.

Inspired by these virtual IMU extraction approaches, DisCaaS extracts facial landmark position data and head motion data from video with OpenFace tool kits. We used machine learning models to conduct prediction of participants' gestures on estimated data. However, since the method of recognizing the nodding gesture based on estimated head motion data did not reach the expected performance, we want to understand the fidelity of gyro data estimated from OpenFace.

[1] Kyushu University
[2] Future University Hakodate

2. Related Work

2.1 Meeting Analysis

Meeting analysis can be based on three aforementioned features: *Participant Appearance*, *Verbal Communication*, and *Nonverbal Communication*.

2.1.1 Participant Appearance

[2, 9, 10] shows the importance of appearance such as age possesses a strong relationship with forgiveness and the number of counteractive behaviors. They also show how forgiveness and counteractive behaviors influence the team's harmony and productivity. Gent et al. [11] proposed a method for estimating the age-based facial aging patterns on FG-Net Aging [12] and the MORPH [13] Database. They extracted facial features with the appearance model [14], after which these features were processed by SVM to obtain the estimated age [15].

2.1.2 Verbal Communication

In verbal communication, context, pronunciation, vocal expression, and Speaking Duration are some import features that impact the quality of a meeting [3, 16, 17]. Therefore, there are some works focusing on detection of these features. For context, Yu and Dang [18] have proposed automatic speech recognition (ASR) to convert speech to text. For pronunciation, Zhang et al. [19] proposed a new end-to-end ASR system based on the hybrid connectionist temporal classification and attention (CTC/attention) architecture to improve advanced automatic pronunciation error detection (APED) algorithms. The new system was a suitable general solution for L1-independent computer-assisted pronunciation training (CAPT). For vocal expression, Zhao et al [20] proposed ROC, which is an online platform system allows ubiquitous access to rate people's communication skills. The rating system served as a feedback system that gradually improved the communication skills of the participants. Regarding the duration of the speech, our previous work DisCaaS applied OpenFace to achieve the detection of the duration of the speech of the participants.

2.1.3 Nonverbal Communication

For nonverbal communication, Pham et al. [21] proposed the estimation of 3D poses from a single RGB camera, which estimates body posture and activities using a camera. Zhao et al.'s ROC speak system include smile detection in their automated feedback process to analyze the smile intensity of the attendants. Zhang et al. proposed eye contact detection using a camera. DisCaaS [22] provide the function of nodding detection by a camera.

2.2 IMU Data Simulation

With the intention of solving the problem of data scarcity in the HAR field, recently researchers proposed methods that extract IMU data from video data. The most representative work is IMUTube [7], which combines multiple methods into a complete pipeline to extract IMU data from the readily available YouTube resource. Inspired by IMUTube, other researchers proposed some modifications based on their method to improve the accuracy of virtual IMU data. For head motion data, these methods provide head pose estimation.

2.2.1 IMUtube

IMUTube aims at converting video data collected from social media platforms into usable virtual sensor (IMU) data. Videos that capture activities of interest are first transformed into 2D-pose skeletons with the SOTA method OpenPose model [23]. After that, the 2D poses are upgraded to the 3D pose with the VideoPose3D model [24]. In order to localize the 3D pose of a frame in the 3D scene, combining information such as camera intrinsic, Ego-motion, and estimated Depth map, each 3D pose is then converted to the full 3D motion presented in 3D space. Finally, by tracking the acceleration and orientation changes of each joint of the body in the 3D space, IMUtube can successfully estimate virtual sensor data from each joint.

2.2.2 Extension work from IMUTube

Extended from this approach, Kwon et al. [25] points out three situations causing common errors when processing videos with IMUTube: occlusion and overlaps, motion blur, and "ghost" recognition. These situations happen when multiple people are in the same scene, the person moves much faster than the frame rate, or there are some objects resembling human structure in the background. To avoid such errors, they replace the bottom-up approach with a top-down approach. They first detect human bounding boxes with a visual person detector such as the YOLO human detector [26] and the AlphaPose model. Within this box, the 2D pose will be generated, which solves the problem of "ghost" recognition by removing the unrelated object behind the scene. For occlusion problems, they exploited occlusion and self-occlusion detection in the scene, and after that, they re-segment the whole sequence into unoccluded 2D-pose clips which are then further analyzed. They introduce a tuple used for fast movement detection. By setting a threshold on the changing range of shape, size, and position of the bounding box of a person in the frame sequence, the pipeline is capable of discarding the frame, which may cause errors.

2.2.3 CROMOSim

Based on the idea of simulating virtual sensor data from IMUTube, Hao et al. proposed their own pipeline named Cross-modality Inertial Measurement Simulator (CROMOSim). With the observation that the IMUTube approach is confined to skeleton body representation, they intended to develop a new framework that supports arbitrary user-specified placement and orientation of target sensors provided with high fidelity virtual data. To achieve this objective, unlike IMUTube, CROMOSim introduces 3D skinned multi-person linear (SMPL) models [27] to represent 3D human body poses. The SMPL model generates a 3D human body with a fine-grained full-body tri-mesh. In such a model, the virtual IMU sensor can capture realistic motion reflecting soft-tissue dynamics in a specific location.

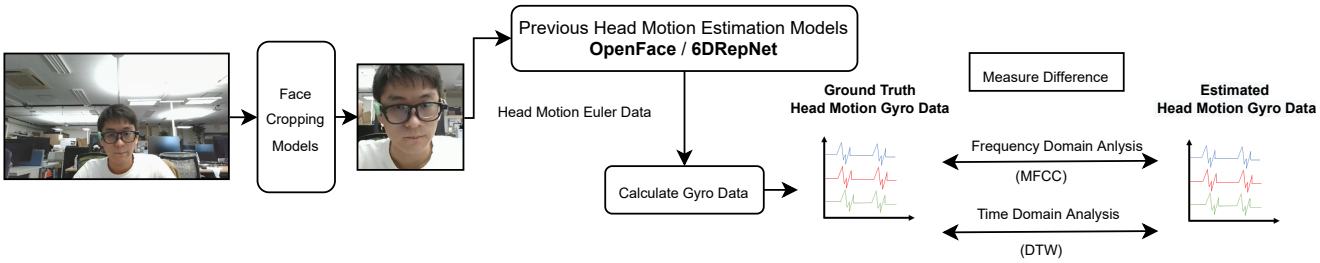


Fig. 1: Workflow of this study. The input Video is first preprocessed by face cropping models. After that, the video data is used to estimate head motion data by head motion estimation models. The output estimated motion data is then be compared with the ground truth data collected by IMU sensor.

Their proposed system constructs a 3D human body with two aspects: reconstruction of human global displacement and rotation and estimation of 3D in-place human motion and body shape. First, with the combination of Robust CVD [28] and OpenPose [23], they located the joint position of the person in the frame. After that, the SOTA method is used to directly estimate realistic 3D human poses and shapes. VIBE [29] is used to extract 3D body poses in a global frame and shape parameters from video data, which is then used in the generation of SMPL body meshes. Due to the noisy result of the previous step, the direct calculation of accelerations and angular velocities on SMPL models could be erroneous. Thus, they designed a neural network model to estimate the virtual IMU data.

These previous papers provide the SOTA solution for IMU simulation and its optimization methods, but all focus on body movement IMU data simulation. In this regard, we want to know the performance of the current head motion estimation model performance on estimating virtual data from video.

3. Proposed Method

This section introduces the process of data creation, cropping videos, and head motion data extraction.

Before feeding video data to head motion estimation models, We first exploit the face tracking function in the work of Syncnet [30] to crop frames in videos to frames only focus on the participant's face. As for head motion estimation models, we choose one model from each of the landmark-free methods (6DRepNet) and landmark-based methods (OpenFace). Finally, we performed a performance analysis on the result data in both the frequency domain and the time domain. Mel Frequency Cepstral Coefficients (MFCC) is used for frequency-domain analysis. As for time domain analysis, dynamic time warping (DTW) is applied. The workflow is shown in Fig. 1.

3.1 Data Creation

As Fig. 2 shows, the participant is asked to wear glasses equipped with an IMU sensor and performs nodding gestures randomly under instruction. We use the standard sensor,

(a) Front view (b) Side view

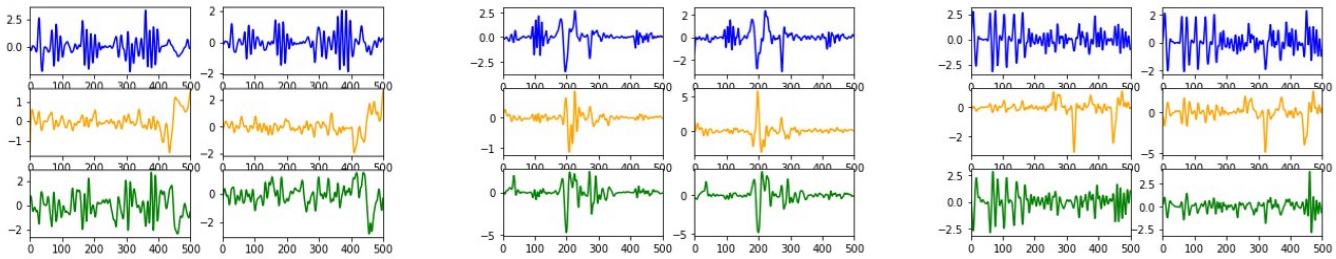
Fig. 2: View of sensor setting

the Metamotions 10-axis IMU sensor[*1]. Since the sampling frequency of the virtual motion data by head motion estimation models is dependent on the video, which is usually about 30 fps, we set our gyro-data sampling rate at 25Hz, which is supported by Metamotions. In the meantime, we record the video with webcams. The video data is then downsampled to 25Hz.

3.2 Face Tracking

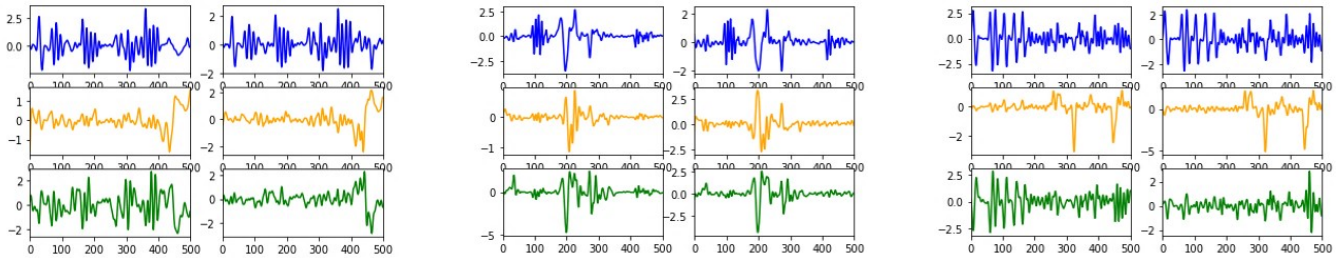
Facetrack aims to solve the common lip-sync problem, which usually happens in TV broadcasting. Facetrack provided a language-independent and speaker-independent solution – SyncNet, which uses only the video and the audio streams. Their main contribution is introducing the ConvNet Architecture and a complete data processing pipeline that does not require labeled data. SyncNet contains two asymmetric streams for audio and video. For the audio stream, the input MFCC type data is first encoded as a heatmap image. The VGG-M layer architecture is then applied to process the heatmap image. For the Visual stream, the input frames are transformed into gray-scaled images which contain the mouth regions of speakers. Chung and Zisserman's architecture [31] is applied to extract features from a video. The contrastive loss is used as a loss function, which ensures that the output of the audio and video networks are similar for genuine pairs and different for false pairs. The SyncNet processing pipeline is based on Chung and Zisserman [31]. Shot boundaries are determined by comparison between color histograms of consecu-

[*1] METAMOTIONR: A wearable device that offers real-time and continuous monitoring of motion and environmental sensor data – `https://mbientlab.com/metamotionr/`



(a) Subject 1 (b) Subject 2 (c) Subject 3

Fig. 3: Waveform of data from Sensor and OpenFace



(a) Subject 1 (b) Subject 2 (c) Subject 3

Fig. 4: Waveform of data from Sensor and 6DRepNet

tive frames [32]. After the HOG-based face detection [33] is performed, frames of detected face are then grouped using a KLT tracker [34]. Here we apply this preprocess method in our study.

3.3 OpenFace 2.0

OpenFace 2.0 is a tool designed for researchers in facial behavior analysis. OpenFace 2.0 provides functions such as facial landmark detection, head pose estimation, gaze estimation, and facial expression prediction. For facial landmark detection, OpenFace 2.0 utilize Convolution Experts Constrained Local Model (CE-CLM) [35], which includes the Point Distribution Model (PDM) to capture landmark shape variations and patch experts to model local appearance variations of each landmark. With the detected facial landmark, the accurate head pose can be estimated by solving the perspective-n-point problem [36]. In order to estimate eye gaze, OpenFace 2.0 detect eyelids, iris, and the pupil by exploiting a Constrained Local Neural Field (CLNF) landmark detector [37]. The detected pupil and the location of the eye are used to compute the eye gaze vector for each eye. As for facial expression recognition, OpenFace 2.0 uses the AU recognition framework by Baltrušaitis et al [38]. Linear kernel SVM are applied; they demonstrated that, despite their outdatedness, it is still competitive with other deep learning methods.

The output of OpenFace 2.0 includes landmarks location of eyes and faces, eye gaze vector and angles, several facial action units, head position and facing angle, etc. Because our work's objective is to predict the nodding motion of the attendants, here we only use the estimated head facing angle, which is presented as 3-axis vector (poseRx, poseRy, poseRz).

3.4 6DRepNet

Unlike previous work such as OpenFace 2.0, recent work applies landmark-free methods to estimate head pose. Landmark-based methods have the disadvantage that head motion angle detection is dependent on the correctness of landmark prediction, so they often suffer from the erroneous prediction of landmark position due to occlusion and extreme rotation in the video. In such an aspect, researchers provide different solutions based on deep neural networks to extract head motion data. 6DRepNet adopted the landmark-free method in their work. Previously, Zhou et al. [39], observed that the representation of rotation with four or fewer dimensions causes the discontinuity. However, most existing work applies quaternions or Euler angles as representations of rotation. Therefore, instead of applying quaternions and Euler angle representation, they used the rotation matrix. The matrix Representation has the orthogonality constraint $RR^T = I$. To enforce the output keeping its orthogonality, they follow the approach of Zhou et al. [39], which reduces the rotation matrix to a 6D rotation representation. Their work shows better results compared to other landmark free works both on the AFLW2000 [40] dataset and BIWI [41] dataset.

4. Evaluation

In this section, we will evaluate the performance of current head movement estimation models.

Since Openface and 6DRepNet target the human facing direction, their output data are both Euler angles. To extract gyro data from these outputs, We calculate the gyro data from Euler angle data by simple subtraction between each sample. The result gyro data is very noisy, so we filter the gyro data with a low-pass filter. Finally, we perform an

Table 1: DTW distance (Senor - OpenFace)

subject ID	x-axis	y-axis	z-axis	Avg
1	11.291	10.722	14.260	12.091
2	6.336	19.382	7.506	11.074
3	8.725	13.500	13.639	11.954

Table 2: DTW distance (Senor - 6DRepNet)

subject ID	x-axis	y-axis	z-axis	Avg
1	7.969	6.608	18.599	11.058
2	5.714	15.516	6.072	9.100
3	6.939	10.771	13.946	10.552

analysis on data from both the model and the ground truth data we collected from the IMU sensor in the time domain and frequency domain.

4.1 Time Domain Analysis

Figure 3 shows the waveform of the collected gyro data from the sensor and the estimated data from OpenFace. For each subfigure, the left is gyro data from the sensor and the right is from OpenFace. Similarly, Figure 4 shows the waveform of the collected gyro data from the sensor and estimated data from 6DRepNet. Both of the figures contain data performed by 3 different subjects. It is easy to see that the estimated data from both models successfully remain the pattern of movement in every case. The data from 6DRepNet show a more noisy result compared to OpenFace. However, although 6DRepNet seems to be noisy in the waveform, it shows better performance when we compute DTW, as shown in Tables 1, 2. This is reasonable since OpenFace often suffers from occlusion error compared to the 6DRepNet method.

4.2 Frequency Domain Analysis

For the frequency domain, we select Mel frequency cepstral coefficients (MFCCs) as our measure method. Librosa library[*2] is applied to calculate Mel coefficients of the gyro data. We choose the first 12 coefficients from the 20 coefficients, which is the default output setting of librosa library. The result is shown in Fig. 5 and Fig.6. 6DRepNet still shows better performance compared to the result of OpenFace. We also observe that the model's performance on the Y-axis (yaw) is worse compared to other axes (roll and pitch).

5. Conclusion

In this paper, we evaluate the accuracy of the gyro data estimated by two methods, the landmark-based method (OpenFace) and the landmark-free method (6DRepNet), by comparing with the data from the IMU sensor. The result shows that these existing models are capable of correctly capturing the head motion of the user from video. Based on this result, we will propose a generative compensation model that is expected to take the output of these existing models and generate gyro data with higher fidelity.

[*2] https://librosa.org/doc/main/index.html

Acknowledgments This work was partially supported by JSPS KAKENHI (JP20H04177) and Initiative for Life Design Innovation (iLDi) Platform for Society 5.0.

References

[1] Flynn, J.: 27 INCREDIBLE MEETING STATISTICS https://www.zippia.com/advice/meeting-statistics/.
[2] Schulte, E. M., Lehmann-Willenbrock, N. and Kauffeld, S.: Age, forgiveness, and meeting behavior: A multilevel study, *Journal of Managerial Psychology* (2013).
[3] McDorman, T.: Implementing existing tools: Turning words into actions–Decision-making processes of regional fisheries management organisations (RFMOs), *The International Journal of Marine and Coastal Law*, Vol. 20, No. 3, pp. 423–457 (2005).
[4] Scheflen, A. E.: The significance of posture in communication systems, *Psychiatry*, Vol. 27, No. 4, pp. 316–331 (1964).
[5] Watanabe, K., Soneda, Y., Matsuda, Y., Nakamura, Y., Arakawa, Y., Dengel, A. and Ishimaru, S.: Discaas: Micro behavior analysis on discussion by camera as a sensor, *Sensors*, Vol. 21, No. 17, p. 5719 (2021).
[6] Baltrušaitis, T., Zadeh, A., Lim, Y. C. and Morency, L.-P.: OpenFace 2.0: Facial behavior analysis toolkit, *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, IEEE, pp. 59–66 (2018).
[7] Kwon, H., Tong, C., Haresamudram, H., Gao, Y., Abowd, G. D., Lane, N. D. and Ploetz, T.: IMUTube: Automatic extraction of virtual on-body accelerometry from video for human activity recognition, *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 4, No. 3, pp. 1–29 (2020).
[8] Hao, Y., Wang, B. and Zheng, R.: CROMOSim: A Deep Learning-based Cross-modality Inertial Measurement Simulator, *arXiv preprint arXiv:2202.10562* (2022).
[9] Karremans, J. C. and Van Lange, P. A.: Forgiveness in personal relationships: Its malleability and powerful consequences, *European Review of Social Psychology*, Vol. 19, No. 1, pp. 202–241 (2008).
[10] Kauffeld, S. and Lehmann-Willenbrock, N.: Meetings matter: Effects of team meetings on team and organizational success, *Small Group Research*, Vol. 43, No. 2, pp. 130–158 (2012).
[11] Geng, X., Zhou, Z.-H. and Smith-Miles, K.: Automatic age estimation based on facial aging patterns, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 29, No. 12, pp. 2234–2240 (2007).
[12] Liu, K., Gao, Z., Lin, F. and Chen, B. M.: FG-Net: Fast Large-Scale LiDAR Point Clouds Understanding Network Leveraging Correlated Feature Mining and Geometric-Aware Modelling, *arXiv preprint arXiv:2012.09439* (2020).
[13] Ricanek, K. and Tesafaye, T.: Morph: A longitudinal image database of normal adult age-progression, *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, IEEE, pp. 341–345 (2006).
[14] Edwards, G. J., Lanitis, A., Taylor, C. J. and Cootes, T. F.: Statistical models of face images—Improving specificity, *Image and Vision Computing*, Vol. 16, No. 3, pp. 203–211 (1998).
[15] Huerta, I., Fernández, C., Segura, C., Hernando, J. and Prati, A.: A deep analysis on age estimation, *Pattern Recognition Letters*, Vol. 68, pp. 239–249 (2015).
[16] Shrivastava, S. and Prasad, V.: Techniques to Communicate in Virtual Meetings Amidst the New Normal — A Consideration, *Wutan Huatan Jisuan Jishu*, Vol. 16, pp. 73–92 (2020).
[17] Knowlton, G. E. and Larkin, K. T.: The influence of voice volume, pitch, and speech rate on progressive relaxation training: application of methods from speech pathology and audiology, *Applied Psychophysiology and Biofeedback*, Vol. 31, No. 2, pp. 173–185 (2006).
[18] Yu, D. and Deng, L.: *Automatic speech recognition*, Vol. 1, Springer (2016).
[19] Zhang, L., Zhao, Z., Ma, C., Shan, L., Sun, H., Jiang, L., Deng, S. and Gao, C.: End-to-end automatic pronunciation error detection based on improved hybrid ctc/attention architecture, *Sensors*, Vol. 20, No. 7, p. 1809 (2020).
[20] Zhao, R., Li, V., Barbosa, H., Ghoshal, G. and Hoque, M. E.: Semi-automated 8 collaborative online training module for

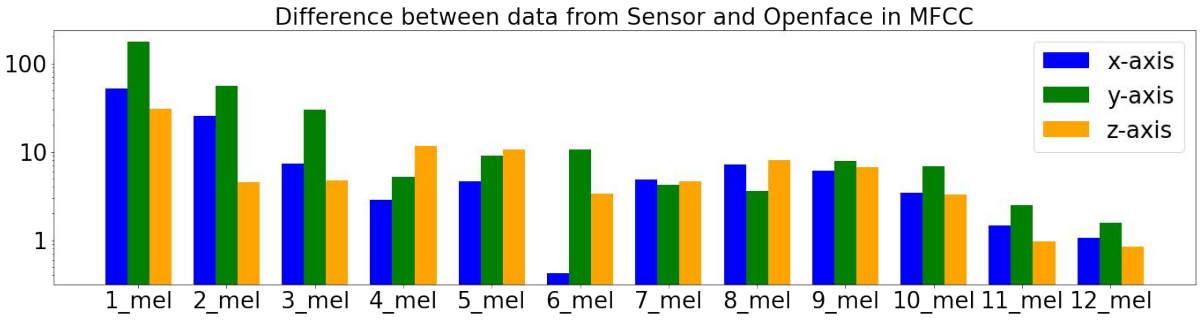


Fig. 5: MFCCs difference between data from Sensor and OpenFace

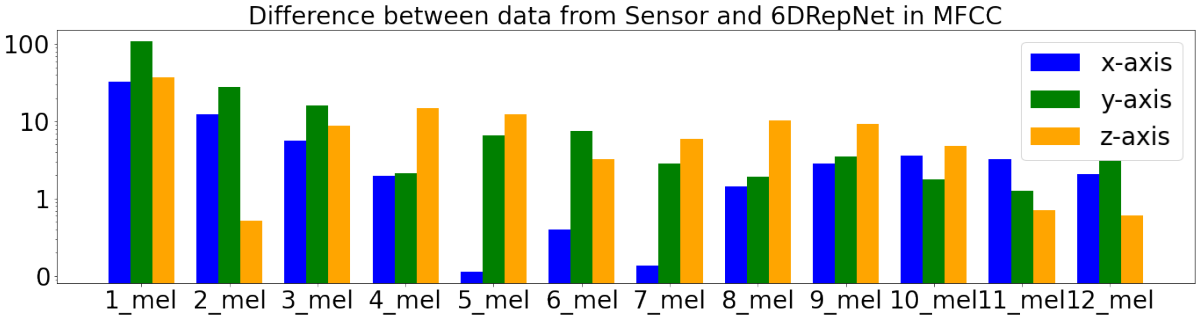


Fig. 6: MFCCs difference between data from Sensor and 6DRepNet

improving communication skills, *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 1, No. 2, pp. 1–20 (2017).

[21] Pham, H. H., Salmane, H., Khoudour, L., Crouzil, A., Velastin, S. A. and Zegers, P.: A unified deep framework for joint 3d pose estimation and action recognition from a single rgb camera, *Sensors*, Vol. 20, No. 7, p. 1825 (2020).

[22] Zhang, X., Sugano, Y. and Bulling, A.: Everyday eye contact detection using unsupervised gaze target discovery, *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, pp. 193–203 (2017).

[23] Cao, Z., Simon, T., Wei, S.-E. and Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7291–7299 (2017).

[24] Pavllo, D., Feichtenhofer, C., Grangier, D. and Auli, M.: 3D human pose estimation in video with temporal convolutions and semi-supervised training, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7753–7762 (2019).

[25] Kwon, H., Wang, B., Abowd, G. D. and Plötz, T.: Approaching the Real-World: Supporting Activity Recognition Training with Virtual IMU Data, *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 5, No. 3, pp. 1–32 (2021).

[26] Redmon, J. and Farhadi, A.: Yolov3: An incremental improvement, *arXiv preprint arXiv:1804.02767* (2018).

[27] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G. and Black, M. J.: SMPL: A skinned multi-person linear model, *ACM transactions on graphics (TOG)*, Vol. 34, No. 6, pp. 1–16 (2015).

[28] Kopf, J., Rong, X. and Huang, J.-B.: Robust consistent video depth estimation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1611–1621 (2021).

[29] Kocabas, M., Athanasiou, N. and Black, M. J.: Vibe: Video inference for human body pose and shape estimation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5253–5263 (2020).

[30] Chung, J. S. and Zisserman, A.: Out of time: automated lip sync in the wild, *Workshop on Multi-view Lip-reading, ACCV* (2016).

[31] Chung, J. S. and Zisserman, A.: Lip reading in the wild, *Asian conference on computer vision*, Springer, pp. 87–103 (2016).

[32] Lienhart, R.: Reliable transition detection in videos: A survey and practitioner's guide, *International Journal of Image and Graphics*, Vol. 1, No. 03, pp. 469–486 (2001).

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

[34] Tomasi, C. and Kanade, T.: Selecting and tracking features for image sequence analysis, *submitted to Robotics and Automation* (1992).

[35] Zadeh, A., Chong Lim, Y., Baltrušaitis, T. and Morency, L.-P.: Convolutional experts constrained local model for 3D facial landmark detection, *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 2519–2528 (2017).

[36] Hesch, J. A. and Roumeliotis, S. I.: A direct least-squares (DLS) method for PnP, *2011 International Conference on Computer Vision*, IEEE, pp. 383–390 (2011).

[37] Baltrušaitis, T., Robinson, P. and Morency, L.-P.: Constrained local neural fields for robust facial landmark detection in the wild, *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 354–361 (2013).

[38] Baltrušaitis, T., Mahmoud, M. and Robinson, P.: Cross-dataset learning and person-specific normalisation for automatic action unit detection, *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Vol. 6, IEEE, pp. 1–6 (2015).

[39] Zhou, Y., Barnes, C., Lu, J., Yang, J. and Li, H.: On the continuity of rotation representations in neural networks, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5745–5753 (2019).

[40] Zhu, X., Lei, Z., Yan, J., Yi, D. and Li, S. Z.: High-fidelity pose and expression normalization for face recognition in the wild, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 787–796 (2015).

[41] Fanelli, G., Dantone, M., Gall, J., Fossati, A. and Van Gool, L.: Random forests for real time 3D face analysis, *International Journal of Computer Vision*, Vol. 101, No. 3, pp. 437–458 (2013).