S & M 3020

Quantitative Evaluation System for Online Meetings Based on Multimodal Microbehavior Analysis

Chenhao Chen,[1*] Yutaka Arakawa,[1] Ko Watanabe,[2] and Shoya Ishimaru[2]

[1]Kyushu University, 744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan
[2]University of Kaiserslautern & DFKI GmbH Trippstadter Str. 122, 67663 Kaiserslautern, Germany

(Received May 2, 2022; accepted July 4, 2022)

Keywords: online meeting, neural network, smile detection, head pose estimation, active speaker detection

Maintaining a positive interaction is the key to a healthy and efficient meeting. Aiming to improve the quality of online meetings, we present an end-to-end neural-network-based system, named MeetingPipe, which is capable of quantitative microbehavior detection (smiling, nodding, and speaking) from recorded meeting videos. For smile detection, we build a neural network framework that consists of an 18-layer residual network for feature representation, and a self-attention layer to explore the correlation between each receptive field. To perform nodding detection, we obtain head rotation data as the key nodding feature. Then we use a gated recurrent unit followed by a squeeze-and-excitation mechanism to capture the temporal information of nodding patterns from head pitch angles. In addition, we utilize TalkNet, an active speaker detection model, which can effectively recognize active speakers from videos. Experiments demonstrate that with K-fold cross validation, the F1 scores of the smile, nodding, and speaking detection are 97.34, 81.26, and 94.90%, respectively. The processing can be accelerated with multiple GPUs due to the multithread design. The code is available at https://github.com/humanophilic/MeetingPipe.

1. Introduction

As a mainstream communication medium, online meetings play a key role in our working lives. It is estimated that 11 million meetings are held in the United States every day.(1) They take up an inordinate amount of working time, especially during the current COVID-19 pandemic. As the number of meetings increases, the quality and value of meetings have a direct impact on working efficiency. According to the publications in social science and human–computer interaction,(2) role (e.g., age and appearance),(3) verbal information (e.g., speech content and context),(4,5) and nonverbal information (e.g., body gesture and facial expression) can have a considerable influence on the behaviors of meeting participants.(6–9) To conclude, maintaining a positive interaction is the key to a healthy and efficient meeting. However, in online meetings,

[*]Corresponding author: e-mail: chen.chenhao.419@s.kyushu-u.ac.jp

https://doi.org/10.18494/SAM3959

ISSN 0914-4935 © MYU K.K.
https://myukk.org/

people are less likely to be aware of the emotional changes of others as well as their own behaviors than in face-to-face meetings due to the noncontact interaction.

Aiming to create a harmonious environment for discussion and improve the quality of online meetings, we present an end-to-end neural-network-based system, named MeetingPipe, which can quantitatively perform microbehavior detection (smiling, nodding, and speaking) from recorded meeting videos. In this work, we select the above three behavior patterns since they are the most frequently seen during meetings and are independent of other factors (e.g., although facial direction is considered as a significant indicator, it varies due to a series of factors such as the position of cameras, the number of displays, etc., and it is difficult to tell whether one is focusing on the meeting content according to his/her facial direction in a visual scene). MeetingPipe is currently capable of microbehavior detection, which makes meeting participants aware of their natural behaviors as well as those of others during online meetings. To realize meeting evaluation and improvement, we plan to implement an evaluation function that scores a meeting on the basis of detected data and a coaching function that provides users with objective analysis and suggestions in future work. As a result, meeting participants can notice their problems (e.g., fewer smiling, nodding, and speaking) during meetings and are expected to behave more actively according to the given analysis and suggestions. In other words, the current system can be considered as the first step of meeting evaluation. Note that MeetingPipe is designed for post-meeting reviews. It takes recorded meeting videos as the input and cannot be applied as an extension program to any online meeting applications.

The overall pipeline is illustrated in Fig. 1. Since the three core detection models are designed to process single-face dynamics in a video stream, the faces of each meeting participant will be cropped and tracked at the frontend. Then the following detection modules will detect and quantify smiling, nodding, and speaking patterns from the obtained face videos. For some cases in which meeting participants wear masks, although faces can be normally detected, the mouth, as the most significant feature of smiling and speaking, will be hidden by masks. Therefore, the smile and speaking model will suffer from that and cannot give a satisfactory prediction.

Considering the operating time in practical use, we focus on the floating-point operations per second (FLOPs) and the time complexity when designing the neural network structure. For example, we use techniques such as depthwise convolution and make each performance-based module shallow to increase the speed. In addition, we design the process schedule to be parallel

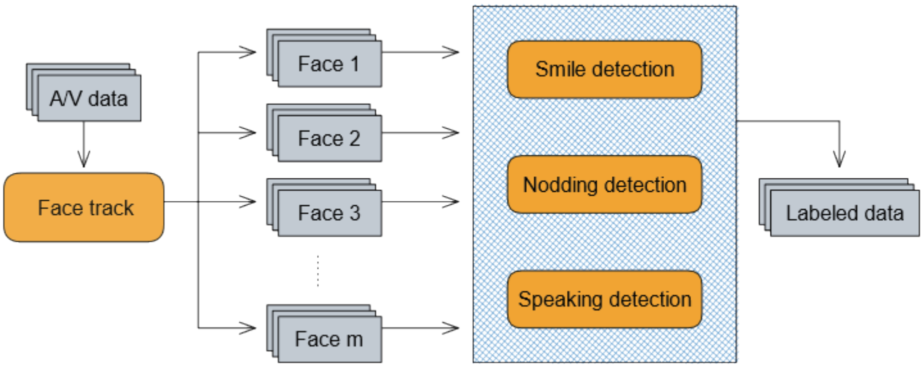


Fig. 1. (Color online) Pipeline of MeetingPipe.

so that the system can deal with multiple faces simultaneously using multiple GPUs, which greatly improves the overall efficiency.

To the best of our knowledge, MeetingPipe is the first open-source tool capable of quantitative microbehavior detection (smiling, nodding, and speaking) for online meetings. Compared with existing systems, it has the following advantages:

(1) Platform compatibility

Unlike most existing meeting analysis systems that request audio-visual data from specific online meeting platforms, such as Zoom, MeetingPipe has no such restriction, which enables users to choose any platform flexibly as long as meeting videos can be recorded.

(2) Multimodal

In addition to audio data, we also use video streams to explore visual information. By quantifying smiling, nodding, and speaking, we can evaluate meetings from facial expressions, body movements, and speech using a multimodal method, which is expected to process more information than a single-modal method.

(3) Quantitative detection

MeetingPipe performs quantitative detection of the target indicators. In other words, the specific number of the target microbehaviors will be recorded. It is considered that quantified values are clearer and more objective than qualitative measurements.

The rest of the paper is organized as follows. In Sect. 2, we discuss related work. In Sect. 3, we give a detailed description of our system. In Sect. 4, we report our experimental results. Finally, we give a conclusion in Sect. 5.

2. Related Work

There have been studies related to meeting-based group discussions, where meeting summarization, the generation of a summary from meeting transcriptions, is a task of great interest in the natural language processing (NLP) field. Some excellent models have been proposed for summarizing online/offline meetings from a textual perspective. For example, HMNet proposed by Zhu *et al.* can generate effective meeting summaries from transcriptions.(10) However, there have been much fewer studies on meeting evaluation. Despite the fact that text is a significant carrier of meeting information, we consider that the evaluation feedback of online meetings also benefits a lot from audio-visual information.

Samrose proposed a fully automated online collaboration platform, CoCo, which analyzes audio-visual data to measure various meeting parameters including participation, attitude, and shared smiles.(11) However, as an online meeting platform, it requests full access to the audio-visual stream from the user side while a meeting is in progress, and an evaluation report about the meeting parameters (participation, attitude, etc.) can be generated afterwards based on the obtained audio-visual data. In other words, the system cannot be applied to meeting videos recorded from other online meeting platforms.

On the basis of the cognitive factor that nodding and speaking correspond to the up-and-down rotation of the head and the actions of lips, respectively, Watanabe *et al.* proposed a single-modal system, DisCaaS, for online meeting evaluation that can quantitatively detect nodding and speaking patterns from recorded meeting videos.(2) With the landmark information output by OpenFace,(12) head pitch data and the distance between landmark points on the lips are taken as nodding and speaking features, respectively. Then a random forest classifier is used for prediction. It is theoretically possible to detect nodding and speaking patterns from landmark points. However, the continuous bias caused by the unavoidable error of 3D measurement technology strongly interferes with accurate classification, making the system difficult to be applied in practical use.

In addition, several meeting support applications, such as JamRoll and ailead,(13,14) have been developed so far. By collaborating with mainstream online meeting platforms, these applications are given full permissions to access audio-visual sources from the user side of online meeting platform. Compared with CoCo, they are not proposed as online meeting applications, but extension software capable of meeting evaluation that supports existing online meeting applications. Most of them are implemented with practical functions such as active speaker detection (ASD), keyword summarization, etc. However, these meeting support applications cannot be applied to scenarios where some recorded meeting videos need reviews. They strongly rely on the real-time meeting data from online meeting applications. As a summary, the implemented functions of the above meeting evaluation tools are organized in Table 1.

3. MeetingPipe

MeetingPipe is designed as an end-to-end pipeline that takes recorded meeting videos as the input and detects whether and when a person is smiling, nodding, and speaking at the frame level. As illustrated in Fig. 1, it consists of a face tracking module at the frontend, which tracks and crops all the detected faces from raw videos, and three modules responsible for smiling, nodding, and speaking detection from the tracked faces, respectively. Since a single GPU can deal with one face in a process, it is possible to concurrently process multiple faces with multiple GPUs by a multiprocessing method. In other words, the number of meeting participants and the number of GPUs used have a considerable influence on the operation time.

3.1 Face tracking

Face detection is the task of detecting faces and distinguishing them from other objects in a visual scene. Based on accurate face detection, face tracking aims to track all the detected faces in a video stream so that each face dynamic can be captured. We perform face tracking at the

Table 1
Comparison of meeting evaluation tools.

Tool	Real-time	Platform compatibility	Transcription (keyword)	Speaker detection	Emotion analysis	Facial expression	Body movement	Coaching (scoring)
JamRoll(16)			√	√	√			√
ailead(17)			√	√				
CoCo(5)	√			√	√	√		√
DisCaaS(6)		√		√			√	
MeetingPipe		√		√		√	√	

frontend since the following three detection models are designed to process single-face dynamics in a video stream. Then we use the Intersection over Union (IoU), also known as the Jaccard index,(15) which is a statistic used for gauging the similarity of sample sets, to measure the overlapping area between individual faces. For two faces from adjacent frames, it is expected that same faces (overlapping area $A_P \geq A$) can be distinguished from different faces (overlapping area $A_N < A$) with a threshold A. As a result, face videos of each participant can be obtained.

MeetingPipe takes recorded meeting videos of multiple people as the input and performs face tracking on each participant. For a meeting of m participants, m videos of individual faces are expected to be generated. On the other hand, the system performance extremely relies on the face detection results. Considering head position variations during online meetings, we selected to apply Single Shot Scale-Invariant Face Detector (S3FD) to our system owing to its effectiveness.(16) The convolutional neural network (CNN)-based architecture makes S3FD robust to object scale and rotation so that nonfrontal faces can also be detected accurately. However, for input videos of high resolution or too many participants, face tracking will be more time-consuming.

3.2 Smile detection

Smile, which directly reveals a positive emotional state, is the most common detection target in facial expression recognition tasks. With the rapid development of CNN and Transformer,(17) facial expression recognition is no longer a challenging problem. As the most frequently seen facial expressions during group discussions, we selected smile as one of the detection indicators in our system.

Smile detection can be considered as a subtask of facial expression recognition. The Haar-feature-based cascade classifier is an effective object detection method based on a machine learning approach,(18) which is a common method for smile detection tasks. However, the Haar cascade classifier is trained to only be sensitive to smile in frontal faces and cannot cope with complex situations that arise during practical use. In addition, despite a series of excellent neural network models proposed for facial expression recognition capable of smile detection, most of them were built with traditional CNN architectures. We consider that the smile detection performance can be improved by employing the latest technologies such as the attention mechanism.

In this paper, we propose a novel smile detection model as illustrated in Fig. 2(a). It takes single-face images as the input and detects smile patterns frame by frame. It starts with an 18-layer residual network (ResNet) for feature representation,(19) followed by a 1D convolution (Conv1D) layer to reduce the feature dimension. As we obtain a sequence of feature embeddings, inspired by the Vision Transformer (ViT),(20) a randomly initialized class token is placed at the beginning, which is expected to collect local and global information from the other feature embeddings through a multi-head self-attention layer. Finally, using a fully connected layer, we can make a binary classification of smile or nonsmile pattern.

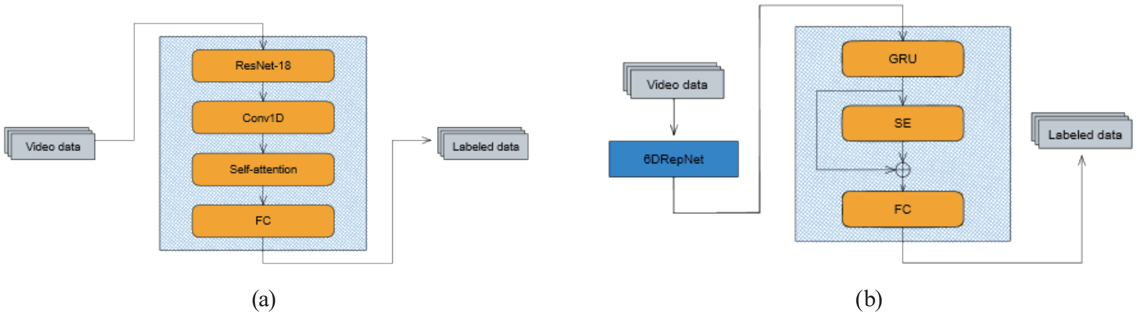


Fig. 2. (Color online) Structure of (a) smile detection module and (b) nodding detection module.

3.3 Nodding detection

Nodding detection is a downstream task of head pose estimation (HPE) that aims to recognize nodding activity in visual scenes. Since nodding is a gesture in which the head is tilted in alternating up and down arcs along the sagittal plane, among yaw, pitch, and roll, pitch is considered as the main evidence of nodding. As a consequence, we recognize nodding patterns from head pitch data. Specifically, we used 6DRepNet,(21) which is a robust HPE model proposed by Hempel *et al.*, to learn the full range of head poses. Then we built a neural-network-based framework for nodding detection, which is expected to explore temporal information from the obtained head pitch data.

Figure 2(b) illustrates the structure of our nodding detection module. It takes single-face videos as the input, from which the time series of head pitch data can be generated by 6DRepNet. The backbone part is designed as a sequence-to-one architecture, which takes head pitch data as the input and outputs a nodding prediction. In detail, a two-layer gated recurrent unit (GRU) is used to explore the temporal information of nodding,(22) followed by a squeeze-and-excitation (SE) block that adaptively recalibrates channelwise feature responses by explicitly modeling interdependences between channels.(23) With a fully connected layer, we seek to distinguish nodding and non-nodding behaviors.

3.4 Speaking detection

Generally, whether a human is speaking is judged on the basis of the following: (1) Does audio exist? (2) Does the mouth of the target person move? (3) Does the mouth movement dynamically match the audio content? Therefore, a multimodal method is commonly used to detect speaking activities. In our system, we use TalkNet, proposed by Tao *et al.*,(24) to explore the audio-visual relationship and perform effective speaking detection.

4. Experiments and Results

4.1 Dataset

To the best of our knowledge, there are few datasets related to group discussion (especially for nodding recognition), which motivates us to build a new meeting dataset. According to past studies,(2,25,26) we have conducted research on meeting evaluation since 2019. The meeting dataset is collected and labeled with consent of ethics review. The annotation is especially a time-consuming work since subtle examples of the target behaviors are difficult to notice from the recorded meeting videos. Our dataset contains video data collected from online and offline meetings. Online meeting videos are recorded from mainstream online meeting applications (e.g., Google Meet, Zoom, etc.), and offline meeting videos are obtained by the great work of Soneda *et al.*(25,26) They created a corpus on human-to-human multimodal communication in group discussions by designing offline meetings where participants were equipped with multiple sensors to detect their microbehavior, which further expands our meeting dataset. In detail, we have currently collected 295 videos of 40 participants with a total video length of 21.7 h, out of which 190 videos of 13.2 h were labeled. For the above three detection models, our meeting dataset contains 7.92 h of smile data, 4.19 h of nodding data, and 12.9 h of speaking data. Note that all videos are 25 frames per second (fps).

In addition to our meeting dataset, we apply some other existing datasets in the experiments. The Denver Intensity of Spontaneous Facial Action (DISFA) dataset is widely used in facial expression recognition tasks.(27) Owing to the high number of smiles, DISFA is applied to the training process of our smile model, and the trained smile model is tested on our meeting dataset. Since there are few datasets related to nodding recognition, we train and test the nodding model on our dataset. Given the pretrained TalkNet model, which has been trained on the benchmark ASD dataset AVA-ActiveSpeaker,(28) we test the performance on our dataset.

4.2 Implementation details

We construct MeetingPipe fully using the PyTorch library. For the smile detection model, the input face images are reshaped to (224, 224). The distance between the prediction and ground truth is computed using the binary cross-entropy loss. The learning process was driven by the Adam optimizer with a learning rate of 10^{-4}. For the nodding detection model, the obtained head pitch data are split into 20 frames (800 ms) with an overlap of 50%, which is experimented to be the optimal window length, and we use the binary cross-entropy loss and the Adam optimizer for training. For the speaking detection model, the input data consist of video data and Mel-frequency cepstral coefficient (MFCC) image data. The video resolution is reshaped to 224 × 224 and the duration is 25 frames. The dimension of MFCC is 13 and the audio source corresponds to the same interval on the timeline.

MeetingPipe was tested on the Ubuntu 20.04 operating system. All experiments are conducted on two NVIDIA RTX A6000 GPUs and one 36-core i9-10980XE CPU.

4.3 Experimental results

As mentioned above, after the training process, we test the three models on their test datasets. The details of the test datasets are summarized in Table 2.

Note that all the test data are sampled from our meeting dataset described above. The experimental results obtained using K-fold cross validation (K = 5) are reported in Table 3.

Table 2
Details of the test datasets.

Module	Data volume		Data shape
	Positive	Negative	
Smiling	28512	28942	(224, 224, 1)
Nodding	3771	3816	(20, 1)
Speaking	9288	9342	V:(25, 224, 224, 3) A:(100, 13, 1)

Table 3
Test results of each detection module.

Module	MeetingPipe				DisCaaS(6)
	Accuracy	Precision	Recall	F1-score	F1-score
Smiling	0.9734	0.9745	0.9723	0.9734	N/A
Nodding	0.8158	0.8523	0.7742	0.8126	0.6400
Speaking	0.9472	0.9852	0.9153	0.9490	0.7200

We observe that the smile and speaking models achieve 97.34 and 94.9% F1 scores on their test dataset, respectively. The results outperform the nodding model by over 10%. To explore the reasons, we visualize the HPE results and find that there exists a slight and continuous bias on head rotation predictions even for non-nodding patterns. This indicates that the measured head movement curves are always oscillatory. In other words, it is difficult to distinguish a slight nodding pattern from a non-nodding pattern. It greatly limits the nodding recognition performance since in most cases people tend to nod slightly. In addition, compared with another two models, the nodding model was only trained and tested on our meeting dataset. Thus, a small data volume is also considered as one of the causes. For comparison, we test the meeting evaluation system DisCaaS on the same dataset. From the results, we can observe that MeetingPipe outperforms DisCaaS by about 20% in all tests.

We further perform ablation experiments on nodding window length. To study the influence between different nodding window lengths, we experiment with different window lengths of 5, 10, 20, and 50 that amounts to 0.2, 0.4, 0.8, and 2 s, respectively, and the results are reported in Table 4.

It can be concluded that a short window length, e.g., 5, is insufficient to fully capture the temporal feature of nodding activity. As the window length increases, the F1 score improves from 62.32 to 81.26%. On the other hand, it can also be concluded from the table that a very large window length has a negative effect on nodding detection.

Then we test MeetingPipe on a 304.76 s meeting video involving four participants. To test videos with different resolutions, we resize the raw resolution from 1920 × 1080 to 1280 × 720. The detailed operating time is shown in Table 5.

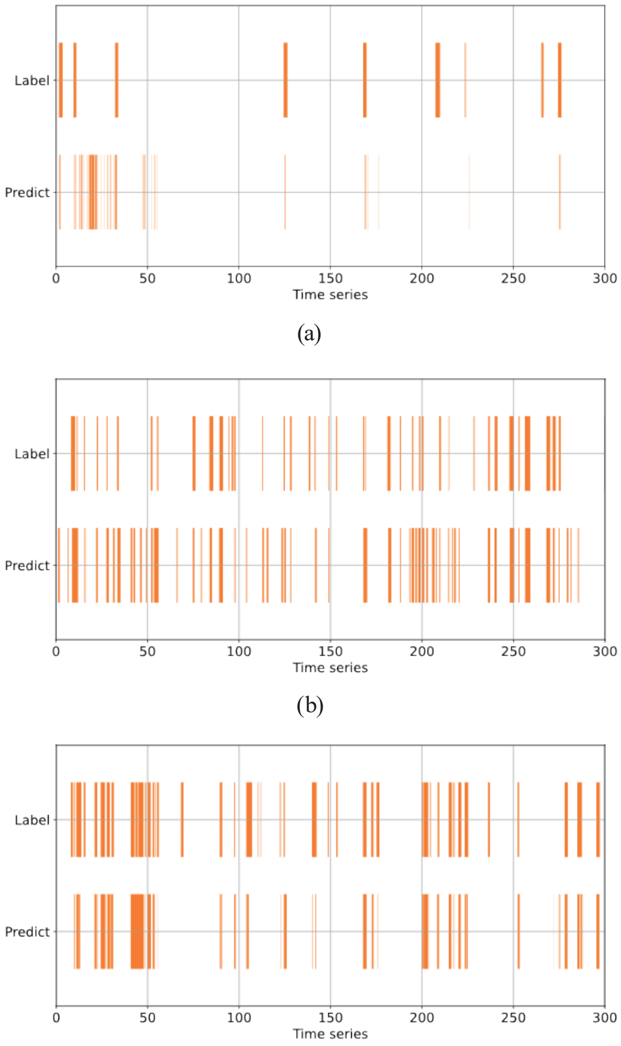
From the above table, we observe that both video resolution and GPU number affect operating time considerably. It can be concluded that (1) low-resolution videos consume relatively fewer computational resources and take less time to process. (2) It is more efficient to use multiple GPUs. Therefore, it is suggested to record meeting videos with a relatively low resolution (e.g., 1280 × 720) and use more GPUs. Note that the detection performance cannot be guaranteed on excessively blurred videos. For a more visual demonstration, the prediction results of smiling, nodding, and speaking for one participant on the above 5 min video are shown in Fig. 3.

Table 4
Performance for different nodding window lengths.

Window length (frames)	F1 score
5	0.6232
10	0.7176
20	0.8126
50	0.7658

Table 5
Operating time on test videos.

		Operating time (s)			
Module	Resolution	1080p		720p	
	GPU number	1	2	1	2
Face tracking		370.85	364.96	163.42	169.98
Behavior detection		741.34	565.60	532.18	445.65
Total		1112.19	930.56	695.6	615.63



(a)

(b)

(c)

Fig. 3. (Color online) Prediction results of (a) smiling, (b) nodding, and (c) speaking.

5. Conclusions

In this work, we built an end-to-end quantitative evaluation system named MeetingPipe for online meetings. It can effectively detect and quantify three common microbehavior indicators, namely, smiling, nodding, and speaking, during online meetings. In addition, we built and labeled our meeting dataset manually. It was concluded from our experimental results that MeetingPipe can achieve satisfactory performance for practical use. As future work, we plan to implement additional functions such as semantic analysis and keyword extraction for more comprehensive meeting analysis.

Acknowledgments

This work was supported by JSPS KAKENHI (JP18H03233), Grand Challenge Open Research Project of MEXT Innovation Platform for Society 5.0 (JPMXP0518071489), and JST (Establishment of University Fellowships Towards the Creation of Science Technology Innovation; JPMJFS2132).

References

1 J. Allen, S. Rogelberg, and J. C. Scott: Qual. Prog. **41** (2008) 48.

2 K. Watanabe, Y. Soneda, Y. Matsuda, Y. Nakamura, Y. Arakawa, A. Dengel, S. Ishimaru: Sensors **21** (2021) 5719. https://doi.org/10.3390/s21175719

3 E. M. Schulte, N. Lehmann-Willenbrock, and S. Kauffeld: J. Manag. Psychol. **28** (2013) 928. https://doi.org/10.1108/jmp-06-2013-0193

4 S. Shrivastava and V. Prasad: WHJJ **16** (2020) 73.

5 G. E. Knowlton and K. T. Larkin: Appl. Psychophysiol. Biofeedback **31** (2006) 173.

6 A. E. Scheflen: Psychiatry **27** (1964) 316. https://doi.org/10.1080/00332747.1964.11023403

7 A. Mehrabian: Psychol. Bull. **71** (1969) 359. https://doi.org/10.1037/h0027349

8 S. Centorrino, E. Djemai, A. Hopfensitz, M. Milinski, and P. Seabright: Evol. Hum. Behav. **36** (2015) 8. https://doi.org/10.1016/j.evolhumbehav.2014.08.001

9 L. S. Bohannon, A. M. Herbert, J. B. Pelz, and E. M. Rantanen: Displays **34** (2013) 177. https://doi.org/10.1016/j.displa.2012.10.009

10 C. Zhu, R. Xu, M. Zeng, and X. Huang: arXiv preprint arXiv:2004.02016 (2020). https://doi.org/10.48550/arXiv.2004.02016

11 S. Samrose: ACM Int. Joint Conf. and Int. Symp. Pervasive and Ubiquitous Computing and Wearable Computers (2018) 510–515.

12 B. Amos, B. Ludwiczuk, and M. Satyanarayanan: CMU School of Comput. Sci. **6** (2016) 20.

13 JamRoll: https://jam-roll.webempath.ai/ (accessed April 2022).

14 Ailead: https://www.ailead.app/ (accessed April 2022).

15 A. H. Myrphy: Weather Forecasting **11** (1996) 3. https://doi.org/10.1175/1520-0434(1996)011%3C0003:TFAASE%3E2.0.CO;2

16 S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li: Proc. 2017 IEEE Int. Conf. Computer Vision (IEEE, 2017) 192–201.

17 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, and I. Polosukhin: Advances Neural Information Processing Systems (2017) 30. https://doi.org/10.48550/arXiv.1706.03762

18 P. Viola and M. Jones: Proc. 2001 IEEE Computer Society Conf. Computer Vision and Pattern Recognition (IEEE, 2001). https://doi.org/10.1109/CVPR.2001.990517

19 K. He, X. Zhang, S. Ren, and J. Sun: Proc. 2016 IEEE Computer Vision and Pattern Recognition Conf. (IEEE, 2016) 770–778. https://doi.org/10.48550/arXiv.1512.03385

20 A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, and N. Houlsby: arXiv preprint arXiv:2010.11929 (2020). https://doi.org/10.48550/arXiv.2010.11929

21 T. Hempel, A. A. Abdelrahman, and A. Al-Hamadi: arXiv preprint arXiv:2202.12555 (2022). https://doi.org/10.48550/arXiv.2202.12555

22 K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio: arXiv preprint arXiv:1406.1078 (2014). https://doi.org/10.48550/arXiv.1406.1078
23 J. Hu, L. Shen, and G. Sun: Proc. 2018 IEEE Computer Vision and Pattern Recognition Conf. (IEEE, 2018) 7132–7141. https://doi.org/10.48550/arXiv.1709.01507
24 R. Tao, Z. Pan, R. K. Das, X. Qian, M. Z. Shou, and H. Li: Proc. 29th ACM Int. Conf. Multimedia (2021) 3927–3935. https://doi.org/10.48550/arXiv.2107.06592
25 Y. Soneda, Y. Matsuda, Y. Arakawa, and K. Yasumoto: M3B corpus: Proc. ACM Int. Joint Conf. Pervasive and Ubiquitous Computing and Proc. ACM Int. Symp. Wearable Computers (2019) 825–834. https://doi.org/10.1145/3341162.3345588
26 Y. Soneda, Y. Matsuda, Y. Arakawa, and K. Yasumoto: Proc. 27th Int. Conf. Computers in Education (2019) 466–471.
27 S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn: IEEE Trans. Affective Computing **4** (2013) 151. https://doi.org/10.1109/T-AFFC.2013.4
28 J. Roth, S. Chaudhuri, O. Klejch, R. Marvin, A. Gallagher, L. Kaver, and C. Pantofaru: Proc. 2019 IEEE Computer Vision Workshop Conf. (IEEE, 2019) 3718–3722. https://doi.org/10.1109/ICCVW.2019.00460

About the Authors

Chenhao Chen received his B.S. degree from Nanjing University of Science and Technology, China, in 2017. He joined the double M.E. degree program of Nanjing University of Science and Technology, China and Fukuoka Institute of Technology, Japan from 2017 to 2020 and received his double M.E. degrees in 2020. Since 2021, he has been a Ph.D. student at Kyushu University, Japan. His research interests include computer vision and neural networks. (chen.chenhao.419@s.kyushu-u.ac.jp)

Yutaka Arakawa received his B.E., M.E., and Ph.D. degrees from Keio University, Japan, in 2001, 2003, and 2006, respectively. From 2006 to 2013, he was an assistant professor at Keio University and Kyushu University, Japan. From 2013 to 2019, he was an associate professor at Nara Institute of Science and Technology, Japan. Since 2019, he has been a professor at Kyushu University. His current research interests include human activity recognition and behavior change support systems. (arakawa@ait.kyushu-u.ac.jp)

Ko Watanabe received his B.E. degree in mechanical engineering from Tokyo University of Agricultural and Technology, Japan, in 2017. He then received his M.E. degree from Nara Institute of Science and Technology, Japan in 2019. He is currently a Ph.D. candidate in computer science at the University of Kaiserslautern. His current research focuses on investigating technologies that augment human intellect. (ko.watanabe@dfki.uni-kl.de)

Shoya Ishimaru received his B.E. and M.E. degrees in engineering from Osaka Prefecture University, Japan in 2014 and 2016, respectively. He received his Ph.D. degree in engineering with a summa cum laude award from the University of Kaiserslautern, Germany in 2019. Since 2021, he has been a junior professor at the University of Kaiserslautern. His research focuses on investigating technologies that augment human intellect.
(ishimaru@cs.uni-kl.de)