The 36th Annual Conference of the Japanese Society for Artificial Intelligence, 2022

Quantifying a Multi-person Meeting based on Multi-modal Micro-behavior Analysis

Chenhao Chen[*1] Kosuke Tokuhara[*2] Yutaka Arakawa[*3] Ko Watanabe[*4]
Shoya Ishimaru[*4]

[*1*2*3] Kyushu University [*4] University of Kaiserslautern & DFKI GmbH

In this paper, we present an end-to-end online meeting quantifying system, which can exactly detect and quantify three micro-behavior indicators, speaking, nodding, and smile, for online meeting evaluation. For active speaker detection (ASD), we build a multi-modal neural network framework which consists of audio and video temporal encoders, audio-visual cross-attention mechanism for inter-modality interaction, and a self-attention mechanism to capture long-term speaking evidence. For nodding detection, based on the WHENet framework proposed in the research field of head pose estimation (HPE), we can estimate the head pitch angles as the nodding feature. Then we build a gated recurrent unit (GRU) network with squeeze-and-excitation (SE) module to recognize nodding movement from videos. Finally, we utilize a Haar cascade classifier for smile detection. The experimental results using K-fold Cross Validation show that the F1-score of each detection module achieves 94.9%, 79.67% and 71.19% respectively.

1. Introduction

In recent years, policies that recommend reductions in working hours have been actively carried out to implement the work style reform in Japan. Especially during the COVID-19 period, online meeting has gradually become a mainstream communication form that takes up a considerable part of the working time. Problems of *what is a successful meeting* and *how to further improve the meeting quality* remain to be researched. In this paper, we present an end-to-end online meeting quantifying system to deal with the above problems. Based on multi-modal methods, we mainly detect and quantify three micro-behavior indicators, speaking, nodding, and smile, for online meeting evaluation.

For speaking detection, in other words active speaker detection (ASD), it seeks to detect who is speaking in a visual scene of one or more speakers[6]. There have been deep learning solutions that extract audio and video features to make binary classification. Inspired by the great progress of ASD, we build a multi-modal neural network framework which takes audio and video data as inputs. Through a cross-attention module, the inter-modality interaction between audio and video data can be used to study whether the target is speaking at the frame level.

For nodding detection, it is based on the previous study of head pose estimation (HPE). From OpenFace[1] to the latest WHENet[10], the existing HPE model is efficient enough to detect the target head pose from all viewpoints. Thus, we utilize the WHENet, which meets or beats state-of-the-art methods for frontal HPE, to extract the head pitch angles. In view of the temporal dynamics of video flow, we build a gated recurrent unit (GRU) network to study the temporal features of the head pitch angles and make binary classification at last.

Contact: chen.chenhao.419@s.kyushu-u.ac.jp

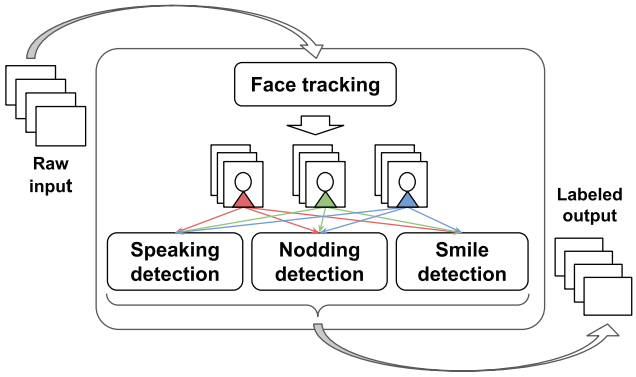

Figure 1: An overview of the online meeting quantifying system

Despite much progress in emotion recognition, we only perform smile detection since smile is the most common face expression during a meeting other than natural state, for speaking and nodding as well. For smile detection, we simply use a Haar cascade classifier which can efficiently recognize smile expression from video flows.

By quantifying an online meeting with the above three indicators, combined with subjective feedback from participants, it is possible to evaluate an online meeting and further improve the meeting quality. That is expected to be significant to our work.

The rest of the paper is organized as follows. In section 2, we discuss the related work. In section 3, we give a detailed description of the whole meeting quantifying system. In section 4 , we report the experiments and results. Finally, section 5 will give a conclusion.

2. Related work

There actually have been prior studies on the research of meetings so far.

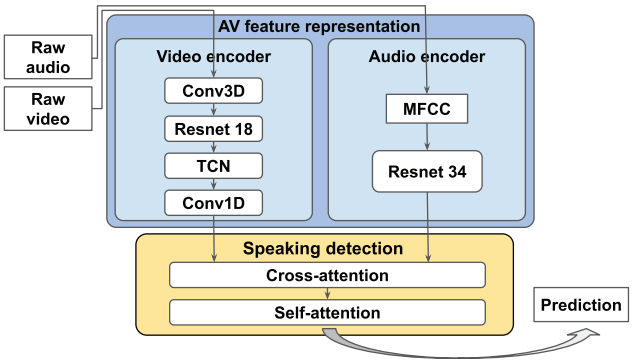

Figure 2: The structure of speaking detection module

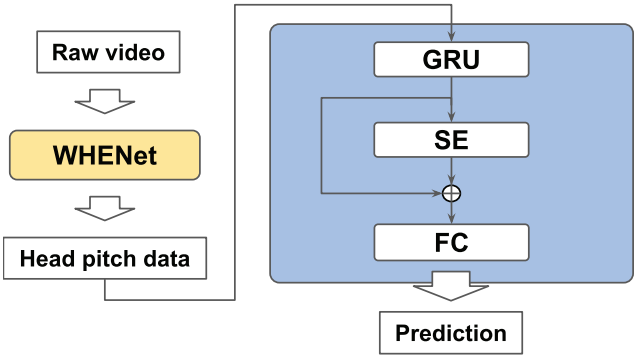

Figure 3: The structure of nodding detection module

Ohnishi[5] et al. proposed a system that structures a meeting log by detecting and tagging actions using acceleration sensors. However, the facts that equipped with wearable sensors is inconvenient and costly cannot be ignored. It is assumed that the system can only be used in certain specific meetings, which limits the generality to a great extent.

As a prior study of this paper, Watanabe[9] et al. proposed an online meeting quantifying system based on random forest, which can detect speaking and nodding actions from video flows. Where OpenFace is used to extract 68 3D face landmarks and head pose at frame level so that the distance between upper and lower lips can be used as the speaking feature, and the head pitch angles can be used as the nodding feature. In the ideal situation, it is possible to make use of the distance between lips to recognize speaking action. However, whereas the unavoidable slight deviations caused by OpenFace which have great interference to ASD, the system cannot be actually applied in practice.

3. Methods

We present an end-to-end online meeting quantifying system that consists of three detection modules, speaking, nodding, and smile. In this section, we would like to give a detailed description on each module.

The whole system framework is illustrated in Figure 1. In terms of the different identity labels of the detected faces between adjacent video frames, a face tracking module is contained at the frontend, so that the following detection modules will only deal with one target in video scenes.

3.1 Speaking detection

Inspired by TalkNet proposed by Tao[8], we build the speaking detection module illustrated in Figure 2. The frontend contains a video encoder and an audio encoder to obtain visual-audio feature representations from the raw data.

The video encoder aims to extract long-term features from facial/mouth dynamics. A 3D convolution block followed by an 18-layer ResNet[2] is expected to explore the spatial features within each frame. Based on that, the following temporal convolution network[4] (TCN) focus on the temporal information across frames, which is followed by a 1D convolution layer to reduce the feature dimension.

It is observed that Mel-frequency Cepstral Coefficients (MFCCs) take into account human perception of sensitivity at appropriate frequencies and are thus suitable for speech-related research. We build a 34-layer ResNet that takes MFCC feature as the input. it is expected to generate the sequence of audio embeddings by the audio encoder.

Given pairs of video and audio 128-dimension embeddings, both of them characterize the spatial-temporal information of the raw inputs. Because of the difference of facial/mouth dynamics between the speaking and non-speaking actions, we can recognize whether the target is speaking by exploring the inter-modality interaction between the given video and audio embeddings. Due to the great progress of the attention mechanism, we would like to utilize a cross-attention module rather than simply compute the cosine similarity or Euclidean distance. Finally, we add a self-attention module to give a prediction.

3.2 Nodding detection

Generally, one relies on the observed bobbing up and down of the head to determine whether the target is nodding. Based on this cognitive finding, we can perform nodding detection with the head pitch angle. Nodding detection can be considered as an application of HPE. In this paper, we select to use the latest WHENet that achieves the state-of-the-art performance of HPE. Here we highlight the method we use to recognize the nodding action from the obtained head pitch data.

As we can see in Figure 3, given the nodding feature generated by WHENet, the nodding detection module is designed as a sequence-to-one architecture. It is designed to be shallow due to the feature complexity, which consists of a 2-layer GRU followed with a squeeze-and-excitation (SE) block that adaptively re-calibrates channel-wise feature responses by explicitly modelling inter-dependencies between channels[3]. Despite the fact that the nodding detection module is not complex enough to be worried about vanishing gradient problem, a skip connection structure is proven to achieve better results in experiments. And likewise, we add a fully-connected layer at last to give a prediction.

3.3 Smile detection

Smile is the most common detection indicator in facial expression recognition tasks, which directly reveals a positive

emotional state. As the most frequently seen facial expression in the meeting, a smile detection module is added to our meeting quantifying system. Haar feature-based cascade classifier is an effective object detection method based on machine learning approach. It can be effectively applied to various fields including smile detection. To be specific, we take advantage of MediaPipe, which offers open source machine learning solutions for live and streaming media, to perform accurate face detection from video frames. Based on that, a Haar cascade smile classifier can be applied to capture the smiling mouth from the detected faces.

4. Experiments

4.1 Dataset

To the best of our knowledge, there are few benchmark datasets for meeting analysis, especially for nodding detection. That leads to the idea of building a new dataset. We have intentionally recorded online meeting videos from 2020 so far, and extracted and tagged the target scenes, including speaking, nodding, smile, etc. That can be time consuming since the target micro-behavior needs to be labeled at the frame level by human. Besides, thank Soneda[7] et al. for their great work. They design face-to-face offline meetings to collect meeting data where all participants are equipped with multiple sensors to detect their micro-behavior, which further expands our meeting dataset.

Until now, we have collected 295 videos of 40 participants, and the total length can be up to 21.7 hours. Out of that 190 videos of 13.2 hours have been labeled. After converting the videos to 25 frames-per-second (fps), our meeting dataset consists of 12.9 hours of speaking data, 4.19 hours of nodding data and 7.92 hours of smile data. In detail, one speaking sample is actually composed of a pair of video and audio data, where the video data refers to a tracked face video clip. Video data is specified with the shape of [25,224,224,3] that in turn represents the window length, width, height, and channel, while the audio data refers to the 13-dimension MFCC feature with a window length of 100. The shape of nodding data is uniformly set to [20,] that means the nodding window length is 20 frames. Despite the fact that we do not need to train the Haar cascade classifier, the smile data are segmented with a window length of 25 frames for the following tests. On the other hand, we add the corresponding negative data to maintain the data balance.

4.2 Implementation details

The meeting quantifying system is constructed using Pytorch library. Since we perform a face track operation at the frontend, the input of the three detection modules is fixed to the same spatial size.

For the speaking module, we improve the pre-trained model in [8] that has already been trained on the benchmark ASD dataset AVA-ActiveSpeaker[6] by performing fine-tuning on our meeting speaking data. During the process, a decay schedule for learning rate is applied and the initial value is 10^{-4}. The distance between the softmax output and the ground truth label is computed by the cross-entropy loss, which is minimized by the Adam optimizer.

For the nodding module in Figure 3, the learning rate is set to 10^{-3} without a decay schedule. Likewise, the learning process is driven by the cross-entropy loss and the Adam optimizer.

For the smile module, there are several parameters that can directly affect the smile detection performance. Notably, the scaleFatcor and minNeighbors are set to 1.8 and 20, respectively.

As for hardware devices, all experiments are conducted on two NVIDIA RTX A6000 GPUs and one 36-core i9-10980XE CPU.

4.3 Results

We prepare test datasets for the three detection modules so that the practical performance of each can be recorded, the details are summarized in Table 1.

Table 1: Data distribution of test datasets

Dataset	Label		Shape
	Positive	Negative	
Speaking	9288*	9342	[25, 224, 224, 3](V) [100, 13, 1](A)
Nodding	3771*	3816	[20,]
Smile	28512	28942	[25, 224, 224, 3]

As mentioned in subsection 4.1, we prepare 12.9 hours of speaking data, 4.19 hours of nodding data, and 7.92 hours of smile data. The flagged data means the test data 20% sampled from the total data. Note that all smile data is used for test. The results on the test data is summarized in Table 2, where K-fold Cross Validation method is used.

Table 2: Test results of each module

Module	Accuracy	Precision	Recall	F1-score
Speaking	0.9472	0.9852	0.9153	0.9490
Nodding	0.8028	0.8438	0.7546	0.7967
smile	0.7509	0.6198	0.8358	0.7119

The speaking and nodding module can be considered as binary classification problem since the Softmax function is applied to the last layer, which outputs the probability distribution that the input is recognized to be positive or negative. Whereas the Haar cascade classifier detects a smile at each frame image, so we consider that for over 50% frames where smile is detected, the input video clip is recognized to be positive.

We observe that the F1-score of the speaking module achieve 94.9%, which quite outperforms the others. For the nodding module, it is inevitable that the head pitch data obtained from the WHENet slightly offset from the true value when Euler angles are predicted. Figuratively, that means the data waveform is always oscillatory even the target is not nodding. It greatly limits the nodding detection performance. As for the smile module, the Haar cascade classifier is trained to only detect smile in frontal faces. In

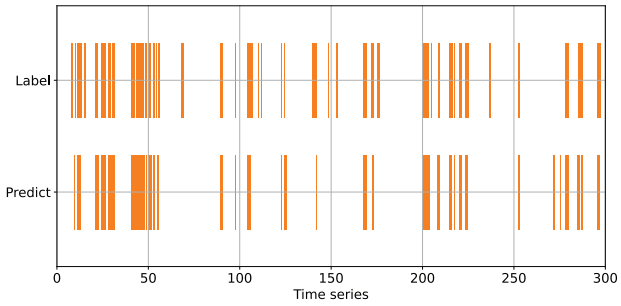

Figure 4: Ground truth and prediction of speaking

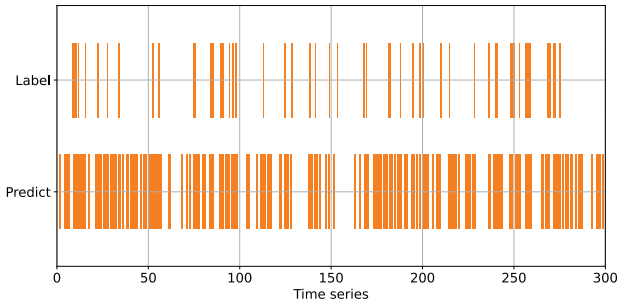

Figure 5: Ground truth and prediction of nodding

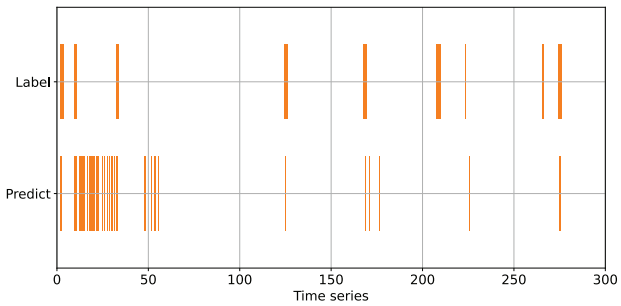

Figure 6: Ground truth and prediction of smile

other words, the smile can hardly be detected if the frontal face is not facing towards the camera.

For a more visual representation of the system's performance, we actually apply the whole meeting quantifying system to an online meeting video. The quantifying results of one participant illustrate in Figure 4, 5 and 6.

5. Conclusion

In this work, we build an end-to-end online meeting quantifying system that can detect and quantify three micro-behavior indicators, speaking, nodding, and smile, for meeting evaluation. The system consists of three detection modules, of which the speaking and nodding module is built based on the neural network, and the smile module utilizes the Haar cascade classifier. In addition, we manually build our meeting dataset to support the necessary training requirements. Experimental results show that the F1-score of each module achieve 94.9%, 79.67% and 71.19% respectively. It can be concluded that our system can be practically applied to online meeting evaluation.

In the future, we would like to further improve the detection performance of nodding and smile. Furthermore, we also consider adding other significant detection indicators for better meeting evaluation effects.

Acknowledgment

This work was supported by JSPS KAKENHI (JP18H03233) and Grand Challenge Open Research Project in MEXT Innovation Platform for Society 5.0 (JPMXP0518071489).

References

[1] Brandon Amos, Bartosz Ludwiczuk, Mahadev Satyanarayanan, et al. "Openface: A general-purpose face recognition library with mobile applications". In: *CMU School of Computer Science* 6.2 (2016), p. 20.

[2] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[3] Jie Hu, Li Shen, and Gang Sun. "Squeeze-and-excitation networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7132–7141.

[4] Xin Li et al. "Deep concept-wise temporal convolutional networks for action localization". In: *Proceedings of the 28th ACM International Conference on Multimedia*. 2020, pp. 4004–4012.

[5] Ayumi Ohnishi et al. "A method for structuring meeting logs using wearable sensors". In: *Internet of Things* 5 (2019), pp. 140–152.

[6] Joseph Roth et al. "Ava active speaker: An audio-visual dataset for active speaker detection". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 4492–4496.

[7] Yusuke Soneda et al. "M3B corpus: Multi-modal meeting behavior corpus for group meeting assessment". In: *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. 2019, pp. 825–834.

[8] Ruijie Tao et al. "Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection". In: *Proceedings of the 29th ACM International Conference on Multimedia*. 2021, pp. 3927–3935.

[9] Ko Watanabe et al. "Discaas: Micro behavior analysis on discussion by camera as a sensor". In: *Sensors* 21.17 (2021), p. 5719.

[10] Yijun Zhou and James Gregson. "Whenet: Real-time fine-grained estimation for wide range head pose". In: *arXiv preprint arXiv:2005.10353* (2020).