Visual-only Voice Activity Detection using Human Motion in Conference Video

Keisuke Yamazaki[1], Satoshi Tamura[2], Yuuto Gotoh[3] and Masaki Nose[3]
[1]*Graduate School of Natural Science and Technology, Gifu University, Gifu, Japan*
[2]*Faculty of Engineering, Gifu University, Gifu, Japan*
[3]*Ricoh Company, Ltd., Kanagawa, Japan*

Keywords: Voice Activity Detection, Human Motion, Speaker Diarization, Dynamic Image, Multi-modal Transfer Module, Conference Video Processing.

Abstract: In this paper, we propose a visual-only Voice Activity Detection (VAD) method using human movements. Although audio VAD is commonly used in many applications, it has a problem it is not robust in noisy environments. In such the cases, multi-modal VAD using speech and mouth information is effective. However, due to the current pandemic situation, people wear masks causing we cannot observe mouths. On the other hand, utilizing a video capturing the entire of a speaker is useful for visual VAD, because gestures and motions may contribute to identify speech segments. In our scheme, we firstly obtain dynamic images which represent motion of a person. Secondly, we fuse dynamic and original images using Multi-Modal Transfer Module (MMTM). To evaluate the effectiveness of our scheme, we conducted experiments using conference videos. The results show that the proposed model has better than the baseline. Furthermore, through model visualization we confirmed that the proposed model focused much more on speakers.

1 INTRODUCTION

We have many meetings for discussion and decision making. Minutes are then often taken, which are useful to confirm the contents of meetings later. It is still common to take minutes manually, however, it costs significantly. Therefore, in recent years, a lot of research works have been done to automate the process of taking minutes. In order to take minutes automatically, it is necessary to develop technology for speaker diarization, which recognizes "who spoke when," and speech recognition and lip reading, which recognize "what was said." In this study, we focus on Voice Activity Detection (VAD) to recognize "when a person spoke," that is one of the speaker diarization elements. Although VAD has been investigated for decades, there is a problem that the VAD accuracy is often degraded by the existence of background noise or cross talks. In order to further improve the accuracy, multi-modal VAD using face movies in addition to speech data is considered to be an effective method; Visual VAD (VVAD) utilizes mouth movements to detect speech segments, and multi-modal VAD incorporates audio and visual information in an early- or a late-fusion manner.

The spread of COVID-19 threat has drastically changed the situation. Most meetings are held online, and attendees usually wear masks. This causes new issues for speaker diarization. Audio data are compressed and distorted so that they could be transmitted on the internet, resulting a mismatch between the data and a VAD model. Wearing masks affects acoustic characteristics, and more crucially, mouths are hidden making conventional VVAD difficult. To overcome the last problem, we have investigated another perspective. Here we focus on video movies capturing the entire of a speaker, and utilizing the whole movements of the speaker. This is based on the discovery that individuals unintentionally synchronize their nonverbal and linguistic behavior during social interactions (N.Latif et al., 2014). We consider that if we could obtain features effective for VAD from body movements and gestures, VVAD could be still possible even when a speaker's mouth is not visible.

In this paper, we try to develop our VVAD based on the above discussion. As a preliminary step, we perform visual-only VAD using videos in which a mask is artificially worn to a speaker, and verify the effectiveness. To the best of my knowledge, this paper is the first attempt to perform VVAD on a speaker

Yamazaki, K., Tamura, S., Gotoh, Y. and Nose, M.
Visual-only Voice Activity Detection using Human Motion in Conference Video.
DOI: 10.5220/0010829200003122
In *Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2022)*, pages 570-577
ISBN: 978-989-758-549-4; ISSN: 2184-4313

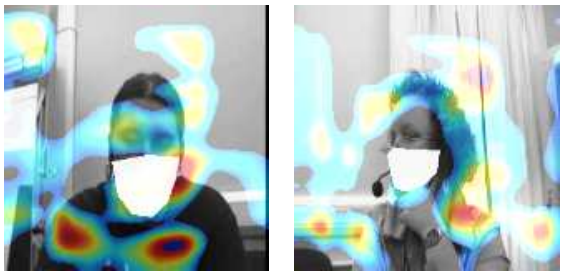

Figure 1: Visualization examples in our preliminary experiments. Which area a VVAD model focused on is illustrated.

wearing a mask. We firstly build a VAD model using a simple 3D Convolutional Neural Network (3DCNN). Next, we conduct visualization to find the areas of interest in the prediction results, to examine the interpretability of the model. Examples of the visualization results are shown in Figure 1. Based on the preliminary results, we propose a novel VVAD framework using two modalities: video data and dynamic images. Employing dynamic images is expected to focus much more on human gestures.

The rest of this paper is organized as follows. At first, we describe related works in Section 2. Next, Section 3 introduces our method including techniques used in this work. Experiments, results and discussion are described in Section 4. Finally, conclusion and future work are presented in Section 5.

2 RELATED WORKS

2.1 Visual-only VAD

Recently, VVAD has been studied actively. There are two main types of VVAD methods: using facial information such as lip movements and head movements, and using human appearance like gestures and body movements.

In terms of using facial information, (A.Aubrey et al., 2007) proposed two VVAD methods. The first one is applying Active Appearance Model to lip regions to represent lip shape and texture as a vector, and then modeling the features by hidden Markov model. Another one is applying a spatio-temporal filter which models human retina behaviors and calculates the change to classify voice activity. In (B.Joosten et al., 2015), the authors proposed a VVAD method based on spatio-temporal gabor filters. The authors applied the method to a mouth, a head or the entire video frame, and found that the mouth regions gave the best result. One of deep-learning models, Long Short-Term Memory (LSTM), was trained using face images detected by CNN, and speech and non-speech intervals were estimated by audio VAD in (K.Stefanov et al., 2017). The LSTM model was able to predict speech activities using video frame only, and gave a good result on average.

The study of VVAD using human appearance is more relevant to this paper. (B.G.Gebre et al., 2014b) proposed a VVAD method using Motion History Image (MHI). MHI is a method of representing motion in a video consisting of multiple frames into a single image. The method showed the potential of VVAD using human motion information. (M.Cristani et al., 2011) introduced a method obtaining optical flows from a video, and encoded its energy and "complexity" using an entropy-like measure. Their experiment showed the usefulness of using gestures for VAD. Note that the dataset used in the paper was top-view. In order to address the VVAD task with cues from whole body motion, some methods (optical flow, RGB-dynamic image, and combination of these) for representing human movements were compared in (M.Shahid et al., 2019a). They showed that dynamic images gave better results on average for the dataset they used. Subsequently, (M.Shahid et al., 2019b) proposed a method combining RGB-dynamic images and unsupervised domain adaptation technique, and improved results in same dataset used in (M.Shahid et al., 2019a). In (M.Shahid et al., 2021), the authors extended their work and proposed a method applying the segmentation task to VAD. They generated class activation maps using features extracted from RGB-dynamic images, and created labels for each speech/non-speech/background region. Using the labels, they performed supervised learning to segment RGB-dynamic images into three classes.

2.2 Multi-modal VAD

Some researches of VAD using audio and visual information were reported. (K.Hoover et al., 2017) introduced a system that associates faces with voices in a video by fusing information from the audio and visual signals. They connected speech and face by clustering speech segments and face images detected in the video, and selecting the face clusters that co-occur most often with the speech clusters. (B.G.Gebre et al., 2014a) proposed a speaker diarization method using speech and human motion (gestures) based on the assumption that people who are speaking are often performing gestures. They created a speaker model from speech samples corresponding to the gestures.

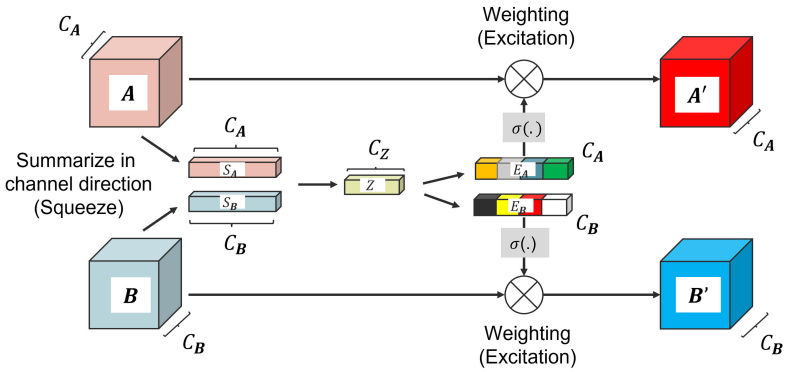

Figure 3: An overview of MMTM.

3 METHODOLOGY

In this paper, we propose a method to fuse features of two modalities, dynamic images and video data, in the middle layer in order to make the model focus on human motion much more. In this section, we at first introduce dynamic images (H.Bilen et al., 2016), followed by Multi-Modal Transfer Module (MMTM) (H.R.V.Joze et al., 2020) which fuses features of multiple modalities. After that, we describe details of our scheme.

3.1 Dynamic Image

There are several ways to represent the information of human movement. Optical Flow is a typical method and is often used to represent motion information. In other cases, Motion History Image (HMI) is used in (B.G.Gebre et al., 2014b). Recently, a new representation method, dynamic image was proposed in (H.Bilen et al., 2016).

A dynamic image is an image that combines motion information of a video composed of multiple frames into a single image. Figure 2 shows an example of dynamic images. In dynamic images, motion in the video is emphasized, and the information in the non-moving parts is removed. Furthermore, dynamic image has the potential to fine-tune existing CNN models because it converts motion information into a single image. Experiments in (M.Shahid et al., 2019a) show that dynamic image performs better on average than the other methods.

The generation of dynamic images is based on the concept of rank pooling, which can be obtained via the parameters of a ranking function that encodes the temporal evolution of video frames. However, rank pooling requires optimization of the ranking function. In this paper, we use dynamic images obtained by

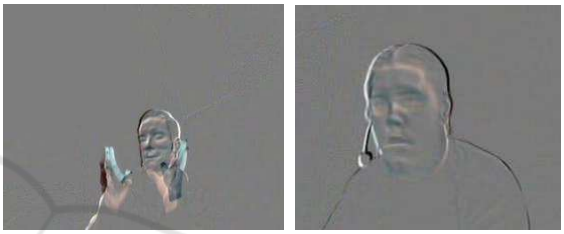

Figure 2: A dynamic image example.

approximate rank pooling, which is a more efficient method among rank pooling schemes. These algorithms are described in detail in (H.Bilen et al., 2016).

3.2 Multi-modal Transfer Module (MMTM)

An MMTM (H.R.V.Joze et al., 2020) is a method that, accepting multiple modalities as input, fuses features of each modality in the middle layers of a neural network model. Since the MMTM can have various layers and be implemented with minimal changes in the network structure, it can be used in many fields such as gesture recognition, speech enhancement, action recognition, and so on.

An overview of MMTM is shown in Figure 3. An MMTM consists of a squeeze step which summarizes a feature map in a channel, and an excitation step which weights each channel. Let us describe $\boldsymbol{A}$ and $\boldsymbol{B}$ in Figure 3 as feature maps extracted from each modality. The number of channels in $\boldsymbol{A}$ and $\boldsymbol{B}$ are represented as $C_{\boldsymbol{A}}$ and $C_{\boldsymbol{B}}$, respectively. For squeezing, a pooling method such as Global Average Pooling (GAP) (M.Lin et al., 2014) can be used. The squeezed values of $\boldsymbol{A}$ and $\boldsymbol{B}$ are denoted as $S_{\boldsymbol{A}}$ and $S_{\boldsymbol{B}}$, respectively. Next, in the excitation step,Then, the squeezed features are combined and adapt two full connection layers. Assuming that the fully connected feature map is Z and the number of dimensions of Z is denoted by

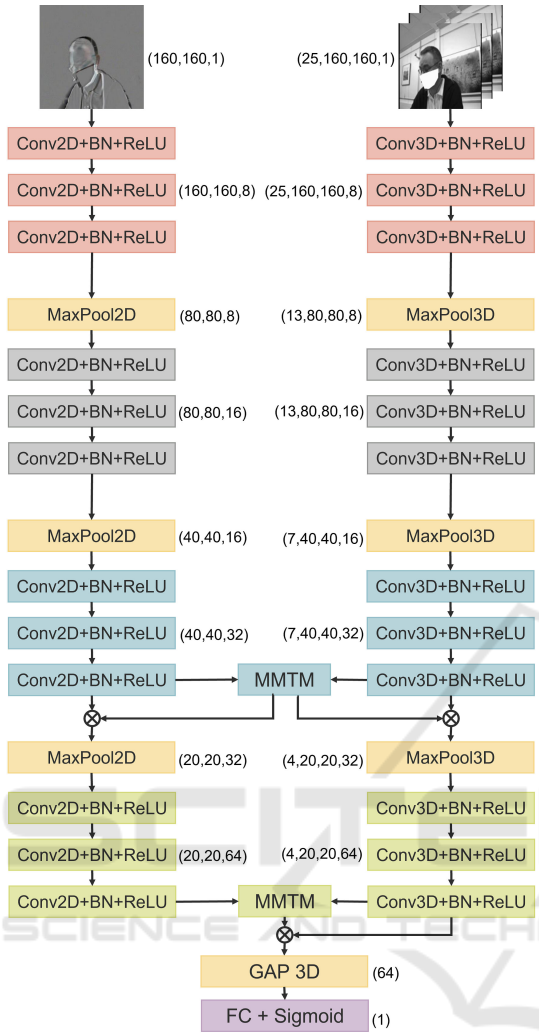

Figure 4: Details of our proposed model.

C_Z. Let the output feature maps for each modality are represented as $E_{\boldsymbol{A}}$ and $E_{\boldsymbol{B}}$ respectively. Finally, the sigmoid function(σ) is applied to the output of the full connection layer to convert the values in each dimension from 0 to 1. The value of each dimension can be regarded as the importance for each channel of the feature map. Features in the dimension of which value is close to 1 are retained, and the ones close to 0 are removed. Multiplying the importance by the original feature map adjusts the value for each channel of the feature maps. Assuming adjusted feature maps as $\boldsymbol{A}'$ and $\boldsymbol{B}'$ respectively. Further details can be found in (H.R.V.Joze et al., 2020). In this way, MMTM can adjust features extracted from one modality using information from the other modalities.

3.3 Our Proposed Model

In (M.Shahid et al., 2019b), fine tuning of ResNet50 (K.He et al., 2016) with dynamic images and VVAD were performed. However, the performance was insufficient for the dataset used in this study, which will be introduced later: about 65% for data without masks and about 60% for data with masks. Therefore, we propose another technique to fuse the video data stream and dynamic image stream much more efficiently.

Our proposed model is shown in Figure 4. In the figure, we denote a convolutional layer as 'Conv' and a batch normalization as 'BN'. In each of the video data and dynamic image streams, a set of three convolutional layers followed by a max pooling layer is repeated four times, resulting 12 convolutional layers. All convolutional layers have the same kernel size $= (3 \times 3 \times 3)$ and stride=1. In addition, zero padding is also performed. Features of the whole video are finally extracted from the video data stream, while from the dynamic image stream features focusing only on body movements in the video are extracted. After the last convolutional layers of the third and fourth layer sets respectively, feature maps before pooling layers are given to an MMTM block, and the output weights are multiplied by the feature map. Note that MMTM blocks are inserted based on the paper (H.R.V.Joze et al., 2020). We then perform GAP on the feature map of the video data stream by adapting MMTM, in order to use dynamic images as supplemental information. Finally, classification is performed to identify speech and non-speech, through full connection. Figure 5 shows the details of the MMTM block.

4 EXPERIMENTS

In this paper, we performed VVAD using a baseline 3DCNN model and our proposed model. We also visualize the areas of interest in each model.

4.1 Dataset

4.1.1 AMI Corpus

In this experiment, we used the AMI Corpus (J.Carletta et al., 2005), a multi-modal conference dataset. The AMI Corpus is a dataset consisting of synchronized audio and video data of each participant. The corpus also has videos of the entire conference room, projection mapping, whiteboard, and other information. In this paper, we carried out experiments using videos of each participant. Examples of

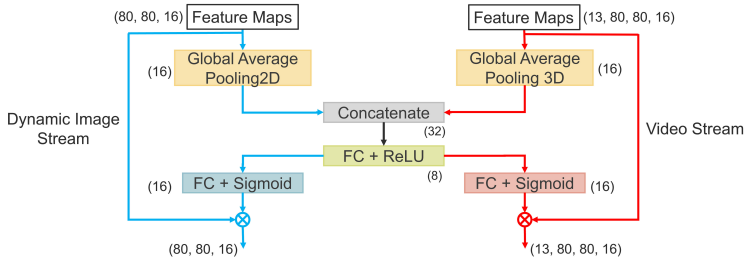

Figure 5: Details of MMTM block (note that here the number of filters in the feature map is 16).

Figure 6: Examples of AMI Corpus videos.

Figure 7: Examples of mask-wearing data.

participant videos are shown in Figure 6. The video size is 352×288, and the frame rate is 25 fps.

In the AMI Corpus, there are several scenario meetings which were prepared for the purpose of dataset creation, as well as natural non-scenario meetings. In the scenario conference, there were many scenes where each participant stood in front of the projection mapping, and people were often missing from the video. In the non-scenario meetings, participants were less likely to leave their seats because the meetings were generally interactive, and we can observe much more conversations among participants. Only the non-scenario conferences were used in this work.

4.1.2 Mask-wearing Data

As mentioned in Section 1, for VVAD, it is needed to investigate the influence that attendees in a conference wear masks. However, as long as we know there is no dataset in which we can observe mask-wearing people. Therefore, we created simulation data in which a speaker's mouth was artificially covered in a mask-like shape. Dlib (D.E.King, 2009) was firstly applied to obtain face landmarks if a speaker's face was found in an image. Using the obtained landmarks, we then generated a white-filled polygon to hide the mouth by OpenCV (G.Bradski, 2000). We finally created a video of a person wearing a mask by repeatedly applying this process to all frames in the video.

Figure 7 depicts examples of mask-wearing data.

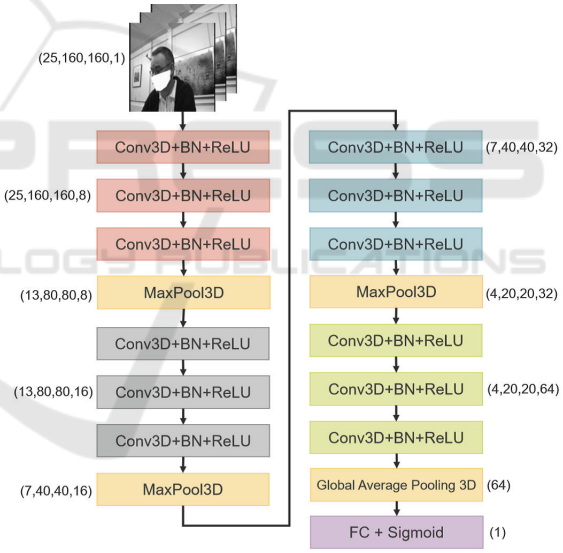

Figure 8: A baseline model structure.

4.1.3 Preprocessing

Because in a video a speaker generally looked small compared to the video size, we adjusted the size of the person in all the videos to be approximately the same by cropping the videos. According to speech activities, we labeled each frame in a video: 0 for non-speech or 1 for speech. The input data for a video stream was a set of 25 consecutive frames. We employed video data only when the same label was found and face detection was succeeded among 25 consecutive frames. To keep the continuity of the

Table 1: Data specification.

Data	# data	# attendees
Train	42,880	20
Validation	5,600	4
Test	29,928	17

Table 2: A confusion matrix.

		Prediction	
		non-speech	speech
Label	non-speech	TP	FN
	speech	FP	TN

data, the frame shift length of adjacent data was set to 20 frames. Regarding dynamic images, one image was created for every 25 frames. Both data were resized into 160×160, grayscaled, and normalized. Since there was a large difference in the number of preprocessed data between the two classes, the data of the larger class was randomly deleted so that the numbers of non-speech and speech data per video would be the same.

4.2 Experimental Setup

Using the preprocessed data, we trained a baseline model and our proposed model to test their accuracy. Details of the baseline model are shown in Figure 8. The proposed model is shown in Figure 4, and the squeeze method for MMTM is GAP.

We can see several meeting rooms in the AMI corpus. In this study, we used meeting videos recorded in two rooms among them. In order to conduct open experiments, the training and evaluation data were divided so that their meeting rooms were different. Similarly, the participants of the meeting were different between the training and evaluation data.

Details of the number of data are shown in Table 1. In addition, to verify the effect of wearing masks, the same experiment was conducted for videos without a mask. The optimization method for both the baseline model and the proposed model was Adam, and the learning rate was set to 0.001.

4.3 Metrics

4.3.1 Accuracy

We evaluated the baseline and our method by classification accuracy:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (1)$$

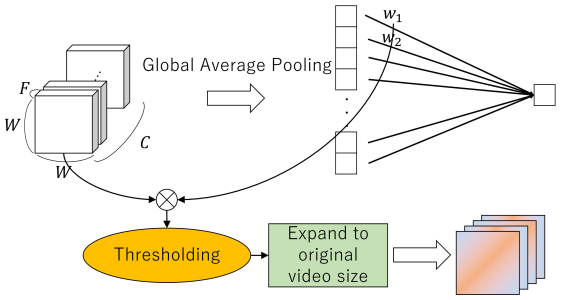

Figure 9: An overview of the Saliency tubes.

Table 3: VVAD classification accuracy (%).

Model	w/o mask	w/ mask
Baseline	84.78	74.98
Proposed	**86.07**	**75.14**

Table 4: A confusion matrix of the baseline with masks.

		Prediction	
		non-speech	speech
Label	non-speech	11,576	3,388
	speech	3,940	11,024

Table 5: A confusion matrix of our model with masks.

		Prediction	
		non-speech	speech
Label	non-speech	10,932	4,032
	speech	3,124	11,840

True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) in Equation 1 are defined in Table 2. We conducted the same experiments three times, and the average accuracy was computed as the result.

4.3.2 Visualization

As a visualization method, we used the Saliency tubes (A.Stergiou et al., 2019) method. Saliency tubes creates a heat map by multiplying the weights of the output layer by each filter of the feature map output from the last convolutional layer. We can finally visualize the areas of the input data that contribute to class prediction. An overview of the Saliency tubes is shown in Figure 9.

4.4 Results

The VVAD results of the baseline model and our proposed model are shown in Table 5. Table 4 and Table 5 show confusion matrices of the baseline model and our proposed model respectively, when they were trained and evaluated on the mask-wearing data.

First, according to Table 5, it is observed that the accuracy of the data without masks is roughly 10% higher than that with masks. This means that mouth information includes crucial cues for identification of voice activity. On the other hand, the accuracy with masks is still acceptable, indicating the effectiveness of using human gestures for VVAD. In both cases, visual-only VAD has enough performance which can compensate audio VAD degradation. Therefore, as mentioned already, a multi-modal VAD combining speech and lip animation seems to be effective.

Next, from the comparison of the baseline and proposed models, our model achieved slightly higher performance. We conducted the statistical test, finding that there was a significant difference between the baseline and our proposed model in the data without masks ($p < 0.01$). According to Table 4 and Table 5, we can see that in the baseline model sometimes misclassification from speech to non-speech occurred, while our proposed model had much more errors that non-speech segments were recognized as speech. This result suggests that employing dynamic images as supplementary information could make the model more sensitive to motion, and robust against the existence of masks. We checked the misrecognized data, and found that in those mislabelled segments human movements were often observed due to posture changes. VAD is usually performed also as a preprocessing of speech recognition, errors from speech to non-speech are crucial; once a segment is categorized into non-speech, the segment is no longer used for speech recognition and utterances in the segment are missing. From this standpoint, our model seems better than the baseline, thanks to dynamic images.

We also carried out visualization by Saliency tubes, to discuss how those models worked to the mask-wearing data. Figure 10 and Figure 11 illustrate the visualization results of the baseline model and the proposed model, respectively. Figure 10 indicates that the baseline model still focused on the background information as well as attendee's movements. In contrast, from Figure 11, it is confirmed that our proposed model correctly focused on attendee's face, head and body parts. These results suggest that adopting dynamic images has two functions: to reduce the effect of background information and to increase the importance of information about speaker's movements.

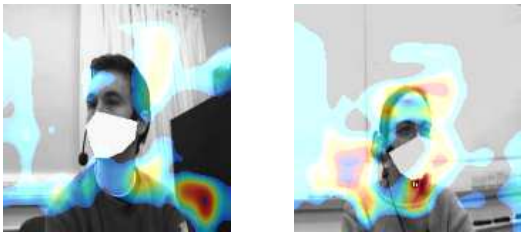

Figure 10: Visualization results of the baseline model.

Figure 11: Visualization results of our proposed model.

5 CONCLUSION

In this paper, we proposed visual-only VAD using human movements, as a preliminary step to multi-modal VAD using audio and visual information. Assuming the current pandemic situation, we focused on conference videos with participants virtually wearing masks to cover their mouths. In our proposed method, we employed not only raw video images but also dynamic images, so that our VAD model could focus much more on attendees by providing motion information as supplementary information. We chose the MMTM architecture that efficiently fuses two modalities: features from the motion information and original images. We conducted evaluation experiments also using the 3DCNN-based baseline model. As a result, the accuracy of our proposed model could be improved compared to the baseline, and our model could also work in the mask-wearing environments. According to visualization results, it is also confirmed that our model could focus much more on humans and their movements for VAD.

As our future work, it is necessary to perform multi-modal VAD using audio and video information. In noisy environments, when a mouth is not visible, the accuracy is expected to be improved by using both audio and the entire video. The effectiveness of multi-modal VAD needs to be confirmed by comparing audio-only VAD. We will also try to improve the performance of VVAD by combining state-of-the-art VVAD methods with our proposed approach. In this paper, we generated and used simulated artificial data in which participants wore masks, because there is no corpus having participants wearing real masks. One possible future attempt is to train and test our model with speakers who actually wear masks.

REFERENCES

A.Aubrey, B.Rivet, Y.Hicks, L.Girin, J.Chambers, and C.Jutten (2007). Two novel visual voice activity detectors based on appearance models and retinal filtering. In *Proc. 2007 15th European Signal Processing Conference (EUSIPCO)*, pages 2409–2413.

A.Stergiou, G.Kapidis, G.Kalliatakis, C.Chrysoulas, R.Veltkamp, and R.Poppe (2019). Saliency tubes: Visual explanations for spatio-temporal convolutions. In *Proc. 2019 IEEE International Conference on Image Processing (ICIP)*, pages 1830–1834.

B.G.Gebre, P.Wittenburg, S.Drude, M.Huijbregts, and T.Heskes (2014a). Speaker diarization using gesture and speech. In *Proc. INTERSPEECH 2014*, pages 582–586.

B.G.Gebre, P.Wittenburg, T.Heskes, and S.Drude (2014b). Motion history images for online speaker/signer diarization. In *Proc. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1537–1541.

B.Joosten, E.Postma, and E.Krahmer (2015). Voice activity detection based on facial movement. In *Journal on Multimodal User Interfaces volume 9*, pages 183–193.

D.E.King (2009). Dlib-ml: A machine learning toolkit. In *Journal of Machine Learning Research (JMLR) vol.10*, page 1755–1758.

G.Bradski (2000). The opencv library. In *Dr.Dobb's Journal of Software Tools*.

H.Bilen, B.Fernando, E.Gavves, and A.Vedaldi (2016). Dynamic image networks for action recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3034–3042.

H.R.V.Joze, A.Shaban, M.L.Iuzzolino, and K.Koishida (2020). Mmtm: Multimodal transfer module for cnn fusion. *Proc. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13286–13296.

J.Carletta, S.Ashby, S.Bourban, M.Flynn, M.Guillemot, T.Hain, J.Kadlec, V.Karaiskos, W.Kraaij, M.Kronenthal, G.Lathoud, M.Lincoln, A.L.Masson, I.Mccowan, W.Post, D.Reidsma, and P.Wellner (2005). The ami meeting corpus: A pre-announcement. In *Proc. Machine Learning for Multimodal Interaction (MLMI)*, pages 28–39.

K.He, X.Zhang, S.Ren, and J.Sun (2016). Deep residual learning for image recognition. In *Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

K.Hoover, S.Chaudhuri, C.Pantofaru, M.Slaney, and I.Sturdy (2017). Putting a face to the voice: Fusing audio and visual signals across a video to determine speakers. In *arXiv:1706.00079*.

K.Stefanov, J.Beskow, and G.Salvi (2017). Vision-based active speaker detection in multiparty interactions. In *Proc. Grounding Language Understanding (GLU)*, pages 47–51.

M.Cristani, A.Pesarin, A.Vinciarelli, M.Crocco, and V.Murino (2011). Look at who's talking: Voice activity detection by automated gesture analysis. In *Proc. AmI Workshops 2011*, pages 72–80.

M.Lin, Q.Chen, and S.Yan (2014). Network in network. In *Proc. International Conference for Learning Representations (ICLR)*.

M.Shahid, C.Beyan, and V.Murino (2019a). Comparisons of visual activity primitives for voice activity detection. In *Proc. Image Analysis and Processing – ICIAP 2019*, pages 48–59.

M.Shahid, C.Beyan, and V.Murino (2019b). Voice activity detection by upper body motion analysis and unsupervised domain adaptation. In *Proc. 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 1260–1269.

M.Shahid, C.Beyan, and V.Murino (2021). S-vvad: Visual voice activity detection by motion segmentation. In *Proc. 2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2331–2340.

N.Latif, A.V.Barbosa, E.Vatiokiotis-Bateson, M.S.Castelhano, and K.G.Munhall (2014). Movement coordination during conversation. *PLOS ONE*, pages 1–10.