Comparisons of Visual Activity Primitives for Voice Activity Detection

Muhammad Shahid*[1], Cigdem Beyan*[1], and Vittorio Murino[1,2]

[1] Pattern Analysis and Computer Vision, Istituto Italiano di Tecnologia, Genova, Italy
[2] Department of Computer Science, University of Verona, Verona, Italy
{Muhammad.Shahid,Cigdem.Beyan,Vittorio.Murino}@iit.it

Abstract. Voice activity detection (VAD) with solely visual cues have usually performed by detecting lip motion, which is not always feasible. On the other hand, visual activity (e.g., head, hand or whole body motion) is also correlated with speech, and can be used for VAD. Convolutional Neural Networks (CNNs) have demonstrated significantly good results for many applications including visual activity-related tasks. It can be possible to exploit CNN's effectiveness to visual-VAD when whole body visual activity is used. The way visual activity is represented (called visual activity primitives) to be given to a CNN as input, might be important to perform an effective VAD. Some primitives might result in better detection and provide consistent VAD performance such that the detector works equally well for all speakers. This is investigated, for the first time, in this paper. Regarding that, we compare visual activity primitives quantitatively in terms of the overall performance and the standard deviation of the performance, and qualitatively by visualizing the discriminative image regions determined by CNN trained to identify VAD classes. We perform a data-driven VAD with a person-invariant training i.e., without using any labels or features of the test data. This is unlike the state-of-the-art (SOA), which realizes a person-specific VAD with hand-crafted features. Improved performances with much lower standard deviation as compared to SOA are demonstrated.

Keywords: voice activity detection · visual activity · dynamic images · optical flow · social interactions.

1 Introduction

Voice Activity Detection (VAD) consists in automatically detecting "Who is Speaking and When" in an audio/video recording. Automatic VAD contributes various applications of human-human interaction analysis, human-computer (robot) interaction and also many industrial applications. As an example, for analysis of human-human interactions, VAD can be used to extract speaking turn-based nonverbal features (e.g., the length of the speech, the length of the overlapping

* Equal contribution

speech, etc.), which later on can be used to detect personality traits (e.g. [17]), dominance (e.g., [15]) or emergent leaders (e.g., [1]). Performing an accurate VAD can allow a robot (or a computer) to reply to a specific interlocutor when there is more than one person in a human-robot interaction environment [5]. Video conferencing systems can utilize VAD to present the video of the speaking person only during multi-person meetings. Additionally, an effective VAD can improve video navigation and retrieval, speaker model adaptation to enhance speaker recognition, and speaker attributed speech-to-text transcription [9].

Traditionally, VAD is performed by processing audio only, which is typically called speaker diarization [21]. On the other hand, multimodal approaches, normally referred to as active speaker detection [22, 6], have become popular, mostly adopting video and audio modalities. Multimodal approaches have either modelled the speech and visual cues such as facial, body cues jointly (e.g., [6]) or have performed audio speaker diarization while video has used to track/localize/associate a person to a speech (e.g., [11]). There are relatively few studies that have performed VAD based on video-based cues only (called visual-VAD in this paper). In fact, VAD with solely visual cues can be very desirable when the audio is not available due to technical or privacy related reasons. There can also be cases that the task of distinguishing voices robustly becomes very challenging such as in social gatherings, where much background noise is present. In such conditions, an effective visual-VAD can compensate audio speaker diarization.

The majority of the studies on visual-VAD have been performed based on lip motion detection, e.g., [18, 16, 12, 4]. Facial expressions [20], hand movement [14, 9, 4], head activity [9], and visual focus of attention (VFOA) [14] are other cues that have been utilized. On the other hand, visual activity cues extracted from whole body (without specifically focusing on a certain body part such as hands or head) [10, 7] can result in very effective VAD. For instance, whole upper body activity cues outperformed lip motion cues in [4].

There are diverse way to detect/represent the visual activity of a person to perform visual-VAD. For example, in [14], a combination of motion vectors, DCT (discrete cosine transform) coefficients and residual coding bit-rate were used. In [10], motion history images (MHI) were utilized. Optical flow has been another popular method to represent the visual activity as applied in [7]. Recently, in [4, 5], improved trajectory features that comprise of a concatenation of Histogram of Oriented Gradients (HOG), Histogram of Flow (HoF) and Motion Boundary Histogram (MBH) features were used. These examples all resulted in handcrafted visual activity features. On the other hand, deep learning models, such as Convolutional Neural Networks (CNN) have demonstrated state-of-the-art results for many research problems, including activity recognition and localization (e.g., [8, 2]), which are highly related to visual activity detection and representation. Therefore, there is no reason not to exploit the effectiveness of CNNs to visual-VAD. However, the way visual activity is initially represented (called visual activity primitives, from now on in this paper) to be fed to CNNs for training, can be critical to perform an effective VAD. In detail, some primitives

can result in better detection performance on average as compared to others, or can perform more consistent VAD performance so that the detector can work equally well for any speaker.

In this study, we compare the most popular visual activity primitives by modeling them with CNN for video-based VAD, which has never been addressed before. This comparative analysis is performed not only quantitatively but also qualitatively allowing us to better show why some primitives are performing better than others. Another contribution of this work is presenting improved performances as compared to the state-of-the-art (SOA) visual-VAD methods. The results obtained are also more stable such that the detection performances are equally good for all persons. The way we perform VAD is data-driven, does not use either labels or features belonging to the test data, thus, supports person-invariant training, i.e., it is not requiring model re-training for each new person. This is advantageous as compared to SOA presenting person-specific visual-VAD methods with hand-crafted visual activity features.

The rest of this paper is organized as follows. In Section 2, existing video-based VAD approaches are reviewed and the main differences between our work and theirs are highlighted. In Section 3, the details of the visual activity primitives and the way CNN fine-tuning is applied are described. The experimental setup is illustrated in Section 4 with a brief description of the dataset used. Subsequently, in Section 5, we compare the quantitative visual-VAD results of different visual activity primitives with the results of SOA, while qualitative comparisons are also performed among visual activity primitives. Finally, conclusive remarks and future work are sketched in Section 6.

2 Related Work

VAD solely based on video-based cues can be categorized in terms of the body parts investigated such as: face-based approaches that includes lip motion, head activity, face gestures, visual focus of attention (VFOA) etc., body-based methods, which contain hand gestures, full body motion, upper body motion, etc., or composition of these two categories.

As an earlier work on video-based VAD, in [18] the results of face detection, skin color, skin texture and mouth motion sensors have been combined and a Bayes Net model has been applied. In [16], facial movements corresponds to mouth, head and entire face have been extracted by Spatiotemporal Gabor filters, while mouth region gave the best VAD results. Haider et al. [12] analyzed the performance of head movement vs. head and lip movements together, and lip movement vs. lip and head movements together for speaker-dependent, speaker-independent or hybrid human-machine multiparty interactive dialogue settings. The results in that study [12] showed that head movement contributes to VAD significantly such that it outperforms lips movement except speaker-independent setting, and in overall, the fusion of head and lips movements perform the best. As seen, lip motion-based VAD is popular e.g.; [18, 16, 12] and effective. However, existing techniques are limited as detecting lip motion is not always possible.

For instance, when speaker presents a profile view to the camera or the camera resolution is low, or the speaker is far away from the camera or the speaker's lips is occluded by her hands, facial features detectors fail to detect the lips.

Hung et al. [14] analyzed the correlation between gaze and hand activities and speaking status given the assumptions that; the speaker is the one who moves most, and group's gaze (detected in terms of VFOA) is more likely to be on the speaker than on others. In that study [14], the visual activity of hands were detected by Discrete Cosine Transform (DCT) coefficients and residual coding bit-rate, while VFOA was determined by a Bayesian approach. The output features were tested with supervised and unsupervised learning in small group meeting datasets, and the results approved the assumptions regarding VFOA and hand motion. By using the same small group meeting dataset with [14], Gebre et al. [10] proposed using motion history images (MHI) as a likelihood measure of speaking activity, which resulted in promising performance as compared to [14], although only one type of cue was used. Detecting VFOA, head motion, body activity, lip motion and face is relatively less challenging in the meeting datasets [14, 10]. For instance, the detection of VFOA is drastically robust when there are individual cameras capturing each person specifically at close distance and in the meeting datasets [14, 10], the cameras are always static, there are more than one cameras capturing participants from their frontal view, and the places of the cameras are known by the participants.

On the other hand, Cristani et al. [7] performed visual-VAD for surveillance scenarios where the camera is located in a more distant place as compared to the meeting or human-machine interactive dialogue environments. In that method [7], a local video descriptor, which extracts the optical flow of human body, and encodes optical flow energy and complexity using an entropy-like measure was applied. Although, the results presented in [7] were successful, it is important to highlight that, the dataset they used has a top-view that already diminishes the possibility of occlusions and also the frames that the region of interests overlap (i.e., inter-person occlusions) were discarded from their analyses.

Directional audio information was used to label improved trajectory features extracted from upper body tracks of people as speaking or not-speaking in [4]. These labels were used for the training of an SVM to perform visual-VAD. Improved trajectories obtained for each 15 consecutive frames, pooled by a fisher vector representation were represented by the spatio-temporal features i.e.; the mean pixel location of the trajectory, and Histogram of Gradients (HoG), Histogram of Flow (HoF) and Motion Boundary Histogram (MBH). Chakravarty et al. [5] extended that scheme [4] to an online learning setting, starting from a generic VAD, which gradually adapts itself to a specific person. One drawback of that study [5] is, performing person-specific VAD, which requires training data for each new person. Additionally, even though, [5] performed person-specific VAD, the results were still fluctuated, such that VAD was performed well for some persons, while for others highly insufficient results were obtained.

More recently, deep learning-based feature extraction has become common for visual-VAD as well. For instance, in [20], face features have been extracted

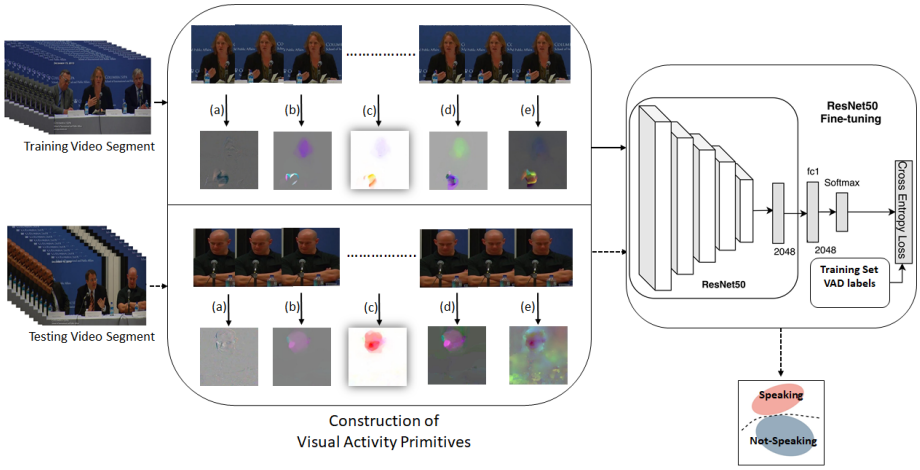

Fig. 1. The overall illustration of the methodology. See text for details.

from AlexNet, then Long Short-Term Memory (LSTM) has been used to model the temporal dependencies between face features over time, which was used to perform VAD in real-time multiparty interactions. That study is different than ours as focusing on face features and also limited due to requiring tightly cropped face images.

3 Methodology

The methodology applied to compare visual activity primitives is illustrated in Figure 1. During training, for each consecutive 10 RGB video frames, visual activity primitives: *a)* optical flow image (OFI) as proposed in [3], *b)* OFI as presented in [23], *c)* dynamic image (DI) as proposed in [2], *d)* the combination of OFI [3] with DI and *e)* the combination of OFI [23] with DI are obtained. For each type of image, a ResNet50 model is fine-tuned with the VAD labels (speaking or not-speaking). Given a test video, the same type of primitive whichever ResNet50 model is fine-tuned with, is obtained and, softmax is used to perform classification (end-to-end). Alternatively, the fine-tuned ResNet50 model is used to extract features, which are given to a Support Vector Machine (SVM) trained with the same training data ResNet50 is fine-tuned. The predicted label corresponds to the test video frames, those the test optical flow images or dynamic images (or the combinations of both) are constructed from.

3.1 Visual Activity Primitives

A video segment having 10 frames is given as an example in Figure 2 with the five visual activity primitives obtained from it. This 10 frames are equal to: one RGB dynamic image (DI), three optical flow images (OFI) obtained as in [3], three OFI obtained as in [23] and optical flow based dynamic images i.e.; one DI image for each optical flow method. These primitives are described as follows.

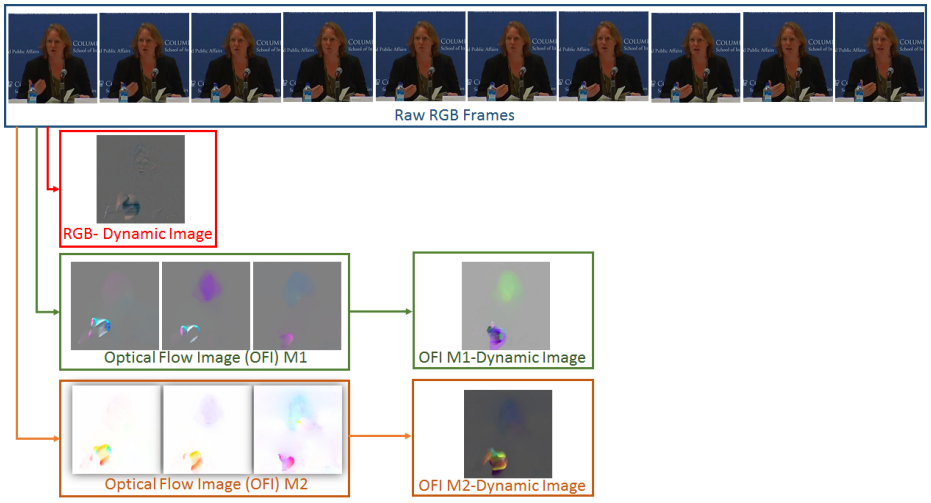

Fig. 2. Visual activity primitives: 1) RGB-dynamic image, 2) optical flow image M1; refers to [3], 3) optical flow image M2; refers to [23], 4) dynamic image obtained from optical flow image M1 and 5) dynamic image obtained from optical flow image M2. This example shows a video segment composed of 10 frames while the person is speaking.

Optical Flow Image (OFI) [3]: The main objective of the optical flow methods is to calculate a flow field by estimating the motion of pixels between two images. In this study, for all optical flow methods, this is performed for every F_i and F_{i+3} frames, such that F_{i+1} and F_{i+2} are discarded from the calculation. In other words, for every 30 frames (equals to 1 second for the dataset used), we obtain 10 OFIs. We discard the frames F_{i+1} and F_{i+2} to be able to better represent the motion because the image differencing applied to the consecutive frames showed that, the motion between two successive frames are very small, i.e. not creating informative flow images.

Brox et al. [3] presents a variational approach that applies a coarse-to-fine warping strategy to combine three assumptions: the gradient constancy, the grey-value constancy and the spatio-temporal smoothness constraint of the optical flow estimation. The gradient constancy deals with the aperture problem while the grey value constancy assumption makes the method robust against grey value changes. The spatio-temporal smoothness constraint allows to estimate the displacement of a pixel only locally by taking the interaction between neighbouring pixels into account. Given these, there are three parameters to be set: the weight between the grey value and the gradient constancy assumption, the smoothness parameter and the Gaussian convolution parameter to pre-process the input images. In our experiments smoothness parameter is 80, weight is 5 and Gaussian parameter is 0.9, which are empirically found. Once the optical flow is computed as described, we obtain the flow RGB images such that the first two channels are obtained from x and y flow values, respectively. The x and y flow values are centered around 128 and, then they are multiplied by a scalar such that they fall between 0-255. The third channel is created with the flow magnitude.

Optical Flow Image (OFI) [23]: This method is also a variational method, uses total variation regularization with L1-norm and applies point-wise thresholding strategy. Its objective is to preserve the edges and discontinuities in flow field while being robust against to the illumination changes, occlusions and noise. For visualization purpose, the optical flow field in x and y directions are normalized in the range of [-1, 1], which is further converted into HSV color space such that hue (H) indicates the direction, saturation (S) is represented by magnitude of flow field and value (V) is fixed to 255. Then, the optical flow images are obtained by converting them from HSV color space to RGB space.

Dynamic Image (DI) [2]: The objective of dynamic image [2] is to obtain a compact representation of a video sequence summarizing the appearance and dynamics of it. DI discards the static pixels such as background pixels and focuses on the object in an action. Construction of a DI contains rank-pooling that encodes the temporal evolution of the frames. The resulting DI can be used to fine-tune any CNN model. Herein, DIs are obtained from RBG data (i.e., raw video frames) or from OFIs extracted as described above.

3.2 ResNet50 Fine-tuning

Training a CNN from scratch might not be effective if the size of the data is limited. In this case, an alternative way is to fine-tune a pre-trained CNN model. Given the better performance of ResNet50 as compared to many other architectures [13], all the analysis regarding visual activity primitives are applied by fine-tuning ResNet50 (pre-trained on ImageNet dataset). During fine-tuning, a fully-connected layer having 2048 neurons is added after the final convolution layer. Its weights are randomly initialized and are updated during training. The weights of convolution layers are not updated. This model is trained with an end-to-end manner while cross entropy loss function, Adam optimizer, and $10e^{-5}$ learning rate are applied for 20 epochs.

The training data (more details are given in Section 4) is highly imbalanced such that there are a lot more not-speaking segments than speaking segments, which can mislead the classification task and result in poor performance. To overcome this, the training data in each batch (in total 128 samples) is balanced such that equal amount of randomly selected speaking and not-speaking samples (64 samples for each) are used. Furthermore, data augmentation is also applied such that some randomly selected training images are horizontally flipped and/or a 64x64 randomly selected patch is replaced with the mean value of the images, which can be observed as a dropout in input layer.

3.3 Classifier Learning and Inference

The ResNet50 fine-tuning of all visual activity primitives are applied with Softmax, which are used to classify the test data as speaking or not-speaking as well. Additionally, we apply a linear SVM for the best performing visual activity primitives to perform more fair comparisons with SOA [5]. The SVM kernel parameter C is taken as 10^k while $k = \{-4, -3, -2, -1, 0, 1, 2, 3, 4\}$.

Table 1. F1-scores (%). AVG and STD stand for average and standard deviation of F1-scores of all speakers, respectively. W, OFI, DI mean window size, optical flow image and dynamic image, respectively. The best results are emphasized in bold-face.

Method	Bell	Bollinger	Lieberman	Long	Sick	AVG	STD	Details
[5]	82.90	65.80	73.60	86.90	81.80	78.20	8.45	W=10, SVM
[5]	90.30	69.00	82.40	96.00	89.30	85.40	10.36	W=100, SVM
[6]	**93.70**	83.40	86.80	**97.70**	86.10	**89.54**	5.94	W=10
OFI [3]	84.01	69.25	68.8	53.31	68.19	68.71	9.71	Softmax
OFI [23]	85.63	81.73	80.12	69.36	70.83	77.53	6.35	Softmax
RGB-DI	86.07	**93.30**	**91.88**	73.62	86.34	**86.24**	6.94	Softmax
RGB-DI	86.34	**93.78**	**92.34**	76.09	86.25	**86.96**	6.24	SVM
OFI [3]-DI	84.08	72.27	80.57	60.01	68.89	73.164	8.56	Softmax
OFI [23]-DI	**89.97**	86.56	85.15	82.46	85.43	**85.91**	**2.44**	Softmax
OFI [23]-DI	89.16	88.82	85.82	81.39	85.97	**86.23**	**2.79**	SVM

4 Experimental Setup

The visual activity primitives are compared using publicly available dataset, called Columbia [5], which contains a 87 minutes-long video (frame rate: 30 frames per second) of a panel discussion. The field of the view of the camera changes to focus on smaller groups of panelist at a time. Following SOA, we only focus on the parts of the video where there is more than one person in the frame and discard any person in the margins of the video. This results in 5 speakers (Bell, Bollinger, Lieberman, Long, Sick) out of 7, while 2-3 speakers are visible per frame. In order to compare our results with SOA [5], we use the VAD labels (speaking/not-speaking) belonging to these 5 persons for each video frame. As per the performed analyses, the whole upper body motion of each speaker is used (in other words, the entire body parts that are visible). Finally, leave-one-person-out cross validation with F1-score as the evaluation metric is used still for comparative purposes [5].

5 Results

The best SOA results [5], [6] and the best result of each visual activity primitive with Softmax are given in Table 1. For the best performing visual activity primitives, their results with SVM are also given. As seen, the average performance of the visual activity primitives: RGB-DI, and OFI [23]-DI are the best out of all primitives and they also perform better than visual modality based SOA [5] method, no matter Softmax or SVM is used. Among all the SOA based on multi-modality, the method in [6] performed best as it uses audio and lip based visual information. The lip based visual information is not always reliable if the subject is more expressive through body motion. As shown in Table 1 the performance of RGB-DI for Bollinger and Lieberman is quite high as compared to multi-modality based SOA method [6], where in case Long (subject), it is the opposite. The performance of SOA [5] is highly dependent on the choice of window size (W) of temporal continuity algorithm that is based on the heuristics that if a person is speaking it is more likely that she will continue speaking for a

while rather than stop speaking. Using temporal continuity largely corrected the mis-classification results, but it is not clear how the window size of the temporal continuity should be selected to obtain accurate VAD results. Given that we create dynamic images for each 10 consecutive frames, it can be fairer to compare the performances with SOA [5] while W is equal to 10. In this case, all visual activity primitives except OFI [3] and OFI [3]-DI, perform better than SOA [5].

Better average visual-VAD performance is definitely very important but having low VAD standard deviation (STD) of all speakers while still performing well on average, is also significant. In detail, the performance of SOA [5] has fluctuations such that it performs well for some persons (e.g., Long: 86.90%), while performs highly worse for some others (e.g., Bollinger: 65.89%). This can be observed from the high STD values, 8.45% and 10.36% as well. In other words, this means that SOA [5] is not able to overcome domain-shift problem such that the distributions of training data and the test data are different from each other, which results in poorer VAD performance for some speakers. Domain-shift problem is highly possible for visual-VAD given that the way people moves while speaking varies a lot from person to person, resulting in dissimilar visual activity representations, as also mentioned in the psychology literature. On the other hand, the performance of any visual activity primitives is more consistent showing the superiority of $fc1$ features of ResNet50 as compared to the features of SOA. Especially, OFI [23]-DI is able to detect speaking and not-speaking video segments equally well for every speaker.

5.1 Qualitative Analysis

Given 4 video segments, each composed of 10 video frames, two of them having the ground-truth label (GT) "speaking" and other two having GT "not-speaking", we visualize the class activation maps in Figure 3 using the approach in [19] for the ResNet50 fine-tuned for VAD using visual activity primitives separately. Grad-CAM [19] is used to localize class-discriminative regions while they are overlaid with the intermediate raw RBG frame of the corresponding video segment in Figure 3.

For the video segments having GT=speaking, it is expected that head and hand motions are detected as the body of the person is more stable. Out of all, OFI [3] (M1) is weaker to detect these motions, while RGB-DI and OFI [23](M2)-DI localize the hands and head motions the best. For video segments having GT=not-speaking, in the first one, the person is slightly raising her hands up, whereas in the second one, the person is drinking water. RGB-DI and especially OFI M2-DI are still good at detecting the motions and more importantly, they are able to differentiate these types of motions from the motions during speech, i.e., they classify the frames correctly. However, OFI-M1 localizes other parts of the image such as background or the area close to person's shoulder, where the motion is very subtle to allow the correct classification of these frames. These results are in line with the quantitative results, showing that RGB-DI and OFI [23](M2)-DI are better to localize the motion correlated with speech.

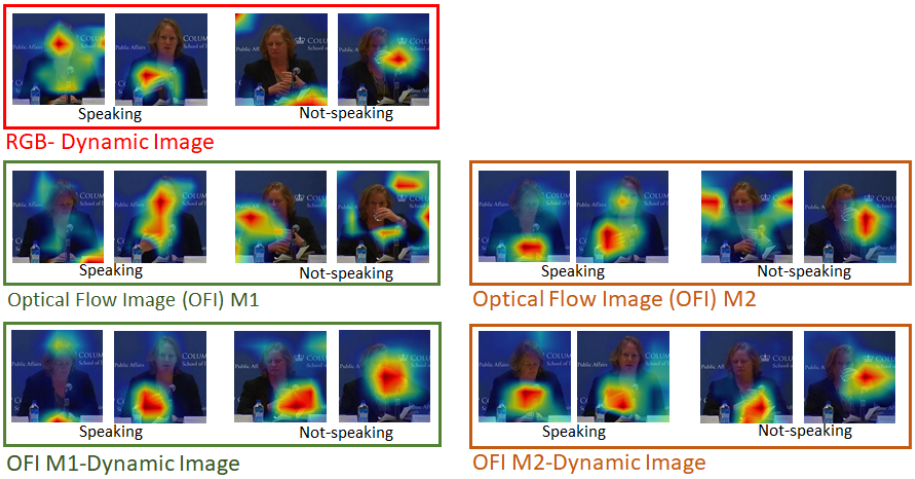

Fig. 3. The visualization of the class-discriminative regions overlaid with the intermediate raw RGB frame of the video segments when ResNet50 trained with visual activity primitives separately is used. Red regions in the heat map correspond to the high scores for the ground-truth class. M1 refers to [3] and M2 refers to [23].

6 Conclusions

We have addressed video-based voice activity detection (VAD) task with cues from whole body motion with a data-driven person-invariant setting. A detailed analysis was realized to compare the visual activity primitives representing the body motion, which are fed into CNNs to learn an effective VAD model. Some visual activity primitives resulted in better detection on average, while performing equally well for all speakers. Our detection results are also better on average and more consistent than the current literature.

As future work, a novel, effective way of combining these visual activity primitives will be investigated to perform visual-VAD in more complex scenarios such as in crowd or multiparty egocentric video streams, after construction of new benchmark datasets.

References

1. Beyan, C., Capozzi, F., Becchio, C., Murino, V.: Prediction of the leadership style of an emergent leader using audio and visual nonverbal features. IEEE Trans. on Multimedia **20**(2), 441–456 (2018)
2. Bilen, H., Fernando, B., Gavves, E., Vedaldi, A., Gould, S.: Dynamic image networks for action recognition. In: IEEE CVPR (2016)
3. Brox, T., Bruhn, A., Papenberg, N., Weickert, J.: High accuracy optical flow estimation based on a theory for warping. In: ECCV. pp. 25–36 (2004)
4. Chakravarty, P., Mirzaei, S., Tuytelaars, T., hamme, H.V.: Who's speaking?: Audio-supervised classification of active speakers in video. In: ACM ICMI. pp. 87–90 (2015)

5. Chakravarty, P., Tuytelaars, T.: Cross-modal supervision for learning active speaker detection in video. In: IEEE ECCV. pp. 285–301 (2016)
6. Chung, J.S., Zisserman, A.: Learning to lip read words by watching videos. CVIU **173**, 76–85 (2018)
7. Cristani, M., Pesarin, A., Vinciarelli, A., Crocco, M., Murino, V.: Look at who's talking: Voice activity detection by automated gesture analysis. In: International Joint Conference on Ambient Intelligence. pp. 72–80 (2011)
8. Donahue, J., Hendricks, L.A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Darrell, T., Saenko, K.: Long-term recurrent convolutional networks for visual recognition and description. In: IEEE CVPR. pp. 2625–2634 (2015)
9. Gebre, B.G., Wittenburg, P., Heskes, T.: The gesturer is the speaker. In: IEEE ICASSP. pp. 3751–3755 (2013)
10. Gebre, B.G., Wittenburg, P., Heskes, T., Drude, S.: Motion history images for online speaker/signer diarization. In: IEEE ICASSP. pp. 1537–1541 (2014)
11. Gebru, I.D., Ba, S., Li, X., Horaud, R.: Audio-visual speaker diarization based on spatiotemporal bayesian fusion. IEEE Trans. PAMI **40**(5), 1086–1099 (2018)
12. Haider, F., Campbell, N., Luz, S.: Active speaker detection in human machine multiparty dialogue using visual prosody information. In: IEEE GlobalSIP. pp. 1207–1211 (2016)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE CVPR. pp. 770–778 (2016)
14. Hung, H., Ba, S.O.: Speech/non-speech detection in meetings from automatically extracted low resolution visual features. In: IEEE ICASSP (2010)
15. Jayagopi, D.B., Hung, H., Yeo, C., Gatica-Perez, D.: Modeling dominance in group conversations using nonverbal activity cues. IEEE Transs. on Audio, Speech, and Language Processing **17**(3), 501–513 (2009)
16. Joosten, B., Postma, E., Krahmer, E.: Voice activity detection based on facial movement. Journal on Multimodal User Interfaces **9**(3), 183–193 (2015)
17. Lepri, B., Subramanian, R., Kalimeri, K., Staiano, J., Pianesi, F., Sebe, N.: Employing social gaze and speaking activity for automatic determination of the extraversion trait. In: ACM ICMI-MLMI (2013)
18. Rehg, J.M., Murphy, K.P., Fieguth, P.W.: Vision-based speaker detection using bayesian networks. In: IEEE CVPR. vol. 2, pp. 110–116 (1999)
19. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: IEEE ICCV (Oct 2017)
20. Stefanov, K., Beskow, J., Salvi, G.: Vision-based active speaker detection in multiparty interaction. In: Int. Workshop Grounding Language Understanding. pp. 47–51 (2017)
21. Tranter, S.E., Reynolds, D.A.: An overview of automatic speaker diarization systems. IEEE Trans. on Audio, Speech, and Language Processing **14**(5), 1557–1565 (2006)
22. Vajaria, H., Sarkar, S., Kasturi, R.: Exploring co-occurence between speech and body movement for audio-guided video localization. IEEE Trans. CSVT **18**(11), 1608–1617 (2008)
23. Zach, C., Pock, T., Bischof, H.: A duality based approach for realtime tv-l1 optical flow. In: DAGM Conference on Pattern Recognition. pp. 214–223 (2007)