This CVPR 2020 workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Improved Active Speaker Detection based on Optical Flow

Chong Huang
University of California, Santa Barbara
Santa Barbara, CA, 93106
chonghuang@umail.ucsb.edu

Kazuhito Koishida
Microsoft Corporation
Redmond, WA, 98052
kazukoi@microsoft.com

Abstract

Active speaker detection refers to the task of inferring which (if any) of the visible people in a video is/are speaking. Existing methods based on audiovisual fusion are often confused by factors such as non-speaking facial motions, varied illumination, and low-resolution recording. To address these problems, we propose a robust active speaker detection model by incorporating the dense optical flow to strengthen the visual representation of the facial motion. These audio and visual features are processed by a two-stream embedding network, and the embeddings are fed into a prediction network for the binary speaking/non-speaking classification. To improve the learning efficiency of the entire network, we design a multi-task learning strategy to train the network. The proposed method is evaluated on the most challenging audiovisual speaker detection benchmark, the AVA-ActiveSpeaker dataset. The results demonstrate that optical flow can improve the performance of neural networks when combined with raw pixels and audio signal. It is also shown that our method consistently outperforms the state-of-the-art method [22] in terms of both the area under the receiver operating characteristic curve (+4.4%) and the balanced accuracy (+5.28%).

1. Introduction

Active speaker detection has received significant attention in speech-based interactive applications [24, 5, 2, 13]. Although these methods have achieved high precision for a clear frontal face, two key factors may impact the detection performance in real-world applications: First, these methods assume that the visible face landmarks (e.g., lip contour) are available, but the more complex situations (e.g., non-speaking facial motions, varied illumination, and low-resolution recording) may result in the failure of detecting the landmarks. Second, if the labels (speaking/non-speaking) of a certain identity are imbalanced, the model tends to learn the connection between the label and the person identity. Therefore, we aim to incorporate more robust features to address the above problems.

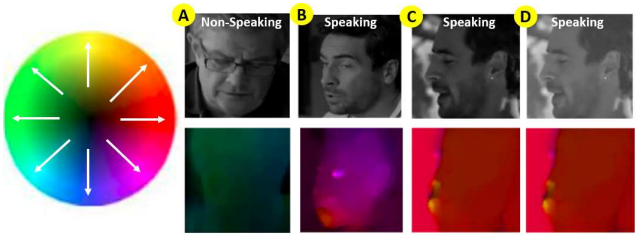

Figure 1. The gray-scale face image and the visualization of dense optical flow. We color the optical flow per pixel based on its magnitude and direction. A) The non-speaking face appears to have a consistent motion pattern on the entire face region. B) The speaking face has an obviously different motion pattern on the mouth region from other regions. C) and D) The face with different illuminations generates the same optical flow.

The visual perception theory [11] demonstrates that the human visual cortex contains two pathways: the ventral stream (which performs object recognition) and the dorsal stream (which recognizes motion), but existing active speaker detection methods mainly learn the visual embedding only from a stream of raw images. Therefore, we aim to advance the representative ability of the visual embedding by incorporating a motion stream.

Optical flow [17, 10] is an efficient representation of visual motion because it can measure the spatiotemporal variations between two subsequent images. The dense optical flow [10] describes the motion vectors of all pixels of the image. As Fig. 1 shows, the dense optical flow has several advantages over gray-scale images: 1) the dense optical flow illustrates the unique motion pattern (i.e., open or closed) of the speaking mouth; 2) the dense optical flow can capture the subtle motion of the textureless facial skin (e.g., cheek); 3) the face generates consistent optical flow under different illuminations; and 4) the optical flow removes the identity-related visual details, and it can avoid the network mislearning the correlation between the identity and the speaking label.

In this work, we propose an end-to-end active speaker detection system (see Fig. 3) by fusing raw face images,

dense optical flow and audio signal. To this end, we apply a two-stream network architecture to process the audio and visual inputs separately. In particular, we design two visual embedding strategies to fuse the optical flow and the raw face images. The visual and audio embeddings are concatenated and fed into a prediction network for binary speaking/non-speaking classification. To improve the learning efficiency of each network, we design a multi-task learning strategy with a combination of the overall classification loss, the intermediate classification loss from visual embedding networks, and the contrastive loss between visual and audio embeddings. The experimental results on the AVA-ActiveSpeaker dataset demonstrate that our proposed method achieves higher accuracy than the state-of-the-art method [22].

We introduce the proposed method in Sec. II. We present the experimental results in Sec. III. Finally, we give the conclusion and future work in Sec. IV.

2. Related Work

Active speaker detection based on vision: There is a large family of work on the active speaker detection problem. When audio information is unavailable, methods such as [13, 9, 4] rely on visual information from the face, lip and upper body. Everingham et al. [9] assumed the motion in the lip area implies speech, and the motion of facial landmarks along the video was used to determine the speaker. However, this method will suffer other face/mouth motions, such as eating and yawning. Besides lip motion, the movements of the head, upper body and hands of active speakers are important cues. Punarjay et al. [4] used the dense trajectory within upper bodies to detect the active speaker. Although this method achieves nearly perfect accuracy for clear frontal faces, it is at risk of failing to generalize to more complex situations, such as low resolution.

Active speaker detection based on audiovisual fusion: Compared to visual-only methods, the audio information provides an important clue for speaker identification, as speech rhythm and word pronunciation are closely related to facial motion and mouth shape. Chung et al. [5] learned the visual/audio embedding features by minimizing audio-video synchronization error, and predicted the active speaker by thresholding the distance between visual and audio embedded features. Joseph et al. [22] proposed an end-to-end multimodal active speaker detection framework with a two-stream convolution network for audiovisual feature extraction followed by a recurrent neural network for classification. However, given a stack of raw images as visual input, it is not intuitive for the network to learn the subtle differences between neighboring frames.

Application of optical flow in speech processing: The effectiveness of the optical flow has been investigated in many speech-related tasks. Mase et al. [19] discovered that the velocity of lip motions can be measured from optical flow data, which allows muscle action to be estimated. Pauses in muscle action result in zero velocity of the flow and are used to locate word boundaries. The pattern of muscle action is used to recognize the spoken words. Aubrey et al. [3] proposed a new method of voice activity detection using solely optical flow in the form of a speaker's mouth region. Wollmer et al. [26] combines the deep neural network and optical flow to model continuous emotions in an audiovisual affect recognition framework. However, these methods assume that the each subject has a clear front face and the subject's lips are clearly visible. If the face suffers from low resolution and occlusion, it is not feasible to detect the contour of the lip, which poses great challenges for the feature extraction and classifier design.

3. Proposed Method

3.1. Problem Definition

We aim at a learner $\dot{p} = f_\theta(\dot{I}, \dot{o}, \dot{a})$ to determine the active speaker, where θ is a set of model parameters. $\dot{I}$, $\dot{o}$ and $\dot{a}$ represents the image sequence, optical flow sequence and audio feature, respectively. Let $\dot{y}$ denote the binary labels (speaking/non-speaking) of image sequences; the problem is defined as the minimization of the binary classification error, which can be written as the cross-entropy loss:

$$Loss = -\dot{y}\log(f_\theta(\dot{I}, \dot{o}, \dot{a})). \tag{1}$$

In practice, we sample the video feature (face images and optical flow) with 20 fps and the audio feature with 100 fps.

3.2. Feature Representation

In this subsection, we introduce the feature representation of the audio signal, face images and optical flow.

Audio signal: The audio signal at 16 kHz is converted with a 25 ms FFT window to 64 mel-frequency bands, each of which is a representation of the short-term power spectrum of a sound on a non-linear mel scale of frequency. Each audio feature contains 48 frames of the mel-spectrogram.

Face images: We crop the grayscale image within the face bounding box as the feature. We utilize the entire face image instead of the mouth region because occlusion and low-resolution recording will affect the mouth detection.

Optical flow: We apply dense optical flow to extract the subtle motion of the face (see Fig. 2). We choose Gunner-Farnebacks algorithm [10] to calculate the dense optical flow for its efficient computation. The Gunner-Farnebacks algorithm [10] estimates the motion between two consecutive frames based on polynomial expansion. Firstly, each neighborhood of both frames is approximated by quadratic polynomials. Afterwards, considering these quadratic polynomials, a new signal is constructed by a

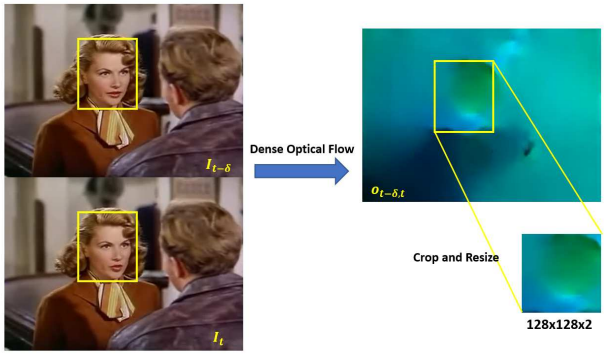

Figure 2. The facial optical flow extraction.

global displacement. Finally, this global displacement is calculated by equating the coefficients in the quadratic polynomials' yields.

Compared with the sparse optical flow (e.g, the Lucas-Kanade method [17]), the dense optical flow in the method [10] can extract the motion pattern of each pixel, even if the pixel lies on the smooth skin region. To capture more obvious displacement, we calculate the dense optical flow between the fifth previous frame and the current frame. Then we crop the dense optical flow with the face bounding box of the current image as the optical flow feature.

3.3. Network Architecture and Training

In this subsection, we describe the proposed network architecture and loss function for training.

3.3.1 Network architecture

A desired learning model should effectively fuse spatial and temporal information from multiple inputs. To this end, we follow an early fusion strategy to design our model, including a visual embedding network, an audio embedding network and a prediction network, as shown in Fig. 3 (a).

We design our model based on an early fusion strategy (i.e., combining audio and visual embeddings before prediction). We will introduce the network architecture (see Fig. 3) in terms of visual embedding network, audio embedding network and prediction network.

Visual Embedding Network: The input format to the visual network is a stack of 10 grayscale face images (i.e., $[I_{t-9}, I_{t-8}, ..., I_t]$) and 5 dense optical flows (i.e., $[o_{t-9,t-4}, o_{t-8,t-3}, ..., o_{t-5,t}]$), where $o_{t-5,t}$ refers to the optical flow from I_{t-5} to I_t. The face images and dense optical flows are resized by 128×128 (pixels). We design the following two methods to fuse the face image and optical flow streams.

Visual-Coupled Embedding: There is coupled spatial information between the facial appearance and the motion pattern. For example, the dense optical flow on the lip region has the 'up-and-down' pattern. As Fig. 3 (b) shows, we stack the face images and dense optical flow together as an input (i.e., the input dimension is $128\times128\times(10+5\times2)$. Given this input, we extract a 128-d visual embedding by using a modified MobileNet model [14]. In particular, we replace the last fully-connected layer in the original MobileNet with a global average pooling (GAP) layer [16] for visualizing the class activation map (detailed in Sec.4.4.5).

Table 1. Layer parameters of visual-coupled embedding network. Input size: width × height × depth. Filter shape: filter width × filter height × input channels × output channels. Conv: convolutional layer, Conv dw: depth-wise convolutional layer, GAP: global average pooling.

Type/Stride	Filter Shape	Input Size
Conv / s2	3x3x20x32	128x128x20
Conv dw / s1	3x3x32	64x64x32
Conv / s1	1x1x32x64	64x64x32
Conv dw / s2	3x3x64	64x64x64
Conv / s1	1x1x64x64	32x32x64
Conv dw / s1	3x3x64	32x32x64
Conv / s1	1x1x64x64	32x32x64
Conv dw / s2	3x3x64	32x32x64
Conv / s1	1x1x64x64	16x16x64
Conv dw / s1	3x3x64	16x16x64
Conv / s1	1x1x64x128	16x16x128
Conv dw / s2	3x3x128	16x16x128
Conv / s1	1x1x128x128	8x8x128
GAP	N/A	8x8x128

Independent Embedding: An alternative embedding method is to use two independent networks to process the face images and dense optical flow separately (see Fig. 3 (c)). The face and optical flow embeddings are concatenated as the visual embedding. Due to the limited availability of large amounts of annotated image data in the past, this two-stream architecture has been widely used in action recognition with pre-trained models such as FaceNet [23] and FlowNet [8].

In this work, each stream of independent embedding networks is implemented with the same network configuration (with different input sizes) as the visual-coupled embedding network. To fairly compare both visual embedding networks, we do not introduce any additional datasets to pre-train the model.

Audio Embedding Network: The input format is a sequence of mel-spectrogram bands (64×48×1), which is computed over the preceding 0.48 seconds of audio. We apply the audio embedding network in the method [22] to generate a 128-dim audio embedding.

Prediction Network: For each 10-frame clip, we concatenate both the visual and audio embeddings to form a composite feature as the input of the prediction network.

We design the prediction network based on a sequence-to-sequence (seq2seq) model [25], including an encoder and a decoder. The encoder and decoder are based on a long short-term memory (LSTM) [12] network with 512 units, while the decoder is followed by a 2-dim fully connected

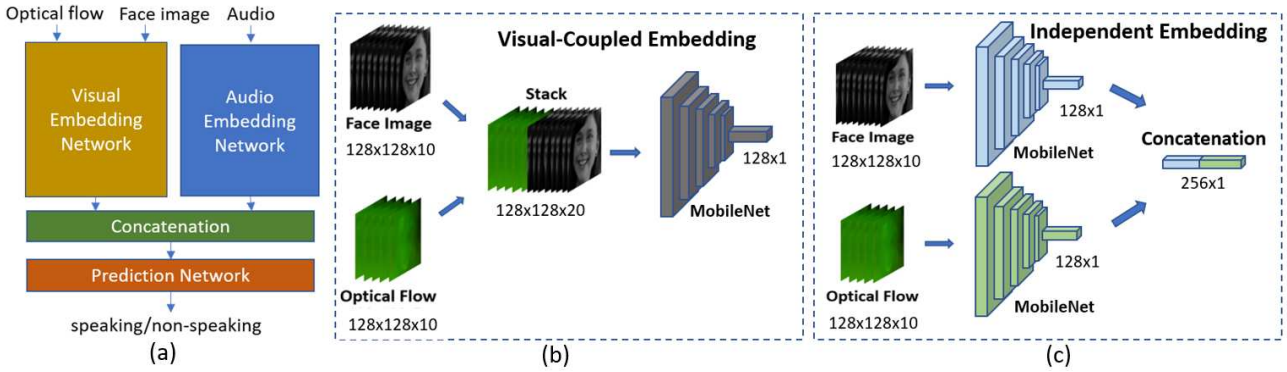

Figure 3. (a) The end-to-end active speaker detection framework. Two architectures of visual embedding network: (b) the visual-coupled embedding and (c) the independent embedding.

layer and a softmax layer. The decoder produces a sequence of prediction probability after receiving the state vector of the encoder conditioned on M inputs (we set M as 50 in this work). Each probability of the prediction sequence is calculated from an overlapping sliding window with a 10-frame length and 1-frame stride.

3.3.2 Loss Function

Considering that our model consists of two embedding networks and a prediction network, we use multi-task learning to improve the learning efficiency of each network. The loss function is designed from three aspects:

1) To minimize the prediction error, we define a cross-entropy loss between the predictions and labels:

$$L_f = -\dot{y}\log(\dot{p}) + \lambda\|w\|^2, \tag{2}$$

where the regularization hyperparameter λ is set as 0.01.

2) To improve the visual embedding (e_v), we minimize the visual-based classification error as follows:

$$L_v = -\dot{y}\log(h(e_v(\dot{I}, \dot{o}))), \tag{3}$$

where h is 2-dim fully-connected layer followed by a softmax layer to convert to probability.

3) To learn the audiovisual synchronization, we use the contrastive loss to minimize the distance between synchronized visual and audio embeddings, and maximize the distance for the non-synchronized pair.

$$\begin{aligned} L_c = (1-\dot{y})\frac{1}{2}\|e_v(\dot{I}, \dot{o}) - e_a(\dot{a})\|^2 \\ + \dot{y}\frac{1}{2}\max(0, D - \|e_v(\dot{I}, \dot{o}) - e_a(\dot{a})\|)^2, \end{aligned} \tag{4}$$

where e_v and e_a refers the visual embedding and audio embedding, respectively. The margin D is set as 1.0.

Based on the above discussion, the overall loss function can be written as follows:

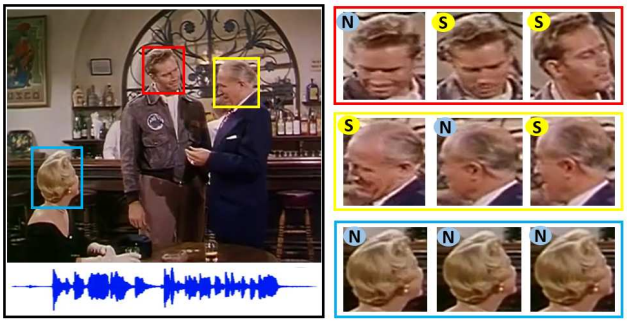

Figure 4. The Example of the AVA Active Speaker detection dataset. Left: The video, audio (waveform visualized below the frame) and bounding boxes of each face. Right: Each face over the frames is annotated with whether or not it is speaking (S: speaking; N: non-speaking).

$$L = L_f + \alpha L_v + \beta L_c, \tag{5}$$

where both α and β are set as 0.4.

4. Experiments

In this section, we describe the dataset and the implementation details, followed by the experimental results.

4.1. Dataset

The AVA Active Speaker detection dataset (AVA-ActiveSpeaker) was used to evaluate our method. The dataset contains temporally labeled face tracks in video, where each face instance is labeled as speaking or not.

Compared with other datasets (e.g., Columbia [4] and VoxCeleb [20]), the AVA-ActiveSpeaker dataset has much more labeled data, including 3.65 million face frames (38.5 hours), including 113 training videos (28,108 face tracks) and 32 testing videos (7,900 face tracks). In addition, the AVA-ActiveSpeaker dataset is very challenging due to low resolution (e.g., people in the distance) or occlusion (e.g., profile faces). 44.6% labeled faces have a size that is

smaller than 100 pixels wide, and 48.2% of the face mouth region cannot be detected by the state-of-the-art face landmark detection library Dlib [1].

4.2. Implementation Details

All the models in the experiments were trained under the same conditions. We cropped the labeled tracks using a 3-second (60-frame) sliding window with 1-second overlap. For tracks less than 3 seconds, we included them entirely.

Our implementation was based on the tensorflow1.13 toolbox and the model training was executed on a NVIDIA GT 1080 with 12 GB memory. The network was trained with batch normalization [15]. We utilized Adamax [6] to perform the optimization, with a learning rate of 0.002. We stopped training after 50 epochs.

4.3. Metrics

We adopted "area under the Receiver Operating Characteristic curve (AUC)" as a metric. Additionally, we used the balanced accuracy at a fixed threshold T to evaluate a model's performance. Because speaking faces only occupied 28% in the training set and 24% in the testing set, we chose a lower threshold to determine whether the face was the speaker. In this experiment, we set T as 0.3.

To compare the visual-coupled embedding and independent embedding, we utilize the ratio R between intraclass and interclass distance (see Eq. 6) to analyze the distribution of the embedding space. The more discriminative feature should be more separable with lower ratio value R.

$$R = \frac{intraclass(speaking) + intraclass(nonspeaking)}{interclass(speaking, nonspeaking)}$$
$$intraclass(A) = \frac{1}{N_A N_A} \sum_{i=1}^{N_A} \sum_{j=1}^{N_A} d(X_A^i, X_A^j) \quad (6)$$
$$interclass(A, B) = \frac{1}{N_A N_B} \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} d(X_A^i, X_B^j)$$

where $d(X_A^i, X_B^j)$ refers to the Euclidean distance between the ith embedding sample from class A and the jth embedding sample from class B.

4.4. Experimental Results

4.4.1 Performance on different features

Tab. 2 shows the performance of different features in the independent embedding network without contrastive loss. We abbreviate the facial image, optical flow and audio signal as F , O and A, respectively. It appears to be a trend: F+O+A $>$ F+A $>$ O+A $>$ F+O $>$ F $>$ O, which demonstrates that the optical flow can provide an additional (and important) clue for detection.

Table 2. The performance (AUC) among different features

F	O	F+A	O+A	F+O	F+A+O
0.8146	0.8014	0.8914	0.8597	0.8257	**0.9042**

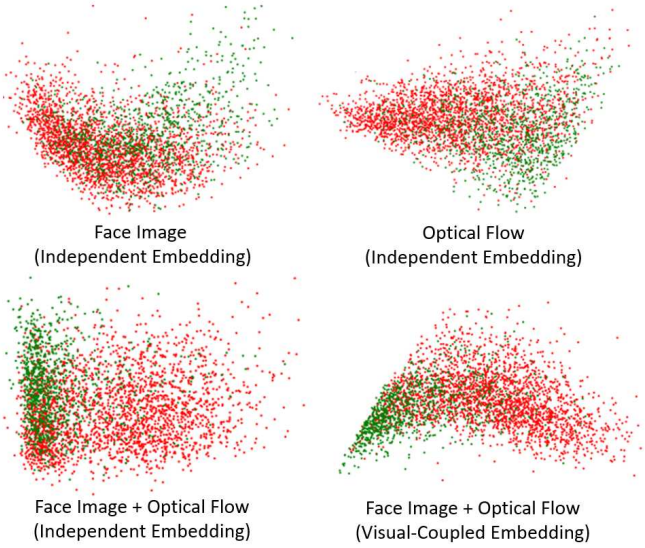

Figure 5. The distribution of the embedded feature after PCA. Each green/red point is an embedded feature for a speaking/non-speaking face, respectively.

4.4.2 Visual embedding network

We compare the models between independent embedding (IE) and visual-coupled embedding (VCE) in Tab. 3. Similar to Sec.4.4.1, we do not consider contrastive loss in model training. Although an independent embedding network has more parameters, a visual-coupled embedding network can improve by 0.55% (F+O) and 1.56% (F+O+A) because it can learn the spatial correlation between raw image and optical flow.

Table 3. The performance (AUC) between independent embedding and visual-coupled embedding.

(F+O, IE)	(F+O, VCE)	(F+O+A, IE)	(F+O+A, VCE)
0.8257	**0.8312**	0.9042	**0.9198**

To visualize the distribution of visual-coupled embedding and independent embedding, we randomly select 3800 embedded features from the test set and project them to the 2D space (see Fig. 5). Tab. 4 quantifies the separability of different embedding features in terms of ratio between intraclass and interclass distance. We can see that the fusion of optical low and face image appears to be more separable than the single-modality feature. Meanwhile, the visual-coupled embedding performs better than the independent embedding (R=1.2153 vs. 1.5219), proving that the visual-coupled embedding can generate more discriminative features. This can explain why the visual embedding network can achieve more accurate classification.

Table 4. The ratio between intraclass and interclass distance in terms of different embedding features.

(F, IE)	(O, IE)	(F+O, IE)	(F+O, VCE)
2.0742	2.4142	1.5219	**1.2153**

4.4.3 Contrastive loss function

We evaluate the influence of the contrastive loss (CL) in Tab. 5. It shows that the contrastive loss can consistently improve the prediction accuracy by 1%∼2%, which proves our assumption that the contrastive loss is beneficial to learn synchronization information between audio and visual features. In particular, the model (F+O+A, VCE, with CL), which has the visual-coupled embedding network with the optical flows trained on the contrastive loss, achieves the best performance among all the baselines.

Table 5. The performance (AUC) between the model with/without contrastive loss.

Features	IE		VCE	
	w/o CL	with CL	w/o CL	with CL
F+A	0.8914	**0.9095**	-	-
O+A	0.8597	**0.8701**	-	-
F+O+A	0.9042	**0.9237**	0.9198	**0.9317**

4.4.4 Comparison against the baseline [22]

We also compare our methods with the best baseline (recurrent model) in Joseph et al. [22] in Tab. 6. Note that this baseline model has the same architecture as one of our models (F+A, IE, w/o CL) but it is trained with an additional cross-entropy loss for an audio embedding network. It appears that the baseline performs worse than our equivalent model by 0.4% (AUC) and 2.15% (balanced accuracy), where the degradation can be explained by the fact that non-speaking and speaking faces may appear in the same video while sharing the same audio track.

Table 6. The performance (AUC and balanced accuracy) between our methods and the baseline [22].

Metrics	F+A, baseline [22]	F+A, IE, w/o CL	F+O+A, VCE, with CL
AUC	0.8874	0.8914	**0.9317**
balanced accuracy	0.8164	0.8379	**0.8692**

The model (F+O+A, VCE, with CL) achieves absolute 4.4% (AUC) and 5.28% (balanced accuracy) improvements over the baseline. In addition, we evaluate the baseline and our best model against different face sizes and orientations. We adopt the method in [18] to obtain the face orientation. Tab. 7 and Tab. 8 show that our proposed method consistently performs better than the baseline, even with a small face (i.e., the face whose width is less than 40 pixels) and side face (i.e., the orientation is larger than 60°). These results indicate the robustness of our model in adverse conditions.

Figure 6. The snapshots with different face sizes and orientations: (a) Face width: 318 pixels (red) vs 63 pixels (yellow); (b) Face orientations: 11°(red) vs 83°(yellow).

Table 7. The balanced accuracy over the face sizes

Model	Face Width (pixel)			
	[0-40)	[40-80)	[80-120)	[120+)
(F+A, baseline [22])	0.7351	0.7921	0.8301	0.8579
(F+O+A, VCE, with CL)	**0.7761**	**0.8279**	**0.8734**	**0.9181**

Table 8. The balanced accuracy over the face orientations

Model	Face Orientation (degree)			
	[0 ,20)	[20 ,40)	[40 ,60)	[60 ,90)
(F+A, baseline [22])	0.8627	0.8539	0.8318	0.8022
(F+O+A, VCE, with CL)	**0.9139**	**0.9029**	**0.8739**	**0.8419**

4.4.5 Visualizing sound sources

To understand which part of the face region contributes to the prediction, we use the same technique in [21] to visualize an activation map of the "speaking" class.

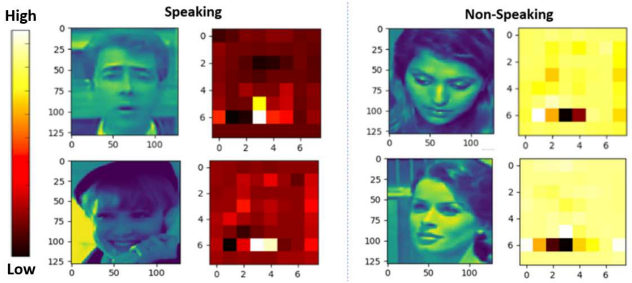

Figure 7. Class activation map of the class "speaking". We scale its range per image to compensate for the wide range of values.

Fig. 7 illustrates the class activation maps generated from the visual-coupled embedding network of our best model. The class activation maps of speaking faces show that the region with the highest response lies on the mouth region. Meanwhile, the mouth region of non-speaking faces shows lower responses than the other facial regions. The results demonstrate that our model can localize the mouth region and analyze it to distinguish between active and non-active speakers.

5. Conclusion and Future Work

In this study, we presented an improved active speaker detection framework by fusing face images, dense optical flow and audio streams. The multi-task learning allowed us to optimize the entire network effectively in an end-to-end manner. Our experimental results on the AVA-ActiveSpeaker dataset demonstrated that our method achieves significant accuracy improvements over the baseline method, particularly by the optical flow feature.

As we have shown, the performance of the model degrades with decreasing image resolution, which inspires us to use curriculum learning to improve the training process. There is still much room for improvements in our framework by using more sophisticated network architectures (e.g., VGG) and loss functions (e.g., ArcFace loss [7]). In addition, our model can be extended to other applications, such as lip reading, by incorporating the optical flow and changing the loss function.

References

[1] http://http://dlib.net.

[2] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 609–617, 2017.

[3] AJ Aubrey, Yulia Alexandrovna Hicks, and JA Chambers. Visual voice activity detection with optical flow. *IET image processing*, 4(6):463–472, 2010.

[4] Punarjay Chakravarty and Tinne Tuytelaars. Cross-modal supervision for learning active speaker detection in video. In *European Conference on Computer Vision*, pages 285–301. Springer, 2016.

[5] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pages 251–263. Springer, 2016.

[6] Matthieu Courbariaux, Itay Hubara, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. *arXiv preprint arXiv:1602.02830*, 2016.

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *arXiv preprint arXiv:1801.07698*, 2018.

[8] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.

[9] Mark Everingham, Josef Sivic, and Andrew Zisserman. Hello! my name is... buffy"–automatic naming of characters in tv video. In *BMVC*, volume 2, page 6, 2006.

[10] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis*, pages 363–370. Springer, 2003.

[11] Melvyn A Goodale and A David Milner. Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1):20–25, 1992.

[12] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[13] Ken Hoover, Sourish Chaudhuri, Caroline Pantofaru, Malcolm Slaney, and Ian Sturdy. Putting a face to the voice: Fusing audio and visual signals across a video to determine speakers. *arXiv preprint arXiv:1706.00079*, 2017.

[14] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[15] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. pages 448–456, 2015.

[16] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

[17] Bruce D Lucas, Takeo Kanade, et al. An iterative image registration technique with an application to stereo vision. 1981.

[18] Satya Mallick. Head Pose Estimation using OpenCV and Dlib. *https://www.learnopencv.com/head-pose-estimation-using-opencv-and-dlib/*.

[19] Kenji Mase and Alex Pentland. Automatic lipreading by optical-flow analysis. *Systems and Computers in Japan*, 22(6):67–76, 1991.

[20] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017.

[21] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018.

[22] Joseph Roth, Sourish Chaudhuri, Ondrej Klejch, Radhika Marvin, Andrew Gallagher, Liat Kaver, Sharadh Ramaswamy, Arkadiusz Stopczynski, Cordelia Schmid, Zhonghua Xi, et al. Ava-activespeaker: An audio-visual dataset for active speaker detection. *arXiv preprint arXiv:1901.01342*, 2019.

[23] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[24] Kalin Stefanov, Jonas Beskow, and Giampiero Salvi. Vision-based active speaker detection in multiparty interaction. In *Grounding Language Understanding GLU2017 August 25, 2017, KTH Royal Institute of Technology, Stockholm, Sweden*, 2017.

[25] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[26] Martin Wöllmer, Moritz Kaiser, Florian Eyben, BjöRn Schuller, and Gerhard Rigoll. Lstm-modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing*, 31(2):153–163, 2013.