Supplementary Material
S-VVAD: Visual Voice Activity Detection by Motion Segmentation

Muhammad Shahid[1,2], Cigdem Beyan[1], Vittorio Murino[1,3,4]
{shahid.muhammad, cigdem.beyan, vittorio.murino}@iit.it
[1]Pattern Analysis and Computer Vision, Istituto Italiano di Tecnologia, Italy
[2] DITEN, Università degli Studi di Genova, Italy
[3]Department of Computer Science, Università di Verona, Italy
[4]Huawei Technologies Ltd., Ireland Research Center, Ireland

1. The S-VVAD

The code of S-VVAD is publicly available and can be found in github.com/IIT-PAVIS/S-VVAD.

For two different video segments, examples of dynamic images, class activation maps (CAMs) [3] and masks, which are generated during the training of S-VVAD, are given in Figure 1. For both of the video segments, the panelist in the middle is the one who is speaking while others are not speaking.

2. Qualitative Results on Columbia Dataset [2]

Figure 2 and 3 show example localization results corresponding to persons, which were correctly detected as speaking and not-speaking, respectively. Given that trained Fully Convolutional Network (FCN) takes dynamic images as the input, we show the localization results by imposing them on the corresponding dynamic images.

These results show that S-VVAD is able to differentiate the body motion of the speakers and non-speakers. Some activities resulting in body motion are gesticulating, changing the head pose, opening a bottle, drinking, note taking, etc. Additionally, S-VVAD is able to distinguish body motions from the background motion occurring due to camera movement.

3. Qualitative Results on Modified Columbia Dataset

In Figure 4, example localization results obtained when S-VVAD was applied to Modified Columbia dataset are shown. The localization results (red for not-speaking, green for speaking) were imposed on the middle RGB video frame (recall that one dynamic image is obtained from 10 consecutive RGB video frames). Additionally, in that figure, we plotted the predicted bounding boxes (red for not-speaking, green for speaking), which were correctly detected given the ground-truth (intersection-over-union (IoU) > 0.5, see main paper for more detail).

As seen, there are some cases (columns 6, 7, 11 and 12), which the localization was performed correctly, (i.e., S-VVAD was able to detect the body motion associated with speech activity correctly) but the predicted bounding boxes were not plotted. This is because the predicted bounding boxes do not supply the IoU rule applied. For all other cases, VAD and corresponding localization were correctly performed.

4. Qualitative Results on RealVAD Dataset [1]

Experiments on RealVAD dataset [1] includes the cross-dataset analysis which is: training the S-VVAD with the whole Modified Columbia dataset (i.e., the images are composed of two or three persons) and testing the trained S-VVAD on the entire video frames of RealVAD dataset [1]. On the other hand, the baseline method [1] *i)* was trained and tested on RealVAD dataset (same-set analysis) while both training and testing images include a single person at a time and *ii)* was trained on Columbia dataset (i.e., the images composed of single person) and tested on RealVAD dataset's images composed of single person. These experiments are illustrated in Figure 5.

Following that, the localization results of the S-VVAD for the aforementioned cross-dataset analysis are also given in Figure 6. Even though there is a domain gap between Modified Columbia and RealVAD datasets, the results show that S-VVAD is able to localize the body motion of the speakers and non-speakers as well as being able to distinguish body motions from the background motion, e.g., the ones occurring when the person(s) in the back-row moves.

References

[1] C. Beyan, M. Shahid, and V. Murino. RealVAD: A real-world dataset and a method for voice activity detection by body mo-

tion analysis. *IEEE Transactions on Multimedia*, Early Access.

[2] P. Chakravarty and T. Tuytelaars. Cross-modal supervision for learning active speaker detection in video. In *ECCV*, pages 285–301, 2016.

[3] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE ICCV*, pages 618–626, 2017.

Figure 1. Images obtained during the training of the S-VVAD: dynamic images, class activation maps, and masks. The black, white and grey pixels of the generated masks correspond to the background, speaking and not-speaking, respectively.

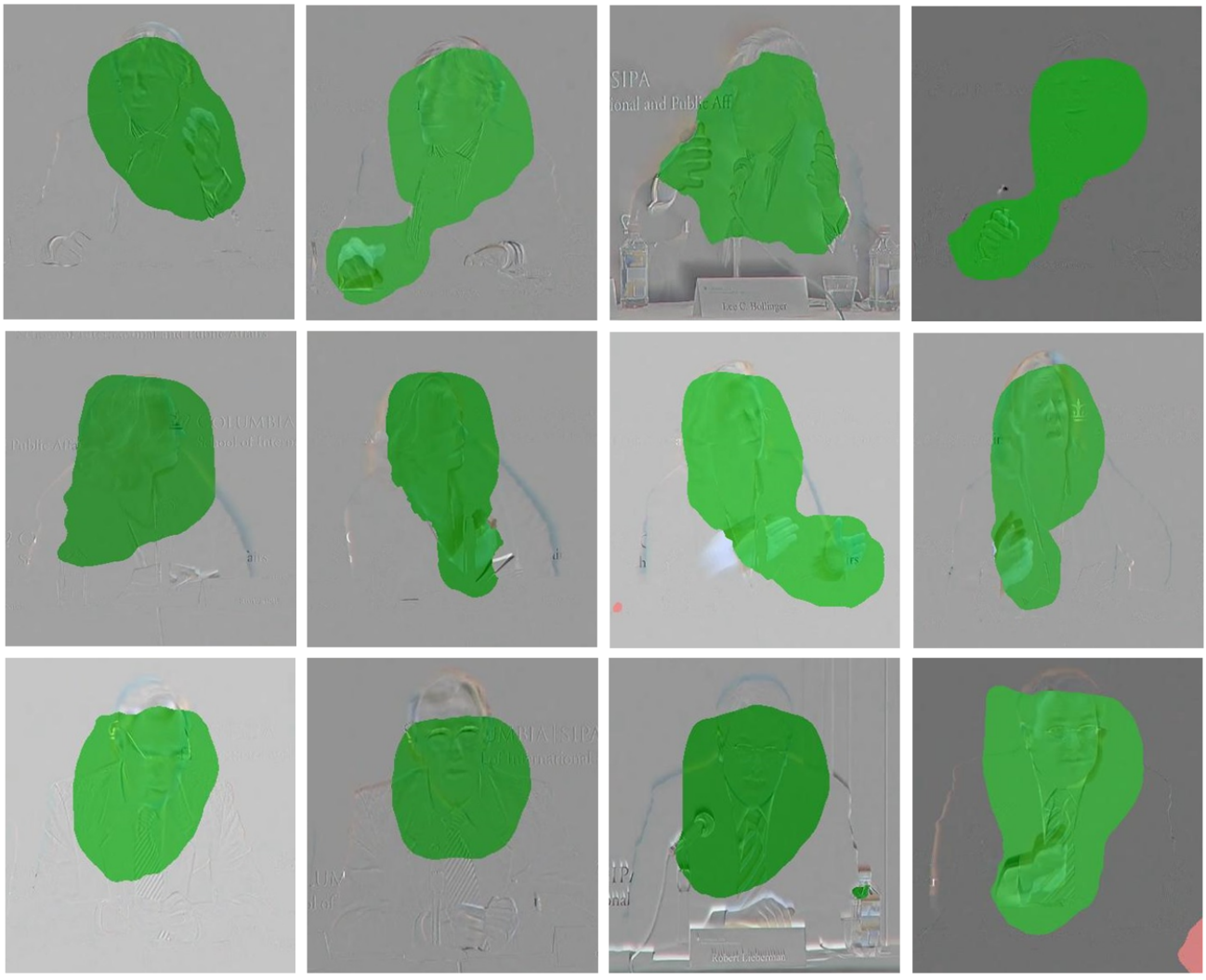


Figure 2. Example localization results (in green) for the persons (implying the body motion belong to a person) correctly detected as speaking when S-VVAD was applied to Columbia dataset.

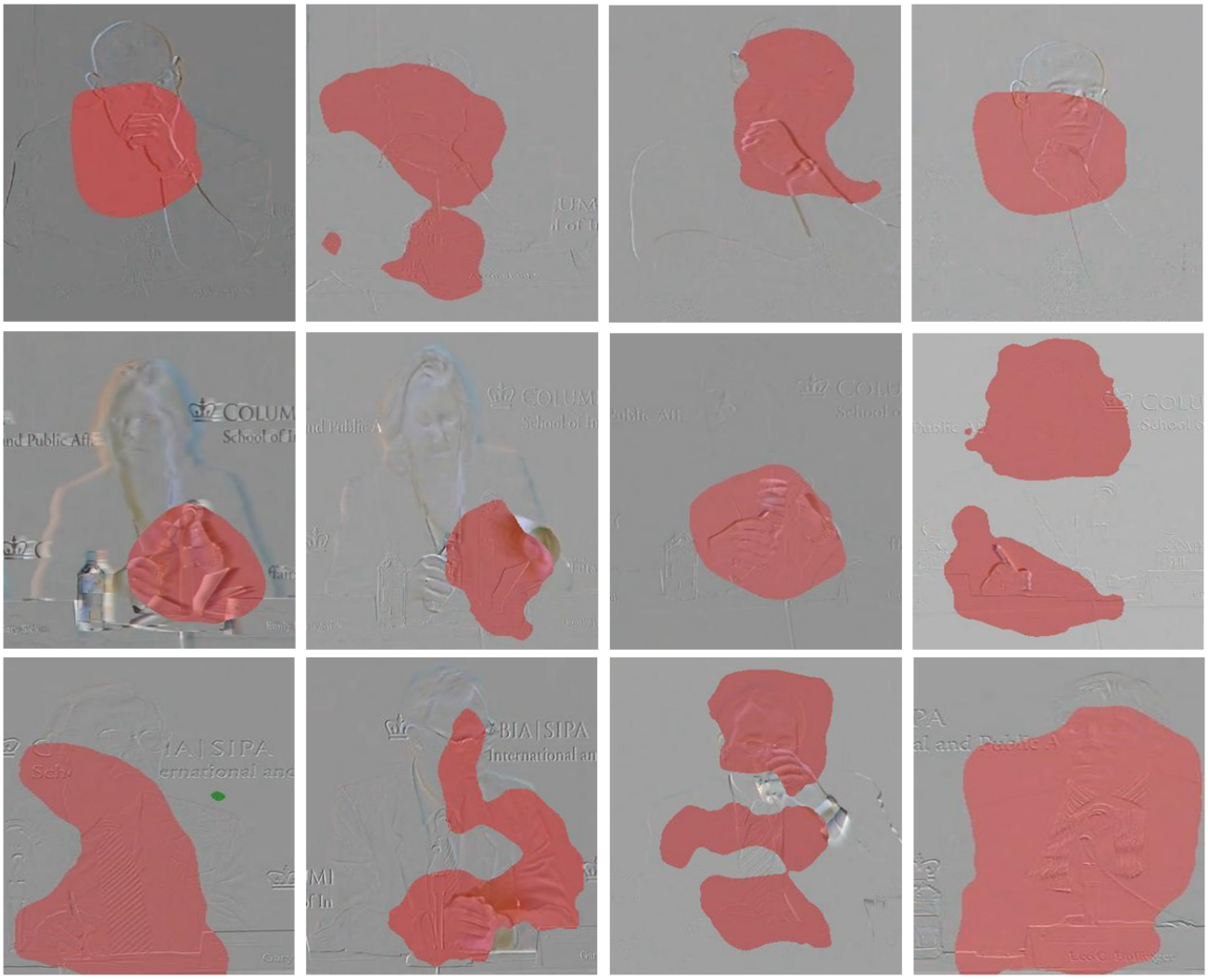
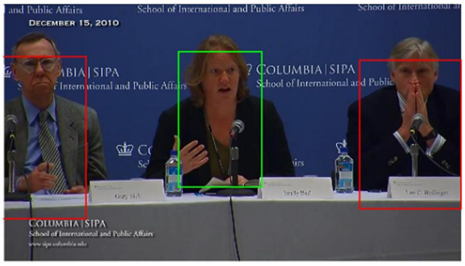
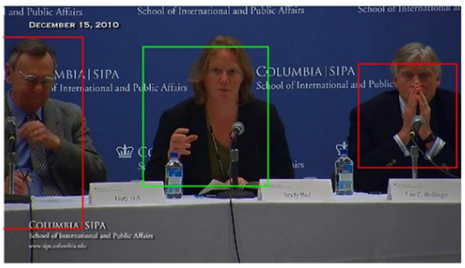
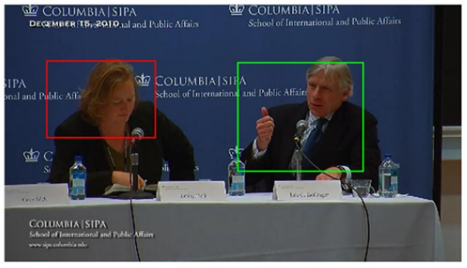


Figure 3. Example localization results (in red) for the persons (implying the body motion belong to a person) correctly detected as not-speaking when S-VVAD was applied to Columbia dataset.







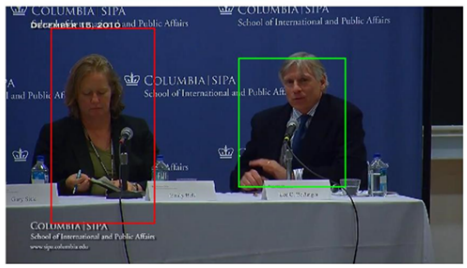
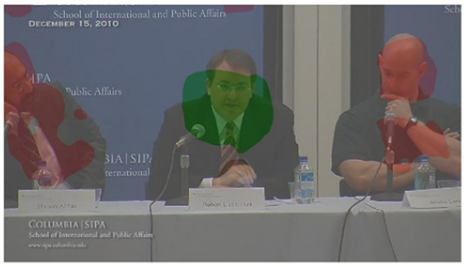



Figure 4. Example localization results (red for not-speaking, green for speaking) imposed on the middle RGB video frame (left). The corresponding predicted bounding boxes (red for not-speaking, green for speaking), which are correctly detected (right). All these results were obtained when S-VVAD was applied to Modified Columbia dataset.

Figure 5. Experiments on RealVAD dataset [1]. Baseline refers to the method in [1].

Figure 6. Example localization results (red for not-speaking, green for speaking) imposed on the middle RGB video frame for the participants having VAD ground-truth for that specific video frame. All these results were obtained when S-VVAD was applied to ReadVAD dataset within cross-dataset setting (see text for details).