Who's Speaking? Audio-Supervised Classification of Active Speakers in Video

Punarjay Chakravarty
ESAT-PSI
KU Leuven
iMinds MMT
pchakrav@esat.kuleuven.be

Sayeh Mirzaei
Department of Electrical Engineering
Amirkabir University of Technology
sayeh.mirzaei@esat.kuleuven.be

Tinne Tuytelaars
ESAT-PSI
KU Leuven
iMinds MMT
tinne.tuytelaars@esat.kuleuven.be

Hugo Van hamme
ESAT-PSI
KU Leuven
hugo.vanhamme@esat.kuleuven.be

ABSTRACT

Active speakers have traditionally been identified in video by detecting their moving lips. This paper demonstrates the same using spatio-temporal features that aim to capture other cues: movement of the head, upper body and hands of active speakers. Speaker directional information, obtained using sound source localization from a microphone array is used to supervise the training of these video features.

Categories and Subject Descriptors

H.5.1 [**Multimedia Information Systems**]: Audio input/output; I.2.10 [**Vision and Scene Understanding**]: Video analysis; I.4.8 [**Scene Analysis**]: Sensor fusion

General Terms

Algorithms, Human Factors

Keywords

Ambient Intelligence & Smart Environments; Human-Robot Interaction; Multimodal Fusion and Integration

1. INTRODUCTION

Existing research for active speaker detection [2, 3, 6] has focussed on tracking the face and correlating lip movement with speech. While this method will work for high quality, high resolution frontal shots of people, it will not always be ideal when the speaker presents a profile view to the camera or her hands occlude her lips while speaking or when she is too far from the camera for her lips to be detected by facial feature detectors.

ICMI '15, November 09-13, 2015, Seattle, WA, USA

DOI: http://dx.doi.org/10.1145/2818346.2820780 ...$15.00.

This research tries to use spatio-temporal action recognition features, also called dense trajectory features [10] to pick up subtle clues accompanying a person's speech, like the movement of the head and hands. We use audio to supervise the training of the video features. Microphone pairs are standard on a lot of modern laptops, and we use a pair of laptop-embedded microphones for localizing the bearing direction of the active speaker, which is associated with an upper body detection in video. Audio is used to label dense trajectory features extracted from within upper body tracks of people as speaking vs non-speaking, and these labelled examples are used to train an active speaker classifier using video alone.

The problem of determining who is speaking in video is useful for a number of applications. It could be a starting point for video diarization, the process of annotating speakers in video. Human-Computer Interaction (HCI) systems would benefit from determining who is speaking, so that the robot or computer can respond to the specific interlocuter when more than one person is present in the system's interaction environment. Video conferencing systems could use active speaker detection to highlight one amongst a group of meeting attendees at a table and only transmit the video of the person who is doing the talking.

There has been some research on combining features detected from audio with features detected from video to both identify a specific speaker, and for speech recognition. An early example of this is by Cutler et al. [2], who use a single microphone and camera system to detect the correlation between audio and video data corresponding to speech. Correlation between Mel-cepstrum coefficients in the audio signal and frame differencing around the lips on successive frames are used to identify specific utterances in the speech.

Audio and video features have also been fused to distinguish between speakers. Li et al. [6] used Canonical Correlation Analysis (CCA) to find the commonality between synced audio and visual features. CCA calculates basis vectors that are used to project audio and video feature vectors onto a common subspace, following which the concatenated feature vector is used to train a classifier to distinguish between speakers.

Video annotation using speaker detection was demonstrated by Everingham et al. [3]. They annotate a TV series, Buffy

the Vampire Slayer, by matching the script to the subtitles. Characters are tracked using similarities in face and clothing and the detection of active speakers is used as an additional cue to link speakers in the script to faces in the video. Frame-differencing in a rectangular window around the lip features in a tracked face is used to determine movement of the lips, and faces with lips that move above a threshold value are determined to be speaking.

The detection of active speakers has been used for automatic editing of classroom video. Hariharan et al. [5] use a microphone array to localize a questioner with a Time Difference of Arrival (TDOA) algorithm, and combine this with video detection of the raised hand of the questioner.

Zhang et al. [11] describe a meeting room video conferencing system that automatically detects the speaker and concentrates the transmitted video on the speaker. A circular mic array and a panoramic camera system are used to localize active speakers around a table. Instead of having separate audio and video classifiers, they feed information from both sources into an Adaboost classifier. Weak classifiers - functions of the audio sound source localization and image-based frame differencing are used in a boosting framework to determine the region of interest of an active speaker in panoramic video.

More recently, Cech et al. [1] have used audio-visual cues to focus the attention of Nao, a child-sized humanoid robot equipped with a microphone array and a stereo camera pair, on specific speakers in its environment. They also use TDOA to get the azimuth and elevation of the active speaker, which is then coupled with face detections in the stereo camera pair.

In contrast, our work uses audio to supervise the training of spatio-temporal features from the face and upper body that accompany a person's speech. Section 2 describes the setup of the system used for capturing the audio-visual data, the audio sound source localization and the video feature training, followed by experimental results and analysis.

2. SYSTEM DESCRIPTION

Experimental Setup.

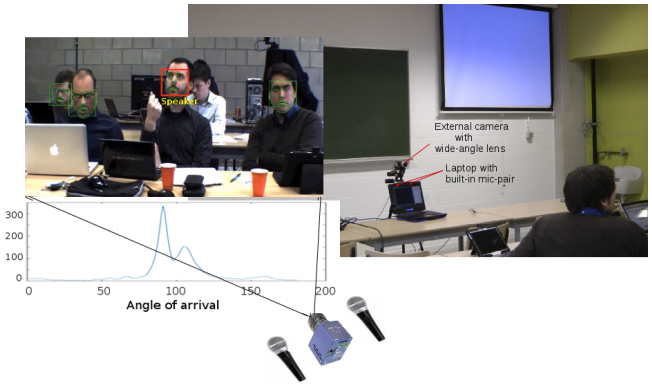

Figure 1: Hardware: wide angle camera & microphone pair in laptop. GCC-PHAT metric plotted against bearing of targets. Largest peak above a threshold is associated with face/upper body detections, and this face/upper body is identified as speaking.

The audio-visual (one camera, one microphone pair) recordings were conducted in a natural setting - a student presentation session at university. 7 recordings were made, for 7 different presentations, to a jury of examiners. Each recording comprised of roughly 25 minutes of presentation, followed by 5 minutes of questions. A laptop with its in-built microphone pair was used for recording the audio, along with an external camera with a wide-angle lens, to capture the video. The setup was pointed towards the jury, with the 3 members in the first row dominating the view. The audio-based speaker bearing estimation software was run in real-time and the video was stored frame-by-frame, along with a bearing angle of the speaker for each frame. The detection and tracking of the upper bodies, the extraction of the dense trajectory features from within the upper body bounding boxes for each track, and the training of the classifier was done off-line. The following paragraphs explain these algorithms in more detail.

Audio Sound Source Localization.

The audio signal at the microphones is a mixture of several different sound sources, due to speakers speaking at the same time, background noise or reverberation effects. The separation of these sound sources is called Blind Source Separation (BSS), which is traditionally done by clustering of the Fourier Transform coefficients in each Time Frequency (TF) cell of the Short Term Fourier Transform (STFT) of the signals. However, this requires two assumptions - the number of sources be known in advance, and signal sparsity, i.e., the source represenations do not overlap in the TF domain. The number of sources is not always known, and the sparsity assumption is violated in the presence of noise or when sources overlap. We use the approach proposed by Mirzaei et al. [7], where an angular spectrum is derived for estimating the number of sources as well as the direction of arrival of the sources. A non-linear function of the Generalized Cross Correlation - Phase transform (GCC-PHAT) between the two signals is calculated in each TF bin, against all angles of arrival of the source signal with respect to the microphone baseline direction. The dominant peak in this function above a threshold gives the direction of the sound source (Figure 1). This method is particularly useful when the microphone baseline is small compared to the distance to the sound sources, as is the case in our setup.

Video Person Tracking.

An upper body detector from Ferrari et al. [4] is run on each frame of the video and a multi-target tracker is used to cluster these detections in space and time. A track is initialized at persistent detections, and is updated with a new detection at each frame. Detections are associated with tracks based on proximity and bounding box size. A simple alpha filter update rule suffices in this case - the track's updated position is based on a weighted linear combination of the previous state and the new observation (detection) associated with it. A track not associated with any detections for a pre-defined number of frames is deleted.

Dense Trajectories for Active Speaker Classification.

Dense Trajectories pooled by a Fisher vector (FV) [9] representation, and subsequently trained with Support Vector Machine (SVM) classifiers, is a standard, state-of-the-art pipeline used for action recognition [10], and we use the same for active speaker detection. Dense trajectory features in-

side the upper body bounding boxes of each person track are categorized as speak and non-speak trajectories, as classified by audio sound source separation. To this end, the sound source separation gives, for each frame, the bearing value of the highest intensity sound at that frame. These bearing values are associated with the upper body detections, and a detection is marked as speaking/non-speaking depending on this association. A simple linear transformation (requiring no special calibration) is used between the bearing range of the audio and the image width - this transforms a bearing value to a column index.

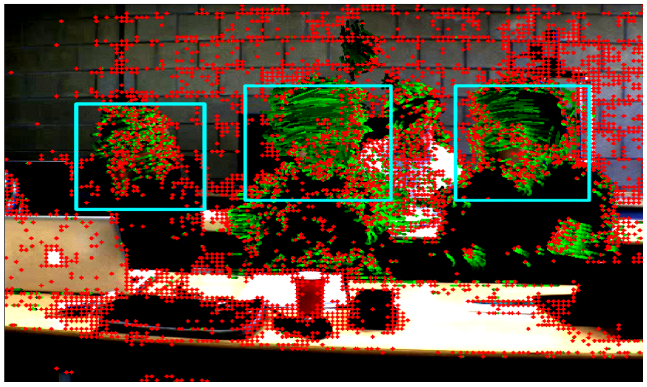

Figure 2: Dense Trajectories within upper bodies

Each dense trajectory track is calculated from 15 successive frames and has associated with it the mean pixel location of the trajectory, and Histogram of Gradients (HoG), Histogram of Flow (HoF) and Motion Boundary Histogram (MBH) features [10]. The feature vectors are each reduced to half their original dimensionality using Principal Components Analysis (PCA). For the FV encoding, we use a Gaussian Mixture Model (GMM) with 256 components. Trajectories that start at the same frame and are inside the bounding box of the upper body of a track are pooled using the FV representation. FVs for the HoG, HoF and MBH features are separately calculated and concatenated to give a 101,376 dimensional FV representing all the features along all trajectories within a sequence of 15 contiguous upper body detections, and the label of the starting frame of these trajectories - speaking/non-speaking is used as the label for the FV. FVs from speaking frames are used as positive training samples and those from non-speaking frames are used as negative training samples, and a linear SVM classifier is trained using these samples. Intra-class L2-normalization of the FV and a final power and L2 normalization of the whole vector are performed before the classification step. These are techniques that have been shown to significantly boost the performance of action recognition classifiers and are considered best practice [8]. Figure 3 illustrates the steps in the training pipeline.

Results.

7 audio-visual recordings are used: each recording has roughly 25 minutes of presentation by the student (during which the jury mostly remains silent, but occasionally whispers amongst themselves), and 5 minutes of questioning by the jury. The active speaker classifier is trained using Leave-One-Out-Cross-Validation (LOOCV) - trained on 6 presentations and tested on the 7th, and this is repeated 7 times,

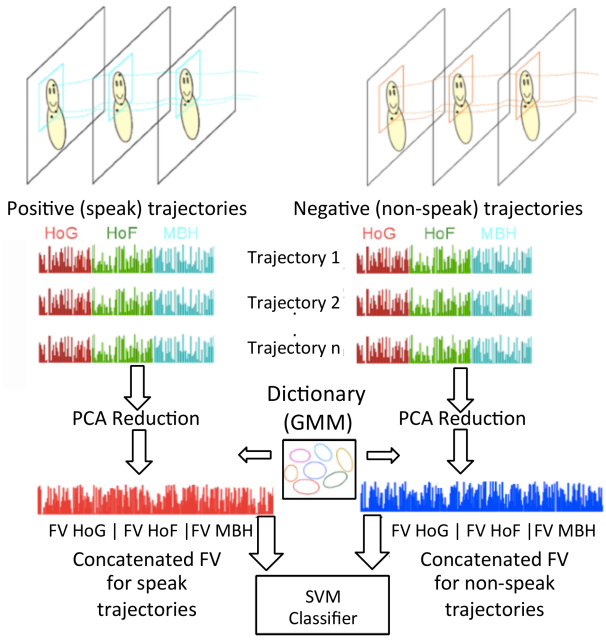

Figure 3: Active speaker classifier training pipeline

based on the recording that is used for testing.

The speaking/non-speaking labelling generated by the audio sound source localization is noisy - see column 1 in the table in Figure 4. This is corrected manually to obtain ground truth labels for speak/non-speak frames in all the videos. Only the 3 speakers sitting in the first row are considered in the experiments, as speakers behind them are partly or fully occluded. The video-based spatio-temporal dense trajectories classifier is trained using both the cleaned up (ground truth) and noisy frame labels from audio - let's call these results video-clean, and video-noisy respectively. Finally, a baseline method, based on lip movement detection (video-lips) within tracked faces [3] is used for comparison. The Receiver Operating Characteristic (ROC) curves (Figure 4), plots of the true positive rate (TPR) vs the false positive rate (FPR), show the results of the video-clean and video-noisy classifiers. It can be seen that even using the noisy training data, the video-based classifier can distinguish speaking vs non-speaking better than the baseline lip-movement based method. Furthermore, the mean Equal Error Rate (EER) for video-noisy is 0.66 and for video-clean is 0.69 - this shows that the supervision from audio is sufficient to train the video classifier, with slight improvement using training with ground truth. Both the audio and lip-movement based active speaker classifiers present a binary classification result, as opposed to a real-valued score for the video-based classifiers video-clean and video-noisy. Consequently, only a single (TPR, FPR) pair is available for audio and video-lips, shown as dots and stars respectively in the ROC curves of Figure 4. One anomalous result is the low EER value for video-clean, fold 5, and at the time of writing, it is not clear why this is so. When a speaker is silent he often sits with his fingers obscuring his lips, resulting in false-positives and poor performance for the video-lips baseline method.

fold	audio	video-clean	video-noisy	video-lips
1	(64,7)	(75,24)	(72,28)	(27,21)
2	(77,2)	(75,24)	(72,28)	(30,19)
3	(50,3)	(72,27)	(65,35)	(40,25)
4	(39,4)	(71,29)	(64,36)	(37,24)
5	(60,2)	(54,46)	(70,29)	(19,25)
6	(17,3)	(67,32)	(61,39)	(34,21)
7	(26,2)	(71,29)	(61,39)	(30,19)

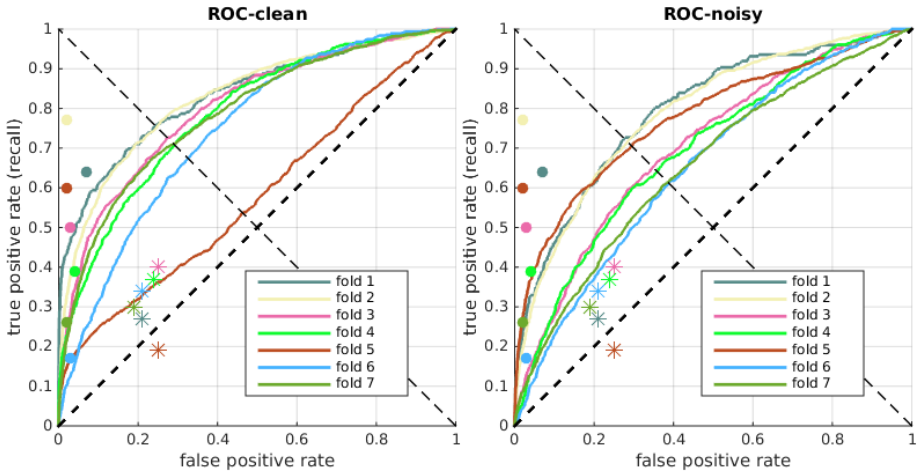

Figure 4: Left: True and False Positive rates (TPR, FPR) in % for audio, spatio-temporal video classifiers (video-clean and video-noisy) and lip-movement based methods (video-lips) across the 7 cross-validation folds. Middle: ROC curves for video-clean. Right: ROC curves for video-noisy. Circles and stars show (TPR, FPR) pairs for audio and video-lips.

3. CONCLUSIONS

This work presents to our knowledge, the first attempt at detecting active speakers in video using action recognition features. Whereas earlier work has used the movement of the lips for detecting speakers in video, our work trains a classifier on spatio-temporal action recognition features to pick up face and upper body movements and gesticulations accompanying a person's speech.

Our other contribution is the use of audio sound source localization (from a microphone array) to supervise the labelling of speaking vs non-speaking parts of the video. We use a microphone pair in a laptop to capture the sound data, and the bearing of the speaker calculated from the time difference of arrival of the sound is associated with upper bodies detected and tracked in video. Upper body tracks and dense trajectory features within them are labelled as speaking and non-speaking and the samples, quantized and pooled with a Fisher vector representation are fed into a SVM classifier. We obtain a mean Equal Error Rate (EER) of 0.66 with supervision from audio over 7 experiments (3.5 hours of audio-visual data), tested using Leave One Out Cross Validation (LOOCV). Training on ground truth labels give a slight improvement - an EER of 0.69.

Future work will focus on training the classifier in a single audio channel setting, which is the case for the vast majority of videos available. Voice Activity Detection will tell us whether or not there is speech in a frame, and the particular person that is speaking will be learnt as a latent variable in a Latent SVM framework. Methods for fusing audio and video for more reliable detection will also be investigated.

4. ACKNOWLEDGMENTS

This work has been supported by the KU Leuven GOA project CAMETRON.

5. REFERENCES

[1] J. Cech, R. Mittal, A. Deleforge, J. Sanchez-Riera, X. Alameda-Pineda, and R. Horaud. Active-speaker detection and localization with microphones and cameras embedded into a robotic head. In *IEEE-RAS International Conference on Humanoid Robots*, 2013.

[2] R. Cutler and L. Davis. Look who's talking: Speaker detection using video and audio correlation. In *IEEE International Conference on Multimedia and Expo (ICME)*, volume 3, pages 1589–1592, 2000.

[3] M. Everingham, J. Sivic, and A. Zisserman. Taking the bite out of automatic naming of characters in TV video. *Image and Vision Computing*, 27(5), 2009.

[4] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, June 2008.

[5] B. Hariharan, H. S, and U. Gopalakrishnan. Multi speaker detection and tracking using audio and video sensors with gesture analysis. In *Tenth International Conference on Wireless and Optical Networks (WOCN)*, pages 1–5, 2013.

[6] D. Li, C. M. Taskiran, N. Dimitrova, W. Wang, M. Li, and I. K. Sethi. Cross-modal analysis of audio-visual programs for speaker detection. In *IEEE 7th Workshop on Multimedia Signal Processing, MMSP*, pages 1–4, 2005.

[7] S. Mirzaei, H. Van hamme, and Y. Norouzi. Blind audio source separation of stereo mixtures using bayesian non-negative matrix factorization. In *Signal Processing Conference (EUSIPCO)*, pages 621–625, Sept 2014.

[8] X. Peng, L. Wang, X. Wang, and Y. Qiao. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. *CoRR*, abs/1405.4506, 2014.

[9] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, pages 143–156, Berlin, Heidelberg, 2010.

[10] H. Wang and C. Schmid. Action Recognition with Improved Trajectories. In *ICCV*, pages 3551–3558, Sydney, Australia, Dec. 2013.

[11] C. Zhang, P. Yin, Y. Rui, R. Cutler, P. A. Viola, X. Sun, N. Pinto, and Z. Zhang. Boosting-based multimodal speaker detection for distributed meeting videos. *IEEE Transactions on Multimedia*, 10(8):1541–1552, 2008.