Audio-Visual Face Detection for Tracking in a Meeting Room Environment

Mark Barnard, Wenwu Wang and Josef Kittler
Centre for Visions, Speech and Signal Processing (CVSSP)
University of Surrey
GU2 7XH, UK.
Email: {mark.barnard, w.wang, j.kittler}@surrey.ac.uk

Syed Mohsen Naqvi and Jonathon Chambers
Advanced Signal Processing Group
Loughborough University
Loughborough, Leicester, UK.
{s.m.r.naqvi, j.a.chambers}@lboro.ac.uk

***Abstract*—A key task in many applications such as tracking or face recognition is the detection and localisation of a subject's face in an image. This can still prove to be a challenging task particularly in low resolution or noisy images. Here we propose a robust method for face detection using both audio and visual information. We construct a dictionary learning based face detector using a set of distinctive and robust image features. We then train a support vector machine classifier using sparse image representations produced by this dictionary to classify face versus background. This is combined with the azimuth angle of the speaker produced by an audio localisation system to constrain the search space for the subject's face. This increases the efficiency of the detection and localisation process by limiting the search area. However, more importantly, the audio information allows us to know *a priori* the number of subjects in the image. This greatly reduces the possibility of false positive face detections. We demonstrate the advantage of this proposed approach over traditional face detection methods on the challenging AV16.3 dataset.**

I. Introduction

A crucial step in any automatic tracking system is the detection and localisation of the object to be tracked in the initial frame. In speaker tracking this can be seen as a face detection problem. Face detection has been extensively researched over the last 20 years and a number of comprehensive reviews have been published [1], [2]. Here we focus on the task of face detection as the initialisation step in a system for tracking people in an indoor meeting room environment.

The initialisation of visual tracking systems through automatic face detection is still very much an open problem. In the majority of cases the object to be tracked is simply manually selected in the initial frame of the sequence [3], [4], [5], [6], [7], [8]. In some cases a prior template or model is used to search in the initial frame for the object, for example a prior colour template of a face is used by [9], [10], [11]. Alternatively, a common face detection algorithm such as that proposed by Viola and Jones [12] can be used, as in the case of Naqvi et al. [13]. However, these methods require an exhaustive search of the initial frame and also if the number of objects to be tracked is not known *a priori* they can lead to false positive object detections.

Currently, one of the most effective methods of object recognition in still images is dictionary learning or bag of visual words. Dictionary learning [14] has shown state-of-the-art performance in many object recognition comparisons such as the PASCAL Visual Object Class challenge [15] and the ImageCLEF Visual Concept Detection challenge 2010 [16]. Recently, dictionary learning methods have also been applied to the problem of face recognition, with success under varying conditions [17], [18], [19]. In this work we propose the use of the dictionary based face model. First we create a dictionary using K-means clustering, then use Soft Assignment (SA) methods [20] to generate histograms or coefficient vectors. While this method of SA for histogram generation improves results it can be computationally expensive, so we further use a method of limiting the number of dictionary codewords used for histogram generation, so-called Locality constrained Soft Assignment (LcSA) [21]. These sparse histograms are then used to train a linear Support Vector Machine (SVM) classifier to discriminate face/head from background. Here we demonstrate the use of this dictionary based face model in the task of face detection.

A small number of publications have proposed the fusion of audio and video information for face detection. Zhang et al. [22] present an audio-visual speaker detection method using feature level audio-visual fusion and boosting to select the most distinctive features. In the feature level fusion potentially any small error in the audio tracker will seriously degrade the performance of the tracker. We propose a novel initialisation method by using the audio Direction of Arrival (DOA) angle for each speaker to constrain the search area for the visual tracker. This high level fusion of audio and visual information helps reduce the effect of noise.

We show that even the DOA from a noisy audio tracker combined with our general dictionary learning based classifier can be used for effective initial face localisation in a tracking system. Our primary method of face detection is a visual detector based on dictionary learning and a discriminative SVM classifier. To improve the robustness of the visual detection audio DOA is used to constrain the search area for the visual face detection. In addition to constraining the search area audio information also provides us *a priori* with the number of speakers in the room, thus greatly reducing the number of false positive face detections.

In the following section we discuss the extraction of a set of distinctive and robust features for face detection. In Section III and Section IV we explain our method of dictionary construction and classifier training for visual face detection. Section V presents the audio tracking method used for estimating the speaker's DOA. The fusion of audio and visual information to constrain the search area for the face is described in Section VI. In Section VII we present a set of face detection experiments on the challenging AV16.3 database. Finally, we present our conclusions in Section VIII.

II. Feature Extraction

We define a feature vector as, $\mathbf{f} = \{f_1, f_2, \dots, f_M\}^T \in \mathbb{R}^M$, where M is the feature dimension and T is a transpose. In order to train and test our visual face detection model we need to extract a set of distinctive and robust features. We extract two common types of visual features, SIFT and colour histogram features, respectively. SIFT descriptors, introduced by Lowe [23], are histograms of gradient orientation, which have been shown to be highly distinctive and also robust to affine image transformation in the task of object recognition [24]. We extract a set of standard grey-scale SIFT descriptors each of dimension M_s, densely sampled with a horizontal step size of $I_w/3$ and a vertical step size of $I_h/3$ from each image patch as shown in Figure 1, where I_w and I_h are the image patch width and height respectively. Specifically, we extract a series of 12×12 pixel squares from the image patch, each of which, centred at a sampling point, is then divided into 16 regions with 8 gradient orientations quantisation in each region. This gives us a 128 dimensional SIFT feature vector, $\mathbf{f} = \{f_1, f_2, \dots, f_{Ms}\}^T \in \mathbb{R}^{Ms}$ where $Ms = 128$. We extract nine SIFT feature vectors for each image patch as shown by the white crosses, in the image patch in Figure 1.

We also extract a colour histogram of each image patch. Colour histograms have many advantages in tracking applications being rotation and partially scale invariant, robust to partial occlusions and easy to calculate. We transform the image from the RGB colour space to HSV space and form typically a $Mc = 100$ bin histogram of Hue values, and this provides a degree of robustness to changes in illumination. This produces a 100 dimensional vector of colour histogram features $\mathbf{f} = \{f_1, f_2, \dots, f_{Mc}\}^T \in \mathbb{R}^{Mc}$ where $Mc = 100$. These features are then concatenated to give a combined feature vector of both SIFT and Hue histogram feature vectors, $\mathbf{f} = \{f_1, f_2, \dots, f_{Msc}\}^T$, where $Msc = Ms + Mc = 228$.

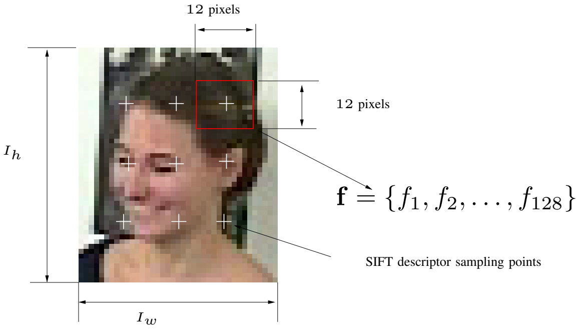

12 pixels
12 pixels
I_h
$\mathbf{f} = \{f_1, f_2, \dots, f_{128}\}$
SIFT descriptor sampling points
I_w

Fig. 1. Extraction of densely sampled SIFT features from an image patch.

III. Dictionary Construction and Classifier Training

In this section we discuss the construction of a dictionary using the combined feature vector described in the previous section. This is followed by a description of how this dictionary can be used to generate compact and sparse image representations and using these representations to train the SVM classifier. An outline of this training procedure can be seen in Figure 2.

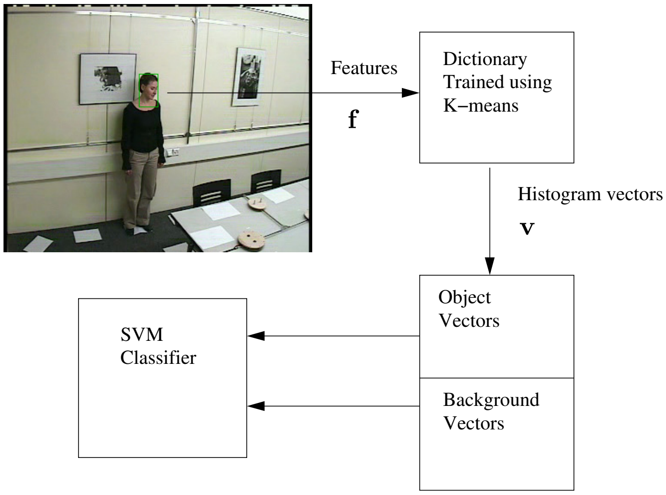


Fig. 2. Training an SVM classifier using labelled histogram representations generated from a dictionary.

The method described in Section II provides us with a number of feature vectors. These form the column vectors of a matrix $\mathbf{F}$ defined as $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_{\bar{L}}] \in \mathbb{R}^{M \times \bar{L}}$ where each vector $\mathbf{f}_l, l = 1, \dots, \bar{L}$ is an M dimensional feature vector from the training set and $\bar{L}$ is the number of vectors in the training set.

A dictionary $\mathbf{D}$ is defined as a matrix, $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_U] \in \mathbb{R}^{M \times U}$ where U is the total number of visual words in the dictionary. This set of vectors is capable of providing a succinct representation of the feature vectors in $\mathbf{F}$ and are often described as *visual codewords* or *atoms*. In order to construct a dictionary from the training data in $\mathbf{F}$ we use the K-means clustering algorithm. The visual words $\mathbf{d}_u, u = 1, \dots, U$ in the dictionary $\mathbf{D}$ are the cluster centres produced by the K-means algorithm. We need to balance the number of visual words in the dictionary between being able to discriminate the object we wish to recognise and becoming over-specified on the training set, i.e. so-called over-fitting.

Our goal in constructing the dictionary $\mathbf{D}$ is to create a compact representation of an image, or image patch. In order to achieve this we define a vector $\mathbf{v}$ as, $\mathbf{v} = \{v_1, \dots, v_U\} \in \mathbb{R}^U$. This vector is a histogram or coefficient vector representing an image patch in the training set based on the visual codewords in the dictionary, which is populated using a soft voting technique, as discussed next.

A. Histogram Generation Based on Soft Assignment (SA)

One of the simplest forms of histogram generation for image representation employs a vector quantisation method known as Hard Assignment (HA). For each visual codeword $\mathbf{d}_u$ in the dictionary $\mathbf{D}$ the u^{th} bin of the histogram $\mathbf{v}$ is simply assigned according to

$$v_u = \frac{1}{L} \sum_{l=1}^{L} \begin{cases} 1 & \text{if } \mathbf{d}_u = \arg\min_{\mathbf{d} \in \mathbf{D}} (\mathbb{E}(\mathbf{d}, \mathbf{f}_l)) \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $\mathbb{E}(\mathbf{d}, \mathbf{f}_l)$ is the Euclidean distance from the visual codeword $\mathbf{d}$ to the feature vector $\mathbf{f}_l$ and L is the number of feature vectors extracted from an individual image patch.

Although this simple histogram generation method can generate sparse and compact image representations, recent results in object recognition show that histogram generation methods based on SA provide much better performance over the HA [20], [25]. SA uses the codeword uncertainty method presented in [20] where the entry in the histogram $\mathbf{v}$ for the u^{th} visual codeword $\mathbf{d}$ is given by

$$\mathbf{v}_u = \frac{1}{L}\sum_{l=1}^{L} \frac{G_\sigma(K(\mathbf{d}_u, \mathbf{f}_l))}{\sum_{j=1}^{U} G_\sigma(K(\mathbf{d}_j, \mathbf{f}_l))}, \tag{2}$$

where L is the number of descriptors in the image, $K(\mathbf{d}_u, \mathbf{f}_l)$ is the Euclidean distance between the u^{th} dictionary atom $(\mathbf{d}_u)$ and the l^{th} feature vector $\mathbf{f}_l$. G is a Gaussian kernel centred on the visual codeword $\mathbf{d}_u$ with smoothing factor σ. The variance of the Gaussian kernel G is given by σ^2 and so the value of σ can be seen as a control on the sparsity of the histogram $\mathbf{v}_u$.

B. Approximate Locality-constrained Soft Assignment (LcSA)

One drawback of the SA method described in the previous section is the high degree of complexity involved in histogram generation. The histogram generation process in SA based methods involves the whole set of the dictionary codewords, as can be seen in equation (2). This can render the recognition process very expensive particularly for dictionaries with large numbers of visual codewords. To address this limitation of SA, we adopt the notion of locality in coding by constraining codeword activations to the most relevant few.

We define the locality around the feature vector $\mathbf{f}_l$, as the region of the dictionary space containing the c nearest codewords to the feature vector $\mathbf{f}_l$, determined by the Euclidean distance. Specifically, we constrain SA to activate only the c nearest codewords of feature vectors as in [26], [27] when computing the histogram bin assignments. We refer to this variant of SA as approximate Locality-constrained SA (LcSA) [21]. Hence, LcSA obtains an *approximate* locality-constrained solution rather than a fully analytical one [21], and also achieves local smoothness and sparsity. Limiting the histogram generation in equation (2) to be based on only this subset of, c, codewords $\mathbf{D}_l^c$ yields:

$$\mathbf{v}_u = \begin{cases} \frac{1}{L}\sum_{l=1}^{L} \frac{G_\sigma(K(\mathbf{d}_u, \mathbf{f}_l))}{\sum_{j=1}^{U} G_\sigma(K(\mathbf{d}_j, \mathbf{f}_l))}, & \text{if } \mathbf{d}_u \in \mathbf{D}_l^c \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

IV. Classifier Training

We have a number of histogram vectors with each being a sparse representation of an image patch in the training set, $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N]$, where N is the total number of image patches in the training set. These histograms which are produced by the processes described in Section III-B are then used as labelled training data to train an SVM classifier. Due to the sparsity of the histograms produced by these methods the two classes, face/head and background, are more likely to be linearly separable in a high dimensional space. This is confirmed in our experimental results where a binary linear SVM is used for classification.

V. Audio Tracker

We use an audio tracker for audio source detection and localisation, and this tracker was developed and tested on multi-party speech in a smart meeting room environment [28]. Audio tracking is performed in a two-step process known as the SAM-SPARE-MEAN method. The first step consists of a sector based combined detection and localization. In this step the space around a circular microphone array is divided into a number of sectors. The frequency spectrum is also discretised into a number of frequency bins and for each sector and frequency bin it is determined whether there is at least one active audio source present.

At each time frame for each sector and frequency bin a sector activity measure, SAM, is estimated, this is the posterior probability that at least one audio source is active within that sector and frequency bin. This measure of activeness is then compared to a threshold to determine whether there is an active source in that sector. In the second step a point based search is conducted in each of the sectors labelled as having at least one active source. The localization uses a parametric approach [29], the location parameters are optimized with respect to a cost function such as SRP-PHAT [30]. While this method does perform reasonably well the output of the DOA of the audio source can be very noisy at times.

VI. Audio-visual face detection

Here we propose exploiting the audio DOA information and our general face model in order to detect faces in video frames. Specifically, we use the direction of the speaker given by the audio tracker described in Section V in order to automatically detect the location of the faces of the speakers in the room. While the DOA angle produced by the tracker can be noisy we can use it for two very important pieces of prior information: the number of speakers and the general direction of each speaker. This prior information can be used to constrain the number of faces to be recognised and the area of the image to search for these faces. We project a line in three dimensions from the centre of the microphone array to a point (a, b, z), where a is equal to the distance from the centre of the microphone array to the wall of the room in metres, denoted as R (which is 1.75 meters in our experiment, as shown on Figure 3), z can be estimated as the height of a human speaker, typically chosen as 1.80 metres in our experiment, and b is calculated according to

$$b \quad = \quad \tan(\phi \times \frac{\pi}{180}) \cdot R \tag{4}$$

where ϕ is the azimuth angle (in degrees) of the speaker with respect to the circular microphone array shown in Figure 3.

The image is sampled at intervals along this line, as shown in Figure 4 in order to detect a face. To account for noise in the DOA angle we sample a distribution of image patches at a number of sampling points along the line, as shown in Figure 4. We define the initial image patch as $I_0 = \{x_0^1, y_0^1, x_0^2, y_0^2\}$ where x_0^1, y_0^1 and x_0^2, y_0^2 are the coordinates of the top left and bottom right corners of the image patch centred on the sampling point. We then generate a number of image patches in order to sample the area around the sampling point, we define this collection of T points as $\mathbf{I} = \{I_0, I_1 \ldots I_T\}$. The t^{th} image patch is propagated according to the dynamic model

$$I_t = I_0 + s, \quad (5)$$

where s, is a random variable with 2D Gaussian distribution with zero mean. Hence the image patches are propagated based on the value of I_0 and a certain amount of additive white Gaussian noise added in order to model the uncertainty of the DOA estimation.

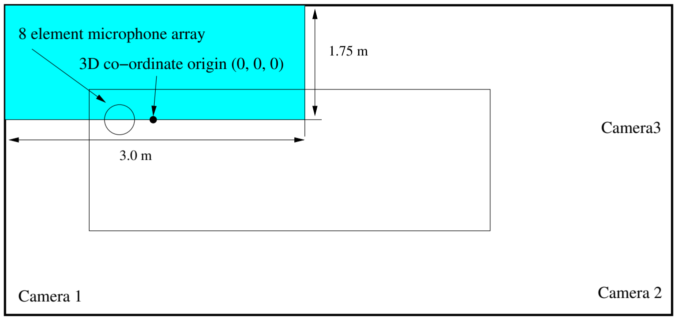


Fig. 3. Layout of room used for audio-visual recordings. The shaded area indicates the performance area for the subjects.

So for each sampling point on the line estimated using the audio DOA we distribute a number of image patches each of which is tested with our discriminative SVM classifier. We then select the image patch with the highest likelihood from the SVM classifier as containing the subject's face. The sampling points and the initial face positions for a single person sequence can be seen in Figure 4.

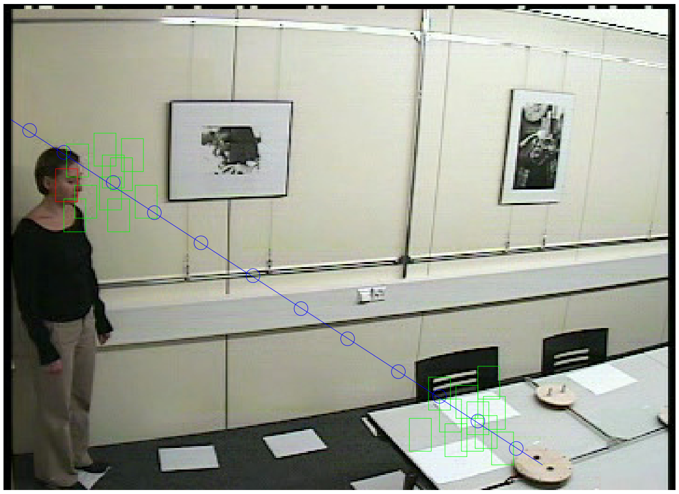


Fig. 4. Face detection process using audio DOA to constrain the search space for the subject's face.

VII. Experiments

A. Experimental Set-up

The data used in our experiments consists of sequences from the AV16.3 dataset [31]. This data was recorded at the IDIAP research institute in 2004, in a smart meeting room environment using three calibrated cameras and a single eight element omnidirectional circular microphone array. The AV16.3 dataset contains 10 annotated sequences; we selected four of these annotated sequences to form the training set (sequences 2, 3, 5 and 6). The variability of appearance in the training data was maximised by combining data from all three cameras to train a single model. Examples of selected face and background image patches can be seen in Figure 5.

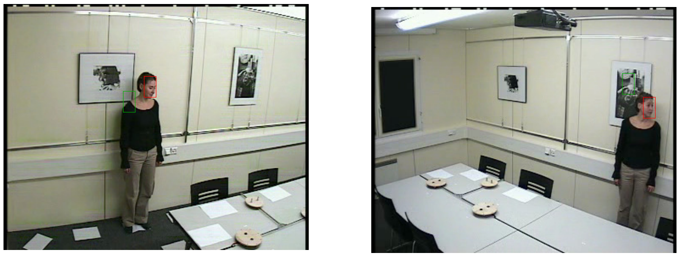


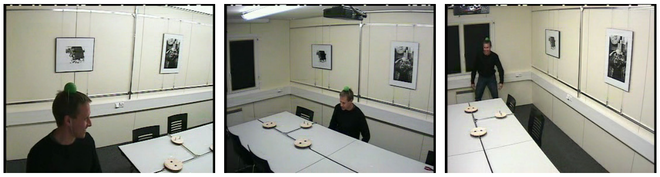
Fig. 5. The selection of positive, red rectangle, and negative, green rectangle, training examples.

The layout of the smart meeting room with the locations of the three cameras and audio microphone array can be seen in Figure 3. The sequences feature subjects moving within the field of view of the three cameras and speaking continuously. The shaded area in Figure 3 indicates the area within which the speakers move. Each video frame is a colour image of 288x360 pixels. Within the data the scale of the face/head may vary from approximately 50×70 pixels to 8×12 pixels, this can be seen in Figure 6. The illumination changes within the meeting room can also be seen in Figure 6.

We set the number of sampling points to be 12 and these are spaced evenly along the line generated from the audio DOA as shown in Figure 4. We also set the number of image patches to be sampled at each sampling point to be 20, this gives a reasonable sampling around the line (note Figure 4 shows less image patches for clarity).

To test our proposed face detection method we take the first frame from a number of sequences where the subject's faces are visible and they are talking. In order to provide a reasonable amount of data we annotated the initial face position on a total of 20 sequences from the AV16.3 dataset, 9 single person and 11 multiple person sequences. Each frame was annotated with a rectangle enclosing the subject's face, this gives us a total of 84 faces in the test set. We compare our method with one of the most common face detection algorithms proposed by Viola and Jones [12] implemented using the OpenCV computer vision library [32]. We used face images from the four training sequences in order to set the thresholds for the Viola-Jones method.

The audio was sampled at 16 kHz using an 8 element



(a) Camera 1 (b) Camera 2 (c) Camera 3

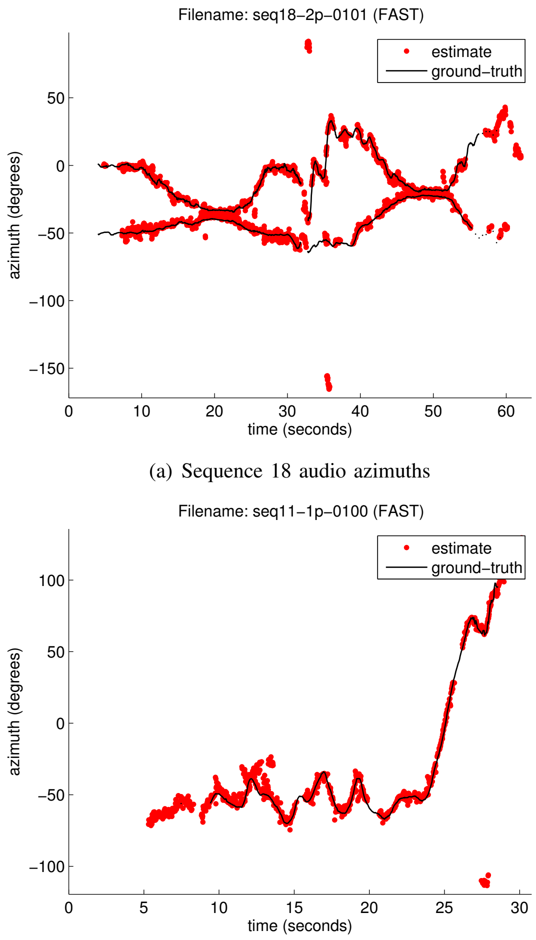
Fig. 6. Three images from sequence 11 from cameras 1, 2, and 3 respectively.

circular microphone array with diameter 10 cm. The following parameters were fixed for all of our audio tracking experiments, time frame windows were 32 ms with an overlap of 16 ms. For the Fast Fourier Transform the number of samples was 512 and the number of histogram bins was 512. For the SRP-PHAT algorithm the number of sectors was fixed at 18 with each sector covering 20 degrees and the speed of sound was fixed at 320 m/s. Further details of the implementation can be found in [28]. The output of the audio tracker for two annotated sequences (sequences 18 and 11) can be seen in Figure 7 this shows that while the azimuth estimation is noisy it can still prove useful in providing an estimate of the DOA for each subject. More importantly it can be seen that it also provides *a piori* an estimate of the number of speakers. Figure 7(a) shows a sequence with two speakers and Figure 7(b) shows a sequence with a single speaker.

To measure the performance of both methods we use precision and recall, where precision is given by $Meas_p = \frac{Num_c}{Num_r}$, where Num_c is the number of correct matches and Num_r is the number of potential faces identified by each method and recall is given by $Meas_r = \frac{Num_c}{Num_a}$ where Num_a is the number of faces in the frame. There are many ways to define what constitutes a correct face detection and this is often linked to the application and dataset being used. Rowley et al. [33] define a correct detection as the centre of the detected face rectangle being less than four pixels from the centre of the annotated rectangle and within 1.2 of the scale of the annotated rectangle. We follow a similar scheme in our experiments, however, we adapt the measure to our particular data and application. We set the criteria for a face detection to be less than a Euclidean distance of 10 pixels from the centre of the annotated face rectangle and a scale within 1.5 of the scale of the annotated rectangle.

The results of this can be seen in Table I, this shows the results in terms of precision and recall for the 84 faces in the initial frames taken from 20 test sequences in the AV16.3 dataset. Figure 8 shows the initial frames for each camera for the multiple subject sequences 18 and 24. The line in the images shows the track used by our proposed audio-visual face detection system and the rectangles show the estimated face location. Figure 8 also demonstrates the robustness of our method to noise in the audio information. In the lower set of images the blue line for the subject is not passing directly through the face, this is due to the subject being taller than average and standing close to the microphone array. Despite this the distribution of the images patches manages to detect the face.

It can be seen from the results that our proposed method performs significantly better than the baseline method. It seems most of the errors in the Viola-Jones approach come from an incorrect estimation of the number of speakers in the initial frame. The low precision of the Viola-Jones method caused the return of a number of false positives due to the overestimate of the number of faces in the frame. Audio information allows us to estimate *a priori* the number of speakers present, this information has the advantage of greatly reducing the rate of false positive detections.



(a) Sequence 18 audio azimuths

(b) Sequence 11 audio azimuths

Fig. 7. Audio azimuth estimated by the audio tracking algorithm for sequences 18 and 11.

Method	Precision ($Meas_p$)	Recall ($Meas_r$)
Viola-Jones	0.6	0.83
Propose AV method	0.97	0.97

TABLE I. COMPARISON BETWEEN OUR PROPOSED AUDIO-VISUAL FACE DETECTION WITH THE VIOLA-JONES METHOD.

VIII. CONCLUSION

We have presented a novel method of combining audio and visual information for robust face detection in a meeting room environment. We have demonstrated that using the audio DOA to constrain the search space can increase the accuracy of our visual face detector. Additionally, the audio tracker provides us with the number of speakers *a priori* thus greatly reducing the chances of a false positive face detection. We compared our proposed method to one of the standard methods for face detection [12], the results showed our method outperformed this baseline method on the challenging AV16.3 dataset.

ACKNOWLEDGMENT

This research was supported by the Engineering and Physical Sciences Research Council of the UK (grant no. EP/H050000/1).

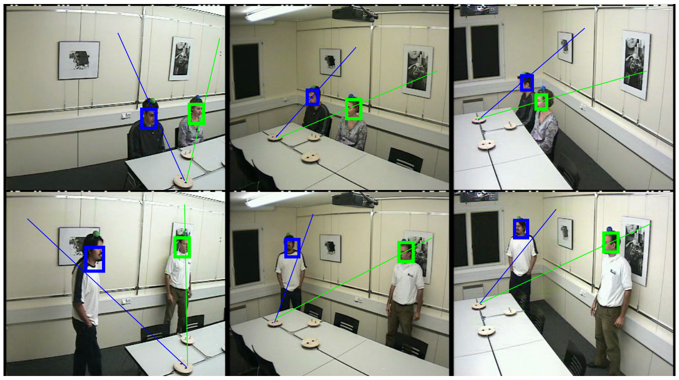


Fig. 8. Images from the testing set featuring multiple people. Blue and green lines show the sampling line for each subject and the rectangles are the position of the image patch with the maximum likelihood of a face for each subject.

References

[1] M. Yang, D. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 34–58, 2002.

[2] C. Zhang and Z. Zhang, "A survey of recent advances in face detection," *Microsoft Research, June*, 2010.

[3] C. Wang and Z. Li, "A new face tracking algorithm based on local binary pattern and skin color information," in *Proceedings of International Symposium on Computer Science and Computational Technology*, 2008, pp. 657–660.

[4] J. Ning, L. Zhang, D. Zhang, and C. Wu, "Robust object tracking using joint color-texture histogram," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 23, no. 7, pp. 1245–1263, 2009.

[5] J. Ye, Z. Liu, and J. Zhang, "A face tracking algorithm based on lbp histograms and particle filtering," in *Proceedings of International Conference on Natural Computation*, 2010, pp. 3550–3553.

[6] C. Shan, Y. Wei, T. Tan, and F. Ojardias, "Real time hand tracking by combining particle filtering and mean shift," in *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, 2004, pp. 669–674.

[7] H. Zhou, Y. Yuan, and C. Shi, "Object tracking using sift features and mean shift," *Computer Vision and Image Understanding*, vol. 113, no. 3, pp. 345–352, 2009.

[8] Y. Tong, Y. Wang, Z. Zhu, and Q. Ji, "Robust facial feature tracking under varying face pose and facial expression," *Pattern Recognition*, vol. 40, no. 11, pp. 3195–3208, November 2007.

[9] D. N. Zotkin, R. Duraiswami, and L. S. Davis, "Joint audio-visual tracking using particle filters," *EURASIP Journal of Applied Signal Processing*, vol. 2002, no. 1, pp. 1154–1164, January 2002.

[10] K. Nummiaro, E. Koller-Meier, and L. J. V. Gool, "An adaptive color-based particle filter," *Image and Vision Computing*, vol. 21, no. 1, pp. 99–110, 2003.

[11] Q. Cai and J. K. Aggarwal, "Tracking human motion using multiple cameras," in *Proceedings of International Conference on Pattern Recognition*, vol. 3, 1996, pp. 68–72.

[12] P. A. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[13] S. M. Naqvi, M. Yu, and J. A. Chambers, "A multimodal approach to blind source separation of moving sources," *IEEE Journal of Selected Topics in Signal Processing*, pp. 895–910, 2010.

[14] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proceedings of IEEE International Conference on Computer Vision*, vol. 2, 2003, p. 1470.

[15] M. A. Tahir, J. Kittler, K. Mikolajczyk, F. Yan, K. Sande, and T. Gevers, "Visual category recognition using spectral regression and kernel discriminant analysis," in *Proceedings of Subspace Workshop in conjunction with International Conference on Computer Vision*, 2009.

[16] M. A. Tahir, F. Yan, M. Barnard, M. Awais, K. Mikolajczyk, and J. Kittler, "The university of surrey visual concept detection system at imageclef 2010: Working notes," in *Proceedings of International Conference on Pattern Recognition*, 2010.

[17] V. M. Patel, T. Wu, S. Biswas, P. J. Phillips, and R. Chellappa, "Dictionary-based face recognition under variable lighting and pose," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 3, pp. 954–965, 2012.

[18] Q. Zhang and B. Li, "Discriminative K-SVD for dictionary learning in face recognition," in *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2691–2698.

[19] V. M. Patel, T. Wu, S. Biswas, P. J. Phillips, and R. Chellappa, "Illumination robust dictionary-based face recognition," in *Proceedings International Conference on Image Processing*, 2011, pp. 777–780.

[20] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J. M. Geusebroek, "Visual word ambiguity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1271–1283, 2010.

[21] P. Koniusz, F. Yan, and K. Mikolajczyk, "Comparison of mid-level feature coding approaches and pooling strategies in visual concept detection," *Computer Vision and Image Understanding*, to appear.

[22] C. Zhang, P. Yin, Y. Rui, C. R., V. P., X. Sun, P. N., and Z. Zhang, "Boosting-based multimodal speaker detection for distributed meeting videos," *IEEE Transactions on Multimedia*, vol. 10, no. 8, pp. 1541–1552, December 2008.

[23] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[24] K. Mikolajczyk and C. Schmid, "Comparison of affine-invariant local detectors and descriptors," in *Proceedings of European Signal Processing Conference*, 2004, pp. 1729–1732.

[25] P. Koniusz and K. Mikolajczyk, "Soft assignment of visual words as linear coordinate coding and optimisation of its reconstruction error," *Proceedings of International Conference on Image Processing*, pp. 2413–2416, 2011.

[26] J. Wang, J. Yang, K. Yu, F. Lu, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3360–3367.

[27] L. Lingqiao, L. Wang, and X. Liu, "In defence of soft-assignment coding," in *Proceedings of International Conference on Computer Vision*, 2011, pp. 2486–2493.

[28] G. Lathoud and M. Magimai-Doss, "A sector-based, frequency domain approach to detection and localization of multiple speakers," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, vol. 3, 2005, pp. 265–268.

[29] G. Lathoud, J. Bourgeois, and J. Freudenberger, "Sector-based detection for hands-free speech enhancement in cars," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 169–184, 2006.

[30] J. DiBiase, "A high-accuracy, low-latency technique for talker localisation in reverberant environments," Ph.D. dissertation, Brown University, Providence, RI, USA, 2000.

[31] G. Lathoud, J. M. Odobez, and D. Gatica-Perez, "AV16.3: an audio-visual corpus for speaker localization and tracking," in *Proceedings of the Machine Learning for Multi-modal Interaction*, 2004, pp. 182–195.

[32] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[33] H. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 23–38, 1998.