Detection and Localization of 3D Audio-Visual Objects Using Unsupervised Clustering

Vasil Khalidov[1], Florence Forbes[1], Miles Hansard[1], Elise Arnaud[1,2] and Radu Horaud[1]
[1]INRIA Rhône-Alpes, 655 avenue de l'Europe, 38334 Montbonnot, France
[2]Université Joseph Fourier, BP 53, 38041 Grenoble Cedex 9, France
{vasil.khalidov, florence.forbes, miles.hansard, elise.arnaud, radu.horaud}@inrialpes.fr

ABSTRACT

This paper addresses the issues of detecting and localizing objects in a scene that are both seen and heard. We explain the benefits of a human-like configuration of sensors (binaural and binocular) for gathering auditory and visual observations. It is shown that the detection and localization problem can be recast as the task of clustering the audio-visual observations into coherent groups. We propose a probabilistic generative model that captures the relations between audio and visual observations. This model maps the data into a common audio-visual 3D representation via a pair of mixture models. Inference is performed by a version of the expectation-maximization algorithm, which is formally derived, and which provides cooperative estimates of both the auditory activity and the 3D position of each object. We describe several experiments with single- and multiple-speaker detection and localization, in the presence of other audio sources.

Categories and Subject Descriptors

I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*Sensor Fusion*; I.5.3 [**Pattern Recognition**]: Clustering—*Algorithms*

General Terms

Algorithms, Experimentation, Theory

Keywords

Audio-Visual Clustering, Mixture Models, Binaural Hearing, Stereo Vision

1. INTRODUCTION

In most systems that handle multi-modal data, audio and visual inputs are first processed by modality-specific subsystems, whose outputs are subsequently combined. The performance of such procedures in realistic situations is limited in the following ways. Confusion may arise from factors such as background auditory noise, presence of both speech and non-speech multiple audio sources, acoustic reverberations, rapid changes in the visual appearance of an object, varying illumination conditions, visual occlusions, and so forth. The different attempts that have been made to increase robustness are based on the observation that improved object detection and localization can be achieved by integrating auditory and visual information. This is because each modality can compensate for the shortcomings of the other; Simultaneous audiovisual (AV) processing is particularly critical in complex situations such as the ones encountered when distant sensors (microphones and cameras) are used within realistic AV scenarios. This raises two questions: *Where?* – in which mathematical space the AV data fusion should live, and *What?* – which A and V features to select in order to account for an optimal compromise between single- and cross-modality.

Choosing a fusion space.

There are several possibilities. In contrast to the fusion of previous independent processing of each modality [1], the integration could occur at the feature level. In this case audio and video features are concatenated into larger feature-vectors, which are then processed by a single algorithm. However, owing to the very different physical natures of audio and visual stimuli, direct integration is not straightforward. For example, there is no obvious way to associate dense visual maps with sparse sound sources. The approach that we propose in this paper lies between these two extremes. The input features are first transformed into a common representation and the processing is then based on the combined features in this representation. Within this strategy, we identify two major directions depending on the type of *synchrony* being used:

- The first one focuses on *spatial synchrony* and implies combining those signals that were observed at a given time, or through a short period of time, and correspond to the same location. Generative probabilistic models in [2] and [3] for the problem of single speaker tracking achieve this by introducing dependencies of both auditory and visual observations on 2D locations, i.e., in the image plane. Although authors in [2] suggested an enhancement of the model that would tackle the multi-speaker case, it has yet to be implemented. The explicit dependency on the source location in these models can be generalized by the use of particle filters. Such approaches have been used for the task of single speaker tracking [4, 5, 6, 7, 8, 9] and multiple speaker tracking [10, 11, 7, 12, 13]. In the latter case the parameter space grows exponentially as the number of speakers increases, so efficient sampling procedures may be needed, to keep the problem tractable [11, 7].
- The second direction focuses on *temporal synchrony*. It efficiently generalizes the previous approach by making no

ICMI'08, October 20–22, 2008, Chania, Crete, Greece.

a priori assumption on AV object location. Signals from different modalities are grouped if their evolution is correlated through time. The work in [14] shows how the principles of information theory can be used to select those features from different modalities that correspond to the same object. Although the setup consists of a single camera and a single microphone and no special signal processing is used, the model is capable of selecting the speaker among several persons that were visible. Another example of this strategy is described in [15], where matching is performed on the basis of audio and video onsets (times at which sound/motion begins). This model has been shown to work with multiple, as well as with individual, AV objects. Most of these approaches are, however, non-parametric and highly dependent on the choice of appropriate features. Moreover they usually require either learning or ad-hoc tuning of quantities such as window sizes and temporal resolution. They tend to be quite sensitive to artifacts, and may require careful implementation.

Features to be selected.

Some methods rely on complex audio-visual hardware such as microphone arrays, that are calibrated mutually and with respect to one or more cameras [6]. This yields an approximate spatial localization of each audio source. A single microphone is simpler to set up, but it cannot, on its own, provide spatial localization. However, these procedures typically do not treat AV object localization in the true spatial (3D) domain. In contrast, it is argued here that real-world AV data tends to be influenced by the structure of the 3D environment in which it was generated.

Note that two distinct AV objects may project to nearby locations in an image. The more distant object will be partially or totally occluded in this case, and so purely 2D visual information is not sufficient to solve the localization problem. We propose to use a human-like sensor setup that has both binaural hearing and stereoscopic vision. The advantages of using two cameras is twofold. First, the field of view is increased. Second, it allows the extraction of *depth* information through the computation of binocular disparities. Whenever a microphone pair is used, certain audio characteristics, such as interaural time differences (ITD) and interaural level differences (ILD) can be computed as indicators of the 3D position of the sources present in the scene. This type of 3D audio localization plays an important role in some algorithms, such as partitioned sampling [6] (which combines a microphone pair with one camera) and may be a pre-requisite of the data fusion process. An additional advantage of our setup is therefore to allow a more symmetric integration in which neither audio nor vision are assumed to be dominant. We noticed that, so far, there has been no attempt to use visual depth in combination with 3D auditory cues.

In [11] microphone- and camera-arrays are used. A moving AV object is tracked in auditory space and the appropriate camera is selected for further 2D visual analysis. Nevertheless, selecting the appropriate camera to be used in conjunction with a moving object and predicting its visual appearance in a realistic and reliable manner can be quite problematic. The majority of models maintain the image location of a target by supposing that there are no occlusions or by considering them as a special case [11].

The first original contribution of our paper is to embed the problem in the physical 3D space, which is not only natural but has more discriminative power in terms of AV object detection and localization. Typically, it is possible to discriminate between visually adjacent or overlapping objects, provided that we consider them in 3D space. We attempt to combine the benefits of both types of synchronies described above. Our approach makes use of *spatial synchrony*, but unlike the majority of existing models, we perform the binding in 3D space which fully preserves localization information so that the integration is reinforced. At the same time we do not rely on high-level features such as structural templates [11], or on photometric features such as colour models [6]. The fact that we rely on low-level A and V features makes our model more general and less dependent on supervised learning techniques, such as face and speech detectors. We also make use of *temporal synchrony* in the sense that we recast the problem, of how to best combine audio and visual data for 3D object detection and localization, as the task of finding coherent groups of AV observations. The statistical method of choice for solving this problem is cluster analysis.

The second original contribution is to propose a unified framework in which we define a probabilistic generative model that links audio and visual data by mapping them to a common 3D representation. Indeed, the 3D object locations are chosen as a common representation to which both A and V features are mapped, through two mixture models. This approach has a number of inteteresting characteristics: (i) the number of AV objects can be determined from the observed data using statistical model-selection criteria; (ii) a joint probabilistic model, specified through two mixture models which share common parameters, captures the relations between A and V observations; (iii) object localization in 3D within this framework is defined as a maximum likelihood estimation problem in the presence of missing variables, and is carried out by a version of the Expectation Maximization (EM) algorithm which we formally derive; (iv) we show that the model suits well our problem formulation and results into cooperative estimation of both 3D positions of AV objects and detection of auditory activity using procedures that are standard for mixture models, and (v) we evaluate our model within a multiple-speaker detection and localization task, so that each AV object is a person and the auditory activity consists in the speaking state of a person.

2. AUDIO-VISUAL CLUSTERING

The input data consists of M visual observations $\mathbf{f}$, and K auditory observations $\boldsymbol{g}$;

$$\mathbf{f} = \left\{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_m, \ldots, \boldsymbol{f}_M\right\},$$
$$\boldsymbol{g} = \left\{g_1, \ldots, g_k, \ldots, g_K\right\}.$$

This data is recorded over a time interval $[t_1, t_2]$, which is short enough to ensure that the AV objects responsible for $\mathbf{f}$ and $\boldsymbol{g}$ are effectively stationary in space. Then we address the estimation of the AV object *sites*

$$\mathbf{S} = \left\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n, \ldots, \boldsymbol{s}_N\right\},$$

where each $\boldsymbol{s}_n$ is described by its 3D coordinates $(x_n, y_n, z_n)^\top$. Note that in general N is unknown and should be considered as a parameter.

Our acquisition device consists of a stereo pair of cameras and a pair of microphones. A visual observation $\boldsymbol{f}_m$ then is a 3D *binocular coordinate* $(u_m, v_m, d_m)^\top$, where u and v denote the 2D location in the Cyclopean image. This corresponds to a viewpoint halfway between the left and right cameras, and is easily computed from the original image coordinates. The scalar d denotes the binocular disparity at $(u, v)^\top$. Hence, Cyclopean coordinates $(u, v, d)^\top$ are associated with each point $\boldsymbol{s} = (x, y, z)^\top$ in the visible scene. We define a function $\mathcal{F} : \mathbb{R}^3 \mapsto \mathbb{R}^3$ that maps $\mathbf{S}$ onto $\mathbf{f}$, as well as its inverse [16]:

$$\mathcal{F}(\boldsymbol{s}) = \frac{1}{z}(x, y, B)^\top \qquad \mathcal{F}^{-1}(\boldsymbol{f}) = \frac{B}{d}(u, v, 1)^\top, \quad (1)$$

where B is the length of the inter-camera baseline. We note here that cases when d is close to zero correspond to points on very distant objects (for fronto-parallel setup of cameras) from which no 3D structure can be recovered. So it is reasonable to set a threshold and disregard the observations that contain small values of d.

An auditory observation g_k is represented by an auditory disparity, namely the *interaural time difference*, or ITD. To relate a location to an ITD value we define a function $\mathcal{G} : \mathbb{R}^3 \mapsto \mathbb{R}$ that maps $\mathbf{S}$ on $\boldsymbol{g}$:

$$\mathcal{G}(\boldsymbol{s}) = \frac{1}{c}\Big(\|\boldsymbol{s} - \boldsymbol{s}_{M_1}\| - \|\boldsymbol{s} - \boldsymbol{s}_{M_2}\|\Big). \tag{2}$$

Here $c \approx 330\text{ms}^{-1}$ is the speed of sound and $\boldsymbol{s}_{M_1}$ and $\boldsymbol{s}_{M_2}$ are microphone locations in camera coordinates. We notice that each isosurface defined by (2) is represented by one sheet of a two sheet hyperboloid in 3D. So given an observation we can deduce the surface that should contain the source.

We address the problem of AV localization in the framework of *unsupervised clustering*. The rationale is that observations form *groups* that correspond to the different AV objects in the scene. So the problem is recast as a clustering task: an assignment of each observation to one of the clusters should be performed as well as the estimation of cluster parameters, which include the $\boldsymbol{s}_n$'s, the 3D positions of AV objects. To account for the presence of observations that are not related to any AV object, we introduce an additional background (outlier) class. The resulting classes are indexed as $1, \ldots, N, N+1$, the final class being reserved for outliers. Because of the different nature of the observations, clustering is performed via two mixture models respectively in the audio (1D) and video (3D) observation spaces, subject to the common parametrization provided by the positions $\boldsymbol{s}_n$.

In this framework, the observed data are naturally augmented with as many unobserved or missing data, also referred to as *hidden variables*. Thus the complete-data vector consists of an observation and its assignment to one of the $N+1$ groups. We denote by a_m the integer assignment-code for a visual observation $\boldsymbol{f}_m$, and by a'_k the integer assignment-code for an auditory observation g_k. Each observation must be assigned, and hence we have two vectors $\mathbf{a} = \{a_m\}$ and $\mathbf{a}' = \{a'_k\}$ with entries:

$$\begin{aligned} a_m,\, a'_k \,&\in\, \{1,\, \ldots,\, N,\, N+1\}, \\ \text{where}\quad m = 1 \ldots M \quad &\text{and} \quad k = 1 \ldots K. \end{aligned} \tag{3}$$

If the variable a_m takes the value $n \leq N$, then the m^{th} observed visual disparity $\boldsymbol{f}_m$ is attributed to object n. Alternatively, if $n = N+1$, then the disparity is attributed to the outlier class. The auditory assignment variables a'_k follow the same scheme. The observed data are considered as specific realizations of random variables. Here and in what follows we use capital letters for random variables whereas small letters designate their particular realizations.

Perceptual studies have shown that, in human speech perception, audio and video data are treated as class conditional independent [17, 18]. We will further assume that the individual audio and visual observations are also independent given assignment variables. Under this hypothesis, the joint conditional likelihood can be written as

$$P(\mathbf{f}, \boldsymbol{g} \,|\, \mathbf{a}, \mathbf{a}') = \prod_{m=1}^{M} P(\boldsymbol{f}_m | a_m) \prod_{k=1}^{K} P(g_k | a'_k). \tag{4}$$

We use one type of probability distribution to model the AV objects, and a different type to model the outliers. The likelihoods of visual/auditory observations, given that they correspond to an *AV object*, are *Gaussian* distributions whose means respectively $\mathcal{F}(\boldsymbol{s}_n)$ and $\mathcal{G}(\boldsymbol{s}_n)$ depend on the corresponding AV object positions through functions $\mathcal{F}$ and $\mathcal{G}$ defined in (1) and (2):

$$\begin{aligned} &P(\boldsymbol{f}_m \,|\, A_m = n) = \mathcal{N}\big(\boldsymbol{f}_m \,\big|\, \mathcal{F}(\boldsymbol{s}_n), \boldsymbol{\Sigma}_n\big) \\ &= (2\pi)^{-3/2} |\boldsymbol{\Sigma}_n|^{-1/2} \exp\Big(-\|\boldsymbol{f}_m - \mathcal{F}(\boldsymbol{s}_n)\|^2_{\boldsymbol{\Sigma}_n}/2\Big), \end{aligned} \tag{5}$$

$$\begin{aligned} &P(g_k \,|\, A'_k = n) = \mathcal{N}\big(g_k \,\big|\, \mathcal{G}(\boldsymbol{s}_n), \sigma_n^2\big) \\ &= (2\pi)^{-1/2} |\sigma_n|^{-1} \exp\Big(-\big(g_k - \mathcal{G}(\boldsymbol{s}_n)\big)^2 / \big(2\sigma_n^2\big)\Big). \end{aligned} \tag{6}$$

The (co)variances are respectively denoted by $\boldsymbol{\Sigma}_n$ and σ_n^2. The notation $\|\boldsymbol{x}\|^2_{\boldsymbol{\Sigma}}$, used above, represents the Mahalanobis distance $\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{x}$. Similarly, we define the likelihoods for a visual/auditory observation to belong to an *outlier* cluster as *uniform* distributions

$$P\big(\boldsymbol{f}_m \,|\, A_m = N+1\big) = 1/V, \tag{7}$$

$$P\big(g_k \,|\, A'_k = N+1\big) = 1/U, \tag{8}$$

where V and U represent the respective 3D and 1D observed data *volumes* (see Section 4). Our clustering model assumes AV object clusters to have the same distribution type within an observation space, which means that it can be viewed as a standard mixture model, extended by the addition of the outliers class. Although the above distributions are widely applicable, these choices are not enforced by the method described here.

For simplicity, we assume that the assignment variables are independent. More complex choices would be interesting such as defining a random field model to account for more structure within/between the classes. Following [19] the implementation of such models can then be reduced to adaptive implementations of the independent case making it natural to start with

$$P(\mathbf{a}, \mathbf{a}') = \prod_{m=1}^{M} P(a_m) \prod_{k=1}^{K} P(a'_k)\ . \tag{9}$$

The prior probabilities for the video and audio labels are denoted by

$$\pi_n = P(A_m = n) \quad \text{and} \quad \pi'_n = P(A'_k = n), \tag{10}$$

for all $n = 1, \ldots, N+1$. The priors are set to be equal in the absence of a specific prior model.

The posterior probabilities $\alpha_{mn} = P(A_m = n | \boldsymbol{f}_m)$ and $\alpha'_{kn} = P(A'_k = n | g_k)$, can then be calculated using Bayes' theorem. The corresponding expressions for α_{mn} and α'_{kn} are given by:

$$\alpha_{mn} = P\big(A_m = n \,|\, \boldsymbol{f}_m\big) = \frac{\pi_n P\big(\boldsymbol{f}_m | A_m = n\big)}{\sum_{i=1}^{N+1} \pi_i P\big(\boldsymbol{f}_m | A_m = i\big)}, \tag{11}$$

$$\alpha'_{kn} = P\big(A'_k = n \,|\, g_k\big) = \frac{\pi'_n P\big(g_k | A'_k = n\big)}{\sum_{i=1}^{N+1} \pi'_i P\big(g_k | A'_k = i\big)}, \tag{12}$$

where the likelihoods are given by (5 - 8).

To summarize, we formulated our clustering model in terms of the two extended mixture models, bound together through the common parameter space. We denote the concatenated set of parameters by $\boldsymbol{\Theta}$;

$$\begin{aligned} \boldsymbol{\Theta} = \big\{\ &\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N, \\ &\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_N,\ \sigma_1, \ldots, \sigma_N, \\ &\pi_1, \ldots, \pi_{N+1},\ \pi'_1, \ldots, \pi'_{N+1}\ \big\}. \end{aligned} \tag{13}$$

The next step is to devise a procedure that finds the best values for the assignments and for the parameters.

3. ESTIMATION PROCEDURE

Given the probabilistic model defined above, we wish to determine the AV objects that generated the visual and auditory observations, that is to derive values of assignment vectors $\mathbf{a}$ and $\mathbf{a}'$, together with the AV object position vectors $\mathbf{S}$ (which are part of our model unknown parameters). Direct maximum likelihood estimation of mixture models is usually difficult, due to the missing assignments. The Expectation Maximization (EM) algorithm [20] is a general and now standard approach to maximization of the likelihood in missing data problems. The algorithm iteratively maximizes the expected complete-data log-likelihood over values of the unknown parameters, conditional on the observed data and the current values of those parameters. In our clustering context, it provides unknown parameter estimation but also values for missing data by providing membership probabilities to each group. The first problem is how to choose the initial parameter values $\mathbf{\Theta}^{(0)}$ for the algorithm. This question is discussed in Section 4. As soon as the initialization is performed, the algorithm comprises two steps. At iteration q, for current values $\mathbf{\Theta}^{(q)}$ of the parameters, the *E step* consists in computing the conditional expectation:

$$Q(\mathbf{\Theta}, \mathbf{\Theta}^{(q)}) = \sum_{\mathbf{a},\mathbf{a}'} P(\mathbf{a}, \mathbf{a}' | \mathbf{f}, \boldsymbol{g}; \mathbf{\Theta}^{(q)}) \log P(\mathbf{f}, \boldsymbol{g}, \mathbf{a}, \mathbf{a}'; \mathbf{\Theta}) \tag{14}$$

with respect to variables $\mathbf{a}$ and $\mathbf{a}'$, as defined in (3).

The *M step* consists in updating $\mathbf{\Theta}^{(q)}$ by maximizing (14) with respect to the vector $\mathbf{\Theta}$, i.e. in finding $\mathbf{\Theta}^{(q+1)}$ as $\mathbf{\Theta}^{(q+1)} = \operatorname{argmax}_{\mathbf{\Theta}} Q(\mathbf{\Theta}, \mathbf{\Theta}^{(q)})$. We now give detailed descriptions of the E- and M-steps, based on our assumptions.

E-step.

We first rewrite the conditional expectation (14) taking into account decompositions (4) and (9) that arise from independency assumptions. This leads to

$$Q(\mathbf{\Theta}, \mathbf{\Theta}^{(q)}) = Q_{\mathcal{F}}(\mathbf{\Theta}, \mathbf{\Theta}^{(q)}) + Q_{\mathcal{G}}(\mathbf{\Theta}, \mathbf{\Theta}^{(q)}),$$

where the visual and auditory terms in the conditional expectation are as follows;

$$Q_{\mathcal{F}}(\mathbf{\Theta}, \mathbf{\Theta}^{(q)}) = \sum_{m=1}^{M} \sum_{n=1}^{N+1} \alpha_{mn}^{(q)} \log\big(P(\boldsymbol{f}_m \,|\, A_m = n;\, \mathbf{\Theta})\, \pi_n\big),$$

$$Q_{\mathcal{G}}(\mathbf{\Theta}, \mathbf{\Theta}^{(q)}) = \sum_{k=1}^{K} \sum_{n=1}^{N+1} \alpha_{kn}^{\prime(q)} \log\big(P(g_k \,|\, A'_k = n;\, \mathbf{\Theta})\, \pi'_n\big),$$

where $\alpha_{mn}^{(q)}$ and $\alpha_{kn}^{\prime(q)}$ are the expressions in (11) and (12) for $\mathbf{\Theta} = \mathbf{\Theta}^{(q)}$ the current parameter values. In the case of Gaussian distributions, substituting expressions for likelihoods (5) and (6) further leads to eqs. (15) and (16) on the next page.

M-step.

The goal is to maximize (14) with respect to the parameters $\mathbf{\Theta}$ to find $\mathbf{\Theta}^{(q+1)}$. Optimal values for priors π_n and π'_n are easily derived independently of the other parameters by setting the corresponding derivatives to zero and using the constraints $\sum_{n=1}^{N+1} \pi_n = 1$ and $\sum_{n=1}^{N+1} \pi'_n = 1$. The resulting expressions are

$$\pi_n^{(q+1)} = \frac{1}{M} \sum_{m=1}^{M} \alpha_{mn}^{(q)} \quad \text{and} \quad \pi_n^{\prime\,(q+1)} = \frac{1}{K} \sum_{k=1}^{K} \alpha_{kn}^{\prime(q)} \tag{17}$$

for all $n = 1, \ldots, N+1$. The optimization with respect to the other parameters is less straightforward. Using a coordinate system transformation, we substitute variables $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N$ with $\hat{\boldsymbol{f}}_1 = \mathcal{F}(\boldsymbol{s}_1), \ldots, \hat{\boldsymbol{f}}_N = \mathcal{F}(\boldsymbol{s}_N)$. For convenience we introduce the function $h = \mathcal{G} \circ \mathcal{F}^{-1}$ and the parameter set

$$\tilde{\mathbf{\Theta}} = \Big\{ \hat{\boldsymbol{f}}_1, \ldots, \hat{\boldsymbol{f}}_N,\ \mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_N,\ \sigma_1, \ldots, \sigma_N \Big\}.$$

Setting the derivatives with respect to the variance parameters to zero, we obtain the usual empirical variances formulas. Taking the derivative with respect to $\hat{\boldsymbol{f}}_n$ gives

$$\begin{aligned} \frac{\partial Q}{\partial \hat{\boldsymbol{f}}_n} &= \sum_{m=1}^{M} \alpha_{mn} \Big(\boldsymbol{f}_m - \hat{\boldsymbol{f}}_n \Big)^{\top} \mathbf{\Sigma}_n^{-1} \\ &+ \sigma_n^{-2} \sum_{k=1}^{K} \alpha'_{kn} \Big(g_k - h(\hat{\boldsymbol{f}}_n) \Big) \mathbf{\nabla}_n^{\top}, \end{aligned} \tag{18}$$

where the vector $\mathbf{\nabla}_n$ is the transposed product of Jacobians $\mathbf{\nabla}_n = \Big(\frac{\partial \mathcal{G}}{\partial \boldsymbol{s}} \frac{\partial \mathcal{F}^{-1}}{\partial \boldsymbol{f}} \Big)^{\top}_{\boldsymbol{f} = \hat{\boldsymbol{f}}_n}$ which can be easily computed from definitions (1) and (2).

Difficulties now arise from the fact that it is necessary to perform simultaneous optimization in two different observation spaces, auditory and visual. It involves solving a system of equations that contain derivatives of $Q_{\mathcal{F}}$ and $Q_{\mathcal{G}}$ whose dependency on $\boldsymbol{s}_n$ is expressed through $\mathcal{F}$ and $\mathcal{G}$ and is non-linear. In fact, this system does not yield a closed form solution and the traditional EM algorithm cannot be performed. However, setting the gradient (18) to zero leads to an equation of special form, namely the *fixed point equation* (FPE), where the location $\hat{\boldsymbol{f}}_n$ is expressed as a function of the variances and itself. Solution of this equation together with the empirical variances give the optimal parameter set. For these reasons we implemented and tested an M-step that iterates through FPE to obtain $\hat{\boldsymbol{f}}_n$: Nevertheless, we noticed that such solutions thus obtained tend to make the EM algorithm converge to local maxima of the likelihood.

An alternative way to seek for the optimal parameter values is to use a gradient descent-based iteration, for example, the Newton-Raphson procedure. However, the limit value $\tilde{\mathbf{\Theta}}^{(q+1)}$ is not necessarily a global optimizer. Provided that the value of Q is improved at every iteration, the algorithm can be considered as an instance of the Generalized EM (GEM) algorithm [21]. The updated value $\tilde{\mathbf{\Theta}}^{(q+1)}$ can be taken of the form

$$\tilde{\mathbf{\Theta}}^{(q+1)} = \tilde{\mathbf{\Theta}}^{(q)} + \gamma^{(q)} \mathbf{\Gamma}^{(q)} \left[\frac{\partial Q(\mathbf{\Theta}, \mathbf{\Theta}^{(q)})}{\partial \tilde{\mathbf{\Theta}}} \right]^{\top}_{\mathbf{\Theta} = \mathbf{\Theta}^{(q)}}, \tag{19}$$

where $\mathbf{\Gamma}^{(q)}$ is a linear operator that depends on $\tilde{\mathbf{\Theta}}^{(q)}$ and $\gamma^{(q)}$ is a scalar sequence of gains. For instance, for Newton-Raphson procedure one should use $\gamma^{(q)} \equiv 1$ and $\mathbf{\Gamma}^{(q)} = -\left[\frac{\partial^2 Q}{\partial \tilde{\mathbf{\Theta}}^2} \right]^{-1}_{\mathbf{\Theta} = \mathbf{\Theta}^{(q)}}$. The principle here is to choose $\mathbf{\Gamma}^{(q)}$ and $\gamma^{(q)}$ so that (19) defines a GEM sequence. In what follows we concentrate on the latter algorithm, as it is more flexible, and it produces better results in our experiments.

Clustering.

As well as providing parameter estimates, the EM algorithm can be used to determine assignments of each observation to one of the $N+1$ classes. Observations $\boldsymbol{f}_m$ and g_k are assigned, respectively, to classes η_m and η'_k as follows;

$$\eta_m = \operatorname*{argmax}_{n=1,\ldots,N+1} \alpha_{mn} \qquad \text{and} \qquad \eta'_k = \operatorname*{argmax}_{n=1,\ldots,N+1} \alpha'_{kn}.$$

We use this in particular to determine active speakers using the auditory observations assignments η'_k's. For every person we can derive the speaking state by the number of associated observations.

$$Q_{\mathcal{F}}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(q)}) = -\frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{N}\alpha_{mn}^{(q)}\left(\|\boldsymbol{f}_m - \mathcal{F}(\boldsymbol{s}_n)\|^2_{\boldsymbol{\Sigma}_n} + \log\left((2\pi)^3|\boldsymbol{\Sigma}_n|\pi_n^{-2}\right)\right) - \frac{1}{2}\sum_{m=1}^{M}\alpha_{m,N+1}^{(q)}\log\left(V^2\pi_{N+1}^{-2}\right) \quad (15)$$

$$Q_{\mathcal{G}}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(q)}) = -\frac{1}{2}\sum_{k=1}^{K}\sum_{n=1}^{N}\alpha_{kn}^{\prime(q)}\left(\frac{(g_k - \mathcal{G}(\boldsymbol{s}_n))^2}{\sigma_n^2} + \log(2\pi\sigma_n^2\pi_n^{\prime -2})\right) - \frac{1}{2}\sum_{k=1}^{K}\alpha_{k,N+1}^{\prime(q)}\log(U^2\pi_{N+1}^{\prime -2}) \quad (16)$$

The case when all η_k''s are equal to $N+1$ would mean that there is no AV object involved in auditory activity.

4. EXPERIMENTAL RESULTS

Within the task of 3D AV object localization there are three sub-tasks to be solved. First, the *number* of AV objects should be determined. Second, these objects should be *localized* and finally, those that are involved in auditory activity should be *selected*. The proposed probabilistic model has the advantage of providing a means to solve all three sub-tasks at once. There is no need to develop separate models for every particular sub-task, and at the same time we formulate our approach within the Bayesian framework which is rich and flexible enough to suit the requirements. To determine the number of AV objects, we gather a sufficient quantity of audio observations and apply the Bayesian Information Criterion (BIC) [22]. This is a well-founded approach to the problem of model selection, given the observations. The task of localization in our framework is recast into the parameter estimation problem. This gives an opportunity to efficiently use the EM algorithm to estimate the 3D positions. We note here that our model is defined so as to perform well in the case of a single AV object as well as in the multiple AV object case without any special reformulation. To obtain the auditory activity state of an object we use the posterior probabilities of the assignment variables calculated at the E step of the algorithm.

We evaluated the ability of our algorithms to estimate the 3D locations and auditory activity of AV objects on the task of person localization and their speaking-state estimation. We considered two scenarios: a typical 'meeting' situation (M1) and the case of a moving speaking person (TTOS1). The two audio-visual sequences, namely M1 and TTOS1, that we use in this paper are part of a database of realistic AV scenarios described in detail in [23]. A mannequin, with a pair of microphones fixed into its ears and a pair of stereoscopic cameras mounted onto its forehead, served as the acquisition device. The reason for choosing this configuration was to record data from the perspective of a person, i.e. to try to capture what a person would both hear and see while being in a natural AV environment. Each of the recorded scenarios comprises two audio tracks and two image sequences, together with the calibration information. The first sequence (M1) is a meeting scenario, shown on figures 1, 2, and 3-top. There are five persons sitting around a table, but only 3 persons are visible. The second sequence (TTOS1) involves a person walking along a zig-zag trajectory towards the camera while speaking, figure 3-bottom. M1 is 20s long (500 stereo-frames at 25 frames/s) while TTOS1 is 9s long (225 stereo-frames). They were farther split into short-time intervals that correspond to three video frames.

Audio and visual observations were collected within each interval using the following techniques. A standard procedure was used to identify 'interest points' in the left and right images [24]. These features were put into binocular correspondence by comparing the local image-structure at each of the candidate points, as described in [16]. The cameras were calibrated [25] in order to define the $(u, v, d)^\top$ to $(x, y, z)^\top$ mapping (1). Auditory disparities were obtained through the analysis of cross-correlogram of the filtered left and right microphone signals for every frequency band [26]. On an average, there were 1000 visual observations and 9 auditory observations within each time interval.

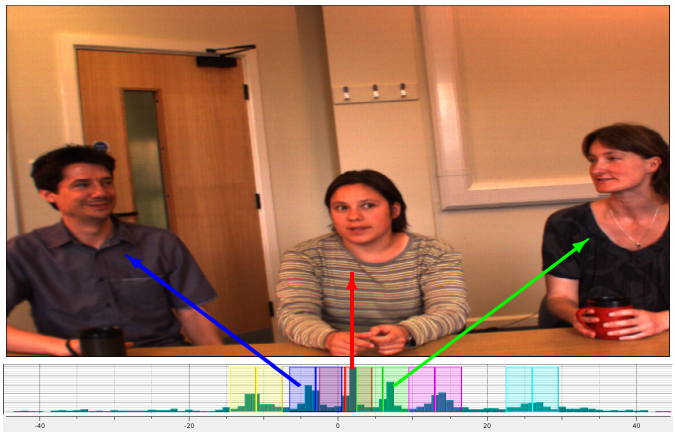

Figure 1: One image from the M1 sequence, the histogram of ITD values, and BIC estimates of the clusters. The transparent rectangles correspond to the variance of each cluster while solid coloured lines correspond to cluster centers.

To determine the number of speakers that were present in the scene, we applied the BIC criterion to the auditory data. Figure 1 shows the results for the observations collected over 20s of the meeting scenario. They are represented as histograms of ITD values together with the estimated clusters in the ITD space, using the Gaussian mixture model outlined above. The transparent coloured rectangles designate the variances of each cluster, while solid coloured lines drawn at their centres are the corresponding cluster centers. Figure 1 shows six detected clusters, which is exactly the number of persons present in the scene: Five persons involved in a meeting (among whom only three are visible), as well as a sixth person who performed a "clap" at the begining of the recording, and then remains present in the room producing sounds sometimes.

The 3D localization and speaking state estimation were performed by the EM algorithm for each time interval. The parameter values from the previous interval served as initial values for the subsequent one. We report here on the results obtained by the versions of the algorithm based on a gradient descent (GD) technique, with $\boldsymbol{\Gamma}$ being block diagonal. We used $\left[-\partial^2 Q/\partial \hat{\boldsymbol{f}}_n^2\right]^{-1}_{\boldsymbol{\Theta}=\boldsymbol{\Theta}^{(q)}}$ as a block for $\hat{\boldsymbol{f}}_n$, so that the descent direction is the same as in Newton-Raphson method. In the examples that we present we adopted the same video variance matrix $\boldsymbol{\Sigma}$ for all the clusters, thus there was one common block in $\boldsymbol{\Gamma}^{(q)}$ that performed linear mapping of the form $\boldsymbol{\Gamma}_{\boldsymbol{\Sigma}}^{(q)}(\cdot) = \left(\sum_{n=1}^{N}\sum_{m=1}^{M}\alpha_{mn}^{(q)}\right)^{-1}\boldsymbol{\Sigma}^{(q)}(\cdot)\,\boldsymbol{\Sigma}^{(q)}$. This direction change corresponds to a step towards the empirical vari-

ance value. Analogous blocks (cells) were introduced for audio variances, though, unlike the visual variances, individual parameters were used. The number of iterations within each M step for GD was chosen to be 1, as further iterations did not yield significant improvements. We tried two types of the gain sequence: with $\gamma^{(q)} \equiv 1$ (classical GD) and $\gamma^{(q)} = \frac{1}{2} + 1/(2(q+1))$ (relaxed GD). By adjusting $\gamma^{(q)}$ one can improve certain properties of the algorithm, such as convergence speed, accuracy of the solution as well as its dynamic properties in the case of parameters changing through time.

We tested the algorithm with the two possible gain sequences described above and very similar results were obtained in terms of likelihood maximization, the classical GD strategy usually converges faster. Nevertheless, there are cases when the moderate behaviour of the relaxed version around the optimal point improves the rate of convergence. This feature of the relaxed GD could prove to be useful in the case of strong noise as well.

Figure 2 shows a typical output of our algorithm applied to a time interval (**see the video provided with the submission and listen to the soundtrack**). The interest points are shown as dots in the left and right images. The 3D visual observations are shown in x, y, z space, below the images. One may notice that after stereo reconstruction there are both inliers and outliers, as well as 3D points that belong to the background. The histogram representation of the ITD observation space is given in the middle. Trnasparent ellipses in the images represent projections of the visual covariances corresponding to 3D clusters. The three 3D spheres (blue, red, and green) correspond to the same visual covariances centered at the cluster centers. Transparent grey spheres surround the current speakers (there are two speakers in this example), also shown as with white circles in the image pair. The small grey square shows the 'ground truth' – the person actually speaking during the time interval. A correct speaker-detection/3D-localization is represented by both a white circle and a grey square. Hence, in this example, one speaker was wrongly detected. (We annotated the original soundtrack as follows: first performed onset detection, then we enriched the results by the offset information and marked the final regions manually). Figure 3 shows four consecutive time-intervals for the M1 sequence (top) and for the TTOS1 sequence (bottom), and the corresponding estimated 3D localization and speech activity. Notice that only one image per frame (and not the stereo-pair) is shown on this figure.

The performance of the algorithm for the two sequences is summarized on Table 1. The first column (*time-int*) shows the total number of time-intervals being considered. The second column (*AV-int*) gives the total number of time-intervals containing AV objects (obtained from the annotated database). The third column (*AV-OK*) gives the total number of time-intervals were auditory objects were correctly detected (which should ideally be equal to the second column). The last two columns show the percentage of 'missed target' (*AV-missed*), i.e. AV objects being marked as non-AV, and the percentage of 'false alarm' (*AV-false*), i.e. non-AV objects being marked as AV.

	time-int	*AV-int*	*AV-OK*	*AV-missed*	*AV-false*
M1	166	89	75	0.16	0.14
TTOS1	76	69	60	0.13	0.43

Table 1: Summary of speech detection for the meeting (M1) and the moving person (TTOS1). The last two columns give some statistics on the probabilities of "missed targets" and "false alarms" (see text for details).

When analyzing these results we noticed that there are three major reasons for the errors to occur. First, the analysis of *silent* regions of the spectrogram gives rise to erroneous ITDs, which are then associated to an AV object. This could have been easily suppressed by filtering out such regions, which would have lead to a smaller percentage of "false alarm" errors (0.07 instead of 0.14 and 0 instead of 0.43, last table column). Second, false alarms in the second case are mainly due to the sound caused by the footsteps while the person is walking. Because the two auditory sources (voice and footsteps) are associated with the same person, they share the same azimuth, and hence they have almost equal ITD's. Moreover, the feet are not visible. Hence, the algorithm fails to extract two distinct AV cluster centers. Third, many errors of "missed target" type occur due to the discretization by time intervals; sometimes only a short fragment of audio gets included into the analysis, which is unsufficient to generate the correct ITD value. So, it would be reasonable to consider an auditory observations distribution on a sequence without any explicit partitioning into intervals. Another means of improving the "missed target" error rate is to introduce some dependency between the observed ITDs. Thus the detection rate could reach 0.92 for M1. It is worth pointing out that one cannot rely on a face detector even in this scenario with rather "favourable" conditions, since the participants turn their heads away from the cameras quite frequently, and would cause a face detector to fail.

Figure 4 shows the results of 3D localization for the TTOS1 sequence. The stereo pair is located at the origin of the 3D coordinate frame. The graph shows the three directions of the person's zig-zag motion: first to the left and towards the camera (colour gradient from blue to green), then to the right and towards the camera (colour gradient from green to red) and then beside the camera to the left into the invisible region (colour gradient from red to brown). The scale of the axes is given in millimeters. We note that for a distant object the estimated location is rather 'noisy' and less precise, but as a person approaches the camera, the behaviour of the estimate becomes more stable. This can be predicted from the formula (1) that gives the dependency of $(x, y, z)^\top$ on $(u, v, d)^\top$. Indeed, objects that are close to the cameras have large values for the d-coordinate that dominate the noise. But as the distance to an object increases, d becomes small and the influence of the noise on z augments. Nevertheless, the amplitude of the observed fluctuations in our case are about $\pm$10cm, so the estimates can be considered to be accurate throughout the sequence. Again, it is worth pointing out that the level of noise in TTOS1 example was very high (see Fig.3), generated by numerous mismatches. But still the clustering technique allowed us to correctly weight the observations, so that the effect of the noisy ones was reduced to minimum.

5. CONCLUSION

We have presented a unified framework that captures the relationships between audio and visual observations, and makes full use of their combination to accurately estimate the number of audio-visual objects, their 3D locations, and their speaking states. Our approach is based on unsupervised clustering, and results in a very flexible and general-purpose model. In particular, it does not depend on any high-level feature detectors, such as faces or speech cues.

The current approach could be extended in the following ways. Firstly, it would be reasonable to abandon the independency assumption for observations within a single modality. This would introduce the notion of *density* of visual observations and the notion of *stream* for auditory observations. In both cases it would improve the quality of the resulting AV clusters. This could be done through

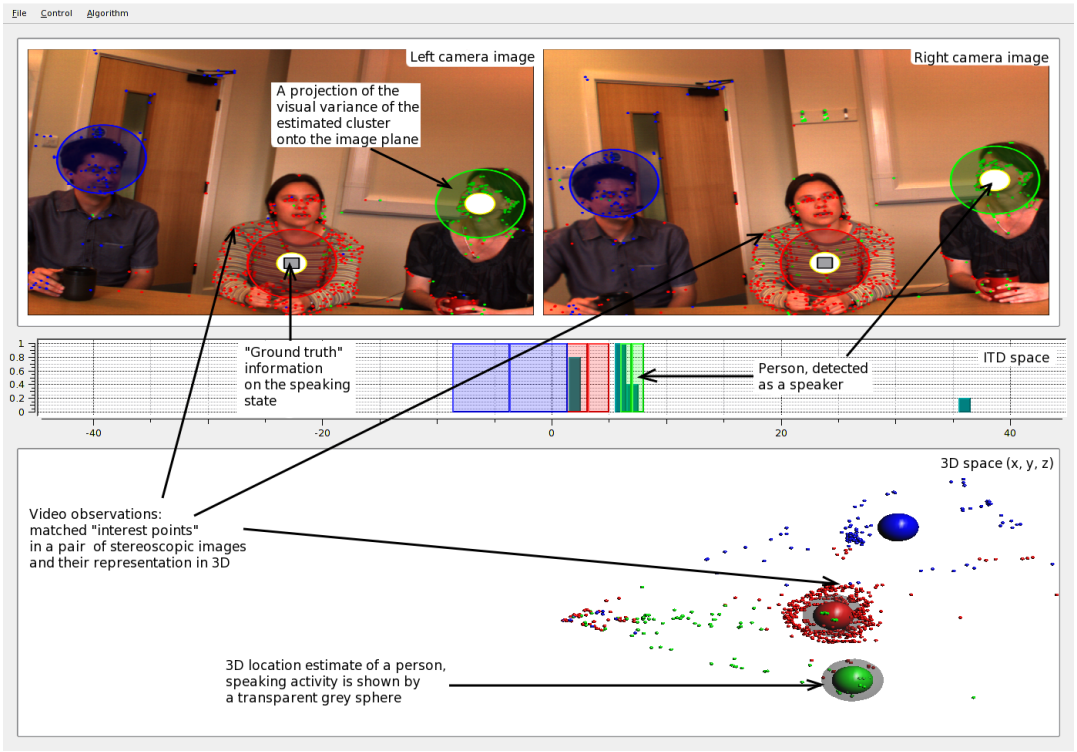

Figure 2: A typical output of the algorithm: stereoscopic image pair, histogram of ITD observation space, and 3D clustering (see text for details).

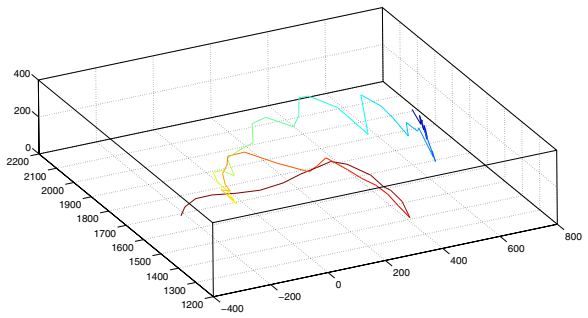

Figure 4: The estimated trajectory of the moving person in the TTOS1 sequence. The zig-zag trajectory eventually moves away from the visual field of view.

the definition of a different missing assignments probability (9). For instance, using a Markov chain model for the auditory assignments and a Markov field model for the visual assignments, the implementation could be derived in a straightforward maner from variational approximations as described in [19]. Secondly, our results show that pre-filtering the spectrogram channels to eliminate low-energy silent regions would also result in a performance increase. Thirdly, our model can immediately be used for the case of dynamic estimation, which would allow us to estimate the number and locations of AV objects online.

Acknowledgements. The authors would like to warmly thank Heidi Christensen for providing the ITD detection software that was used to generate auditory observations in this work. We are grateful to Martin Cooke, Jon Barker, Sue Harding, and Yan-Chen Lu of the Speech and Hearing Group (Department of Computer Science, University of Sheffield) for helpful discussions and comments. We also thank anonymous reviewers for valuable suggestions and remarks. This work has been funded by the European Commission under the POP project (Perception on Purpose), number FP6-IST-2004-027268, `http://perception.inrialpes.fr/POP/`.

6. REFERENCES

[1] M. Heckmann, F. Berthommier, and K. Kroschel. Noise adaptive stream weighting in audio-visual speech recognition. *EURASIP J. Applied Signal Proc.*, 11:1260–1273, 2002.

[2] M. Beal, N. Jojic, and H. Attias. A graphical model for audiovisual object tracking. *IEEE Trans. PAMI*, 25(7):828–836, 2003.

[3] A. Kushal, M. Rahurkar, L. Fei-Fei, J. Ponce, and T. Huang. Audio-visual speaker localization using graphical models. In *Proc. 18th ICPR.*, pages 291–294, 2006.

[4] D. N. Zotkin, R. Duraiswami, and L. S. Davis. Joint audio-visual tracking using particle filters. *EURASIP Journal on Applied Signal Processing*, 11:1154–1164, 2002.

[5] J. Vermaak, M. Ganget, A. Blake, and P. Pérez. Sequential monte carlo fusion of sound and vision for speaker tracking. In *Proc. IEEE ICCV*, pages 741–746, 2001.

[6] P. Perez, J. Vermaak, and A. Blake. Data fusion for visual tracking with particles. *Proc. of IEEE*, 92(3):495–513, 2004.

[7] Y. Chen and Y. Rui. Real-time speaker tracking using particle filter sensor fusion. *Proc. of IEEE*, 92(3):485–494, 2004.

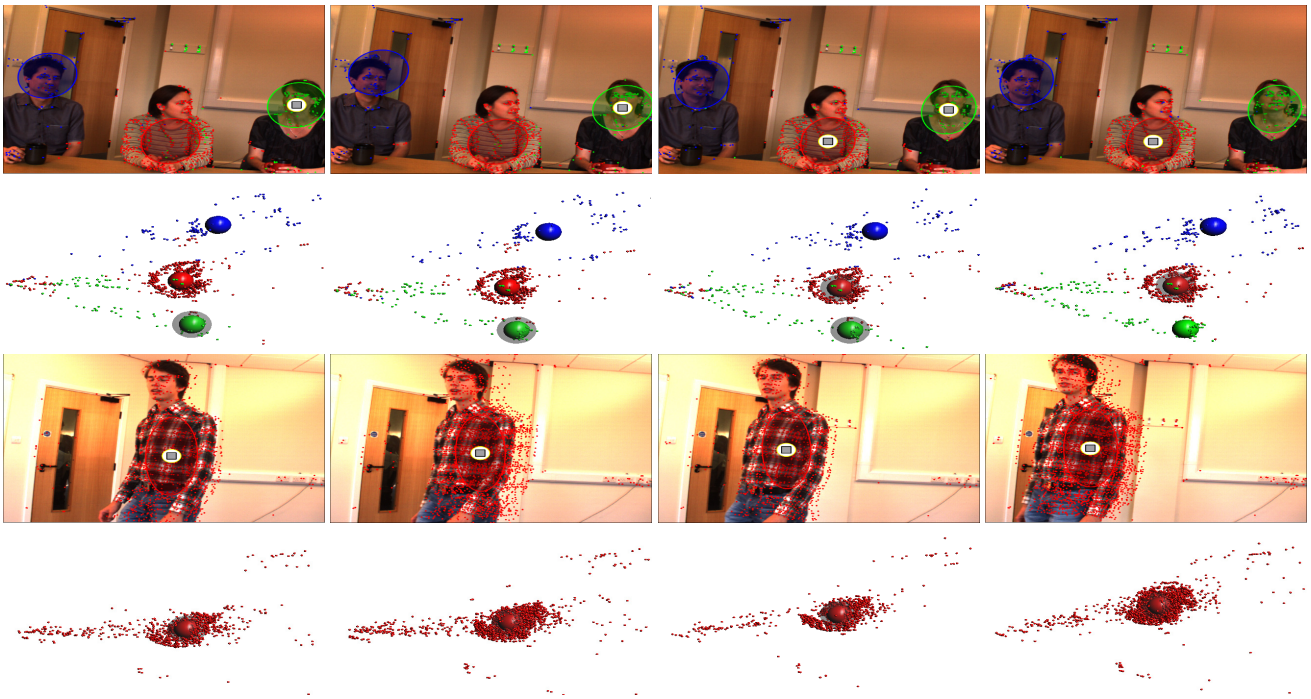

Figure 3: Four consecutive time-intervals from M1 (top) and from TTOS1 (bottom) together with the results of 3D localization and speech activity.

[8] K. Nickel, T. Gehrig, R. Stiefelhagen, and J. McDonough. A joint particle filter for audio-visual speaker tracking. In *Proc. 7th International Conference on Multimodal Interfaces*, pages 61–68, 2005.

[9] T. Hospedales, J. Cartwright, and S. Vijayakumar. Structure inference for Bayesian multisensory perception and tracking. In *Proc. International Joint Conference on Artificial Intelligence*, pages 2122–2128, 2007.

[10] N. Checka, K. Wilson, M. Siracusa, and T. Darrell. Multiple person and speaker activity tracking with a particle filter. In *IEEE Conf. Acoust. Sp. Sign. Proc.*, pages 881–884, 2004.

[11] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan. Audiovisual probabilistic tracking of multiple speakers in meetings. *IEEE Trans. on ASLP*, 15(2):601–616, 2007.

[12] K. Bernardin and R. Stiefelhagen. Audio-visual multi-person tracking and identification for smart environments. In *Proc. 15th International ACM Conference on Multimedia*, pages 661–670, 2007.

[13] R. Brunelli, A. Brutti, P. Chippendale, O. Lanz, M. Omologo, P. Svaizer, and F. Tobia. A generative approach to audio-visual person tracking. In *Multimodal Technologies for Perception of Humans: Proc. 1st International* CLEAR *Evaluation Workshop*, pages 55–68, 2007.

[14] J. Fisher and T. Darrell. Speaker association with signal-level audiovisual fusion. *IEEE Trans. on Multimedia*, 6(3):406–413, 2004.

[15] Z. Barzelay and Y.Y. Schechner. Harmony in motion. In *Proc. of IEEE CVPR*, pages 1–8, 2007.

[16] M. Hansard and R.P. Horaud. Patterns of binocular disparity for a fixating observer. In *Advances in Brain, Vision, & AI, 2nd Int. Symp.*, pages 308–317. Springer, 2007.

[17] J.R. Movellan and G. Chadderdon. Channel separability in the audio-visual integration of speech: A Bayesian approach. In D.G. Stork and M.E. Hennecke, editors, *Speech Reading by Humans and Machines: Models, Systems and Applications*, NATO ASI Series, pages 473–487. Springer, Berlin, 1996.

[18] D.W. Massaro and D.G. Stork. Speech recognition and sensory integration. *American Scientist*, 86(3):236–244, 1998.

[19] G. Celeux, F. Forbes, and N. Peyrard. EM procedures using mean-field approximations for Markov model-based image segmentation. *Pattern Recognition*, 36:131–144, 2003.

[20] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38, 1977.

[21] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[22] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, March 1978.

[23] E. Arnaud, H. Christensen, Y.C. Lu, J. Barker, V. Khalidov, M. Hansard, B. Holveck, H. Mathieu, R. Narasimha, F. Forbes, and R. Horaud. The CAVA corpus: Synchronized stereoscopic and binaural datasets with head movements. In *Proc. of ICMI 2008*, 2008.

[24] C. Harris and M. Stephens. A combined corner and edge detector. In *Proc. 4th Alvey Vision Conference*, pages 147–151, 1988.

[25] Intel OpenCV Computer Vision library. http://www.intel.com/technology/computing/opencv.

[26] H. Christensen, N. Ma, S.N. Wrigley, and J. Barker. Integrating pitch and localisation cues at a speech fragment level. In *Proc. of Interspeech 2007*, pages 2769–2772, 2007.